

Métodos Numéricos con MATLAB

Tercera edición

John H. Mathews

California State University, Fullerton

Kurtis D. Fink

Northwest Missouri State University

Traducción

Pedro José Paúl Escolano
Universidad de Sevilla

Revisión técnica

Antonio Fernández Carrión
Manuel Contreras Márquez
Universidad de Sevilla

PRENTICE HALL

Madrid • México • Santafé de Bogotá • Buenos Aires • Caracas • Lima
Montevideo • San Juan • San José • Santiago • Sao Paulo • White Plains

Find your solutions manual here!

El Solucionario

www.elsolucionario.net



Subscribe RSS



Find on Facebook



Follow my Tweets

Encuentra en nuestra página los Textos Universitarios que necesitas!

Libros y Solucionarios en formato digital

El complemento ideal para estar preparados para los exámenes!

*Los Solucionarios contienen TODOS los problemas del libro resueltos
y explicados paso a paso de forma clara..*

Visítanos para descargarlos GRATIS!

Descargas directas mucho más fáciles...

WWW.ELSOLUCIONARIO.NET

Biology

Investigación Operativa

Computer Science

Physics

Estadística

Chemistry

Matemáticas Avanzadas

Geometría

Termodinámica

Cálculo

Electrónica

Circuitos

Math

Business

Civil Engineering

Economía

Análisis Numérico

Mechanical Engineering

Electromagnetismo

Electrical Engineering

Algebra

Ecuaciones Diferenciales

Find your solutions manual here!

JOHN H. MATHEWS, KURTIS D. FINK

Métodos Numéricos con MATLAB

PRENTICE HALL, Madrid, 2000

ISBN: **84-8322-181-0**

Materia: Análisis Numérico 519.6

Formato 170 x 240

Páginas: 736

JOHN H. MATHEWS, KURTIS D. FINK

Métodos Numéricos con MATLAB

No está permitida la reproducción total o parcial de esta obra
ni su tratamiento o transmisión por cualquier medio o método
sin autorización escrita de la Editorial.

DERECHOS RESERVADOS

© 2000 respecto a la primera edición en español por:

PRENTICE HALL Iberia S.R.L.

C/ Núñez de Balboa, 120

28006 Madrid

ISBN: 84-8322-181-0

Depósito Legal: M- 40.040-1999

Traducido de: Numerical Methods using MATLAB

Copyright© 1999 by Prentice Hall, Inc

ISBN: 0-13-270042-5

Edición en español

Editora: Isabel Capella

Asistente editorial: Ana Isabel García

Editor de producción: Pedro Aguado

Diseño de cubierta: DIGRAF, S. A.

Impreso por:

IMPRESO EN ESPAÑA - PRINTED IN SPAIN

Índice

Prólogo	<i>vii</i>
1 Preliminares	1
1.1 Un repaso al cálculo infinitesimal	2
1.2 Números binarios	14
1.3 Análisis del error	26
2 Resolución de ecuaciones no lineales	45
2.1 Métodos iterativos para resolver $x = g(x)$	46
2.2 Los métodos de localización de raíces	57
2.3 Aproximación inicial y criterios de convergencia	69
2.4 Los métodos de Newton-Raphson y de la secante	77
2.5 Los métodos de Aitken, Steffensen y Muller (opcional)	99
3 Resolución de sistemas lineales	111
3.1 Vectores y matrices	111
3.2 Multiplicación de matrices	120
3.3 Sistemas lineales triangulares	132
3.4 Eliminación gaussiana y pivoteo	137

3.5	Factorización triangular	155
3.6	Métodos iterativos para sistemas lineales	171
3.7	Métodos iterativos para sistemas no lineales (opcional)	183

4 Interpolación y aproximación polinomial 203

4.1	Series de Taylor y cálculo de los valores de una función	205
4.2	Introducción a la interpolación	216
4.3	Interpolación de Lagrange	225
4.4	Polinomio interpolador de Newton	239
4.5	Polinomios de Chebyshev (opcional)	250
4.6	Aproximaciones de Padé	263

5 Ajuste de curvas 273

5.1	Rectas de regresión en mínimos cuadrados	274
5.2	Ajuste de curvas	285
5.3	Interpolación polinomial a trozos	302
5.4	Series de Fourier y polinomios trigonométricos	322

6 Derivación numérica 335

6.1	Aproximaciones a la derivada	336
6.2	Fórmulas de derivación numérica	355

7 Integración numérica 371

7.1	Introducción a la integración numérica	372
7.2	Las reglas compuestas del trapecio y de Simpson	384
7.3	Reglas recursivas y método de Romberg	399
7.4	Integración adaptativa	415
7.5	El método de integración de Gauss-Legendre (opcional)	423

8 Optimización numérica 433

8.1	Minimización de una función	434
-----	-----------------------------	-----

9	Ecuaciones diferenciales ordinarias	463
9.1	Introducción a las ecuaciones diferenciales	464
9.2	El método de Euler	470
9.3	El método de Heun	482
9.4	El método de la serie de Taylor	490
9.5	Los métodos de Runge-Kutta	497
9.6	Métodos de predicción y corrección	515
9.7	Sistemas de ecuaciones diferenciales	529
9.8	Problemas de contorno	539
9.9	El método de las diferencias finitas	548
10	Ecuaciones en derivadas parciales	557
10.1	Ecuaciones hiperbólicas	560
10.2	Ecuaciones parabólicas	570
10.3	Ecuaciones elípticas	582
11	Autovalores y autovectores	601
11.1	El problema de los autovalores	602
11.2	Los métodos de las potencias	615
11.3	El método de Jacobi	629
11.4	Autovalores de matrices simétricas	643
	Apéndice: MATLAB	661
	Referencias temáticas	671
	Bibliografía y referencias	675
	Soluciones de algunos ejercicios	687
	Índice analítico	713

www.elsolucionario.net

www.elsolucionario.net

Prólogo

Este libro proporciona una introducción a los fundamentos del análisis numérico adecuada para estudiantes de matemáticas, informática, física e ingeniería. Se supone que la persona que lee este libro está familiarizada con el cálculo infinitesimal y que ha recibido un curso de programación estructurada. El contenido del texto está organizado de forma modular para que pueda ser ajustado tanto a un curso cuatrimestral como a uno anual. En pocas palabras, el libro contiene material suficiente para que se puedan seleccionar los temas adecuados a las necesidades y los objetivos docentes de cada curso concreto.

Los métodos numéricos son muy útiles e interesantes para estudiantes de diversa procedencia, hecho que tenemos presente a lo largo de todo el libro. Así, hay una amplia variedad de ejemplos y problemas que ayudarán a mejorar las habilidades de los estudiantes tanto en el conocimiento de la teoría como en la práctica del análisis numérico. Los cálculos hechos con un computador se presentan mediante tablas y, cuando sea posible, también mediante gráficas, de manera que sea fácil visualizar e interpretar las aproximaciones numéricas obtenidas. Los programas hechos con el paquete MATLAB son nuestro vehículo de presentación de los algoritmos numéricos subyacentes.

Hemos puesto énfasis en la explicación de por qué los métodos numéricos funcionan y de cuáles son sus limitaciones. Esto constituye un reto que conlleva la necesidad de mantener un equilibrio entre la teoría, el análisis del error y la legibilidad. Presentamos, para cada método, un análisis del error que resulte apropiado para el método en cuestión, pero que, al mismo tiempo, no resulte oscuro para el lector. Damos una deducción matemática de aquellos métodos

que utilizan resultados elementales que debe servir para afianzar la comprensión que cada estudiante tiene de las matemáticas estudiadas hasta el momento. Las tareas de computación con el paquete MATLAB sirven para que los estudiantes tengan la oportunidad de practicar sus habilidades en la computación científica.

Los ejercicios numéricos más cortos pueden realizarse con una calculadora de bolsillo y los más largos pueden llevarse a cabo usando los programas del paquete MATLAB. Queda para el profesorado la labor de usar pedagógicamente los cálculos numéricos para guiar a los estudiantes. Cada cual puede establecer las tareas que sean más apropiadas de acuerdo con los recursos de computación existentes en su centro, pero, en cualquier caso, animamos a los estudiantes a que experimenten con los programas del paquete MATLAB. Estas herramientas pueden ser empleadas para ayudar a los estudiantes a realizar la componente numérica de los ejercicios que deban resolver en el laboratorio.

Esta tercera edición nace de la necesidad de corregir y actualizar los contenidos de la anterior; por ejemplo, hemos añadido el método *QR* al capítulo sobre autovalores y autovectores. Un aspecto nuevo de esta edición es el uso explícito del paquete de programas MATLAB: incluimos un apéndice que contiene una introducción a la sintaxis del paquete MATLAB y hemos añadido ejemplos elaborados con esta herramienta a lo largo del libro, así como programas completos en cada sección. Existe un disquete que puede solicitarse a la editorial.

Nuestra actitud previa era que cualquier lenguaje de programación que los estudiantes supieran utilizar podría resultar adecuado. Sin embargo, teniendo en cuenta que muchos de los estudiantes que se matriculan en este curso no han aprendido todavía ningún lenguaje de programación (salvo los de informática), que el paquete MATLAB se ha convertido en una herramienta para casi todos los campos de la ingeniería y de la matemática aplicada y que sus versiones nuevas han mejorado los aspectos de programación, hemos pensado que los estudiantes emplearán su tiempo de manera más fácil y productiva con esta versión de nuestro libro basada en el paquete MATLAB.

Agradecimientos

Nos gustaría expresar nuestra gratitud a todas las personas que han contribuido con su esfuerzo a las diversas ediciones de este libro. Yo (John Mathews) quiero agradecer, en primer lugar, a mis estudiantes de la California State University en Fullerton; a mis colegas Stephen Goode, Mathew Koshy, Edward Sabotka, Harris Schultz y Soo Tang Tan por su ayuda en la primera edición y, además, a Russell Egbert, William Gearhart, Roneld Miller y Greg Pierce por sus sugerencias para la segunda edición. Le agradezco también a James Friel, Director del Departamento de Matemáticas de la CSUF, el ánimo recibido.

Críticos que hicieron recomendaciones muy útiles para la segunda edición fueron Walter M. Patterson III (Lander College), George B. Miller (Central Connecticut State University), Peter J. Gingo (The University of Akron), Mi-

chael A. Freedman (The University of Alaska, Fairbanks) y Kenneth P. Bube (University of California, Los Ángeles). Por sus recomendaciones para la segunda edición, estamos agradecidos a: Richard Bumby (Rutgers University), Robert L. Curry (U.S. Army); Bruce Edwards (University of Florida) y David R. Hill (Temple University).

En esta tercera edición queremos agradecer sus sugerencias a Tim Sauer, George Mason University; Gerald M. Pitstick, University of Oklahoma; Victor De Brunner, University of Oklahoma; George Trapp, West Virginia University; Tad Jarik, University of Alabama, Huntsville; Jeffrey S. Scroggs, North Carolina State University; Kurt Georg, Colorado State University, y James N. Craddock, Southern Illinois University at Carbondale.

Cualesquiera sugerencias de mejoras o posibles adiciones a este libro son bienvenidas; para ello puede ponerse en contacto directamente con los autores.

John H. Mathews
Mathematics Department
California State University
Fullerton, CA 92634
mathews@fullerton.edu

Kurtis D. Fink
Department of Mathematics
Northwest Missouri State University
Maryville, MO 64468
kfink@mail.nwmissouri.edu

Notas del Traductor

Sobre las traducciones de textos de matemáticas no cuelga la espada de Damocles del dicho “*traductor, traidor*”; en buena parte porque el fondo del asunto, las matemáticas, son un lenguaje universal por sí mismas. Eso no quiere decir que no existan dificultades con algunos términos y expresiones. Particularmente, como lector me hubiera gustado en ocasiones saber qué ha movido a un traductor a elegir una, y no otra, palabra castellana para traducir lo que en inglés tiene ya un vocablo comúnmente aceptado. Cuando me han surgido dudas en la presente traducción, me he atenido al “*Vocabulario Científico y Técnico*” de la Real Academia de Ciencias Exactas, Físicas y Naturales, publicado por Espasa-Calpe (Madrid, 1996); en particular, he empleado la palabra “cercha” como traducción del vocablo inglés “spline”.

Con objeto de que la lectura del texto resulte más fluida, he mantenido ciertos abusos de notación y lenguaje que ya se hacen en la edición inglesa original. Así, por ejemplo, escribiremos “la función $f(x) = \cos(x)$ ” y no “la función $f : \mathbb{R} \rightarrow \mathbb{R}$ definida por $f(x) = \cos(x)$ ”, o bien “el área limitada por la función $f(x)$ para $a \leq x \leq b$ ” y no “el área de la región limitada por la gráfica de la función $f : [a, b] \rightarrow \mathbb{R}$, el eje OX y las rectas verticales de ecuaciones $x = a$ y $x = b$ ”.

Puesto que los algoritmos desarrollados en el libro están escritos en el lenguaje de programación del paquete MATLAB, hemos decidido, de acuerdo

x PRÓLOGO

con la editorial, mantener la notación angloamericana de dicho paquete —ya prácticamente universal en todos los computadores— y usar un punto para separar la parte entera de la parte fraccionaria de los números reales. Así, escribiremos 2.4 y no 2,4, como sería si usáramos la coma decimal de nuestra notación tradicional. Sin embargo, y aun siendo conscientes del abuso que supone, llamaremos “coma decimal” al “punto” y escribiremos, por ejemplo, “aritmética de coma flotante”.

A lo largo de la traducción he ido corrigiendo, de acuerdo con los autores, pequeñas erratas e imprecisiones de la edición inglesa.

Debo agradecer a los autores su colaboración para que la traducción llegara a buen término y a mis compañeros de departamento Manuel D. Contreras Márquez y Antonio Fernández Carrión su ayuda en la detección de erratas y puesta a punto del manuscrito para que pudiera ser procesado con el programa **L^AT_EX**.

Cualesquiera sugerencias y críticas sobre la traducción de este libro son bienvenidas; para ello puede ponerse en contacto directamente con el traductor.

Pedro J. Paúl
Departamento de Matemática Aplicada II
Escuela Superior de Ingenieros
Camino de los Descubrimientos s/n
41092-Sevilla
piti@cica.es

Preliminares

Consideremos la función $f(x) = \cos(x)$, su derivada $f'(x) = -\operatorname{sen}(x)$, y una de sus primitivas $F(x) = \operatorname{sen}(x)$. Estas fórmulas fueron estudiadas en el curso de cálculo infinitesimal. La derivada se usa para determinar la pendiente $m = f'(x_0)$ de la curva $y = f(x)$ en el punto $(x_0, f(x_0))$ y la primitiva se usa para calcular el área limitada por la curva para $a \leq x \leq b$.

La pendiente en el punto $(\pi/2, 0)$ es $m = f'(\pi/2) = -1$ y puede usarse para hallar la recta tangente en dicho punto (véase la Figura 1.1(a)):

$$y_{\tan} = m \left(x - \frac{\pi}{2} \right) + 0 = f' \left(\frac{\pi}{2} \right) \left(x - \frac{\pi}{2} \right) = -x + \frac{\pi}{2}.$$

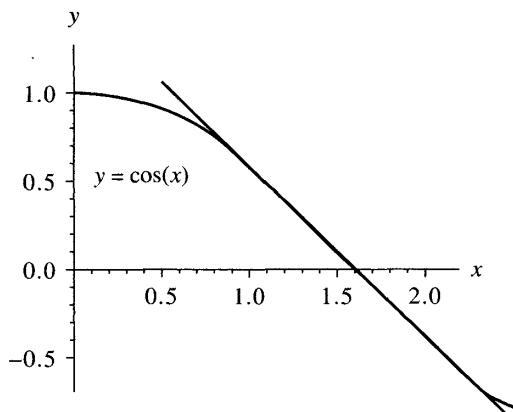


Figura 1.1 (a) La recta tangente a la curva $y = \cos(x)$ en el punto $(\pi/2, 0)$.

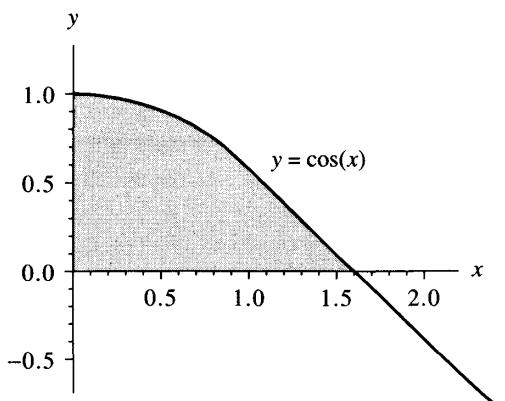


Figura 1.1 (b) El área limitada por la curva $y = \cos(x)$ sobre el intervalo $[0, \pi/2]$.

El área limitada por la curva para $0 \leq x \leq \pi/2$ se calcula mediante una integral (véase la Figura 1.1(b)):

$$\text{área} = \int_0^{\pi/2} \cos(x) dx = F\left(\frac{\pi}{2}\right) - F(0) = \operatorname{sen}\left(\frac{\pi}{2}\right) - 0 = 1.$$

Estos son algunos de los resultados del cálculo infinitesimal que emplearemos en el texto.

1.1 Un repaso al cálculo infinitesimal

Suponemos que la persona que lee este libro conoce la terminología, la notación y los resultados que se explican en un curso típico de cálculo infinitesimal impartido durante el primer año de los estudios universitarios. Esto incluye límites, continuidad, derivadas, integrales, sucesiones y series. A lo largo del libro usaremos los siguientes resultados.

Límites y Continuidad

Definición 1.1. Supongamos que $f(x)$ está definida en un conjunto S de números reales. Se dice que f tiene **límite L** en $x = x_0$, lo que se escribe

$$(1) \quad \lim_{x \rightarrow x_0} f(x) = L,$$

si, dado cualquier $\varepsilon > 0$, existe $\delta > 0$ tal que $|f(x) - L| < \varepsilon$ siempre que $x \in S$ y $0 < |x - x_0| < \delta$. Cuando usamos notación incremental $x = x_0 + h$, la relación (1) se escribe

$$(2) \quad \lim_{h \rightarrow 0} f(x_0 + h) = L.$$

Definición 1.2. Supongamos que $f(x)$ está definida en un conjunto S de números reales y sea $x_0 \in S$. Se dice que f es **continua en $x = x_0$** si

$$(3) \quad \lim_{x \rightarrow x_0} f(x) = f(x_0).$$

Se dice que f es continua en S si es continua en cada punto $x \in S$. Denotaremos por $C(S)$ el conjunto de todas las funciones f que son continuas en S . Cuando S sea un intervalo, digamos $[a, b]$, entonces usaremos la notación $C[a, b]$ ▲

Definición 1.3. Sea $\{x_n\}_{n=1}^{\infty}$ una sucesión de números reales. Se dice que la sucesión tiene límite L , lo que se escribe

$$(4) \quad \lim_{n \rightarrow \infty} x_n = L,$$

si, dado cualquier $\varepsilon > 0$, existe un número natural $N = N(\varepsilon)$ tal que si $n > N$ entonces $|x_n - L| < \varepsilon$. ▲

Cuando una sucesión tiene límite, se dice que es una **sucesión convergente**. Otra notación habitualmente utilizada es “ $x_n \rightarrow L$ cuando $n \rightarrow \infty$ ”. La igualdad (4) es equivalente a

$$(5) \quad \lim_{n \rightarrow \infty} (x_n - L) = 0,$$

así que podemos interpretar la sucesión $\{\varepsilon_n\}_{n=1}^{\infty} = \{x_n - L\}_{n=1}^{\infty}$ como una **sucesión de errores**. El siguiente teorema relaciona los conceptos de continuidad y sucesión convergente.

Teorema 1.1. Supongamos que $f(x)$ está definida en el conjunto S y que $x_0 \in S$. Entonces las siguientes afirmaciones son equivalentes:

- (a) La función f es continua en x_0 .
- (6) (b) Si $\{x_n\}_{n=1}^{\infty} \subset S$ y $\lim_{n \rightarrow \infty} x_n = x_0$, entonces $\lim_{n \rightarrow \infty} f(x_n) = f(x_0)$.

Teorema 1.2 (Teorema del valor intermedio o de Bolzano). Supongamos que $f \in C[a, b]$ y que L es cualquier número entre $f(a)$ y $f(b)$. Entonces existe un número c en (a, b) tal que $f(c) = L$.

Ejemplo 1.1. La función $f(x) = \cos(x - 1)$ es continua en $[0, 1]$ y la constante $L = 0.8 \in (f(0), f(1))$. La solución de $f(x) = 0.8$ en $[0, 1]$ es $c_1 = 0.356499$. De manera similar, $f(x)$ es continua en $[1, 2.5]$ y $L = 0.8$ está entre $f(1)$ y $f(2.5)$. La solución de $f(x) = 0.8$ en $[1, 2.5]$ es $c_2 = 1.643502$. Estos dos casos se muestran en la Figura 1.2. ■

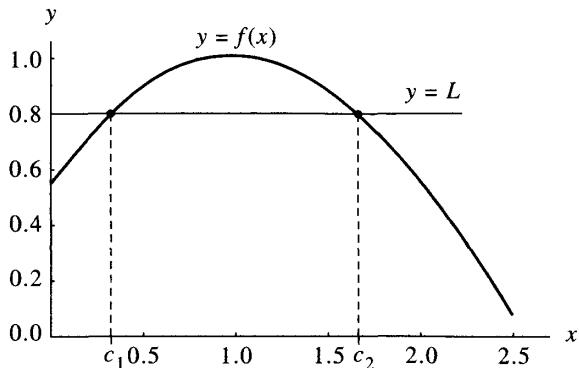


Figura 1.2 El teorema del valor intermedio aplicado a la función $f(x) = \cos(x - 1)$ en $[0, 1]$ y en $[1, 2.5]$.

Teorema 1.3 (Teorema de los valores extremos para una función continua o de Weierstrass). Supongamos que $f \in C[a, b]$. Entonces existen una cota inferior M_1 , una cota superior M_2 y dos números $x_1, x_2 \in [a, b]$ tales que

$$(7) \quad M_1 = f(x_1) \leq f(x) \leq f(x_2) = M_2 \quad \text{para cada } x \in [a, b].$$

A veces se expresa esto mismo escribiendo

$$(8) \quad M_1 = f(x_1) = \min_{a \leq x \leq b} \{f(x)\} \quad \text{y} \quad M_2 = f(x_2) = \max_{a \leq x \leq b} \{f(x)\}.$$

Funciones derivables

Definición 1.4. Supongamos que $f(x)$ está definida en un intervalo abierto que contiene a x_0 . Se dice que f es derivable en x_0 si existe el límite

$$(9) \quad \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

Cuando este límite existe, se denota por $f'(x_0)$ y se llama **derivada** de f en x_0 . Podemos usar notación incremental para expresar este límite de forma equivalente:

$$(10) \quad \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h} = f'(x_0).$$

Recordemos que el número $m = f'(x_0)$ es la pendiente de la recta tangente a la gráfica de la función $f(x)$ en el punto $(x_0, f(x_0))$.

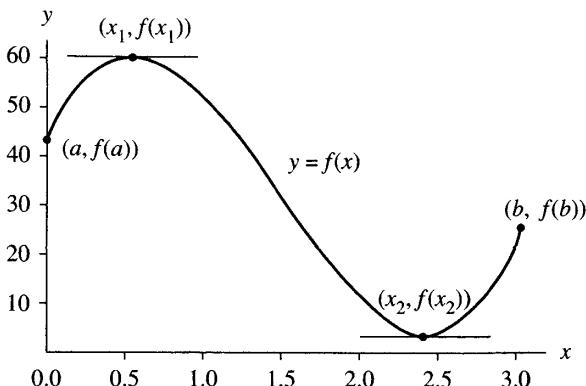


Figura 1.3 El teorema de los valores extremos aplicado a la función $f(x) = 35 + 59.5x - 66.5x^2 + 15x^3$ en el intervalo $[0, 3]$.

Una función que tiene derivada en cada punto de un conjunto S se dice que es **derivable** o **diferenciable** en S . Denotaremos por $C^n(S)$ el conjunto de todas las funciones f tales que f y sus primeras n derivadas son continuas en S . Cuando S sea un intervalo, digamos $[a, b]$, entonces usaremos la notación $C^n[a, b]$. Como ejemplo, consideremos la función $f(x) = x^{4/3}$ en el intervalo $[-1, 1]$. Está claro que $f(x)$ y $f'(x) = (4/3)x^{1/3}$ son continuas en $[-1, 1]$, mientras que $f''(x) = (4/9)x^{-2/3}$ no es continua en $x = 0$. ▲

Teorema 1.4. Si $f(x)$ es derivable en $x = x_0$, entonces $f(x)$ es continua en $x = x_0$.

Se sigue del Teorema 1.3 que una función f derivable en un intervalo cerrado $[a, b]$ alcanza sus valores extremos. Es más, puede deducirse que estos valores extremos los alcanza en los extremos del intervalo o en los puntos críticos (las soluciones de $f'(x) = 0$) que están en el intervalo abierto (a, b) .

Ejemplo 1.2. La función $f(x) = 15x^3 - 66.5x^2 + 59.5x + 35$ es derivable en $[0, 3]$. Las soluciones de $f'(x) = 45x^2 - 123x + 59.5 = 0$ son $x_1 = 0.54955$ y $x_2 = 2.40601$. Los valores máximo y mínimo de f en $[0, 3]$ (véase la Figura 1.3) son:

$$\min\{f(0), f(3), f(x_1), f(x_2)\} = \min\{35, 20, 50.10438, 2.11850\} = 2.11850$$

y

$$\max\{f(0), f(3), f(x_1), f(x_2)\} = \max\{35, 20, 50.10438, 2.11850\} = 50.10438. \blacksquare$$

Teorema 1.5 (Teorema de Rolle). Supongamos que $f \in C[a, b]$ y que $f'(x)$ existe para todo $x \in (a, b)$. Si $f(a) = f(b) = 0$, entonces existe un número c en (a, b) tal que $f'(c) = 0$.

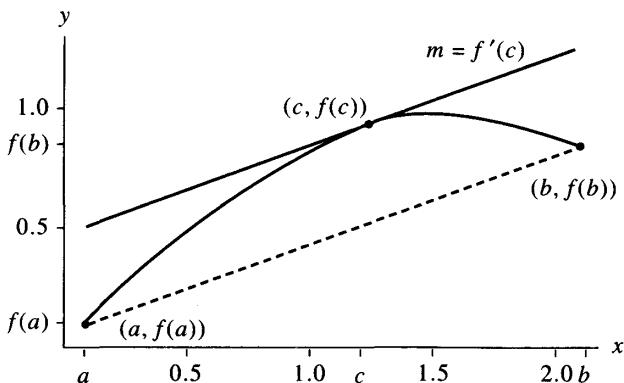


Figura 1.4 El teorema del valor medio aplicado a la función $f(x) = \sin(x)$ en el intervalo $[0.1, 2.1]$.

Teorema 1.6 (Teorema del valor medio o de Lagrange). Supongamos que $f \in C[a, b]$ y que $f'(x)$ existe para todo $x \in (a, b)$. Entonces existe un número c en (a, b) tal que

$$(11) \quad f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Geométricamente hablando, el teorema del valor medio de Lagrange nos dice que hay al menos un número $c \in (a, b)$ tal que la pendiente de la recta tangente a la curva $y = f(x)$ en el punto $(c, f(c))$ es igual a la pendiente de la recta secante que pasa por los puntos $(a, f(a))$ y $(b, f(b))$.

Ejemplo 1.3. La función $f(x) = \sin(x)$ es continua en el intervalo cerrado $[0.1, 2.1]$ y derivable en el intervalo abierto $(0.1, 2.1)$. Entonces, por el teorema del valor medio, existe un número c tal que

$$f'(c) = \frac{f(2.1) - f(0.1)}{2.1 - 0.1} = \frac{0.863209 - 0.099833}{2.1 - 0.1} = 0.381688.$$

La solución de $f'(c) = \cos(c) = 0.381688$ en el intervalo $(0.1, 2.1)$ es $c = 1.179174$. Las gráficas de $f(x)$, de la recta secante $y = 0.381688x + 0.099833$ y de la recta tangente $y = 0.381688x + 0.474215$ se muestran en la Figura 1.4. ■

Teorema 1.7 (Teorema de Rolle generalizado). Supongamos que la función $f \in C[a, b]$, que sus derivadas $f'(x), f''(x), \dots, f^{(n)}(x)$ existen en (a, b) y que $x_0, x_1, \dots, x_n \in [a, b]$. Si $f(x_j) = 0$ para $j = 0, 1, \dots, n$, entonces existe un número c en (a, b) tal que $f^{(n)}(c) = 0$.

Integrales

Teorema 1.8 (Primer teorema fundamental o regla de Barrow). Si f es continua en $[a, b]$ y F es una primitiva cualquiera de f en $[a, b]$ (es decir, $F'(x) = f(x)$), entonces

$$(12) \quad \int_a^b f(x) dx = F(b) - F(a).$$

Teorema 1.9 (Segundo teorema fundamental). Si f es continua en $[a, b]$ y $x \in (a, b)$, entonces

$$(13) \quad \frac{d}{dx} \int_a^x f(t) dt = f(x).$$

Ejemplo 1.4. La función $f(x) = \cos(x)$ verifica las hipótesis del Teorema 1.9 en el intervalo $[0, \pi/2]$, así que por la regla de la cadena se tiene

$$\frac{d}{dx} \int_0^{x^2} \cos(t) dt = \cos(x^2)(x^2)' = 2x \cos(x^2).$$

Teorema 1.10 (Teorema del valor medio para integrales). Supongamos que $f \in C[a, b]$. Entonces existe un número c en (a, b) tal que

$$\frac{1}{b-a} \int_a^b f(x) dx = f(c).$$

El valor $f(c)$ es el valor medio de f en el intervalo $[a, b]$.

Ejemplo 1.5. La función $f(x) = \operatorname{sen}(x) + \frac{1}{3} \operatorname{sen}(3x)$ verifica las hipótesis del Teorema 1.10 en el intervalo $[0, 2.5]$. Una primitiva de $f(x)$ es $F(x) = -\cos(x) - \frac{1}{9} \cos(3x)$. El valor medio de la función $f(x)$ en el intervalo $[0, 2.5]$ es:

$$\begin{aligned} \frac{1}{2.5-0} \int_0^{2.5} f(x) dx &= \frac{F(2.5) - F(0)}{2.5} = \frac{0.762629 - (-1.111111)}{2.5} \\ &= \frac{1.873740}{2.5} = 0.749496. \end{aligned}$$

Hay tres soluciones de la ecuación $f(c) = 0.749496$ en el intervalo $[0, 2.5]$, a saber: $c_1 = 0.440566$, $c_2 = 1.268010$ y $c_3 = 1.873583$. El área del rectángulo de base $b-a = 2.5$ y altura $f(c_j) = 0.749496$ es $f(c_j)(b-a) = 1.873740$. El área del rectángulo vale lo mismo que la integral de $f(x)$ en el intervalo $[0, 2.5]$ que, a su vez, es el área de la región limitada por la curva $y = f(x)$ sobre dicho intervalo; esto puede verse en la Figura 1.5.

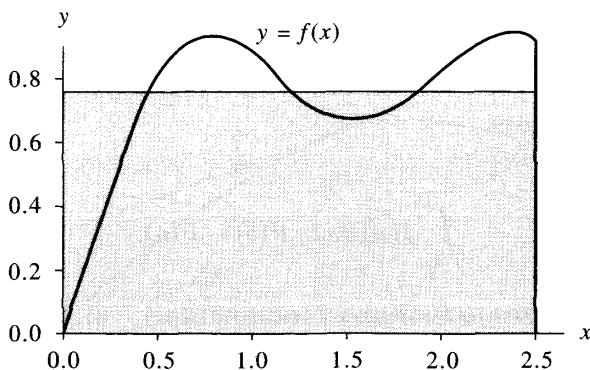


Figura 1.5 El teorema del valor medio para integrales aplicado a $f(x) = \sin(x) + \frac{1}{3} \sin(3x)$ en el intervalo $[0, 2.5]$.

Teorema 1.11 (Teorema del valor medio ponderado para integrales). Supongamos que $f, g \in C[a, b]$ y que $g(x) \geq 0$ para todo $x \in [a, b]$. Entonces existe un número c en (a, b) tal que

$$(14) \quad \int_a^b f(x)g(x) dx = f(c) \int_a^b g(x) dx.$$

Ejemplo 1.6. Las funciones $f(x) = \sin(x)$ y $g(x) = x^2$ verifican las hipótesis del Teorema 1.11 en el intervalo $[0, \pi/2]$. En consecuencia, existe un número c tal que

$$\sin(c) = \frac{\int_0^{\pi/2} x^2 \sin(x) dx}{\int_0^{\pi/2} x^2 dx} = \frac{1.14159}{1.29193} = 0.883631$$

con lo que $c = \sin^{-1}(0.883631) = 1.08356$.

Series

Definición 1.5. Dada una sucesión $\{a_n\}_{n=1}^{\infty}$ denotaremos por $\sum_{n=1}^{\infty} a_n$ la serie de término general a_n . La suma parcial n -ésima de la serie se define como $S_n = \sum_{k=1}^n a_k$ y se dice que la serie **converge** si la sucesión $\{S_n\}_{n=1}^{\infty}$ converge a un límite S llamado **suma** de la serie, o sea,

$$(15) \quad \lim_{n \rightarrow \infty} S_n = \lim_{n \rightarrow \infty} \sum_{k=1}^n a_k = S.$$

Si una serie no converge, entonces se dice que **diverge**. ▲

Ejemplo 1.7. Consideremos la sucesión $\{a_n\}_{n=1}^{\infty} = \left\{ \frac{1}{n(n+1)} \right\}_{n=1}^{\infty}$. Entonces la suma parcial n -ésima es

$$S_n = \sum_{k=1}^n \frac{1}{k(k+1)} = \sum_{k=1}^n \left(\frac{1}{k} - \frac{1}{k+1} \right) = 1 - \frac{1}{n+1}.$$

Por tanto, la suma de la serie es

$$S = \lim_{n \rightarrow \infty} S_n = \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n+1} \right) = 1. \quad \blacksquare$$

Teorema 1.12 (Teorema de Taylor). Supongamos que $f \in C^{n+1}[a, b]$ y sea $x_0 \in [a, b]$. Entonces, para cada $x \in (a, b)$, existe un número $c = c(x)$ (el valor de c depende de x) que está entre x_0 y x y verifica

$$(16) \quad f(x) = P_n(x) + R_n(x),$$

donde

$$(17) \quad P_n(x) = \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k$$

se llama *polinomio de Taylor de grado n de f alrededor de x_0* y

$$(18) \quad R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!} (x - x_0)^{n+1}.$$

Ejemplo 1.8. La función $f(x) = \operatorname{sen}(x)$ verifica las hipótesis del Teorema 1.12. El polinomio de Taylor $P_n(x)$ de grado $n = 9$ de f alrededor de $x_0 = 0$ se obtiene calculando las siguientes derivadas en $x = 0$ y sustituyendo sus valores en la fórmula (17):

$$\begin{aligned} f(x) &= \operatorname{sen}(x), & f(0) &= 0, \\ f'(x) &= \cos(x), & f'(0) &= 1, \\ f''(x) &= -\operatorname{sen}(x), & f''(0) &= 0, \\ f^{(3)}(x) &= -\cos(x), & f^{(3)}(0) &= -1, \\ &\vdots &&\vdots \\ f^{(9)}(x) &= \cos(x), & f^{(9)}(0) &= 1, \end{aligned}$$

$$P_9(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!}.$$

Las gráficas de f y P_9 en el intervalo $[0, 2\pi]$ se muestran en la Figura 1.6. ■

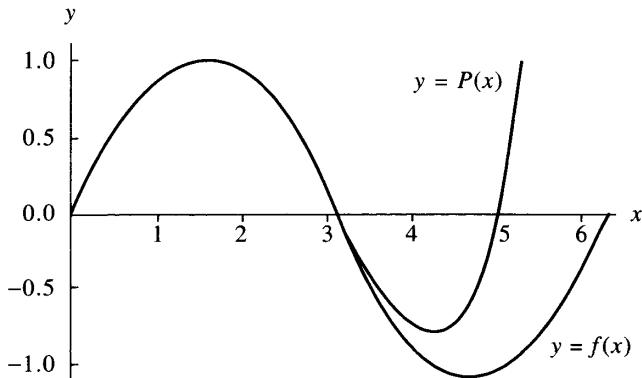


Figura 1.6 Las gráficas de $f(x) = \sin(x)$ y de su polinomio de Taylor

$$P(x) = P_9(x) = x - x^3/3! + x^5/5! - x^7/7! + x^9/9!.$$

Corolario 1.1. Si $P_n(x)$ es el polinomio de Taylor de grado n dado en el Teorema 1.12, entonces

$$(19) \quad P_n^{(k)}(x_0) = f^{(k)}(x_0) \quad \text{para } k = 0, 1, \dots, n.$$

Evaluación de un Polinomio

Supongamos que escribimos un polinomio $P(x)$ de grado n en la forma

$$(20) \quad P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0.$$

El **método de Horner** o **regla de Ruffini** o **división sintética** es una técnica para evaluar polinomios que puede ser vista como una colección de multiplicaciones encajadas. Por ejemplo, un polinomio de quinto grado puede escribirse como una colección de cinco multiplicaciones encajadas

$$P_5(x) = (((((a_5 x + a_4) x + a_3) x + a_2) x + a_1) x + a_0).$$

Teorema 1.13 (Método de Horner o regla de Ruffini para evaluar polinomios). Sea $P(x)$ el polinomio dado por la expresión (20) y sea $x = c$ un número para el que deseamos evaluar $P(c)$.

Pongamos $b_n = a_n$ y calculemos

$$(21) \quad b_k = a_k + c b_{k+1} \quad \text{para } k = n-1, n-2, \dots, 1, 0;$$

Tabla 1.1 Coeficientes b_k para el método de Horner.

x^k	Comparación de (20) y (24)	Obtención de b_k
x^n	$a_n = b_n$	$b_n = a_n$
x^{n-1}	$a_{n-1} = b_{n-1} - cb_n$	$b_{n-1} = a_{n-1} + cb_n$
\vdots	\vdots	\vdots
x^k	$a_k = b_k - cb_{k+1}$	$b_k = a_k + cb_{k+1}$
\vdots	\vdots	\vdots
x^0	$a_0 = b_0 - cb_1$	$b_0 = a_0 + cb_1$

entonces $b_0 = P(c)$. Es más, si definimos

$$(22) \quad Q_0(x) = b_n x^{n-1} + b_{n-1} x^{n-2} + \cdots + b_3 x^2 + b_2 x + b_1$$

y $R_0 = b_0$, entonces

$$(23) \quad P(x) = (x - c)Q_0(x) + R_0,$$

o sea, $Q_0(x)$ es el polinomio cociente de grado $n - 1$ y $R_0 = b_0 = P(c)$ es el resto de la división de $P(x)$ entre $x - c$.

Demostración. En la igualdad (23) sustituimos $Q_0(x)$ por el miembro derecho de la fórmula (22) y R_0 por b_0 para obtener

$$(24) \quad \begin{aligned} P(x) &= (x - c)(b_n x^{n-1} + b_{n-1} x^{n-2} + \cdots + b_3 x^2 + b_2 x + b_1) + b_0 \\ &= b_n x^n + (b_{n-1} - cb_n) x^{n-1} + \cdots + (b_2 - cb_3) x^2 \\ &\quad + (b_1 - cb_2) x + (b_0 - cb_1). \end{aligned}$$

Los números b_k quedan entonces determinados por la fórmula (21) sin más que comparar los coeficientes de x^k en la fórmula (20) y en la igualdad (24), tal y como se muestra en la Tabla 1.1.

La igualdad $P(c) = b_0$ se obtiene fácilmente sustituyendo $x = c$ en la igualdad (23) y usando que $R_0 = b_0$:

$$(25) \quad P(c) = (c - c)Q_0(c) + R_0 = b_0.$$

La fórmula de recursión para b_k dada en (21) es fácil de construir en un computador. Un algoritmo simple es

Tabla 1.2 Tabla de Horner para el proceso de división sintética.

Dato	a_n	a_{n-1}	a_{n-2}	\cdots	a_k	\cdots	a_2	a_1	a_0
c	cb_n	cb_{n-1}	\cdots	cb_{k+1}	\cdots	cb_3	cb_2	cb_1	
	b_n	b_{n-1}	b_{n-2}	\cdots	b_k	\cdots	b_2	b_1	$b_0 = P(c)$
									Resultado

$b(n) = a(n);$
 para $k = n - 1: -1: 0$
 $b(k) = a(k) + c * b(k + 1);$
 fin

Cuando el método de Horner se ejecuta a mano, es más fácil escribir los coeficientes de $P(x)$ en una fila y hacer la suma $b_k = a_k + cb_{k+1}$ en una columna debajo de a_k . Este procedimiento se ilustra en la Tabla 1.2.

Ejemplo 1.9. Uso de la división sintética (método de Horner) para hallar $P(3)$ siendo P el polinomio

$$P(x) = x^5 - 6x^4 + 8x^3 + 8x^2 + 4x - 40.$$

Dato	a_5	a_4	a_3	a_2	a_1	a_0
$c = 3$	1	-6	8	8	4	-40
	1	-3	-1	5	19	17 = $P(3) = b_0$
	b_5	b_4	b_3	b_2	b_1	Resultado

Por tanto, $P(3) = 17$.

Ejercicios

1. (a) Halle $L = \lim_{n \rightarrow \infty} (4n + 1)/(2n + 1)$. Después determine $\{\varepsilon_n\} = \{L - x_n\}$ y halle $\lim_{n \rightarrow \infty} \varepsilon_n$.
 (b) Halle $L = \lim_{n \rightarrow \infty} (2n^2 + 6n - 1)/(4n^2 + 2n + 1)$. Después determine $\{\varepsilon_n\} = \{L - x_n\}$ y halle $\lim_{n \rightarrow \infty} \varepsilon_n$.
2. Sea $\{x_n\}_{n=1}^{\infty}$ una sucesión tal que $\lim_{n \rightarrow \infty} x_n = 2$.
 - Halle $\lim_{n \rightarrow \infty} \sin(x_n)$.
 - Halle $\lim_{n \rightarrow \infty} \ln(x_n^2)$.

funciones alrededor del punto x_0 dado.

- (a) $f(x) = \sqrt{x}$, $x_0 = 1$.
- (b) $f(x) = x^5 + 4x^2 + 3x + 1$, $x_0 = 0$.
- (c) $f(x) = \cos(x)$, $x_0 = 0$.

13. Sean $f(x) = \operatorname{sen}(x)$ y $P(x) = x - x^3/3! + x^5/5! - x^7/7! + x^9/9!$. Pruebe que $P^{(k)}(0) = f^{(k)}(0)$ para $k = 1, 2, \dots, 9$.
14. Use división sintética (el método de Horner) para hallar $P(c)$ en los siguientes casos.
 - (a) $P(x) = x^4 + x^3 - 13x^2 - x - 12$, $c = 3$.
 - (b) $P(x) = 2x^7 + x^6 + x^5 - 2x^4 - x + 23$, $c = -1$.
15. Halle el área media de todos los círculos centrados en el origen cuyo radio está comprendido entre 1 y 3.
16. Supongamos que un polinomio $P(x)$ tiene n raíces reales en el intervalo $[a, b]$. Pruebe que $P^{(n-1)}(x)$ tiene al menos una raíz real en dicho intervalo.
17. Supongamos que f, f' y f'' están definidas en un intervalo $[a, b]$, que $f(a) = f(b) = 0$ y que $f'(c) > 0$ para todo $c \in (a, b)$. Pruebe que existe un número $d \in (a, b)$ tal que $f''(d) < 0$.

Números binarios

Los seres humanos hacemos los cálculos aritméticos usando el sistema numérico decimal (o en base 10). La mayoría de los computadores hacen los cálculos aritméticos usando el sistema numérico binario (o en base 2). Esto puede parecer que no es así porque la comunicación con el computador (datos/resultados) se hace con números en base 10. Ello no quiere decir que el computador use base 10; de hecho, lo que ocurre es que convierte los datos en números en base 2 (o, quizás, en base 16), lleva a cabo los cálculos aritméticos en base 2 y, finalmente, traduce la respuesta a base 10 antes de mostrar el resultado. Hace falta experimentar un poco para verificar este hecho. Un computador con nueve cifras decimales de exactitud dio como respuesta

$$(1) \quad \sum_{k=1}^{100\,000} 0.1 = 9\,999.99447.$$

Aquí se trataba de sumar 100 000 veces el número $\frac{1}{10}$. La respuesta exacta es 10 000 y uno de nuestros objetivos es el comprender la razón de este cálculo aparentemente erróneo. Al final de esta sección, mostraremos que siempre perdemos algo cuando el computador traduce la fracción decimal $\frac{1}{10}$ al sistema binario.

Números binarios

Usamos los números en base 10 para la mayor parte de nuestro quehacer matemático. Por ejemplo, el número 1563 lo podemos expresar de **forma desarrollada** como

$$1563 = (1 \times 10^3) + (5 \times 10^2) + (6 \times 10^1) + (3 \times 10^0).$$

En general, sea N un número natural; entonces existen K cifras a_0, a_1, \dots, a_K tomadas del conjunto $\{0, 1, \dots, 8, 9\}$ tales que N admite el siguiente desarrollo en base 10

$$N = (a_K \times 10^K) + (a_{K-1} \times 10^{K-1}) + \cdots + (a_1 \times 10^1) + (a_0 \times 10^0).$$

Esto nos permite expresar N en notación decimal como

$$(2) \quad N = a_K a_{K-1} \cdots a_2 a_1 a_0_{\text{diez}} \quad (\text{decimal}).$$

Si se sobreentiende que 10 es la base, entonces (2) se escribe como

$$N = a_K a_{K-1} \cdots a_2 a_1 a_0.$$

Por ejemplo, sobreentendemos que $1563 = 1563_{\text{diez}}$.

Usando potencias de 2, el número 1563 se escribe como

$$(3) \quad \begin{aligned} 1563 = & (1 \times 2^{10}) + (1 \times 2^9) + (0 \times 2^8) + (0 \times 2^7) + (0 \times 2^6) \\ & + (0 \times 2^5) + (1 \times 2^4) + (1 \times 2^3) + (0 \times 2^2) + (1 \times 2^1) \\ & + (1 \times 2^0). \end{aligned}$$

Lo que puede comprobarse haciendo las cuentas:

$$1563 = 1024 + 512 + 16 + 8 + 2 + 1.$$

En general, sea N un número natural; entonces existen J cifras b_0, b_1, \dots, b_J , cada una de las cuales es un 0 o un 1, tales que N admite el siguiente desarrollo en base 2

$$(4) \quad N = (b_J \times 2^J) + (b_{J-1} \times 2^{J-1}) + \cdots + (b_1 \times 2^1) + (b_0 \times 2^0).$$

Esto nos permite expresar N en notación binaria como

$$(5) \quad N = b_J b_{J-1} \cdots b_2 b_1 b_0_{\text{dos}} \quad (\text{binaria}).$$

Usando la notación de (5) y la igualdad (3) obtenemos

$$1563 = 11\,000\,011\,011_{\text{dos}}.$$

Observaciones. Usaremos siempre la palabra “dos” como subíndice al final de un número binario. Esto nos permitirá distinguir los números binarios de los números de uso ordinario en base 10. Así 111 significa ciento once mientras que 111_{dos} significa siete.

A menudo se da el caso de que la representación binaria de un número requiere más cifras que su representación decimal. Eso se debe a que las potencias de 2 crecen mucho más lentamente que las potencias de 10.

Un algoritmo eficiente para hallar la representación en base 2 de un número natural N puede deducirse de la igualdad (4): Si dividimos ambos miembros de (4) entre 2, obtenemos

$$(6) \quad \frac{N}{2} = (b_J \times 2^{J-1}) + (b_{J-1} \times 2^{J-2}) + \cdots + (b_1 \times 2^0) + \frac{b_0}{2};$$

así que la cifra b_0 es el resto de la división de N entre 2. A continuación determinamos b_1 ; para ello escribimos (6) como $N/2 = Q_0 + b_0/2$, con lo que

$$(7) \quad Q_0 = (b_J \times 2^{J-1}) + (b_{J-1} \times 2^{J-2}) + \cdots + (b_2 \times 2^1) + (b_1 \times 2^0).$$

Ahora dividimos ambos miembros de (7) entre 2 y obtenemos

$$\frac{Q_0}{2} = (b_J \times 2^{J-2}) + (b_{J-1} \times 2^{J-3}) + \cdots + (b_2 \times 2^0) + \frac{b_1}{2};$$

luego la cifra b_1 es el resto de la división de Q_0 entre 2. Continuando este proceso generamos dos sucesiones $\{Q_k\}$ y $\{b_k\}$ de cocientes y restos, respectivamente. El proceso termina cuando encontramos un número natural J tal que $Q_J = 0$. Estas sucesiones obedecen a las siguientes fórmulas:

$$(8) \quad \begin{aligned} N &= 2Q_0 + b_0 \\ Q_0 &= 2Q_1 + b_1 \\ &\vdots \\ Q_{J-2} &= 2Q_{J-1} + b_{J-1} \\ Q_{J-1} &= 2Q_J + b_J \quad (Q_J = 0). \end{aligned}$$

Ejemplo 1.10. ¿Cómo se prueba que $1563 = 11\ 000\ 011\ 011_{\text{dos}}$?

Empezamos con $N = 1563$ y construimos los cocientes y los restos de acuerdo

con las fórmulas dadas en (8):

$$\begin{aligned}
 1563 &= 2 \times 781 + 1, & b_0 &= 1 \\
 781 &= 2 \times 390 + 1, & b_1 &= 1 \\
 390 &= 2 \times 195 + 0, & b_2 &= 0 \\
 195 &= 2 \times 97 + 1, & b_3 &= 1 \\
 97 &= 2 \times 48 + 1, & b_4 &= 1 \\
 48 &= 2 \times 24 + 0, & b_5 &= 0 \\
 24 &= 2 \times 12 + 0, & b_6 &= 0 \\
 12 &= 2 \times 6 + 0, & b_7 &= 0 \\
 6 &= 2 \times 3 + 0, & b_8 &= 0 \\
 3 &= 2 \times 1 + 1, & b_9 &= 1 \\
 1 &= 2 \times 0 + 1, & b_{10} &= 1.
 \end{aligned}$$

Así que la representación binaria de 1563 es

$$1563 = b_{10}b_9b_8 \cdots b_2b_1b_0_{\text{dos}} = 11\ 000\ 011\ 011_{\text{dos}}.$$

Sucesiones y series

Cuando los números racionales se expresan en forma decimal, se da a menudo el caso de que hace falta una cantidad infinita de cifras. Un ejemplo familiar es

$$(9) \quad \frac{1}{3} = 0.\overline{3}.$$

Aquí el símbolo $\overline{3}$ significa que la cifra 3 se repite sin fin para formar un decimal periódico con infinitas cifras todas iguales a 3 (se sobreentiende que estamos trabajando en base 10). Es más, la intención matemática es que la notación de (9) sea una abreviatura de la serie

$$\begin{aligned}
 S &= (3 \times 10^{-1}) + (3 \times 10^{-2}) + \cdots + (3 \times 10^{-n}) + \cdots \\
 (10) \quad &= \sum_{k=1}^{\infty} 3(10)^{-k} = \frac{1}{3}.
 \end{aligned}$$

Si sólo usamos una cantidad finita de cifras, entonces lo que obtenemos es una aproximación a $1/3$. Por ejemplo, $1/3 \approx 0.333 = 333/1000$. El error de esta aproximación es $1/3000$ ya que, usando (10), podemos comprobar que $1/3 = 0.333 + 1/3000$.

Es importante comprender el desarrollo escrito en (10). Una primera aproximación ingenua sería multiplicar ambos miembros por 10 y restar:

$$\begin{aligned}
 10S &= 3 + (3 \times 10^{-1}) + (3 \times 10^{-2}) + \cdots + (3 \times 10^{-n}) + \cdots \\
 -S &= - (3 \times 10^{-1}) - (3 \times 10^{-2}) - \cdots - (3 \times 10^{-n}) - \cdots \\
 \hline
 9S &= 3 + (0 \times 10^{-1}) + (0 \times 10^{-2}) + \cdots + (0 \times 10^{-n}) + \cdots
 \end{aligned}$$

Por tanto, $S = 3/9 = 1/3$. Los teoremas que justifican este procedimiento para restar dos series infinitas pueden encontrarse en la mayoría de los libros de cálculo infinitesimal. A continuación repasaremos algunos de los conceptos involucrados; para llenar las lagunas se puede acudir a un texto estándar de cálculo infinitesimal.

Definición 1.6 (La serie geométrica). La serie

$$(11) \quad \sum_{n=0}^{\infty} cr^n = c + cr + cr^2 + \cdots + cr^n + \cdots,$$

donde $c \neq 0$ y $r \neq 0$, se llama *serie geométrica* de razón r .

Teorema 1.14 (Convergencia de la serie geométrica). La serie geométrica tiene las siguientes propiedades:

$$(12) \quad \text{Si } |r| < 1, \text{ entonces } \sum_{n=0}^{\infty} cr^n = \frac{c}{1-r}.$$

$$(13) \quad \text{Si } |r| \geq 1, \text{ entonces la serie diverge.}$$

Demostración. La fórmula para sumar una progresión geométrica finita es

$$(14) \quad S_n = c + cr + cr^2 + \cdots + cr^n = \frac{c(1 - r^{n+1})}{1 - r} \quad \text{para } r \neq 1.$$

Para probar (12) observemos que

$$(15) \quad |r| < 1 \quad \text{implica que } \lim_{n \rightarrow \infty} r^{n+1} = 0.$$

Tomando límites en (14) y (15) cuando $n \rightarrow \infty$, obtenemos

$$\lim_{n \rightarrow \infty} S_n = \frac{c}{1 - r} \left(1 - \lim_{n \rightarrow \infty} r^{n+1} \right) = \frac{c}{1 - r}.$$

La expresión (15) de la Sección 1.1 nos dice que este límite prueba (12).

Cuando $|r| \geq 1$ y $r \neq 1$, la sucesión $\{r^{n+1}\}$ no converge y, por tanto, la sucesión $\{S_n\}$ dada en (14) no tiene límite; eso prueba (13) para $r \neq 1$.

Finalmente, si $r = 1$, entonces $S_n = (n+1)c$ y la sucesión $\{S_n\}$ no converge.

La fórmula que aparece en la expresión (12) del Teorema 1.14 es también una forma eficiente de convertir una expresión decimal periódica en una fracción.

Ejemplo 1.11.

$$\begin{aligned} 0.\overline{3} &= \sum_{k=1}^{\infty} 3(10)^{-k} = -3 + \sum_{k=0}^{\infty} 3(10)^{-k} \\ &= -3 + \frac{3}{1 - \frac{1}{10}} = -3 + \frac{10}{3} = \frac{1}{3}. \end{aligned}$$

Fracciones binarias

Las fracciones binarias (base 2) pueden expresarse como sumas en las que aparecen potencias negativas de 2. Si R es un número real tal que $0 < R < 1$, entonces existe una sucesión de cifras $d_1, d_2, \dots, d_n, \dots$, todas ellas en $\{0, 1\}$, tales que

$$(16) \quad R = (d_1 \times 2^{-1}) + (d_2 \times 2^{-2}) + \cdots + (d_n \times 2^{-n}) + \cdots.$$

La expresión del miembro derecho de la igualdad (16) suele expresarse en notación fraccionaria binaria como

$$(17) \quad R = 0.d_1d_2 \cdots d_n \cdots_{\text{dos}}.$$

Hay muchos números reales cuya representación binaria requiere una cantidad infinita de cifras iguales a 1. La fracción $7/10$ puede expresarse como 0.7 en base 10, pero su representación en base 2 requiere una cantidad infinita de unos:

$$(18) \quad \frac{7}{10} = 0.\overline{10110}_{\text{dos}}.$$

Esta fracción binaria es periódica: el grupo de cuatro cifras 0110 se repite sin fin.

Podemos desarrollar ahora un algoritmo eficiente para hallar representaciones en base 2. Si multiplicamos por 2 ambos miembros de la expresión (16), el resultado es

$$(19) \quad 2R = d_1 + ((d_2 \times 2^{-1}) + \cdots + (d_n \times 2^{-n+1}) + \cdots).$$

La cantidad entre paréntesis en el miembro derecho de (19) es un número positivo y menor que 1; por tanto, d_1 es la parte entera de $2R$, lo que denotamos por $d_1 = \text{ent}(2R)$. Para continuar el proceso, tomamos la parte fraccionaria de la igualdad (19) y escribimos

$$(20) \quad F_1 = \text{frac}(2R) = (d_2 \times 2^{-1}) + \cdots + (d_n \times 2^{-n+1}) + \cdots,$$

donde $\text{frac}(2R)$ denota la parte fraccionaria del número real $2R$. Multiplicando por 2 ambos miembros de (20), el resultado es

$$(21) \quad 2F_1 = d_2 + ((d_3 \times 2^{-1}) + \cdots + (d_n \times 2^{-n+2}) + \cdots).$$

Tomando la parte entera en esta igualdad obtenemos $d_2 = \text{ent}(2F_1)$.

El proceso continúa, posiblemente sin fin (si R tiene una representación en base 2 que no es finita ni periódica), y genera de forma recurrente dos sucesiones $\{d_k\}$ y $\{F_k\}$:

$$(22) \quad \begin{aligned} d_k &= \text{ent}(2F_{k-1}), \\ F_k &= \text{frac}(2F_{k-1}), \end{aligned}$$

donde $d_1 = \text{ent}(2R)$ y $F_1 = \text{frac}(2R)$. La representación binaria de R viene dada entonces por la serie convergente

$$R = \sum_{j=1}^{\infty} d_j (2)^{-j}$$

(que es una subserie de la serie geométrica de razón $1/2$).

Ejemplo 1.12. La representación binaria de $7/10$ dada en (18) fue encontrada usando las fórmulas de (22): Sea $R = 7/10 = 0.7$, entonces

$$\begin{array}{lll} 2R = 1.4 & d_1 = \text{ent}(1.4) = 1 & F_1 = \text{frac}(1.4) = 0.4 \\ 2F_1 = 0.8 & d_2 = \text{ent}(0.8) = 0 & F_2 = \text{frac}(0.8) = 0.8 \\ 2F_2 = 1.6 & d_3 = \text{ent}(1.6) = 1 & F_3 = \text{frac}(1.6) = 0.6 \\ 2F_3 = 1.2 & d_4 = \text{ent}(1.2) = 1 & F_4 = \text{frac}(1.2) = 0.2 \\ 2F_4 = 0.4 & d_5 = \text{ent}(0.4) = 0 & F_5 = \text{frac}(0.4) = 0.4 \\ 2F_5 = 0.8 & d_6 = \text{ent}(0.8) = 0 & F_6 = \text{frac}(0.8) = 0.8 \\ 2F_6 = 1.6 & d_7 = \text{ent}(1.6) = 1 & F_7 = \text{frac}(1.6) = 0.6. \end{array}$$

Nótese que $2F_2 = 1.6 = 2F_6$, luego los patrones $d_k = d_{k+4}$ y $F_k = F_{k+4}$ se darán para $k = 2, 3, 4, \dots$. En consecuencia, $7/10 = 0.\overline{10110}_{\text{dos}}$.

Podemos usar la serie geométrica para hallar el número racional en base 10 al que representa un número binario dado.

Ejemplo 1.13. Veamos cómo hallar el número racional en base 10 representado por el número binario $0.\overline{01}_{\text{dos}}$. Escribiéndolo en forma desarrollada:

$$\begin{aligned} 0.\overline{01}_{\text{dos}} &= (0 \times 2^{-1}) + (1 \times 2^{-2}) + (0 \times 2^{-3}) + (1 \times 2^{-4}) + \dots \\ &= \sum_{k=1}^{\infty} (2^{-2})^k = -1 + \sum_{k=0}^{\infty} (2^{-2})^k \\ &= -1 + \frac{1}{1 - \frac{1}{4}} = -1 + \frac{4}{3} = \frac{1}{3}. \end{aligned}$$

Desplazamiento binario

Si queremos hallar un número racional conocida su representación binaria periódica, entonces resulta útil hacer un desplazamiento adecuado de sus cifras binarias. Por ejemplo, supongamos que S viene dado por

$$(23) \quad S = 0.00000\overline{11000}_{\text{dos}}.$$

Multiplicando ambos miembros de (23) por 2^5 desplazamos la coma binaria cinco lugares hacia la derecha y escribimos $32S$ como

$$(24) \quad 32S = 0.\overline{11000}_{\text{dos}}.$$

De forma parecida, multiplicando por 2^{10} ambos miembros de (23) desplazamos la coma binaria diez lugares hacia la derecha y escribimos $1024S$ como

$$(25) \quad 1024S = 11\,000.\overline{11000}_{\text{dos}}.$$

Restando las igualdades (24) y (25) obtenemos $992S = 11\,000_{\text{dos}}$ o, puesto que $11\,000_{\text{dos}} = 24$, también $992S = 24$, luego $S = 8/33$.

Notación científica

Una forma estándar de representar un número real, llamada **notación científica**, consiste en desplazar la coma decimal a la vez que proporcionamos una potencia de 10 adecuada. Por ejemplo

$$\begin{aligned} 0.0000747 &= 7.47 \times 10^{-5}, \\ 31.4159265 &= 3.14159265 \times 10, \\ 9\,700\,000\,000 &= 9.7 \times 10^9. \end{aligned}$$

En química, una constante muy importante es la conocida como número de Avogadro: 6.02252×10^{23} , que es el número de átomos de un elemento que hay en una masa de tantos gramos de ese elemento como indique su peso atómico. En ciencias de la computación, se define $1K = 1.024 \times 10^3$.

Números del computador

Los computadores usan para los números reales una representación binaria en coma flotante normalizada. Esto significa que lo que almacena el computador no es una cantidad matemática x , sino una aproximación binaria a x :

$$(26) \quad x \approx \pm q \times 2^n.$$

El número q se llama **mantisa** y es una expresión binaria finita que verifica la desigualdad $1/2 \leq q < 1$. El número entero n se llama **exponente**.

Un computador sólo utiliza un pequeño subconjunto de números reales. Típicamente, este subconjunto contiene sólo una porción de los números binarios sugeridos por (26) ya que es necesario restringir el número de cifras binarias que puedan tener tanto q como n . Por ejemplo, consideremos el conjunto de todos los números reales positivos de la forma

$$(27) \quad 0.d_1 d_2 d_3 d_4 \dots \times 2^n,$$

Tabla 1.3 Equivalentes decimales para un conjunto de números binarios con una mantisa de 4 cifras y un exponente de $n = -3, -2, \dots, 3, 4$.

Mantisa	Exponente:							
	$n = -3$	$n = -2$	$n = -1$	$n = 0$	$n = 1$	$n = 2$	$n = 3$	$n = 4$
0.1000 _{dos}	0.0625	0.125	0.25	0.5	1	2	4	8
0.1001 _{dos}	0.0703125	0.140625	0.28125	0.5625	1.125	2.25	4.5	9
0.1010 _{dos}	0.078125	0.15625	0.3125	0.625	1.25	2.5	5	10
0.1011 _{dos}	0.0859375	0.171875	0.34375	0.6875	1.375	2.75	5.5	11
0.1100 _{dos}	0.09375	0.1875	0.375	0.75	1.5	3	6	12
0.1101 _{dos}	0.1015625	0.203125	0.40625	0.8125	1.625	3.25	6.5	13
0.1110 _{dos}	0.109375	0.21875	0.4375	0.875	1.75	3.5	7	14
0.1111 _{dos}	0.1171875	0.234375	0.46875	0.9375	1.875	3.75	7.5	15

donde $d_1 = 1$ y d_2, d_3 y d_4 son bien 0, bien 1, y $n \in \{-3, -2, -1, 0, 1, 2, 3, 4\}$. Tenemos ocho elecciones posibles para la mantisa y ocho elecciones posibles para el exponente de (27), lo que nos proporciona un conjunto de 64 números:

$$(28) \quad \{0.1000_{\text{dos}} \times 2^{-3}, 0.1001_{\text{dos}} \times 2^{-3}, \dots, 0.1110_{\text{dos}} \times 2^4, 0.1111_{\text{dos}} \times 2^4\}.$$

Las expresiones decimales de estos 64 números se recogen en la Tabla 1.3. Es importante darse cuenta de que cuando la mantisa y el exponente de (27) se restringen, el computador dispone sólo de un número limitado de valores entre los que elegir para almacenar una aproximación al número real x .

¿Qué ocurriría si pidieramos a un computador con una mantisa de 4 cifras, como la que acabamos de describir, que realizara la operación $(\frac{1}{10} + \frac{1}{5}) + \frac{1}{6}$? Supongamos que el computador redondea todos los números reales al número binario más próximo de los que aparecen en la Tabla 1.3 (mírese en la tabla para comprobar que en cada uno de los pasos siguientes se usa la mejor aproximación).

$$(29) \quad \begin{aligned} \frac{1}{10} &\approx 0.1101_{\text{dos}} \times 2^{-3} = 0.01101_{\text{dos}} \times 2^{-2} \\ \frac{1}{5} &\approx 0.1101_{\text{dos}} \times 2^{-2} = \underline{\underline{0.1101_{\text{dos}} \times 2^{-2}}} \\ \frac{3}{10} &= \underline{\underline{1.00111_{\text{dos}} \times 2^{-2}}}. \end{aligned}$$

El computador debe decidir ahora cómo almacenar el número $1.00111_{\text{dos}} \times 2^{-2}$. Supongamos que se redondea como $0.1010_{\text{dos}} \times 2^{-1}$. El paso siguiente es

$$(30) \quad \begin{aligned} \frac{3}{10} &\approx 0.1010_{\text{dos}} \times 2^{-1} = 0.1010_{\text{dos}} \times 2^{-1} \\ \frac{1}{6} &\approx 0.1011_{\text{dos}} \times 2^{-2} = \underline{\underline{0.01011_{\text{dos}} \times 2^{-1}}} \\ \frac{7}{15} &= \underline{\underline{0.11111_{\text{dos}} \times 2^{-1}}}. \end{aligned}$$

El computador debe decidir ahora cómo almacenar el número $0.11111_{\text{dos}} \times 2^{-1}$. Puesto que suponemos que redondea, almacena $0.10000_{\text{dos}} \times 2^0$. Por tanto, la solución del computador al problema de la suma es

$$(31) \quad \frac{7}{15} \approx 0.10000_{\text{dos}} \times 2^0.$$

El error en el cálculo efectuado por el computador es

$$(32) \quad \frac{7}{15} - 0.10000_{\text{dos}} \approx 0.466667 - 0.500000 \approx -0.033333,$$

que expresado como un porcentaje de $7/15$ es del 7.14%.

Precisión de un computador

Para almacenar los números con una precisión adecuada a las necesidades habituales, los computadores deben trabajar con una aritmética binaria de coma flotante en la que la mantisa disponga de al menos 24 cifras binarias; lo que se traduce en unas siete cifras decimales. Si la mantisa utiliza 32 cifras, entonces se pueden almacenar números de hasta nueve cifras decimales significativas. Volvamos, de nuevo, a la dificultad encontrada en la expresión (1), al comienzo de la sección, cuando se trataba de sumar $1/10$ repetidamente con un computador.

Supongamos que la mantisa q que aparece en (26) contiene 32 cifras binarias, entonces la condición $1/2 \leq q$ implica que la primera cifra es $d_1 = 1$. Por tanto, q es de la forma

$$(33) \quad q = 0.1d_2d_3 \cdots d_{31}d_{32}_{\text{dos}}.$$

Cuando representamos una fracción en forma binaria, lo usual es que esta representación sea periódica; por ejemplo,

$$(34) \quad \frac{1}{10} = 0.\overline{00011}_{\text{dos}}.$$

Cuando usamos una mantisa con 32 cifras, el computador trunca y usa como aproximación interna

$$(35) \quad \frac{1}{10} \approx 0.11001100110011001100110011001100_{\text{dos}} \times 2^{-3},$$

cuyo error, la diferencia entre (34) y (35), es

$$(36) \quad 0.\overline{1100}_{\text{dos}} \times 2^{-35} \approx 2.328306437 \times 10^{-11}.$$

Por este motivo, el computador tiene que cometer un error cuando en (1) le pedimos que sume $1/10$ cien mil veces. Este error en la suma debería ser

al menos de $(100\,000)(2.328306437 \times 10^{-11}) = 2.328306437 \times 10^{-6}$. El error es, de hecho, mucho mayor: habrá ocasiones en que sea necesario redondear la suma parcial; además, conforme la suma crece, los sumandos $1/10$ son pequeños comparados con el tamaño que tenga la suma en ese momento por lo que su contribución se trunca de forma más severa. El efecto compuesto de todos estos errores es lo que produce un error final que vale $10\,000 - 9999.99447 = 5.53 \times 10^{-3}$.

Números del computador en coma flotante

Los computadores disponen de un **modo entero** y de un **modo en coma flotante** para representar números. El modo entero se usa para llevar a cabo operaciones cuyo resultado es con seguridad un número entero, lo que tiene un uso limitado en cálculo numérico. Para las aplicaciones en las ciencias y la ingeniería se utilizan representaciones en coma flotante y debe quedar sobreentendido que el uso de una expresión como (26) establece restricciones sobre el número de cifras usadas para la mantisa q y sobre el rango del exponente n .

Los computadores que usan 32 cifras binarias para representar números reales con precisión simple, reservan 8 cifras para el exponente y 24 para la mantisa (incluyendo los signos); eso les permite representar, aparte del cero, números reales de magnitud comprendida en el rango que va

$$\text{desde } 2.938736E - 39 \quad \text{hasta } 1.701412E + 38$$

(o sea, desde 2^{-128} hasta 2^{127}) con seis cifras decimales de precisión numérica (pues $2^{-23} = 1.2 \times 10^{-7}$).

Los computadores que usan 48 cifras para representar números reales con precisión simple reservan 8 cifras para el exponente y 40 para la mantisa; eso les permite representar números reales de magnitud comprendida en el rango que va

$$\text{desde } 2.9387358771E - 39 \quad \text{hasta } 1.7014118346E + 38$$

(o sea, desde 2^{-128} hasta 2^{127}) con 11 cifras decimales de precisión numérica (pues $2^{-39} = 1.8 \times 10^{-12}$).

Si el computador emplea 64 cifras para representar números reales con precisión doble, entonces puede reservar, por ejemplo, 11 cifras para el exponente y 53 para la mantisa lo que permite representar números reales de magnitud comprendida en el rango que va

$$\text{desde } 5.562684646268003E - 309 \quad \text{hasta } 8.988465674311580E + 307$$

(o sea, desde 2^{-1024} hasta 2^{1023}) con una precisión de unas 16 cifras decimales (ya que $2^{-52} = 2.2 \times 10^{-16}$).

Ejercicios

- Use un computador para realizar las siguientes operaciones de forma acumulada; la intención es que el computador vaya haciendo las substracciones de forma repetida; sin emplear el atajo de la multiplicación.
 - $10\ 000 - \sum_{k=1}^{100\ 000} 0.1$
 - $10\ 000 - \sum_{k=1}^{80\ 000} 0.125$
- Use las relaciones (4) y (5) para convertir los siguientes números binarios en su forma decimal (base 10).
 - 10101_{dos}
 - 111000_{dos}
 - 11111110_{dos}
 - 1000000111_{dos}
- Use las relaciones (16) y (17) para convertir las siguientes fracciones binarias en su forma decimal (base 10).
 - 0.11011_{dos}
 - 0.10101_{dos}
 - 0.1010101_{dos}
 - 0.110110110_{dos}
- Convierta los siguientes números binarios en su forma decimal (base 10).
 - 1.0110101_{dos}
 - $11.0010010001_{\text{dos}}$
- Los números del Ejercicio 4 son aproximadamente $\sqrt{2}$ y π . Halle el error de dichas aproximaciones; es decir, halle
 - $\sqrt{2} - 1.0110101_{\text{dos}}$ (Use que $\sqrt{2} = 1.41421356237309\dots$)
 - $\pi - 11.0010010001_{\text{dos}}$ (Use que $\pi = 3.14159265358979\dots$)
- Siga el Ejemplo 1.10 para convertir los siguientes números en su forma binaria
 - 23
 - 87
 - 378
 - 2388
- Siga el Ejemplo 1.12 para convertir los siguientes números en fracciones binarias de la forma $0.d_1d_2\cdots d_n{}_{\text{dos}}$.
 - $7/16$
 - $13/16$
 - $23/32$
 - $75/128$
- Siga el Ejemplo 1.12 para convertir los siguientes números en fracciones binarias periódicas.
 - $1/10$
 - $1/3$
 - $1/7$
- En las siguientes aproximaciones binarias con siete cifras significativas, halle el error de la aproximación $R - 0.d_1d_2d_3d_4d_5d_6d_7{}_{\text{dos}}$.
 - $R = 1/10 \approx 0.0001100_{\text{dos}}$
 - $R = 1/7 \approx 0.0010010_{\text{dos}}$
- Pruebe que el desarrollo binario $1/7 = 0.\overline{001}{}_{\text{dos}}$ es equivalente a $\frac{1}{7} = \frac{1}{8} + \frac{1}{64} + \frac{1}{512} + \dots$ y use el Teorema 1.14 para justificar dicho desarrollo.
- Pruebe que el desarrollo binario $1/5 = 0.\overline{0011}{}_{\text{dos}}$ es equivalente a $\frac{1}{5} = \frac{3}{16} + \frac{3}{256} + \frac{3}{4096} + \dots$ y use el Teorema 1.14 para justificar dicho desarrollo.
- Pruebe que cualquier número 2^{-N} , siendo N un número natural, puede representarse como un número decimal con N cifras significativas, es decir, $2^{-N} = 0.d_1d_2d_3\cdots d_N$. Indicación. $1/2 = 0.5$, $1/4 = 0.25$, ...

13. Use la Tabla 1.3 para determinar qué ocurre cuando un computador con una mantisa de cuatro cifras lleva a cabo los siguientes cálculos.
- (a) $(\frac{1}{3} + \frac{1}{5}) + \frac{1}{6}$ (b) $(\frac{1}{10} + \frac{1}{3}) + \frac{1}{5}$
 (c) $(\frac{3}{17} + \frac{1}{9}) + \frac{1}{7}$ (d) $(\frac{7}{10} + \frac{1}{9}) + \frac{1}{7}$
14. Pruebe que si sustituimos 2 por 3 en todas las fórmulas de (8), el resultado es un método para hallar la expresión en base 3 de un número natural. Utilice esto para expresar los siguientes números en base 3.
- (a) 10 (b) 23 (c) 421 (d) 1784
15. Pruebe que si sustituimos 2 por 3 en (22), el resultado es un método para hallar la expresión en base 3 de un número positivo R tal que $0 < R < 1$. Utilice esto para expresar los siguientes números en base 3.
- (a) 1/3 (b) 1/2 (c) 1/10 (d) 11/27
16. Pruebe que si sustituimos 2 por 5 en todas las fórmulas de (8), el resultado es un método para hallar la expresión en base 5 de un número natural. Utilice esto para expresar los siguientes números en base 5.
- (a) 10 (b) 35 (c) 721 (d) 734
17. Pruebe que si sustituimos 2 por 5 en (22), el resultado es un método para hallar la expresión en base 5 de un número positivo R tal que $0 < R < 1$. Utilice esto para expresar los siguientes números en base 5.
- (a) 1/3 (b) 1/2 (c) 1/10 (d) 154/625

1.3 Análisis del error

En la práctica del cálculo numérico es importante tener en cuenta que las soluciones calculadas por el computador no son soluciones matemáticas exactas. La precisión de una solución numérica puede verse disminuida por diversos factores, algunos de naturaleza sutil, y la comprensión de estas dificultades puede guiarnos a menudo a desarrollar o a construir algoritmos numéricos adecuados.

Definición 1.7. Supongamos que \hat{p} es una aproximación a p . El **error absoluto** de la aproximación es $E_p = |p - \hat{p}|$ y el **error relativo** es $R_p = |p - \hat{p}|/|p|$, supuesto que $p \neq 0$. ▲

El error absoluto no es más que la distancia entre el valor exacto y el valor aproximado, mientras que el error relativo mide el error entendido como una porción del valor exacto.

Ejemplo 1.14. Vamos a encontrar el error absoluto y el error relativo en los siguientes casos. Sean $x = 3.141592$ y $\hat{x} = 3.14$, entonces el error absoluto es

$$(1a) \quad E_x = |x - \hat{x}| = |3.141592 - 3.14| = 0.001592$$

y el error relativo es

$$R_x = \frac{|x - \hat{x}|}{|x|} = \frac{0.001592}{3.141592} = 0.00507.$$

Sean $y = 1\,000\,000$ e $\hat{y} = 999\,996$, entonces el error absoluto es

$$(1b) \quad E_y = |y - \hat{y}| = |1\,000\,000 - 999\,996| = 4$$

y el error relativo es

$$R_y = \frac{|y - \hat{y}|}{|y|} = \frac{4}{1\,000\,000} = 0.000004.$$

Sean $z = 0.000012$ y $\hat{z} = 0.000009$, entonces el error absoluto es

$$(1c) \quad E_z = |z - \hat{z}| = |0.000012 - 0.000009| = 0.000003$$

y el error relativo es

$$R_z = \frac{|z - \hat{z}|}{|z|} = \frac{0.000003}{0.000012} = 0.25.$$

En el caso (1a) no hay mucha diferencia entre E_x y R_x ; cualquiera de los dos puede usarse para determinar la precisión de \hat{x} . En el caso (1b), el valor de y es del orden de magnitud de 10^6 , el error absoluto E_y es grande y el error relativo R_y es pequeño. En este caso, \hat{y} sería posiblemente considerada como una buena aproximación a y . En el caso (1c), z es del orden de magnitud de 10^{-6} y el error absoluto E_z es el menor de los tres casos. Sin embargo, el error relativo R_z es el mayor; en términos de porcentaje, es de un 25%, por lo que \hat{z} es considerada como una mala aproximación a z . Observemos que conforme $|p|$ se aleja de 1 (creciendo o decreciendo), el error relativo R_p va siendo un indicador de la precisión de la aproximación mejor que el error absoluto E_p . En las representaciones en coma flotante se prefiere trabajar con el error relativo ya que éste está directamente relacionado con la mantisa.

Definición 1.8. Diremos que un número \hat{p} es una aproximación a p con d cifras decimales significativas si d es el mayor número natural tal que

$$(2) \quad \frac{|p - \hat{p}|}{|p|} < \frac{10^{-d}}{2}.$$

Ejemplo 1.15. Vamos a determinar el número de cifras significativas de cada una de las aproximaciones del Ejemplo 1.14.

- (3a) Si $x = 3.141592$ y $\hat{x} = 3.14$, entonces $|x - \hat{x}|/|x| = 0.000507 < 10^{-2}/2$. Por tanto, \hat{x} es una aproximación a x con dos cifras significativas.
- (3b) Si $y = 1\,000\,000$ e $\hat{y} = 999\,996$, entonces $|y - \hat{y}|/|y| = 0.000004 < 10^{-5}/2$. Por tanto, \hat{y} es una aproximación a y con cinco cifras significativas.
- (3c) Si $z = 0.000012$ y $\hat{z} = 0.000009$, entonces $|z - \hat{z}|/|z| = 0.25 < 10^{-0}/2$. Por tanto, \hat{z} es una aproximación a z sin cifras significativas.

Error de truncamiento

La noción de error de truncamiento se refiere normalmente a los errores que se producen cuando una expresión matemática complicada se “reemplaza” por una fórmula más simple. Esta terminología se originó en la sustitución de una función por uno de sus polinomios de Taylor. Por ejemplo, podríamos reemplazar la serie de Taylor

$$e^{x^2} = 1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!} + \cdots + \frac{x^{2n}}{n!} + \cdots$$

por los cinco primeros términos $1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!}$ a la hora de calcular una integral numéricamente.

Ejemplo 1.16. Sabiendo que $\int_0^{1/2} e^{x^2} dx = 0.544987104184 = p$, vamos a determinar la precisión de la aproximación obtenida al reemplazar el integrando $f(x) = e^{x^2}$ por la serie de Taylor truncada $P_8(x) = 1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!}$.

Integrando término a término este polinomio, obtenemos

$$\begin{aligned} \int_0^{1/2} \left(1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!} + \frac{x^8}{4!} \right) dx &= \left(x + \frac{x^3}{3} + \frac{x^5}{5(2!)} + \frac{x^7}{7(3!)} + \frac{x^9}{9(4!)} \right)_{x=0}^{x=1/2} \\ &= \frac{1}{2} + \frac{1}{24} + \frac{1}{320} + \frac{1}{5376} + \frac{1}{110592} \\ &= \frac{2109491}{3870720} = 0.544986720817 = \hat{p}. \end{aligned}$$

Puesto que $10^{-5}/2 > |p - \hat{p}|/|p| = 7.03442 \times 10^{-7} > 10^{-6}/2$, la aproximación \hat{p} coincide con el valor exacto $p = 0.544987104184$ en cinco cifras significativas. Las gráficas de $f(x) = e^{x^2}$ y de $y = P_8(x)$ y el área limitada por la curva para $0 \leq x \leq 1/2$ se muestran en la Figura 1.7.

Error de redondeo

La representación de los números reales en un computador está limitada por el número de cifras de la mantisa, de manera que algunos números no coinciden exactamente con su representación en el computador. Esto es lo que se conoce como **error de redondeo**. En la sección anterior vimos como el número real $1/10 = 0.\overline{00011}_{\text{dos}}$ se truncaba al almacenarlo en un computador. El número que, de hecho, se guarda en la memoria del computador puede haber sufrido la poda o el redondeo de su última cifra; en consecuencia, y puesto que el computador trabaja con números que tienen una cantidad limitada de cifras, los errores de redondeo se introducen y propagan cuando se hacen varias operaciones sucesivas.

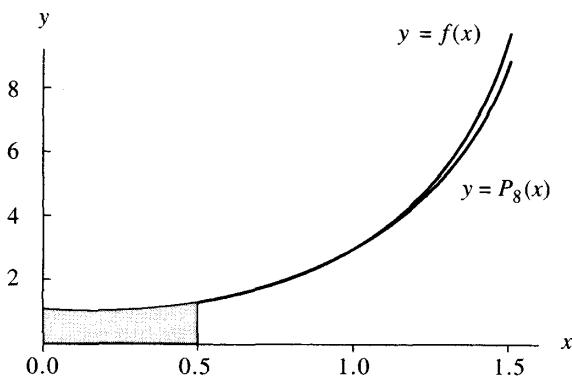


Figura 1.7 Gráficas de $f(x) = e^{x^2}$ y de $y = P_8(x)$ y área limitada por la curva para $0 \leq x \leq 1/2$.

Poda frente a redondeo

Sea p un número real cualquiera expresado de *forma decimal normalizada* como

$$(4) \quad p = \pm 0.d_1d_2d_3\dots d_kd_{k+1}\dots \times 10^n,$$

donde $d_1 \in \{1, 2, \dots, 8, 9\}$ y $d_j \in \{0, 1, \dots, 8, 9\}$ para $j > 1$. Supongamos que k es el número máximo de cifras decimales que se admiten en la aritmética en coma flotante de un computador, entonces el número real p se representa mediante el número de la máquina $fl_{\text{pod}}(p)$ dado por

$$(5) \quad fl_{\text{pod}}(p) = \pm 0.d_1d_2d_3\dots d_k \times 10^n,$$

donde $d_1 \in \{1, 2, \dots, 8, 9\}$ y $d_j \in \{0, 1, \dots, 8, 9\}$ para $1 < j \leq k$. El número de la máquina $fl_{\text{pod}}(p)$ se llama *representación en coma flotante mediante poda o redondeo por debajo* de p . En este caso la k -ésima cifra de $fl_{\text{pod}}(p)$ coincide con la k -ésima cifra de p . Una manera alternativa de usar representaciones con k cifras es la *representación en coma flotante mediante redondeo* $fl_{\text{red}}(p)$, que viene dado por

$$(6) \quad fl_{\text{red}}(p) = \pm 0.r_1r_2r_3\dots r_k \times 10^n,$$

donde $r_1 \in \{1, 2, \dots, 8, 9\}$ y $r_j \in \{0, 1, \dots, 8, 9\}$ para $1 < j \leq k$ se calculan redondeando el número $d_1d_2\dots d_k.d_{k+1}d_{k+2}\dots$ al número entero más próximo (y se sobreentiende que $fl_{\text{red}}(\pm 0.999\dots 9d_{k+1} \times 10^n) = 0.1 \times 10^{n+1}$ si $d_{k+1} \geq 5$). Por ejemplo, el número real

$$p = \frac{22}{7} = 3.142857142857142857\dots$$

tiene las siguientes representaciones en coma flotante con seis cifras significativas

$$\begin{aligned}fl_{\text{pod}}(p) &= 0.314285 \times 10^1, \\fl_{\text{red}}(p) &= 0.314286 \times 10^1.\end{aligned}$$

En situaciones normales la poda y el redondeo los escribiríamos como 3.14285 y 3.14286, respectivamente. Prestemos un poco de atención al caso en que las últimas cifras a redondear son nueves. El número real

$$p = 0.23159963$$

tiene las siguientes representaciones en coma flotante con seis cifras

$$\begin{aligned}fl_{\text{pod}}(p) &= 0.231599, \\fl_{\text{red}}(p) &= 0.231600.\end{aligned}$$

La gran mayoría de los computadores usan la representación en coma flotante mediante redondeo.

Pérdida de cifras significativas

Consideremos los números $p = 3.1415926536$ y $q = 3.1415957341$ que son casi iguales y están ambos expresados con una precisión de 11 cifras decimales. Si calculamos su diferencia $p - q = -0.0000030805$ vemos que, como las seis primeras cifras de p y de q coinciden, su diferencia $p - q$ sólo contiene cinco cifras decimales; este fenómeno se conoce como **pérdida de cifras significativas** o **cancelación** y hay que tener cierto cuidado con él porque puede producir sin que nos demos cuenta una reducción en la precisión de la respuesta final calculada.

Ejemplo 1.17. Vamos a comparar los resultados de calcular $f(500)$ y $g(500)$, usando seis cifras significativas con redondeo, siendo $f(x) = x(\sqrt{x+1} - \sqrt{x})$ y $g(x) = x / (\sqrt{x+1} + \sqrt{x})$. Con la primera función,

$$\begin{aligned}f(500) &= 500 \left(\sqrt{501} - \sqrt{500} \right) \\&= 500(22.3830 - 22.3607) = 500(0.0223) = 11.1500.\end{aligned}$$

Con $g(x)$,

$$\begin{aligned}g(500) &= \frac{500}{\sqrt{501} + \sqrt{500}} \\&= \frac{500}{22.3830 + 22.3607} = \frac{500}{44.7437} = 11.1748.\end{aligned}$$

La segunda función, $g(x)$, es algebraicamente equivalente a $f(x)$, como muestra el siguiente cálculo

$$\begin{aligned}f(x) &= \frac{x(\sqrt{x+1} - \sqrt{x})(\sqrt{x+1} + \sqrt{x})}{\sqrt{x+1} + \sqrt{x}} \\&= \frac{x((\sqrt{x+1})^2 - (\sqrt{x})^2)}{\sqrt{x+1} + \sqrt{x}} \\&= \frac{x}{\sqrt{x+1} + \sqrt{x}}.\end{aligned}$$

La respuesta $g(500) = 11.1748$ tiene un error absoluto menor y es lo que obtendríamos redondeando la respuesta exacta $11.174755300747198\dots$ a seis cifras significativas.

Es digno de estudio el Ejercicio 12 en el que se describe cómo evitar la pérdida de cifras significativas cuando usamos la conocida fórmula para resolver ecuaciones de segundo grado. El siguiente ejemplo ilustra cómo podemos usar una serie de Taylor truncada para evitar el error por pérdida de cifras significativas.

Ejemplo 1.18. Vamos a comparar los resultados de calcular $f(0.01)$ y $P(0.01)$, usando seis cifras significativas con redondeo, siendo

$$f(x) = \frac{e^x - 1 - x}{x^2} \quad \text{y} \quad P(x) = \frac{1}{2} + \frac{x}{6} + \frac{x^2}{24}.$$

La función $P(x)$ es el polinomio de Taylor de grado $n = 2$ de $f(x)$ alrededor de $x = 0$.

Para la primera función tenemos

$$f(0.01) = \frac{e^{0.01} - 1 - 0.01}{(0.01)^2} = \frac{1.010050 - 1 - 0.01}{0.001} = 0.5.$$

Para la segunda

$$\begin{aligned}P(0.01) &= \frac{1}{2} + \frac{0.01}{6} + \frac{0.001}{24} \\&= 0.5 + 0.001667 + 0.000004 = 0.501671.\end{aligned}$$

La respuesta $P(0.01) = 0.501671$ es más exacta y coincide con lo que obtendríamos al redondear la respuesta verdadera $0.50167084168057542\dots$ a seis cifras significativas.

A la hora de evaluar un polinomio, obtenemos a menudo mejores resultados si usamos el Método de Horner de multiplicación encajada

Ejemplo 1.19. Sean $P(x) = x^3 - 3x^2 + 3x - 1$ y $Q(x) = ((x - 3)x + 3)x - 1$. Usando tres cifras significativas con redondeo, vamos a calcular $P(2.19)$ y $Q(2.19)$ y a comparar estas aproximaciones con los valores exactos $P(2.19) = Q(2.19) = 1.685159$;

$$\begin{aligned} P(2.19) &\approx (2.19)^3 - 3(2.19)^2 + 3(2.19) - 1 \\ &= 10.5 - 14.4 + 6.57 - 1 = 1.67. \\ Q(2.19) &\approx ((2.19 - 3)2.19 + 3)2.19 - 1 = 1.69. \end{aligned}$$

Los errores absolutos son 0.015159 y 0.004841, respectivamente. Así que la aproximación $Q(2.19) \approx 1.69$ es mejor. En el Ejercicio 6 se requiere realizar una exploración para ver qué ocurre cerca de la raíz de este polinomio.

Orden de aproximación $O(h^n)$

Está claro que las sucesiones $\{\frac{1}{n^2}\}_{n=1}^{\infty}$ y $\{\frac{1}{n}\}_{n=1}^{\infty}$ son ambas convergentes a cero pero, sin embargo, debemos hacer notar que la primera sucesión converge a cero más rápidamente que la segunda. En los capítulos siguientes introduciremos la terminología y notación necesarias para describir cuánto de rápido converge una sucesión.

Definición 1.9. Se dice que una función $f(h)$ es **de orden** $g(h)$ cuando $h \rightarrow 0$, lo que se denota por $f(h) = O(g(h))$ (lo que se llama notación O mayúscula de Landau), si existen constantes C y c tales que

$$(7) \quad |f(h)| \leq C|g(h)| \quad \text{siempre que } |h| \leq c.$$

Ejemplo 1.20. Consideremos las funciones $f(x) = x^3 + 2x^2$ y $g(x) = x^2$. Puesto que $x^3 \leq x^2$ para $|x| \leq 1$, obtenemos que $x^3 + 2x^2 \leq 3x^2$ para $|x| \leq 1$. Por tanto, $f(x) = O(g(x))$.

La notación $O(\cdot)$ (que también se usa para límites en el infinito) proporciona una forma muy útil de describir la velocidad de crecimiento (o de decrecimiento) de una función en términos de la velocidad de crecimiento (o de decrecimiento) de funciones elementales bien conocidas ($x^n, x^{1/n}, a^x, \log_a x$, etc.).

La velocidad de convergencia de sucesiones puede describirse de forma parecida.

Definición 1.10. Sean $\{x_n\}_{n=1}^{\infty}$ e $\{y_n\}_{n=1}^{\infty}$ dos sucesiones. Se dice que la sucesión $\{x_n\}$ es de orden $\{y_n\}$, lo que denotamos por $x_n = O(y_n)$, si existen constantes C y N tales que

$$(8) \quad |x_n| \leq C|y_n| \quad \text{siempre que } n \geq N.$$

Ejemplo 1.21. $\frac{n^2-1}{n^3} = O\left(\frac{1}{n}\right)$, ya que $\frac{n^2-1}{n^3} \leq \frac{n^2}{n^3} = \frac{1}{n}$ siempre que $n \geq 1$.

A menudo nos encontramos con que una función $f(h)$ se aproxima mediante otra función $p(h)$ y sabemos que una cota del error cometido es $M|h^n|$. Esto nos conduce a la siguiente definición.

Definición 1.11. Supongamos que una función $p(h)$ aproxima a otra $f(h)$ y que existen una constante real $M > 0$ y un número natural n tales que

$$(9) \quad \frac{|f(h) - p(h)|}{|h^n|} \leq M \quad \text{para } h \text{ suficientemente pequeño.}$$

Entonces se dice que $p(h)$ aproxima a $f(h)$ con orden de aproximación $\mathcal{O}(h^n)$, lo que se escribe

$$(10) \quad f(h) = p(h) + \mathcal{O}(h^n).$$

Escribiendo la relación (9) como $|f(h) - p(h)| \leq M|h^n|$, vemos que $\mathcal{O}(h^n)$ ocupa el lugar de la cota del error $M|h^n|$. El siguiente resultado muestra cómo podemos aplicar esta definición a combinaciones simples de dos funciones.

Teorema 1.15. Supongamos que $f(h) = p(h) + \mathcal{O}(h^n)$, $g(h) = q(h) + \mathcal{O}(h^m)$ y sea $r = \min\{m, n\}$. Entonces

$$(11) \quad f(h) + g(h) = p(h) + q(h) + \mathcal{O}(h^r),$$

$$(12) \quad f(h)g(h) = p(h)q(h) + \mathcal{O}(h^r),$$

y

$$(13) \quad \frac{f(h)}{g(h)} = \frac{p(h)}{q(h)} + \mathcal{O}(h^r) \quad \text{supuesto que } g(h) \neq 0 \text{ y que } q(h) \neq 0.$$

Resulta instructivo considerar el caso en que $p(x)$ es la n -ésima aproximación por polinomios de Taylor de $f(x)$; entonces el resto de la fórmula de Taylor se designa simplemente por $\mathcal{O}(h^{n+1})$ y sustituye a todos los términos omitidos, que son el que contiene la potencia h^{n+1} y los de grado superior. El resto de la fórmula de Taylor converge a cero con la misma rapidez que h^{n+1} converge a cero cuando $h \rightarrow 0$, tal y como lo expresa la relación

$$(14) \quad \mathcal{O}(h^{n+1}) \approx Mh^{n+1} \approx \frac{f^{(n+1)}(c)}{(n+1)!}h^{n+1}$$

válida para h suficientemente pequeño. En otras palabras, el término $\mathcal{O}(h^{n+1})$ sustituye a la cantidad Mh^{n+1} , donde M es constante o “se comporta como una constante.”

Teorema 1.16 (Teorema de Taylor). Supongamos que $f \in C^{n+1}[a, b]$. Si x_0 y $x = x_0 + h$ están en $[a, b]$, entonces

$$(15) \quad f(x_0 + h) = \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} h^k + O(h^{n+1}).$$

El siguiente ejemplo sirve para ilustrar los teoremas anteriores. Los cálculos usan las siguientes dos propiedades de la suma (i) $O(h^p) + O(h^p) = O(h^p)$, (ii) $O(h^p) + O(h^q) = O(h^r)$, siendo $r = \min\{p, q\}$, y la propiedad de la multiplicación (iii) $O(h^p)O(h^q) = O(h^s)$, siendo $s = p + q$.

Ejemplo 1.22. Consideremos los desarrollos de Taylor

$$e^h = 1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + O(h^4) \quad \text{y} \quad \cos(h) = 1 - \frac{h^2}{2!} + \frac{h^4}{4!} + O(h^6).$$

Se trata de determinar el orden de la aproximación para su suma y su producto.

Para la suma tenemos

$$\begin{aligned} e^h + \cos(h) &= 1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + O(h^4) + 1 - \frac{h^2}{2!} + \frac{h^4}{4!} + O(h^6) \\ &= 2 + h + \frac{h^3}{3!} + O(h^4) + \frac{h^4}{4!} + O(h^6). \end{aligned}$$

Puesto que $O(h^4) + \frac{h^4}{4!} = O(h^4)$ y $O(h^4) + O(h^6) = O(h^4)$, esto se reduce a

$$e^h + \cos(h) = 2 + h + \frac{h^3}{3!} + O(h^4),$$

y el orden de aproximación es $O(h^4)$.

El producto se trata de manera parecida:

$$\begin{aligned} e^h \cos(h) &= \left(1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + O(h^4)\right) \left(1 - \frac{h^2}{2!} + \frac{h^4}{4!} + O(h^6)\right) \\ &= \left(1 + h + \frac{h^2}{2!} + \frac{h^3}{3!}\right) \left(1 - \frac{h^2}{2!} + \frac{h^4}{4!}\right) \\ &\quad + \left(1 + h + \frac{h^2}{2!} + \frac{h^3}{3!}\right) O(h^6) + \left(1 - \frac{h^2}{2!} + \frac{h^4}{4!}\right) O(h^4) \\ &\quad + O(h^4)O(h^6) \\ &= 1 + h - \frac{h^3}{3} - \frac{5h^4}{24} - \frac{h^5}{24} + \frac{h^6}{48} + \frac{h^7}{144} \\ &\quad + O(h^6) + O(h^4) + O(h^4)O(h^6). \end{aligned}$$

Puesto que $O(h^4)O(h^6) = O(h^{10})$ y

$$-\frac{5h^4}{24} - \frac{h^5}{24} + \frac{h^6}{48} + \frac{h^7}{144} + O(h^6) + O(h^4) + O(h^{10}) = O(h^4),$$

la relación anterior puede simplificarse como

$$e^h \cos(h) = 1 + h - \frac{h^3}{3} + O(h^4),$$

de manera que el orden de aproximación es $O(h^4)$. ■

Orden de aproximación de una sucesión

Las buenas aproximaciones numéricas se suelen conseguir calculando una sucesión de aproximaciones que se acercan más y más a la respuesta deseada. Qué significa ser de orden $O(\cdot)$ para sucesiones lo vimos en la Definición 1.10; la definición de orden de aproximación de una sucesión es análoga a la dada para funciones en la Definición 1.11.

Definición 1.12. Supongamos que $\lim_{n \rightarrow \infty} x_n = x$ y que $\{r_n\}_{n=1}^{\infty}$ es una sucesión tal que $\lim_{n \rightarrow \infty} r_n = 0$. Se dice que $\{x_n\}_{n=1}^{\infty}$ converge a x con orden de aproximación $O(r_n)$ si existe una constante $K > 0$ tal que

$$\frac{|x_n - x|}{|r_n|} \leq K \quad \text{para } n \text{ suficientemente grande.}$$

Esto lo indicamos escribiendo $x_n = x + O(r_n)$, o bien $x_n \rightarrow x$ con orden de aproximación $O(r_n)$. ▲

Ejemplo 1.23. Sean $x_n = \cos(n)/n^2$ y $r_n = 1/n^2$, entonces $\lim_{n \rightarrow \infty} x_n = 0$ con orden de aproximación $O(1/n^2)$. Esto se deduce inmediatamente de la relación

$$\frac{|\cos(n)/n^2|}{|1/n^2|} = |\cos(n)| \leq 1 \quad \text{para todo } n.$$

Propagación del error

Vamos a investigar ahora cómo pueden propagarse los errores en una cadena de operaciones sucesivas. Consideremos la suma de dos números p y q (que son valores exactos) con valores aproximados \hat{p} y \hat{q} cuyos errores son ε_p y ε_q , respectivamente. A partir de $p = \hat{p} + \varepsilon_p$ y de $q = \hat{q} + \varepsilon_q$, la suma es

$$(16) \quad p + q = (\hat{p} + \varepsilon_p) + (\hat{q} + \varepsilon_q) = (\hat{p} + \hat{q}) + (\varepsilon_p + \varepsilon_q).$$

Por tanto, el error en una suma es la suma de los errores de los sumandos.

La propagación del error en una multiplicación es más complicada. El producto es

$$(17) \quad pq = (\hat{p} + \varepsilon_p)(\hat{q} + \varepsilon_q) = \hat{p}\hat{q} + \hat{p}\varepsilon_q + \hat{q}\varepsilon_p + \varepsilon_p\varepsilon_q.$$

Por tanto, si \hat{p} y \hat{q} son mayores que 1 en valor absoluto, los términos $\hat{p}\varepsilon_q$ y $\hat{q}\varepsilon_p$ indican que hay una posibilidad de que los errores originales ε_p y ε_q sean magnificados. Si analizamos los errores relativos, tendremos una percepción más clara de la situación. Reordenando los términos de (17) obtenemos

$$(18) \quad pq - \hat{p}\hat{q} = \hat{p}\varepsilon_q + \hat{q}\varepsilon_p + \varepsilon_p\varepsilon_q.$$

Supongamos que $p \neq 0$ y que $q \neq 0$; entonces podemos dividir (18) entre pq para obtener el error relativo del producto pq :

$$(19) \quad R_{pq} = \frac{pq - \hat{p}\hat{q}}{pq} = \frac{\hat{p}\varepsilon_q + \hat{q}\varepsilon_p + \varepsilon_p\varepsilon_q}{pq} = \frac{\hat{p}\varepsilon_q}{pq} + \frac{\hat{q}\varepsilon_p}{pq} + \frac{\varepsilon_p\varepsilon_q}{pq}.$$

Es más, supongamos que \hat{p} y \hat{q} son buenas aproximaciones de p y q ; entonces $\hat{p}/p \approx 1$, $\hat{q}/q \approx 1$ y $R_p R_q = (\varepsilon_p/p)(\varepsilon_q/q) \approx 0$ (R_p y R_q son los errores relativos de las aproximaciones \hat{p} y \hat{q}). Sustituyendo estas aproximaciones en (19) obtenemos una relación más simple:

$$(20) \quad R_{pq} = \frac{pq - \hat{p}\hat{q}}{pq} \approx \frac{\varepsilon_q}{q} + \frac{\varepsilon_p}{p} + 0 = R_q + R_p.$$

Esto prueba que el error relativo del producto pq es aproximadamente la suma de los errores relativos de las aproximaciones \hat{p} y \hat{q} a los factores.

Es normal que los errores iniciales en los datos se propaguen a lo largo de una cadena de operaciones. Una cualidad deseable de cualquier proceso numérico es que un error pequeño en las condiciones iniciales produzca errores pequeños en el resultado final. Un algoritmo con esta cualidad se llama **estable**; en otro caso, se llama **inestable**. Siempre que sea posible, elegiremos métodos que sean estables. La siguiente definición la usaremos para describir el fenómeno de la propagación de los errores.

Definición 1.13. Supongamos que ε representa un error inicial y que $\varepsilon(n)$ representa el crecimiento de dicho error después de n operaciones. Si se verifica que $|\varepsilon(n)| \approx n\varepsilon$, entonces se dice que el crecimiento es lineal. Si $|\varepsilon(n)| \approx K^n\varepsilon$, entonces se dice que el crecimiento es exponencial. Si $K > 1$, entonces un error exponencial crece cuando $n \rightarrow \infty$ sin que podamos acotarlo; pero si $0 < K < 1$, entonces un error exponencial disminuye a cero cuando $n \rightarrow \infty$. ▲

Los dos ejemplos siguientes muestran cómo un error inicial puede propagarse de manera estable o inestable. En el primer ejemplo se presentan tres algoritmos cada uno de los cuales generaría recursivamente la misma sucesión si se realizasen las cuentas exactamente. Después, en el segundo ejemplo, analizaremos la propagación de pequeños errores en las condiciones iniciales.

Tabla 1.4 La sucesión $\{x_n\} = \{1/3^n\}$ y sus aproximaciones $\{r_n\}$, $\{p_n\}$ y $\{q_n\}$.

n	x_n	r_n	p_n	q_n
0	$1 = 1.0000000000$	0.9999600000	1.0000000000	1.0000000000
1	$\frac{1}{3} = 0.3333333333$	0.3333200000	0.3333200000	0.3333200000
2	$\frac{1}{9} = 0.1111111111$	0.1111066667	0.1110933330	0.1110666667
3	$\frac{1}{27} = 0.0370370370$	0.0370355556	0.0370177778	0.0369022222
4	$\frac{1}{81} = 0.0123456790$	0.0123451852	0.0123259259	0.0119407407
5	$\frac{1}{243} = 0.0041152263$	0.0041150617	0.0040953086	0.0029002469
6	$\frac{1}{729} = 0.0013717421$	0.0013716872	0.0013517695	-0.0022732510
7	$\frac{1}{2187} = 0.0004572474$	0.0004572291	0.0004372565	-0.0104777503
8	$\frac{1}{6561} = 0.0001524158$	0.0001524097	0.0001324188	-0.0326525834
9	$\frac{1}{19683} = 0.0000508053$	0.0000508032	0.0000308063	-0.0983641945
10	$\frac{1}{59049} = 0.0000169351$	0.0000169344	-0.0000030646	-0.2952280648

Ejemplo 1.24. Vamos a probar que los tres esquemas siguientes pueden usarse para generar recursivamente los términos de la sucesión $\{1/3^n\}_{n=0}^{\infty}$ si las operaciones se hicieran exactamente:

(21a)

$$r_0 = 1 \quad \text{y} \quad r_n = \frac{1}{3}r_{n-1} \quad \text{para } n = 1, 2, \dots,$$

(21b)

$$p_0 = 1, \quad p_1 = \frac{1}{3}, \quad \text{y} \quad p_n = \frac{4}{3}p_{n-1} - \frac{1}{3}p_{n-2} \quad \text{para } n = 2, 3, \dots,$$

(21c) $q_0 = 1, \quad q_1 = \frac{1}{3}, \quad \text{y} \quad q_n = \frac{10}{3}q_{n-1} - q_{n-2} \quad \text{para } n = 2, 3, \dots$

La fórmula (21a) es obvia. En (21b) la ecuación en diferencias correspondiente tiene como solución general $p_n = A(1/3^n) + B$, lo que podemos verificar por sustitución directa:

$$\begin{aligned} \frac{4}{3}p_{n-1} - \frac{1}{3}p_{n-2} &= \frac{4}{3} \left(\frac{A}{3^{n-1}} + B \right) - \frac{1}{3} \left(\frac{A}{3^{n-2}} + B \right) \\ &= \left(\frac{4}{3^n} - \frac{3}{3^n} \right) A - \left(\frac{4}{3} - \frac{1}{3} \right) B = A \frac{1}{3^n} + B = p_n. \end{aligned}$$

Tomando $A = 1$ y $B = 0$, lo que corresponde a $p_0 = 1$ y $p_1 = \frac{1}{3}$, generamos la sucesión deseada. En (21c) la ecuación en diferencias correspondiente tiene como

Tabla 1.5 Las sucesiones de errores $\{x_n - r_n\}$, $\{x_n - p_n\}$ y $\{x_n - q_n\}$.

n	$x_n - r_n$	$x_n - p_n$	$x_n - q_n$
0	0.0000400000	0.0000000000	0.0000000000
1	0.0000133333	0.0000133333	0.0000013333
2	0.0000044444	0.0000177778	0.0000444444
3	0.0000014815	0.0000192593	0.0001348148
4	0.0000004938	0.0000197531	0.0004049383
5	0.0000001646	0.0000199177	0.0012149794
6	0.0000000549	0.0000199726	0.0036449931
7	0.0000000183	0.0000199909	0.0109349977
8	0.0000000061	0.0000199970	0.0328049992
9	0.0000000020	0.0000199990	0.0984149998
10	0.0000000007	0.0000199997	0.2952449999

solución general $q_n = A(1/3^n) + B3^n$, lo que también puede verificarse por sustitución:

$$\begin{aligned} \frac{10}{3}q_{n-1} - q_{n-2} &= \frac{10}{3} \left(\frac{A}{3^{n-1}} + B3^{n-1} \right) - \left(\frac{A}{3^{n-2}} + B3^{n-2} \right) \\ &= \left(\frac{10}{3^n} - \frac{9}{3^n} \right) A - (10-1)3^{n-2}B \\ &= A \frac{1}{3^n} + B3^n = q_n. \end{aligned}$$

Tomando $A = 1$ y $B = 0$ lo que corresponde a $q_0 = 1$ y $q_1 = \frac{1}{3}$, generamos la sucesión deseada.

Ejemplo 1.25. Vamos a generar aproximaciones a la sucesión $\{x_n\} = \{1/3^n\}$ usando los esquemas

$$(22a) \quad r_0 = 0.99996 \quad y \quad r_n = \frac{1}{3}r_{n-1} \quad \text{para } n = 1, 2, \dots,$$

$$(22b) \quad p_0 = 1, \quad p_1 = 0.33332, \quad y \quad p_n = \frac{4}{3}p_{n-1} - \frac{1}{3}p_{n-2} \quad \text{para } n = 2, 3, \dots,$$

$$(22c) \quad q_0 = 1, \quad q_1 = 0.33332, \quad y \quad q_n = \frac{10}{3}q_{n-1} - q_{n-2} \quad \text{para } n = 2, 3, \dots$$

En (22a) el error inicial de r_0 es 0.00004 y tanto en (22b) como en (22c) los errores iniciales de p_1 y q_1 son 0.000013. Vamos a investigar la propagación de los errores en cada esquema.

La Tabla 1.4 nos muestra las diez primeras aproximaciones de cada sucesión y la Tabla 1.5 nos muestra los errores de cada fórmula. El error de $\{r_n\}$ es estable

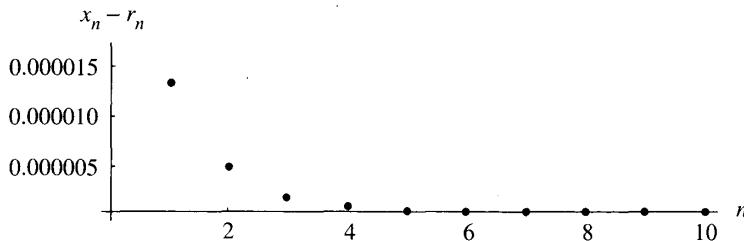


Figura 1.8 Una sucesión de errores $\{x_n - r_n\}$ estable y decreciente.

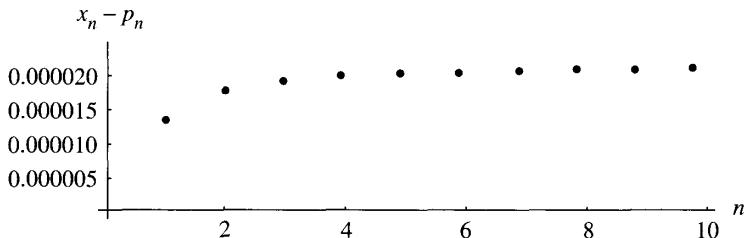


Figura 1.9 Una sucesión de errores $\{x_n - p_n\}$ estable.

y decrece de forma exponencial. El error de $\{p_n\}$ es estable. El error de $\{q_n\}$ es inestable y crece con velocidad exponencial. Aunque el error de $\{p_n\}$ es estable, como los términos verifican que $p_n \rightarrow 0$ cuando $n \rightarrow \infty$, el error acaba dominando a largo plazo y las cifras significativas de los términos posteriores al p_8 no coinciden con las del correspondiente x_n . Las Figuras 1.8, 1.9 y 1.10 muestran los errores de $\{r_n\}$, $\{p_n\}$ y $\{q_n\}$, respectivamente. ■

Incertidumbre en los datos

Los datos de los problemas que se presentan en la realidad contienen incertidumbre o error. Este tipo de error se conoce como ruido y afectará la exactitud de cualquier cálculo numérico que se base en dichos datos. No podemos mejorar la precisión de los cálculos si realizamos operaciones con datos afectados por ruido. Así, si empezamos con datos que contienen d cifras significativas, entonces el resultado de un cálculo con dichos datos debería mostrarse también con d cifras significativas; por ejemplo, supongamos que los datos $p_1 = 4.152$ y $p_2 = 0.07931$ tienen ambos una precisión de cuatro cifras, entonces sería tentador indicar todas las cifras que aparecen en la pantalla de una calculadora al hacer, digamos su suma: $p_1 + p_2 = 4.23131$. Esto no es correcto, no deberíamos obtener conclu-

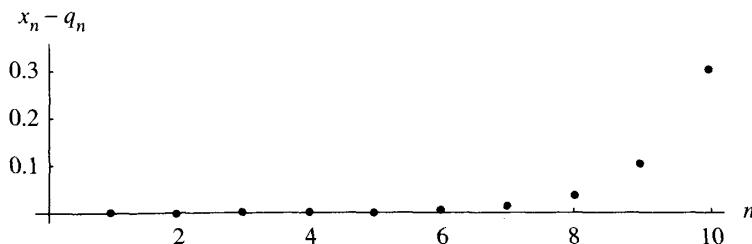


Figura 1.10 Una sucesión de errores $\{x_n - q_n\}$ inestable y creciente.

siones que tengan más cifras significativas que los datos originales. La respuesta adecuada en esta situación es $p_1 + p_2 = 4.231$.

Ejercicios

- En cada uno de los casos siguientes, halle el error absoluto E_x y el error relativo R_x y determine el número de cifras significativas de la aproximación.
 - $x = 2.71828182$, $\hat{x} = 2.7182$
 - $y = 98\,350$, $\hat{y} = 98\,000$
 - $z = 0.000068$, $\hat{z} = 0.00006$
- Complete el siguiente cálculo

$$\int_0^{1/4} e^{x^2} dx \approx \int_0^{1/4} \left(1 + x^2 + \frac{x^4}{2!} + \frac{x^6}{3!}\right) dx = \hat{p}.$$

Determine qué tipo de error se presenta en esta situación y compare su resultado con el valor exacto $p = 0.2553074606$.

- (a) Consideremos los datos $p_1 = 1.414$ y $p_2 = 0.09125$, que vienen dados con una precisión de cuatro cifras significativas. Determine el resultado adecuado en esta situación de la suma $p_1 + p_2$ y el producto $p_1 p_2$.
 (b) Consideremos los datos $p_1 = 31.415$ y $p_2 = 0.027182$, que vienen dados con una precisión de cinco cifras significativas. Determine el resultado adecuado en esta situación de la suma $p_1 + p_2$ y el producto $p_1 p_2$.
- Complete los siguientes cálculos y diga qué tipo de error se presenta en cada situación.

- $\frac{\sin(\frac{\pi}{4} + 0.00001) - \sin(\frac{\pi}{4})}{0.00001} = \frac{0.70711385222 - 0.70710678119}{0.00001} = \dots$
- $\frac{\ln(2 + 0.00005) - \ln(2)}{0.00005} = \frac{0.69317218025 - 0.69314718056}{0.00005} = \dots$

5. La pérdida de cifras significativas se puede evitar a veces reordenando los términos de la función usando una identidad conocida del álgebra o la trigonometría. Encuentre, en cada uno de los siguientes casos, una fórmula equivalente a la dada que evite la pérdida de cifras significativas.

- (a) $\ln(x+1) - \ln(x)$ para x grande.
- (b) $\sqrt{x^2 + 1} - x$ para x grande.
- (c) $\cos^2(x) - \sin^2(x)$ para $x \approx \pi/4$.
- (d) $\sqrt{\frac{1 + \cos(x)}{2}}$ para $x \approx \pi$.

6. Evaluación Polinomial. Sean

$$P(x) = x^3 - 3x^2 + 3x - 1, \quad Q(x) = ((x-3)x+3)x-1, \quad R(x) = (x-1)^3.$$

- (a) Usando aritmética en coma flotante con cuatro cifras y redondeo, calcule $P(2.72)$, $Q(2.72)$ y $R(2.72)$. En el cálculo de $P(x)$, suponga que $(2.72)^3 = 20.12$ y $(2.72)^2 = 7.398$.
 - (b) Usando aritmética en coma flotante con cuatro cifras y redondeo, calcule $P(0.975)$, $Q(0.975)$ y $R(0.975)$. En el cálculo de $P(x)$, suponga que $(0.975)^3 = 0.9268$ y $(0.975)^2 = 0.9506$.
7. Usando aritmética en coma flotante con tres cifras y redondeo, calcule las siguientes sumas (sumando en el orden que se indica):

$$(a) \sum_{k=1}^6 \frac{1}{3^k} \qquad (b) \sum_{k=1}^6 \frac{1}{3^{7-k}}$$

8. Discuta la propagación de los errores en las siguientes operaciones:

- (a) La suma de tres números:

$$p + q + r = (\hat{p} + \varepsilon_p) + (\hat{q} + \varepsilon_q) + (\hat{r} + \varepsilon_r).$$

(b) El cociente de dos números: $\frac{p}{q} = \frac{\hat{p} + \varepsilon_p}{\hat{q} + \varepsilon_q}$.

- (c) El producto de tres números:

$$pqr = (\hat{p} + \varepsilon_p)(\hat{q} + \varepsilon_q)(\hat{r} + \varepsilon_r).$$

9. Dados los desarrollos de Taylor

$$\frac{1}{1-h} = 1 + h + h^2 + h^3 + \mathcal{O}(h^4)$$

y

$$\cos(h) = 1 - \frac{h^2}{2!} + \frac{h^4}{4!} + \mathcal{O}(h^6).$$

Determine el orden de aproximación de su suma y de su producto.

- 10.** Dados los desarrollos de Taylor

$$e^h = 1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + \frac{h^4}{4!} + O(h^5)$$

y

$$\sin(h) = h - \frac{h^3}{3!} + O(h^5).$$

Determine el orden de aproximación de su suma y de su producto.

- 11.** Dados los desarrollos de Taylor

$$\cos(h) = 1 - \frac{h^2}{2!} + \frac{h^4}{4!} + O(h^6)$$

y

$$\sin(h) = h - \frac{h^3}{3!} + \frac{h^5}{5!} + O(h^7).$$

Determine el orden de aproximación de su suma y de su producto.

- 12.** La Fórmula Mejorada para la Resolución de la Ecuación de Segundo Grado.

Supongamos que $a \neq 0$ y que $b^2 - 4ac > 0$ y consideremos la ecuación $ax^2 + bx + c = 0$. Sus raíces pueden hallarse mediante la conocida fórmula

$$(i) \quad x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad \text{y} \quad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

Pruebe que estas raíces pueden calcularse mediante las fórmulas equivalentes

$$(ii) \quad x_1 = \frac{-2c}{b + \sqrt{b^2 - 4ac}} \quad \text{y} \quad x_2 = \frac{-2c}{b - \sqrt{b^2 - 4ac}}.$$

Indicación. Racionalice el numerador de (i). *Observación.* Cuando $|b| \approx \sqrt{b^2 - 4ac}$, hay que proceder con cuidado para evitar la pérdida de precisión por cancelación. Si $b > 0$, entonces x_1 debería ser calculado con la fórmula (ii) y x_2 debería ser calculado con la fórmula (i); mientras que, si $b < 0$, entonces x_1 debería ser calculado usando (i) y x_2 debería ser calculado usando (ii).

- 13.** Use la fórmula adecuada para calcular x_1 y x_2 , tal como se explica en el Ejercicio 12, para hallar las raíces de las siguientes ecuaciones de segundo grado.

- (a) $x^2 - 1000.001x + 1 = 0$
- (b) $x^2 - 10\,000.0001x + 1 = 0$
- (c) $x^2 - 100\,000.00001x + 1 = 0$
- (d) $x^2 - 1\,000\,000.000001x + 1 = 0$

Algoritmos y programas

1. Use los resultados de los Ejercicios 12 y 13 para construir un algoritmo y un programa en MATLAB que calcule las raíces de una ecuación cuadrática en todas las situaciones posibles, incluyendo los casos problemáticos cuando $|b| \approx \sqrt{b^2 - 4ac}$.
2. Siguiendo el Ejemplo 1.25, genere las diez primeras aproximaciones numéricas de cada una de las siguientes ecuaciones en diferencias. En cada caso se introduce un error pequeño; si no hubiera tal error, las tres ecuaciones en diferencias generarían la sucesión $\{1/2^n\}_{n=1}^{\infty}$. Presente sus resultados como en las Tablas 1.4 y 1.5 y las Figuras 1.8, 1.9 y 1.10.
 - (a) $r_0 = 0.994$ y $r_n = \frac{1}{2}r_{n-1}$, para $n = 1, 2, \dots$
 - (b) $p_0 = 1, p_1 = 0.497$ y $p_n = \frac{3}{2}p_{n-1} - p_{n-2}$, para $n = 2, 3, \dots$
 - (c) $q_0 = 1, q_1 = 0.497$ y $q_n = \frac{5}{2}q_{n-1} - q_{n-2}$, para $n = 2, 4, \dots$

2

Resolución de ecuaciones no lineales

Consideremos el problema físico de hallar la porción de una esfera de radio r que queda sumergida al meter la esfera en agua (véase la Figura 2.1). Supongamos que la esfera está construida con una variedad de pino que tiene una densidad de $\rho = 0.638 \text{ gr/cm}^3$ y que su radio mide $r = 10 \text{ cm}$. ¿Cuánto vale la profundidad d a la que está sumergido el polo sur de la esfera?

La masa M_a de agua desplazada cuando la esfera se sumerge es

$$M_a = \int_0^d \pi(r^2 - (x - r)^2) dx = \frac{\pi d^2(3r - d)}{3},$$

y la masa de la esfera es $M_e = 4\pi r^3 \rho / 3$. Aplicando el principio de Arquímedes $M_a = M_e$, obtenemos la siguiente ecuación que debemos resolver:

$$\frac{\pi(d^3 - 3d^2r + 4r^3\rho)}{3} = 0.$$

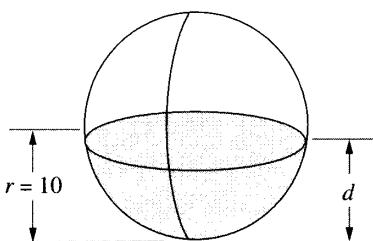
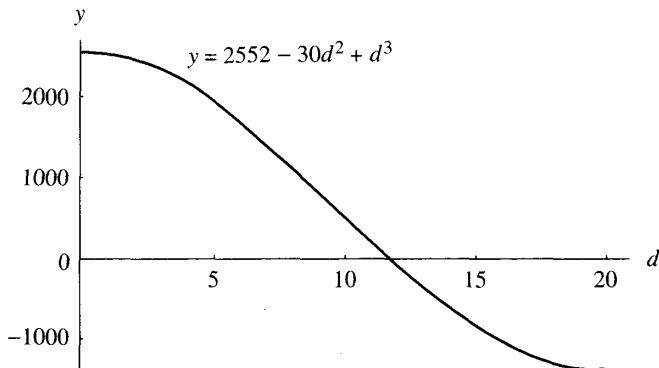


Figura 2.1 Porción de una esfera de radio r sumergida y profundidad d del polo sur.

Figura 2.2 La cúbica $y = 2552 - 30d^2 + d^3$.

En nuestro caso (con $r = 10$ y $\rho = 0.638$) la ecuación es

$$\frac{\pi(2552 - 30d^2 + d^3)}{3} = 0.$$

La gráfica del polinomio cúbico $y = 2552 - 30d^2 + d^3$ se muestra en la Figura 2.2 y en ella podemos ver que la solución está cerca de $d = 12$.

El objetivo de este capítulo es el desarrollo de una variedad de métodos que nos permitan calcular aproximaciones numéricicas a las raíces de una ecuación. Por ejemplo, podríamos usar el método de bisección para obtener las tres raíces $d_1 = -8.17607212$, $d_2 = 11.86150151$ y $d_3 = 26.31457061$. La primera solución d_1 no es una solución aceptable del problema porque d no puede ser negativo. La tercera solución d_3 es mayor que el diámetro de la esfera, así que no es la solución buscada. La raíz $d_2 = 11.86150151$ está en el intervalo $[0, 20]$ y es la solución adecuada. Su magnitud es razonable porque nos dice que sólo se sumerge un poco más de media esfera.

2.1 Métodos iterativos para resolver $x = g(x)$

Una técnica fundamental computación científica es la de *iteración*. Como su propio nombre sugiere, se trata de repetir un proceso hasta que se obtiene un resultado. Se usan métodos iterativos para hallar raíces de ecuaciones, soluciones de los sistemas lineales y no lineales y soluciones de ecuaciones diferenciales. En esta sección estudiaremos el proceso de iteración que consiste en sustituir repetidamente en una misma fórmula el valor previamente obtenido.

Necesitamos una regla, fórmula o función $g(x)$, con la que calcularemos los sucesivos términos, junto con un valor de partida p_0 . Lo que se produce es una

sucesión de valores $\{p_k\}$ obtenida mediante el proceso iterativo $p_{k+1} = g(p_k)$. La sucesión se ajusta al siguiente patrón

$$(1) \quad \begin{array}{ll} p_0 & (\text{valor de partida}) \\ p_1 = g(p_0) & \\ p_2 = g(p_1) & \\ \vdots & \\ p_k = g(p_{k-1}) & \\ p_{k+1} = g(p_k) & \\ \vdots & \end{array}$$

¿Qué nos dice una sucesión interminable de números como ésta? Si los números tienden a un límite, entonces podemos afirmar que tenemos algo entre manos. Pero, ¿qué ocurre si los números divergen o son periódicos? El ejemplo siguiente nos muestra una situación como ésta.

Ejemplo 2.1. El proceso iterativo $p_0 = 1$ y $p_{k+1} = 1.001p_k$ para $k = 0, 1, \dots$ produce una sucesión divergente; sus primeros y centésimo términos son

$$\begin{aligned} p_1 &= 1.001p_0 = (1.001)(1.000000) = 1.001000, \\ p_2 &= 1.001p_1 = (1.001)(1.001000) = 1.002001, \\ p_3 &= 1.001p_2 = (1.001)(1.002001) = 1.003003, \\ &\vdots \qquad \vdots \qquad \vdots \\ p_{100} &= 1.001p_{99} = (1.001)(1.104012) = 1.105116. \end{aligned}$$

Este proceso puede continuarse indefinidamente y es fácil probar que se verifica que $\lim_{n \rightarrow \infty} p_n = +\infty$. En el Capítulo 9 veremos que la sucesión $\{p_k\}$ es una solución numérica de la ecuación diferencial $y' = 0.001y$. Se sabe que la solución de esta ecuación es $y(x) = e^{0.001x}$. De hecho, si comparamos el término centésimo de la sucesión con $y(100)$, vemos que $p_{100} = 1.105116 \approx 1.105171 = e^{0.1} = y(100)$. ■

En esta sección estudiaremos qué tipos de funciones $g(x)$ producen sucesiones $\{p_k\}$ que convergen.

Puntos fijos

Definición 2.1 (Punto fijo). Un *punto fijo* de una función $g(x)$ es un número real P tal que $P = g(P)$. ▲

Geométricamente hablando, los puntos fijos de una función $g(x)$ son los puntos de intersección de la curva $y = g(x)$ con la recta $y = x$.

Definición 2.2 (Iteración de punto fijo). La iteración $p_{n+1} = g(p_n)$ para $n = 0, 1, \dots$ se llama *iteración de punto fijo*. ▲

Teorema 2.1. Supongamos que g es una función continua y que $\{p_n\}_{n=0}^{\infty}$ es una sucesión generada por iteración de punto fijo. Si $\lim_{n \rightarrow \infty} p_n = P$, entonces P es un punto fijo de $g(x)$.

Demostración. Si $\lim_{n \rightarrow \infty} p_n = P$, entonces $\lim_{n \rightarrow \infty} p_{n+1} = P$. Usando esto, la continuidad de g y la relación $p_{n+1} = g(p_n)$, deducimos que

$$(2) \quad g(P) = g\left(\lim_{n \rightarrow \infty} p_n\right) = \lim_{n \rightarrow \infty} g(p_n) = \lim_{n \rightarrow \infty} p_{n+1} = P.$$

Por tanto, P es un punto fijo de $g(x)$.

Ejemplo 2.2. Consideremos la iteración convergente

$$p_0 = 0.5 \quad \text{y} \quad p_{k+1} = e^{-p_k} \quad \text{para } k = 0, 1, \dots$$

Podemos calcular los diez primeros términos

$$p_1 = e^{-0.500000} = 0.606531$$

$$p_2 = e^{-0.606531} = 0.545239$$

$$p_3 = e^{-0.545239} = 0.579703$$

$$\vdots \qquad \vdots$$

$$p_9 = e^{-0.566409} = 0.567560$$

$$p_{10} = e^{-0.567560} = 0.566907.$$

La sucesión converge y, calculando un poco más, se tiene

$$\lim_{n \rightarrow \infty} p_n = 0.567143\dots$$

De esa manera, lo que hemos obtenido es una aproximación al punto fijo de la función $g(x) = e^{-x}$. ■

Los dos siguientes teoremas establecen condiciones para la existencia de un punto fijo y para la convergencia del proceso de iteración de punto fijo.

Teorema 2.2. Supongamos que $g \in C[a, b]$.

- (3) Si la imagen de la aplicación $y = g(x)$ verifica que $y \in [a, b]$ para cada punto $x \in [a, b]$, entonces g tiene un punto fijo en $[a, b]$.
- (4) Supongamos, además, que $g'(x)$ está definida en (a, b) y que $|g'(x)| < 1$ para todo $x \in (a, b)$, entonces g tiene un único punto fijo P en $[a, b]$.

Demostración de (3). Si $g(a) = a$ o $g(b) = b$, entonces la conclusión es cierta. Supongamos, entonces, que los valores $g(a)$ y $g(b)$ verifican que $g(a) \in (a, b]$ y que $g(b) \in [a, b)$. La función $f(x) \equiv x - g(x)$ tiene la siguiente propiedad

$$f(a) = a - g(a) < 0 \quad \text{y} \quad f(b) = b - g(b) > 0.$$

Aplicando el Teorema 1.2, el teorema del valor intermedio, a $f(x)$ y la constante $L = 0$, deducimos que existe un número P en (a, b) tal que $f(P) = 0$. Por tanto, $P = g(P)$ y P es el punto fijo de $g(x)$ que queríamos encontrar.

Demostración de (4). Ahora debemos probar que la solución es única. Supongamos, por el contrario, que existen dos puntos fijos distintos P_1 y P_2 . Aplicando el Teorema 1.6, el teorema del valor medio, deducimos que existe un número $d \in (a, b)$ tal que

$$(5) \qquad g'(d) = \frac{g(P_2) - g(P_1)}{P_2 - P_1}.$$

A continuación, usamos que $g(P_1) = P_1$ y que $g(P_2) = P_2$ para simplificar el miembro derecho de la relación (5) y obtener

$$g'(d) = \frac{P_2 - P_1}{P_2 - P_1} = 1,$$

lo que contradice la hipótesis hecha en (4) de que $|g'(x)| < 1$ en (a, b) . Así que no es posible que existan dos puntos fijos y, por tanto, bajo la condición dada en (4), $g(x)$ tiene un único punto fijo P en $[a, b]$. •

Ejemplo 2.3. Vamos a aplicar el Teorema 2.2 para probar rigurosamente que $g(x) = \cos(x)$ tiene un único punto fijo en $[0, 1]$.

Claramente, $g \in C[0, 1]$. Además, $g(x) = \cos(x)$ es una función decreciente en $[0, 1]$, por lo que su imagen $g([0, 1])$ es $[\cos(1), 1] \subseteq [0, 1]$. Esto prueba que se cumple la condición (3) del Teorema 2.2, así que g tiene un punto fijo en $[0, 1]$. Finalmente, si $x \in (0, 1)$, entonces $|g'(x)| = |-\operatorname{sen}(x)| = \operatorname{sen}(x) \leq \operatorname{sen}(1) \approx 0.8415 < 1$. Por consiguiente, la condición (4) del Teorema 2.2 se cumple y el punto fijo de g en $[0, 1]$ es único. ■

Vamos ahora a establecer el teorema que se suele usar para determinar si el proceso de iteración de punto fijo dado en (1) produce una sucesión convergente o divergente.

Teorema 2.3 (Teorema del punto fijo). Supongamos que (i) $g, g' \in C[a, b]$, (ii) K es una constante positiva, (iii) $p_0 \in (a, b)$ y (iv) $g(x) \in [a, b]$ para todo $x \in [a, b]$. Entonces hay un punto fijo P de g en $[a, b]$.

- (6) Si $|g'(x)| \leq K < 1$ para todo $x \in [a, b]$, entonces P es el único punto fijo de g en $[a, b]$ y la iteración $p_n = g(p_{n-1})$ converge a dicho punto fijo P . En este caso, se dice que P es un punto fijo atractivo.

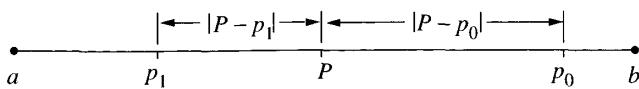


Figura 2.3 La relación entre P , p_0 , p_1 , $|P - p_0|$, y $|P - p_1|$.

- (7) Si $|g'(P)| > 1$ y $p_0 \neq P$ entonces la iteración $p_n = g(p_{n-1})$ no converge a P . En este caso, se dice que P es un punto fijo repulsivo y la iteración presenta divergencia local.

Observación 1. Como g' es continua en un intervalo al que pertenece P , podemos usar hipótesis más simples en (6): que $|g'(P)| < 1$ y que p_0 está suficientemente cerca de P .

Demostración. El apartado (3) del Teorema 2.2 y las hipótesis (i) y (iv) garantizan que g tiene un punto fijo P en $[a, b]$. Para demostrar (6) empecemos notando que, usando (3) y (4), las hipótesis (i)–(iv) implican, por el Teorema 2.2, que el punto fijo de g en $[a, b]$ es único. De (iii) y (iv) se deduce, por inducción, que los puntos $\{p_n\}_{n=0}^{\infty}$ están todos en $[a, b]$. Seguidamente, empezando con p_0 , aplicamos el Teorema 1.6, el teorema del valor medio, para deducir que existe un valor $c_0 \in (a, b)$ tal que

$$(8) \quad \begin{aligned} |P - p_1| &= |g(P) - g(p_0)| = |g'(c_0)(P - p_0)| \\ &= |g'(c_0)||P - p_0| \leq K|P - p_0| < |P - p_0|. \end{aligned}$$

Por tanto, p_1 está más cerca de P que p_0 (véase la Figura 2.3). Razonando de modo análogo tendremos, en general,

$$(9) \quad \begin{aligned} |P - p_n| &= |g(P) - g(p_{n-1})| = |g'(c_{n-1})(P - p_{n-1})| \\ &= |g'(c_{n-1})||P - p_{n-1}| \leq K|P - p_{n-1}| < |P - p_{n-1}|. \end{aligned}$$

Para completar la demostración de (6), probemos que

$$(10) \quad \lim_{n \rightarrow \infty} |P - p_n| = 0.$$

En primer lugar, establecemos por inducción la desigualdad

$$(11) \quad |P - p_n| \leq K^n |P - p_0|.$$

El caso $n = 1$ está ya comprobado en la relación (8). Usando la hipótesis de inducción $|P - p_{n-1}| \leq K^{n-1} |P - p_0|$ y la relación (9), obtenemos

$$|P - p_n| \leq K|P - p_{n-1}| \leq K K^{n-1} |P - p_0| = K^n |P - p_0|.$$

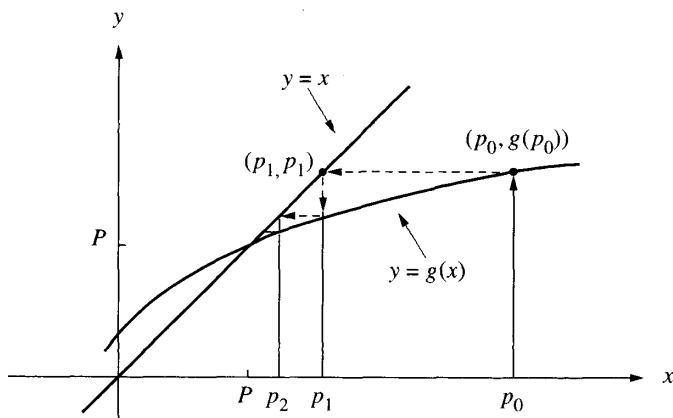


Figura 2.4 (a) Convergencia monótona cuando $0 < g'(P) < 1$.

En consecuencia, la desigualdad (11) se verifica, por inducción, para todo n . Puesto que $0 < K < 1$, el factor K^n converge a cero cuando n tiende a infinito y, por tanto,

$$(12) \quad 0 \leq \lim_{n \rightarrow \infty} |P - p_n| \leq \lim_{n \rightarrow \infty} K^n |P - p_0| = 0.$$

De aquí se deduce que $\lim_{n \rightarrow \infty} |P - p_n| = 0$, con lo cual $\lim_{n \rightarrow \infty} p_n = P$, lo que concluye la demostración del apartado (6) del Teorema 2.3.

Dejamos el enunciado (7) como ejercicio de investigación.

Corolario 2.1. Supongamos que g verifica las hipótesis dadas en el apartado (6) del Teorema 2.3. Entonces las siguientes desigualdades proporcionan cotas del error que se comete cuando usamos p_n como aproximación a P :

$$(13) \quad |P - p_n| \leq K^n |P - p_0| \quad \text{para todo } n \geq 1,$$

y

$$(14) \quad |P - p_n| \leq \frac{K^n |p_1 - p_0|}{1 - K} \quad \text{para todo } n \geq 1.$$

Interpretación gráfica de la iteración de punto fijo

Puesto que buscamos un punto fijo P de la función $g(x)$, es necesario que la curva $y = g(x)$ y la recta $y = x$ se corten en el punto (P, P) . Los dos tipos simples de iteración convergente, monótona y oscilante, se muestran en la Figura 2.4 (a) y (b), respectivamente.

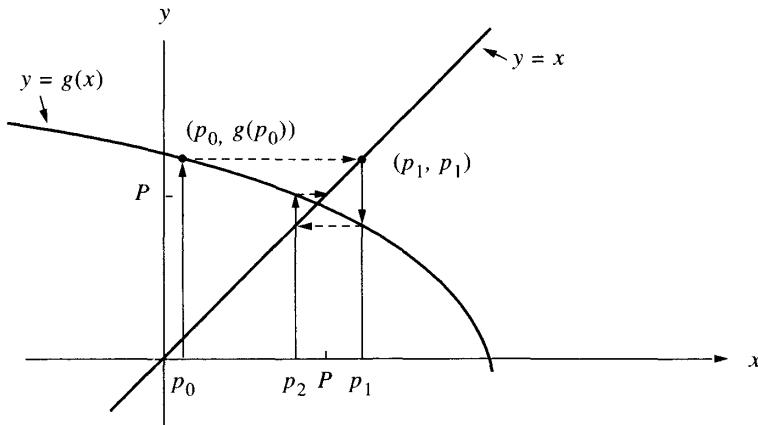


Figura 2.4 (b) Convergencia oscilante si $-1 < g'(P) < 0$.

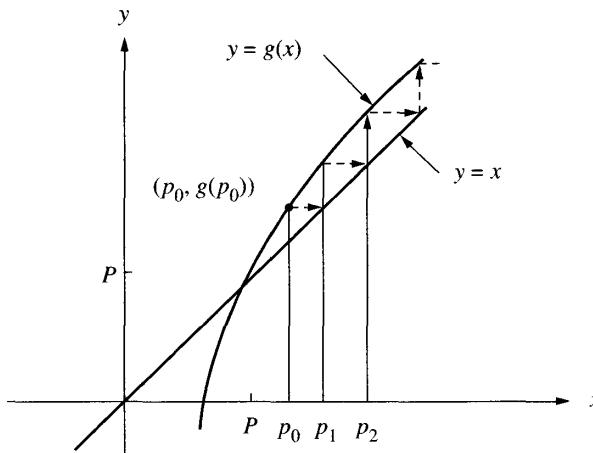


Figura 2.5 (a) Divergencia monótona si $1 < g'(P)$.

Visualicemos el proceso: empezamos en p_0 sobre el eje OX y nos movemos verticalmente hasta el punto $(p_0, p_1) = (p_0, g(p_0))$ que está sobre la curva $y = g(x)$. Entonces nos movemos horizontalmente desde el punto (p_0, p_1) hasta el punto (p_1, p_1) sobre la recta $y = x$. Finalmente nos movemos otra vez verticalmente hasta p_1 sobre el eje OX . Usamos la recursión $p_{n+1} = g(p_n)$ para construir el punto (p_n, p_{n+1}) sobre la curva, entonces un movimiento horizontal nos lleva al punto (p_{n+1}, p_{n+1}) sobre la recta $y = x$ y un movimiento vertical termina ahora en p_{n+1} sobre el eje OX . La situación se muestra en la Figura 2.4.

Si $|g'(P)| > 1$, entonces la iteración $p_{n+1} = g(p_n)$ produce una sucesión que se aleja de P . Los dos tipos simples de iteración divergente, monótona y oscilante, se muestran en la Figura 2.5 (a) y (b), respectivamente.

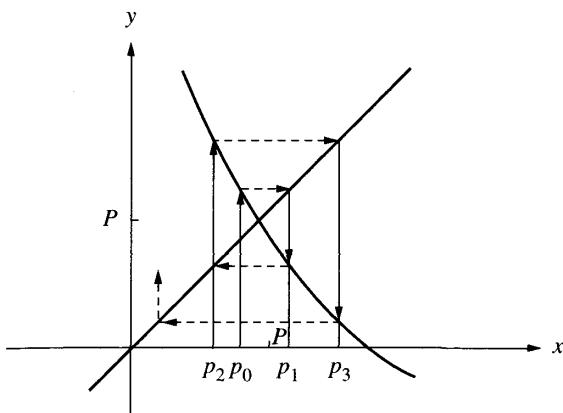


Figura 2.5 (b) Divergencia oscilante si $g'(P) < -1$.

Ejemplo 2.4. Consideremos la iteración $p_{n+1} = g(p_n)$ cuando la función viene dada por $g(x) = 1 + x - x^2/4$. Los puntos fijos pueden determinarse resolviendo la ecuación $x = g(x)$, cuyas soluciones (los puntos fijos de g) son $x = -2$ y $x = 2$. La derivada de la función es $g'(x) = 1 - x/2$ y los dos casos que tenemos que considerar son:

<i>Caso (i):</i>	$P = -2$
Inicio	$p_0 = -2.05$
	$p_1 = -2.100625$
	$p_2 = -2.20378135$
	$p_3 = -2.41794441$
⋮	
	$\lim_{n \rightarrow \infty} p_n = -\infty.$

Como $|g'(x)| > \frac{3}{2}$ en $[-3, -1]$, por el Teorema 2.3, la sucesión no converge a $P = -2$.

<i>Caso (ii):</i>	$P = 2$
Inicio	$p_0 = 1.6$
	$p_1 = 1.96$
	$p_2 = 1.9996$
	$p_3 = 1.99999996$
⋮	
	$\lim_{n \rightarrow \infty} p_n = 2.$

Como $|g'(x)| < \frac{1}{2}$ en $[1, 3]$, por el Teorema 2.3, la sucesión converge a $P = 2$.

El Teorema 2.3 no dice qué ocurrirá si $|g'(P)| = 1$. El siguiente ejemplo se ha construido especialmente para que $\{p_n\}$ converja cuando $p_0 > P$ y diverja cuando $p_0 < P$.

Ejemplo 2.5. Consideremos la iteración $p_{n+1} = g(p_n)$ cuando la función viene dada por $g(x) = 2(x-1)^{1/2}$ para $x \geq 1$. Esta función sólo tiene un punto fijo $P = 2$ y su derivada es $g'(x) = 1/(x-1)^{1/2}$, de manera que $g'(2) = 1$ y no podemos aplicar el Teorema 2.3. Consideraremos los dos casos según que el punto de partida esté a la derecha o a la izquierda de $P = 2$.

Caso (i): Inicio $p_0 = 1.5$,

$$p_1 = 1.41421356$$

$$p_2 = 1.28718851$$

$$p_3 = 1.07179943$$

$$p_4 = 0.53590832$$

$$p_5 = 2(-0.46409168)^{1/2}.$$

Como p_4 está fuera del dominio de $g(x)$, el término p_5 no puede calcularse.

Caso (ii): Inicio $p_0 = 2.5$,

$$p_1 = 2.44948974$$

$$p_2 = 2.40789513$$

$$p_3 = 2.37309514$$

$$p_4 = 2.34358284$$

 \vdots

$$\lim_{n \rightarrow \infty} p_n = 2.$$

Esta sucesión converge muy lentamente a $P = 2$; de hecho, $P_{1000} = 2.00398714$.

Consideraciones sobre el error absoluto y el relativo

En el Ejemplo 2.5, caso (ii), la sucesión converge muy lentamente; después de 1000 iteraciones los tres siguientes términos son

$$p_{1000} = 2.00398714, \quad p_{1001} = 2.00398317 \quad \text{y} \quad p_{1002} = 2.00397921.$$

Esto no debería preocuparnos; después de todo, ¡podríamos calcular unos pocos miles de términos más para encontrar una aproximación mejor! Pero, ¿cuál es el criterio para detener una iteración? Si usamos la diferencia entre dos términos consecutivos, obtenemos

$$|p_{1001} - p_{1002}| = |2.00398317 - 2.00397921| = 0.00000396.$$

Sin embargo, sabemos que el error absoluto de la aproximación p_{1000} es

$$|P - p_{1000}| = |2.00000000 - 2.00398714| = 0.00398714,$$

que resulta ser unas mil veces mayor que $|p_{1001} - p_{1002}|$. Esto prueba que la cercanía entre dos términos consecutivos no garantiza que hayamos obtenido una precisión alta; sin embargo, éste (o bien el error relativo) es usualmente el único criterio del que se dispone y, a menudo, es el que se usa para determinar cuándo debe detenerse un proceso iterativo.

Programa 2.1 (Iteración de punto fijo). Aproximación a una solución de la ecuación $x = g(x)$ mediante la iteración $p_{n+1} = g(p_n)$ realizada a partir de una aproximación inicial p_0 .

```
function [k,p,err,P]=fixpt(g,p0,tol,max1)
% Datos
%      - g es la función de iteración
%      - p0 es el punto de partida
```

```

% - tol es la tolerancia
% - max1 es el número máximo de iteraciones
% Resultados
% - k es el número de iteraciones realizadas
% - p es la aproximación al punto fijo
% - err es la diferencia entre dos términos consecutivos
% - P es la sucesión {pn} completa

P(1)= p0;
for k=2:max1
    P(k)=feval(g,P(k-1));
    err=abs(P(k)-P(k-1));
    relerr=err/(abs(P(k))+eps);
    p=P(k);
    if (err<tol) | (relerr<tol),break;end
end
if k == max1
    disp('se ha excedido el número máximo de iteraciones')
end
P=P';

```

Observación. Para usar este programa fixpt, es necesario introducir la función de iteración $g(x)$ como un archivo, por ejemplo g.m, que pueda ser llamado por fixpt como la cadena de caracteres 'g' (véase el Apéndice sobre el paquete de programas MATLAB).

Ejercicios

- Determine rigurosamente si cada una de las siguientes funciones tiene un único punto fijo en el intervalo dado (siga el Ejemplo 2.3).
 - $g(x) = 1 - x^2/4$ en $[0, 1]$
 - $g(x) = 2^{-x}$ en $[0, 1]$
 - $g(x) = 1/x$ en $[0.5, 5.2]$

- Investigue la naturaleza de la iteración de punto fijo cuando

$$g(x) = -4 + 4x - \frac{1}{2}x^2.$$

- Resuelva $g(x) = x$ y pruebe que $P = 2$ y $P = 4$ son puntos fijos.
- Tome como valor inicial $p_0 = 1.9$ y calcule p_1, p_2 y p_3 .
- Tome como valor inicial $p_0 = 3.8$ y calcule p_1, p_2 y p_3 .
- Halle los errores absolutos E_k y los errores relativos R_k de los valores p_k obtenidos en los apartados (b) y (c).
- ¿Qué conclusiones pueden deducirse usando el Teorema 2.3?

3. En cada uno de los siguientes casos, dibuje la gráfica de $g(x)$, la recta $y = x$ y el punto fijo dado P en un mismo sistema de coordenadas. Usando el valor inicial dado p_0 , calcule p_1 y p_2 y construya figuras similares a las Figuras 2.4 y 2.5. Basándose en su dibujo, determine geométricamente si la iteración de punto fijo correspondiente converge.
- $g(x) = (6 + x)^{1/2}$, $P = 3$ y $p_0 = 7$
 - $g(x) = 1 + 2/x$, $P = 2$ y $p_0 = 4$
 - $g(x) = x^2/3$, $P = 3$ y $p_0 = 3.5$
 - $g(x) = -x^2 + 2x + 2$, $P = 2$ y $p_0 = 2.5$
4. Sea $g(x) = x^2 + x - 4$. ¿Podemos utilizar iteración de punto fijo para hallar las soluciones de la ecuación $x = g(x)$? ¿Por qué?
5. Sea $g(x) = x \cos(x)$. Resuelva $x = g(x)$ y encuentre todos los puntos fijos de g (hay infinitos). ¿Podemos utilizar iteración de punto fijo para hallar las soluciones de la ecuación $x = g(x)$? ¿Por qué?
6. Supongamos que $g(x)$ y $g'(x)$ están definidas y son continuas en (a, b) y sean $p_0, p_1, p_2 \in (a, b)$ con $p_1 = g(p_0)$ y $p_2 = g(p_1)$. Supongamos también que existe una constante K tal que $|g'(x)| < K$. Pruebe que $|p_2 - p_1| < K|p_1 - p_0|$. *Indicación.* Use el teorema del valor medio
7. Supongamos que $g(x)$ y $g'(x)$ son continuas en (a, b) y que $|g'(x)| > 1$ en este intervalo. Pruebe que si un punto fijo P y las aproximaciones iniciales p_0 y $p_1 = g(p_0)$ están en (a, b) , entonces $|E_1| = |P - p_1| > |P - p_0| = |E_0|$. Esto probaría el apartado (7) del Teorema 2.3 (divergencia local).
8. Consideraremos el proceso de iteración de punto fijo con $g(x) = -0.0001x^2 + x$ y $p_0 = 1$.
 - Pruebe que $p_0 > p_1 > \dots > p_n > p_{n+1} > \dots$
 - Pruebe que $p_n > 0$ para todo n .
 - Puesto que la sucesión $\{p_n\}$ es decreciente y está acotada inferiormente, tiene límite. ¿Cuál es ese límite?
9. Consideraremos el proceso de iteración de punto fijo con $g(x) = 0.5x + 1.5$ y $p_0 = 4$.
 - Pruebe que el punto fijo es $P = 3$.
 - Pruebe que $|P - p_n| = |P - p_{n-1}|/2$ para $n = 1, 2, 3, \dots$
 - Pruebe que $|P - p_n| = |P - p_0|/2^n$ para $n = 1, 2, 3, \dots$
10. Consideraremos el proceso de iteración de punto fijo con $g(x) = x/2$.
 - Halle lo que vale $|p_{k+1} - p_k|/|p_{k+1}|$.
 - Discuta lo que ocurriría si usáramos el error relativo como único criterio de parada en el Programa 2.1.
11. Razone por qué es una ventaja tener $g'(P) \approx 0$ en un proceso de iteración de punto fijo.

Algoritmos y programas

1. Use el Programa 2.1 para aproximar los puntos fijos (si es que hay alguno) de cada una de las siguientes funciones. Las respuestas deben tener 12 cifras decimales exactas. Dibuje además una gráfica de cada función y de la recta $y = x$ que muestre claramente los puntos fijos que haya.
- $g(x) = x^5 - 3x^3 - 2x^2 + 2$
 - $g(x) = \cos(\operatorname{sen}(x))$
 - $g(x) = x^2 - \operatorname{sen}(x + 0.15)$
 - $g(x) = x^{x-\cos(x)}$

2.2 Los métodos de localización de raíces

Consideremos un tema de interés familiar. Supongamos que ahorraremos dinero haciendo depósitos mensuales de P euros a una tasa de interés anual I que se compone cada mes; entonces la cantidad total A de euros acumulada después de N depósitos es

$$(1) \quad A = P + P \left(1 + \frac{I}{12}\right) + P \left(1 + \frac{I}{12}\right)^2 + \cdots + P \left(1 + \frac{I}{12}\right)^{N-1}.$$

El primer término del miembro derecho de la expresión (1) es el último depósito; el segundo es el penúltimo depósito cuya contribución, habiendo acumulado un período de interés, es $P \left(1 + \frac{I}{12}\right)$ euros; el antepenúltimo depósito ha acumulado dos períodos de interés y su contribución al total es de $P \left(1 + \frac{I}{12}\right)^2$ euros, y así sucesivamente. Finalmente, el primer depósito, que ha acumulado interés durante $N-1$ períodos, contribuye con $P \left(1 + \frac{I}{12}\right)^{N-1}$ euros al total. Recordemos la fórmula de la suma de los N términos de una progresión geométrica:

$$(2) \quad 1 + r + r^2 + r^3 + \cdots + r^{N-1} = \frac{1 - r^N}{1 - r}.$$

Escribiendo (1) como

$$A = P \left(1 + \left(1 + \frac{I}{12}\right) + \left(1 + \frac{I}{12}\right)^2 + \cdots + \left(1 + \frac{I}{12}\right)^{N-1}\right)$$

y sustituyendo $r = (1 + I/12)$ en (2) obtenemos

$$A = P \frac{1 - (1 + \frac{I}{12})^N}{1 - (1 + \frac{I}{12})},$$

que podemos simplificar para obtener la fórmula del capital acumulado cuando el interés se compone mensualmente:

$$(3) \quad A = \frac{P}{I/12} \left(\left(1 + \frac{I}{12} \right)^N - 1 \right).$$

Usaremos esta fórmula en el siguiente ejemplo, donde tendremos que usarla varias veces para hallar la respuesta.

Ejemplo 2.6. Ahorramos 250 euros al mes durante 20 años y deseamos que el capital acumulado al final de estos 20 años sea de 250 000 euros. ¿A qué tasa de interés I debemos invertir el dinero para obtener ese resultado? Si fijamos $N = 240$, entonces A es función únicamente de I ; o sea, $A = A(I)$. Empezaremos con dos tentativas, $I_0 = 0.12$ e $I_1 = 0.13$, y realizaremos una sucesión de cálculos para ir estrechando el margen de la respuesta. Empezando con $I_0 = 0.12$ obtenemos

$$A(0.12) = \frac{250}{0.12/12} \left(\left(1 + \frac{0.12}{12} \right)^{240} - 1 \right) = 247\,314.$$

Puesto que este valor es menor que el deseado, probamos con $I_1 = 0.13$:

$$A(0.13) = \frac{250}{0.13/12} \left(\left(1 + \frac{0.13}{12} \right)^{240} - 1 \right) = 282\,311.$$

Éste es un poco alto, así que probamos con la media de ambos $I_2 = 0.125$:

$$A(0.125) = \frac{250}{0.125/12} \left(\left(1 + \frac{0.125}{12} \right)^{240} - 1 \right) = 264\,623.$$

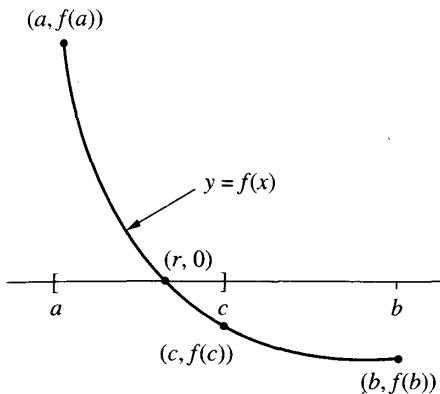
Éste es, de nuevo, un poco alto, por lo que deducimos que el valor deseado debe estar en el intervalo $[0.12, 0.125]$ y probamos, ahora, con su punto medio $I_3 = 0.1225$:

$$A(0.1225) = \frac{250}{0.1225/12} \left(\left(1 + \frac{0.1225}{12} \right)^{240} - 1 \right) = 255\,803.$$

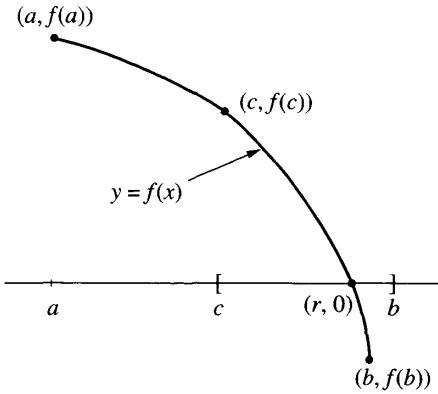
Este valor es mayor que lo que deseamos, así que ahora estrechamos el intervalo a $[0.12, 0.1225]$ y usamos su punto medio $I_4 = 0.12125$ para nuestro último cálculo:

$$A(0.12125) = \frac{250}{0.12125/12} \left(\left(1 + \frac{0.12125}{12} \right)^{240} - 1 \right) = 251\,518.$$

Podemos seguir realizando iteraciones hasta obtener tantas cifras significativas de la solución como deseemos. El propósito en este ejemplo era encontrar el valor de I para el que la función $A(I)$ es igual a un cierto valor especificado L , o sea, hallar una solución de la ecuación $A(I) = L$. La práctica habitual es pasar L al primer miembro y resolver la ecuación $A(I) - L = 0$.



(a) Si $f(a)$ y $f(c)$ tienen signos opuestos, entonces se recorta por la derecha.



(b) Si $f(c)$ y $f(b)$ tienen signos opuestos, entonces se recorta por la izquierda.

Figura 2.6 El proceso de decisión en el método de bisección.

Definición 2.3 (Raíz de una ecuación, cero de una función). Supongamos que $f(x)$ es una función continua. Cualquier número r tal que $f(r) = 0$ se llama **raíz de la ecuación** $f(x) = 0$; también se dice que r es un **cero de la función** $f(x)$. ▲

Por ejemplo, la ecuación $2x^2 + 5x - 3 = 0$ tiene dos raíces reales $r_1 = 0.5$ y $r_2 = -3$, mientras que la función correspondiente $f(x) = 2x^2 + 5x - 3 = (2x - 1)(x + 3)$ tiene dos ceros reales $r_1 = 0.5$ y $r_2 = -3$.

El método de bisección de Bolzano

En esta sección desarrollamos nuestro primer método de localización para hallar ceros de funciones continuas. Debemos empezar con un intervalo de partida $[a, b]$ en el que $f(a)$ y $f(b)$ tengan distinto signo. Entonces, por el Teorema 1.2, el teorema del valor intermedio, la gráfica $y = f(x)$ cruzará el eje OX en un cero $x = r$ que está en dicho intervalo (véase la Figura 2.6). El método de bisección consiste en ir acercando sistemáticamente los extremos del intervalo hasta que obtengamos un intervalo de anchura suficientemente pequeña en el que se localiza un cero. El proceso de decisión para subdividir el intervalo consiste en tomar el punto medio del intervalo $c = (a + b)/2$ y luego analizar las tres posibilidades que pueden darse:

- (4) Si $f(a)$ y $f(c)$ tienen signos opuestos, entonces hay un cero en $[a, c]$.
- (5) Si $f(c)$ y $f(b)$ tienen signos opuestos, entonces hay un cero en $[c, b]$.
- (6) Si $f(c) = 0$, entonces c es un cero.

Si ocurre bien el caso (4), bien el (5), (supondremos de hecho que, como sucede en la mayoría de las aplicaciones prácticas, el caso (6) no se da) entonces hemos encontrado un intervalo la mitad de ancho que el original que contiene una raíz (véase la Figura 2.6). Para continuar el proceso, renombramos el nuevo intervalo más pequeño también como $[a, b]$ y repetimos el proceso hasta que el intervalo sea tan pequeño como deseemos. Puesto que el proceso de bisección genera una sucesión de intervalos encajados, con sus correspondientes puntos medios, usaremos la siguiente notación para tener un registro de los detalles del proceso:

- (7)
- $[a_0, b_0]$ es el intervalo de partida y $c_0 = \frac{a_0+b_0}{2}$ es su punto medio.
 - $[a_1, b_1]$ es el segundo intervalo en el que se localiza un cero y c_1 es su punto medio; el intervalo $[a_1, b_1]$ es la mitad de ancho que $[a_0, b_0]$.
 - Después de llegar al intervalo $[a_n, b_n]$, en el que también se localiza un cero y cuyo punto medio es c_n , se construye el intervalo $[a_{n+1}, b_{n+1}]$, en el que también se sigue localizando un cero, y que mide la mitad que $[a_n, b_n]$.

Dejamos como ejercicio (véase el Ejercicio 8) el demostrar que la sucesión de los extremos izquierdos de estos intervalos es creciente y que la sucesión de los extremos derechos es decreciente; es decir,

$$(8) \quad a_0 \leq a_1 \leq \cdots \leq a_n \leq \cdots \leq b_n \leq \cdots \leq b_1 \leq b_0,$$

donde $c_n = \frac{a_n+b_n}{2}$ y si $f(a_{n+1})f(b_{n+1}) < 0$, entonces

$$(9) \quad [a_{n+1}, b_{n+1}] = [a_n, c_n] \quad \text{o bien} \quad [a_{n+1}, b_{n+1}] = [c_n, b_n] \quad \text{para cada } n.$$

Teorema 2.4 (Convergencia del método de bisección). Supongamos que $f \in C[a, b]$ y que $f(a)$ y $f(b)$ tienen signos distintos. Sea $\{c_n\}_{n=0}^{\infty}$ la sucesión de puntos medios de los intervalos generados por el método de bisección dado en (8) y (9). Entonces existe un número $r \in [a, b]$ tal que $f(r) = 0$ y, además,

$$(10) \quad |r - c_n| \leq \frac{b-a}{2^{n+1}} \quad \text{para } n = 0, 1, \dots,$$

en particular, la sucesión $\{c_n\}_{n=0}^{\infty}$ converge al cero $x = r$; esto es,

$$(11) \quad \lim_{n \rightarrow \infty} c_n = r.$$

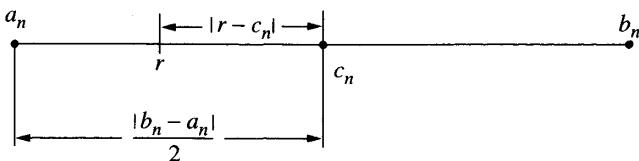


Figura 2.7 La raíz r y el punto medio c_n de $[a_n, b_n]$ en el método de bisección.

Demostración. Observemos que las sucesivas anchuras de los intervalos se ajustan al siguiente patrón:

$$\begin{aligned} b_1 - a_1 &= \frac{b_0 - a_0}{2^1}, \\ b_2 - a_2 &= \frac{b_1 - a_1}{2} = \frac{b_0 - a_0}{2^2}. \end{aligned}$$

Dejamos como ejercicio el probar, usando inducción matemática, que

$$(12) \quad b_n - a_n = \frac{b_0 - a_0}{2^n}.$$

Ahora, la sucesión (a_n) es creciente y está acotada superiormente, luego tiene límite a ; la sucesión (b_n) es decreciente y está acotada inferiormente, luego tiene límite b . Tomando límite cuando $n \rightarrow \infty$ en (12) deducimos que a y b deben coincidir; sea $r = a = b$ (véase el Ejercicio 8). Por continuidad tenemos que $f(r) = 0$, luego r es un cero de f que está en cada intervalo $[a_n, b_n]$. Como el punto medio c_n también está en el intervalo $[a_n, b_n]$, la distancia entre c_n y r no puede ser mayor que la mitad de la anchura de dicho intervalo (véase la Figura 2.7). En consecuencia,

$$(13) \quad |r - c_n| \leq \frac{b_n - a_n}{2} \quad \text{para todo } n.$$

Combinando (12) y (13), resulta

$$(14) \quad |r - c_n| \leq \frac{b_0 - a_0}{2^{n+1}} \quad \text{para todo } n.$$

Un argumento parecido al dado en el Teorema 2.3 muestra que (14) implica que la sucesión $\{c_n\}_{n=0}^{\infty}$ converge a r , lo que completa la demostración. •

Ejemplo 2.7. La función $h(x) = x \operatorname{sen}(x)$ aparece en el estudio de vibraciones forzadas no amortiguadas. Hay que hallar el valor de x que está dentro del intervalo $[0, 2]$ y en el que la función vale $h(x) = 1$ (el ángulo x en la función $\operatorname{sen}(x)$ se mide

Tabla 2.1 Resolución de $x \operatorname{sen}(x) - 1 = 0$ por el método de bisección.

k	Extremo izquierdo, a_k	Punto medio, c_k	Extremo derecho, b_k	Valor de la función, $f(c_k)$
0	0	1.	2.	-0.158529
1	1.0	1.5	2.0	0.496242
2	1.00	1.25	1.50	0.186231
3	1.000	1.125	1.250	0.015051
4	1.0000	1.0625	1.1250	-0.071827
5	1.06250	1.09375	1.12500	-0.028362
6	1.093750	1.109375	1.125000	-0.006643
7	1.1093750	1.1171875	1.1250000	0.004208
8	1.10937500	1.11328125	1.11718750	-0.001216
:	:	:	:	:

en radianes). Usamos el método de bisección para hallar un cero de la función $f(x) = x \operatorname{sen}(x) - 1$. Empezando con $a_0 = 0$ y $b_0 = 2$, calculamos

$$f(0) = -1.000000 \quad \text{y} \quad f(2) = 0.818595,$$

de manera que hay una raíz de $f(x) = 0$ en el intervalo $[0, 2]$. En el punto medio $c_0 = 1$ tenemos que $f(1) = -0.158529$, luego la función cambia de signo en el intervalo $[c_0, b_0] = [1, 2]$.

Para seguir, recortamos el intervalo por la izquierda y ponemos $a_1 = c_0$ y $b_1 = b_0$. El nuevo punto medio es $c_1 = 1.5$ y se tiene $f(c_1) = 0.496242$. Puesto que $f(1) = -0.158529$ y $f(1.5) = 0.496242$, la raíz está en $[a_1, c_1] = [1.0, 1.5]$ y la siguiente decisión es recortar por la derecha y poner $a_2 = a_1$ y $b_2 = c_1$. De esta forma obtenemos una sucesión $\{c_k\}$ que converge a $r \approx 1.114157141$. Vemos los cálculos de los ocho primeros pasos en la Tabla 2.1.

Una de las virtudes del método de bisección es que la fórmula proporciona una estimación predeterminada de la precisión de la solución calculada. En el Ejemplo 2.7 la anchura del intervalo inicial era $b_0 - a_0 = 2$. Supongamos que hubiéramos continuado la Tabla 2.1 hasta la iteración trigésimo primera; entonces, por (10), la cota del error sería $|E_{31}| \leq (2-0)/2^{32} \approx 4.656613 \times 10^{-10}$. Por tanto, c_{31} sería una aproximación a r con nueve cifras decimales de precisión. El número N de bisecciones sucesivas que nos garantizaría que el punto medio c_N es una aproximación a un cero con un error menor que un valor prefijado δ es

$$(15) \quad N = \operatorname{ent} \left(\frac{\ln(b-a) - \ln(\delta)}{\ln(2)} \right)$$

(donde $\text{ent}(x)$ denota la parte entera de un número x). La demostración de esta fórmula queda como ejercicio.

Otro algoritmo popular es el *método de la régula falsa* o *método de la posición falsa*. Una de las razones de su introducción es que la velocidad de convergencia del método de bisección es bastante baja. Como antes, supongamos que $f(a)$ y $f(b)$ tienen distinto signo. En el método de bisección se usa el punto medio del intervalo $[a, b]$ para llevar a cabo el siguiente paso. Suele conseguirse una aproximación mejor usando el punto $(c, 0)$ en el que la recta secante L que pasa por los puntos $(a, f(a))$ y $(b, f(b))$ cruza el eje OX (véase la Figura 2.8). Para hallar el punto c , igualamos dos fórmulas para la pendiente m de la recta L :

$$(16) \quad m = \frac{f(b) - f(a)}{b - a},$$

que resulta de usar los puntos $(a, f(a))$ y $(b, f(b))$, y

$$(17) \quad m = \frac{0 - f(b)}{c - b},$$

que resulta de usar los puntos $(c, 0)$ y $(b, f(b))$.

Igualando las pendientes que aparecen en (16) y (17), tenemos

$$\frac{f(b) - f(a)}{b - a} = \frac{0 - f(b)}{c - b},$$

de donde c puede despejarse fácilmente, obteniéndose

$$(18) \quad c = b - \frac{f(b)(b - a)}{f(b) - f(a)}.$$

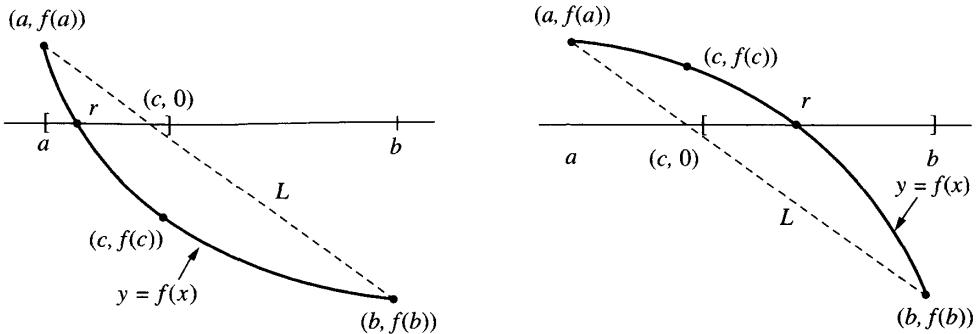
Las tres posibilidades son las mismas que antes:

- (19) Si $f(a)$ y $f(c)$ tienen distinto signo, entonces hay un cero en $[a, c]$.
- (20) Si $f(c)$ y $f(b)$ tienen distinto signo, entonces hay un cero en $[c, b]$.
- (21) Si $f(c) = 0$, entonces c es un cero de f .

Convergencia del método de la régula falsa

La fórmula (18) junto con el proceso de decisión descrito en (19) y (20) se usa para construir una sucesión de intervalos $\{[a_n, b_n]\}$ cada uno de los cuales contiene un cero. En cada paso la aproximación al cero obtenida es:

$$(22) \quad c_n = b_n - \frac{f(b_n)(b_n - a_n)}{f(b_n) - f(a_n)},$$



(a) Si $f(a)$ y $f(c)$ tienen signos opuestos, entonces se recorta por la derecha.

(b) Si $f(c)$ y $f(b)$ tienen signos opuestos, entonces se recorta por la izquierda.

Figura 2.8 El proceso de decisión para el método de la *régula falsi*.

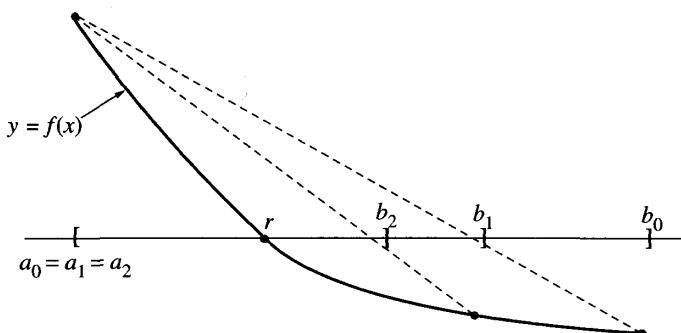


Figura 2.9 Un extremo estacionario en el método de la *régula falsi*.

y puede probarse que la sucesión $\{c_n\}$ converge a un cero r de la función. Sin embargo, hay que prestar atención al siguiente hecho: aunque la anchura del intervalo $b_n - a_n$ se hace más pequeña, es posible que no tienda a cero; si la curva $y = f(x)$ es convexa cerca de $(r, 0)$, entonces uno de los extremos a_n o b_n permanece estacionario y el otro tiende a la solución (véase la Figura 2.9).

Vamos a hallar la solución de $x \operatorname{sen}(x) - 1 = 0$ usando el método de la *régula falsi*; observaremos que converge más rápidamente que el de bisección y veremos también que $\{b_n - a_n\}_{n=0}^{\infty}$ no tiende a cero.

Ejemplo 2.8. Vamos a usar el método de la *régula falsi* para hallar la raíz de $x \operatorname{sen}(x) - 1 = 0$ que está en el intervalo $[0, 2]$ (de nuevo con el ángulo x medido en radianes).

Tabla 2.2 Resolución de $x \operatorname{sen}(x) - 1 = 0$ por el método de la *régula falsi*.

k	Extremo izquierdo, a_k	Punto intermedio, c_k	Extremo derecho, b_k	Valor de la función, $f(c_k)$
0	0.00000000	1.09975017	2.00000000	-0.02001921
1	1.09975017	1.12124074	2.00000000	0.00983461
2	1.09975017	1.11416120	1.12124074	0.00000563
3	1.09975017	1.11415714	1.11416120	0.00000000

Empezando con $a_0 = 0$ y $b_0 = 2$, tenemos $f(0) = -1.00000000$ y $f(2) = 0.81859485$, de manera que hay una raíz en $[0, 2]$. Usando la fórmula (22), tenemos

$$c_0 = 2 - \frac{0.81859485(2 - 0)}{0.81859485 - (-1)} = 1.09975017 \quad \text{y} \quad f(c_0) = -0.02001921.$$

La función cambia de signo en el intervalo $[c_0, b_0] = [1.09975017, 2]$, así que recortamos por la derecha y ponemos $a_1 = c_0$ y $b_1 = b_0$. Usamos otra vez la fórmula (22) para hallar la siguiente aproximación:

$$c_1 = 2 - \frac{0.81859485(2 - 1.09975017)}{0.81859485 - (-0.02001921)} = 1.12124074$$

y

$$f(c_1) = 0.00983461.$$

Ahora $f(x)$ cambia de signo en $[a_1, c_1] = [1.09975017, 1.12124074]$, así que la siguiente decisión es recortar por la derecha y poner $a_2 = a_1$ y $b_2 = c_1$. Estos cálculos se recogen en la Tabla 2.2. ■

El criterio de parada usado en el método de bisección no es útil para el método de la *régula falsi* porque podría producir iteraciones sin fin. En este caso se usan tanto la cercanía entre sí de dos aproximaciones sucesivas $|c_n - c_{n-1}|$ como el tamaño $|f(c_n)|$; así se hace en el Programa 2.3. En la sección 2.3 discutiremos las razones de esta elección.

Programa 2.2 (Método de bisección). Aproximación a una raíz de la ecuación $f(x) = 0$ en el intervalo $[a, b]$. Puede usarse sólo si $f(x)$ es continua y $f(a)$ y $f(b)$ tienen distinto signo.

```
function [c,err,yc]=bisect(f,a,b,delta)
% Datos
% - f es la función, introducida como
% una cadena de caracteres 'f'
% - a y b son el extremo izquierdo y el extremo derecho
```

```

% - delta es la tolerancia
% Resultados
% - c es el cero
% - yc=f(c)
% - err es el error estimado de la aproximación a c

ya=feval(f,a);
yb=feval(f,b);
if ya*yb>0, break, end
max1=1+round((log(b-a)-log(delta))/log(2));
for k=1:max1
    c=(a+b)/2;
    yc=feval(f,c);
    if yc==0
        a=c;
        b=c;
    elseif yb*yc>0
        b=c;
        yb=yc;
    else
        a=c;
        ya=yc;
    end
    if b-a < delta, break, end
end
c=(a+b)/2;
err=abs(b-a);
yc=feval(f,c);

```

Programa 2.3 (Método de la régula falsi o de la posición falsa). Aproximación a una raíz de la ecuación $f(x) = 0$ en el intervalo $[a, b]$. Puede usarse sólo si $f(x)$ es continua y $f(a)$ y $f(b)$ tienen distinto signo.

```

function [c,err,yc]=regula(f,a,b,delta,epsilon,max1)

% Datos
% - f es la función, introducida como una
%     una cadena de caracteres 'f'
% - a y b son el extremo izquierdo y el extremo derecho
% - delta es la tolerancia para el cero
% - epsilon es la tolerancia para el valor de f en el cero
% - max1 es el número máximo de iteraciones

% Resultados
% - c es el cero
% - yc=f(c)

```

```
% - err es el error estimado de la aproximación a c
ya=feval(f,a);
yb=feval(f,b);
if ya*yb>0
    disp('Note: f(a)*f(b)>0'),
    break,
end
for k=1:max1
    dx=yb*(b-a)/(yb-ya);
    c=b-dx;
    ac=c-a;
    yc=feval(f,c);
    if yc==0,break;
    elseif yb*yc>0
        b=c;
        yb=yc;
    else
        a=c;
        ya=yc;
    end
    dx=min(abs(dx),ac);
    if abs(dx)<delta,break,end
    if abs(yc)<epsilon,break,end
end
c;
err=abs(b-a)/2;
yc=feval(f,c);
```

Ejercicios

En los Ejercicios 1 y 2, debe hallar una aproximación de la tasa de interés anual I con la que se conseguiría un capital acumulado total A tras hacer 240 depósitos mensuales. Use los valores de partida que se dan y calcule las tres aproximaciones siguientes mediante el método de bisección.

1. $P = 275$ euros, $A = 250\,000$ euros, $I_0 = 0.11$ e $I_1 = 0.12$.
2. $P = 325$ euros, $A = 400\,000$ euros, $I_0 = 0.13$ e $I_1 = 0.14$.
3. Para cada una de las siguientes funciones, halle un intervalo $[a, b]$ de manera que $f(a)$ y $f(b)$ tengan distinto signo.
 - (a) $f(x) = e^x - 2 - x$
 - (b) $f(x) = \cos(x) + 1 - x$

- (c) $f(x) = \ln(x) - 5 + x$
 (d) $f(x) = x^2 - 10x + 23$

En los Ejercicios 4 a 7, empiece con el intervalo $[a_0, b_0]$ y use el método de la *régula falsi* para calcular c_0, c_1, c_2 y c_3 .

4. $e^x - 2 - x = 0$, $[a_0, b_0] = [-2.4, -1.6]$

5. $\cos(x) + 1 - x = 0$, $[a_0, b_0] = [0.8, 1.6]$

6. $\ln(x) - 5 + x = 0$, $[a_0, b_0] = [3.2, 4.0]$

7. $x^2 - 10x + 23 = 0$, $[a_0, b_0] = [6.0, 6.8]$

8. Denotemos por $[a_0, b_0]$, $[a_1, b_1]$, \dots , $[a_n, b_n]$ los intervalos que se generan en el método de bisección.

(a) Pruebe que $a_0 \leq a_1 \leq \dots \leq a_n \leq \dots$ y que $\dots \leq b_n \leq \dots \leq b_1 \leq b_0$.

(b) Pruebe que $b_n - a_n = (b_0 - a_0)/2^n$.

(c) Sea $c_n = (a_n + b_n)/2$ el punto medio de cada intervalo. Pruebe que

$$\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} c_n = \lim_{n \rightarrow \infty} b_n$$

Indicación. Repase la convergencia de las sucesiones monótonas en su libro de cálculo.

$$c_n = \frac{a_n f(b_n) - b_n f(a_n)}{f(b_n) - f(a_n)}.$$

13. Establezca la fórmula (15) con la que se determina el número de iteraciones necesarias en el método de bisección. *Indicación.* Use que $|b - a|/2^{n+1} < \delta$ y tome logaritmos.

14. El polinomio $f(x) = (x - 1)^3(x - 2)(x - 3)$ tiene tres ceros: $x = 1$ de multiplicidad 3 y $x = 2$ y $x = 3$, cada uno de ellos de multiplicidad 1. Si a_0 y b_0 son números reales tales que $a_0 < 1$ y $b_0 > 3$, entonces $f(a_0)f(b_0) < 0$. En consecuencia, en el intervalo $[a_0, b_0]$ el método de bisección convergerá a uno de los tres ceros. Si $a_0 < 1$ y $b_0 > 3$ se eligen tales que $c_n = \frac{a_n+b_n}{2}$ no es igual a 1, ni a 2 ni a 3 para ningún $n \geq 1$, entonces el método de bisección nunca convergerá a ¿cuál o cuáles ceros? ¿por qué?

15. Si un polinomio $f(x)$ tiene un número impar de ceros en un intervalo $[a_0, b_0]$ y cada uno de los ceros es de multiplicidad impar, entonces $f(a_0)f(b_0) < 0$ y el método de bisección convergerá a uno de los ceros. Si $a_0 < 1$ y $b_0 > 3$ se eligen de manera que $c_n = \frac{a_n+b_n}{2}$ no es igual a ninguno de los ceros de $f(x)$ para ningún $n \geq 1$, entonces el método de bisección nunca convergerá a ¿cuál o cuáles ceros? ¿por qué?

Algoritmos y programas

1. Halle una aproximación (exacta hasta la décima cifra decimal) a la tasa de interés I con la se conseguiría un capital acumulado total de 500 000 euros si se realizaran 240 depósitos mensuales de 300 euros.
2. Se construye una esfera de madera de radio $r = 15$ cm con una variedad de roble blanco que tiene una densidad $\rho = 0.710$ gr/cm³. ¿Cuánto de la esfera quedará sumergido (preciso hasta la octava cifra decimal) cuando la ponemos en agua?
3. Modifique los Programas 2.2 y 2.3 para obtener como resultado una matriz análoga a las Tablas 2.1 y 2.2, respectivamente (o sea, la primera fila de la matriz debería ser $[0 \ a_0 \ c_0 \ b_0 \ f(c_0)]$).
4. Use su programa anterior para aproximar las tres raíces positivas más pequeñas de la ecuación $x = \tan(x)$ (exactas hasta la octava cifra decimal).
5. Un plano corta una esfera de radio unidad en dos trozos de manera que uno de ellos tiene un volumen triple que el volumen del otro. Determine la distancia x del plano al centro de la esfera (con una precisión de 10 cifras decimales).

3 Aproximación inicial y criterios de convergencia

Los métodos de localización dependen de la determinación de un intervalo inicial $[a, b]$ en el que $f(a)$ y $f(b)$ tengan signo distinto. Una vez encontrado este intervalo, no importa lo grande que sea, podremos empezar a iterar hasta que encontremos una raíz con la precisión deseada. Por esta razón se dice que estos métodos son ***globalmente convergentes***. Sin embargo, si $f(x) = 0$ tiene varias raíces en $[a, b]$, entonces debemos encontrar un intervalo de partida distinto para hallar cada raíz y no suele ser fácil hallar estos intervalos más pequeños en los que el signo de $f(x)$ cambia.

En la Sección 2.4 desarrollaremos el método de Newton-Raphson y el método de la secante para resolver $f(x) = 0$. Estos dos métodos requieren, como garantía de su convergencia, que el punto inicial esté cerca de la raíz, por lo que se dice que son ***localmente convergentes***. A cambio, suelen converger más rápidamente que los métodos globales, de manera que existen algoritmos

híbridos que empiezan con un método de convergencia global y, cuando las iteraciones nos han llevado cerca de la raíz, cambian a un método de convergencia local.

Si el cálculo de las raíces de una ecuación es una parte de un proyecto más amplio, entonces se sugiere tomárselo con calma y empezar por dibujar la gráfica de la función. A la vista de la curva, podemos tomar decisiones basadas en su aspecto (concavidad, pendiente, conducta oscilatoria, extremos locales, puntos de inflexión, etc.); pero, lo que es todavía más importante, si la gráfica permite conocer, o al menos estimar, las coordenadas de los puntos, entonces podemos determinar la localización aproximada de las raíces y usar estas aproximaciones como valores iniciales de nuestros algoritmos.

Debemos, no obstante, proceder con cautela. Los paquetes informáticos utilizan programas gráficos más o menos sofisticados. Supongamos que usamos un computador para dibujar la curva $y = f(x)$ con x en $[a, b]$. Típicamente, el intervalo se divide en N partes iguales, cuyos extremos son $N + 1$ puntos equiespaciados $a = x_0 < x_1 < \dots < x_N = b$, y luego se calculan los valores $y_k = f(x_k)$. Entonces cada pareja de puntos adyacentes (x_{k-1}, y_{k-1}) y (x_k, y_k) , para $k = 1, 2, \dots, N$, se unen en el dibujo mediante una línea recta o una curva adecuada. Tiene que haber un número suficiente de puntos, de tal manera que no perdamos ninguna raíz que pueda estar en una porción de la curva donde, por ejemplo, la función cambie muy rápidamente. Si $f(x)$ es continua y dos puntos adyacentes (x_{k-1}, y_{k-1}) y (x_k, y_k) están situados en lados distintos del eje de abscisas, entonces el teorema del valor intermedio garantiza que hay al menos una raíz en el intervalo $[x_{k-1}, x_k]$. Pero si hay una raíz, o incluso varias raíces próximas entre sí, en el intervalo $[x_{k-1}, x_k]$ y los puntos adyacentes (x_{k-1}, y_{k-1}) y (x_k, y_k) están en el mismo lado del eje de abscisas, entonces el dibujo realizado por el computador no nos indicará que en dicho intervalo hay raíces; este dibujo no será una buena representación de la gráfica de la función f . Este fenómeno suele darse con funciones que tienen raíces muy próximas entre sí, funciones cuya gráfica parece que sólo toca al eje OX pero no lo cruza; una raíz doble sería un ejemplo de esta situación, otro ejemplo de este fenómeno sería una raíz muy cercana a una asíntota vertical de la gráfica de la función. Es necesario que tengamos en cuenta estas características de una función a la hora de aplicar un algoritmo numérico de cálculo de raíces.

Finalmente, volviendo al ejemplo citado antes, cerca de dos raíces próximas entre sí o de una raíz doble en el intervalo $[x_{k-1}, x_k]$, puede ocurrir que la gráfica de la curva entre los puntos (x_{k-1}, y_{k-1}) y (x_k, y_k) ofrecida por el computador no cruce el eje de abscisas. Si $|f(x_k)|$ es menor que una tolerancia ε prefijada (o sea, $f(x_k) \approx 0$), entonces podemos pensar que x_k es una raíz aproximada. Pero podría ocurrir que los valores de f estuvieran muy cerca de cero para un rango amplio de su variable x cerca de x_k , con lo que x_k podría no estar cerca de una raíz. Por ello es usual añadir, en esta situación, el requisito de que la pendiente cambie de signo cerca de (x_k, y_k) ; esto es, que $m_{k-1} = \frac{y_k - y_{k-1}}{x_k - x_{k-1}}$ y $m_k = \frac{y_{k+1} - y_k}{x_{k+1} - x_k}$

Tabla 2.3 Localización aproximada de las raíces.

x_k	Valores de la función		Diferencias en las y		Cambios significativos en $f(x)$ o $f'(x)$
	y_{k-1}	y_k	$y_k - y_{k-1}$	$y_{k+1} - y_k$	
-1.2	-3.125	-0.968	2.157	1.329	f cambia de signo en $[x_{k-1}, x_k]$
-0.9	-0.968	0.361	1.329	0.663	
-0.6	0.361	1.024	0.663	0.159	
-0.3	1.024	1.183	0.159	-0.183	
0.0	1.183	1.000	-0.183	-0.363	
0.3	1.000	0.637	-0.363	-0.381	
0.6	0.637	0.256	-0.381	-0.237	
0.9	0.256	0.019	-0.237	0.069	
1.2	0.019	0.088	0.069	0.537	

tengan distinto signo. Como $x_k - x_{k-1} > 0$ y $x_{k+1} - x_k > 0$, no es necesario usar los cocientes de las diferencias, basta comprobar si las diferencias $y_k - y_{k-1}$ e $y_{k+1} - y_k$ tienen signos distintos, en cuyo caso aceptamos que x_k es una raíz aproximada. Desafortunadamente, esto no garantiza que este valor de partida produzca una sucesión convergente; por ejemplo, si la gráfica de f tiene un mínimo (o un máximo) local que está muy próximo a cero, entonces es posible que x_k sea catalogado como un cero aproximado de f cuando $f(x_k) \approx 0$, aunque quizás x_k no está cerca de ninguna raíz.

Ejemplo 2.9. Vamos a localizar aproximadamente las raíces de $x^3 - x^2 - x + 1 = 0$ en el intervalo $[-1.2, 1.2]$. Para ello, por ejemplo, tomamos $N = 8$ y miramos los cálculos en la Tabla 2.3.

Las tres abscisas que debemos tener en cuenta son $-1.05, -0.3$, y 0.9 . Como $f(x)$ cambia de signo en $[-1.2, -0.9]$, el punto medio -1.05 es una raíz aproximada; de hecho, $f(-1.05) = -0.210$.

Aunque la pendiente cambia de signo cerca de -0.3 , vemos que $f(-0.3) = 1.183$; luego -0.3 no está cerca de ninguna raíz. Finalmente, la pendiente cambia de signo cerca de 0.9 y $f(0.9) = 0.019$, de manera que 0.9 es una raíz aproximada (véase la Figura 2.10). ■

Comprobación de la convergencia

Podemos usar una gráfica para ver la localización aproximada de una raíz, pero debemos usar un computador para calcular un valor p_n que sea una solución aceptable. Normalmente empleamos un procedimiento iterativo que genera una

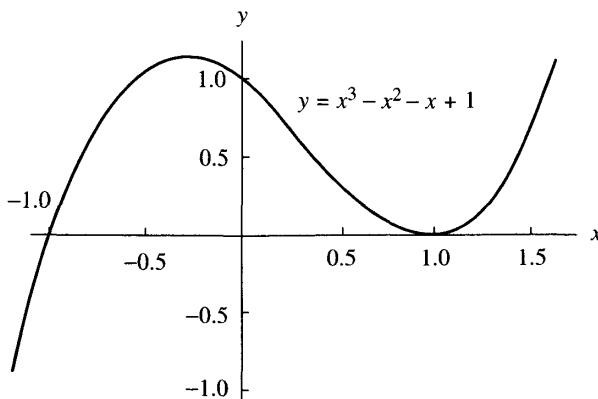


Figura 2.10 Gráfica del polinomio cúbico

$$y = x^3 - x^2 - x + 1.$$

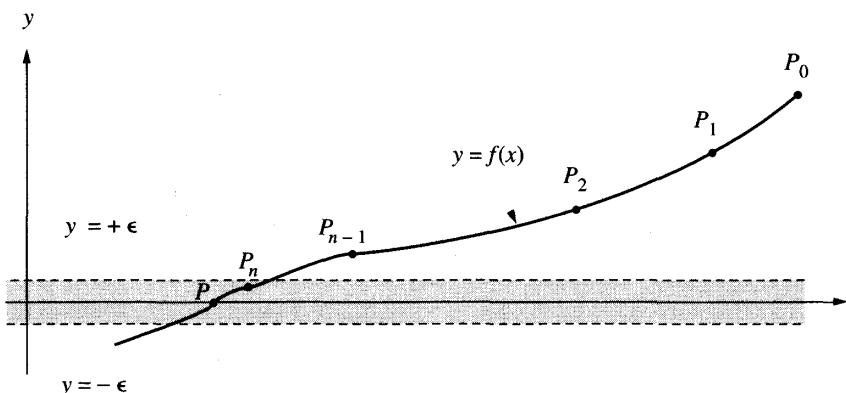


Figura 2.11 (a) Banda de convergencia horizontal para localizar una solución de $f(x) = 0$.

sucesión $\{p_k\}$ convergente a una raíz p , así que es imprescindible diseñar por adelantado un criterio, o estrategia, de parada para que el computador detenga las iteraciones cuando haya obtenido una aproximación suficientemente precisa. Puesto que nuestro objetivo es resolver $f(x) = 0$, el valor final p_n debería verificar que $|f(p_n)| < \varepsilon$.

Podemos proporcionar un valor de la tolerancia ε para el tamaño de $|f(p_n)|$ y el proceso iterativo producirá puntos $P_k = (p_k, f(p_k))$ hasta que el último punto P_n esté en la banda horizontal comprendida entre las rectas de ecuaciones $y = +\varepsilon$ e $y = -\varepsilon$, como se muestra en la Figura 2.11(a). Este criterio es útil si lo que deseamos es resolver una ecuación $h(x) = L$ aplicando un algoritmo de cálculo de ceros a la función $f(x) = h(x) - L$.

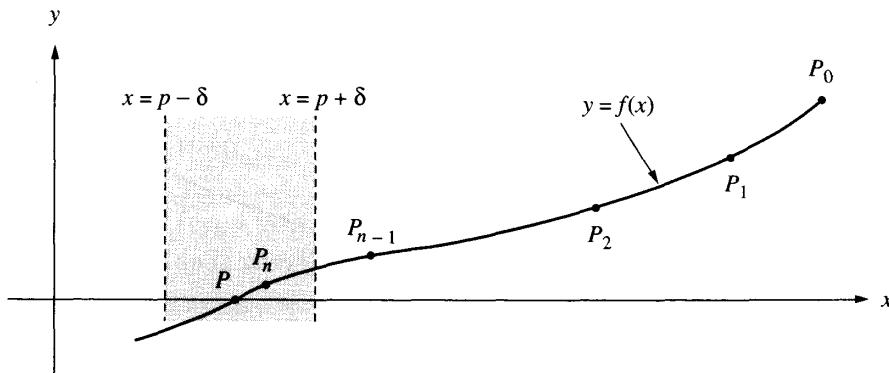


Figura 2.11 (b) Banda de convergencia vertical para localizar una solución de $f(x) = 0$.

Otro criterio de parada involucra las abscisas. Tratamos de determinar si la sucesión $\{p_k\}$ converge y, para ello, dibujamos dos rectas verticales de ecuaciones $x = p + \delta$ y $x = p - \delta$ a cada lado de $x = p$, y detenemos el proceso cuando P_n está entre ambas rectas, como se muestra en la Figura 2.11(b).

Aunque es un criterio aceptable, tal y como está es imposible de utilizar porque supone conocida la solución p que buscamos. Lo que hacemos es adaptar la idea que subyace y detener las iteraciones cuando dos valores consecutivos p_{n-1} y p_n están suficientemente cerca o cuando coinciden en sus M primeras cifras significativas.

Algunas veces nos conformaremos con que $p_n \approx p_{n-1}$ y otras con que $f(p_n) \approx 0$. Razonemos de forma precisa para entender las consecuencias de nuestra elección. Si requerimos $|p_n - p| < \delta$ y $|f(p_n)| < \varepsilon$, el punto P_n estará situado en la región rectangular que rodea la solución $(p, 0)$ y se muestra en la Figura 2.12(a). Si estipulamos que $|p_n - p| < \delta$ o bien $|f(p_n)| < \varepsilon$, el punto P_n podría estar situado en cualquier lugar de la región formada por la unión de las bandas vertical y horizontal que se muestra en la Figura 2.12(b). Los tamaños de las tolerancias δ y ε son cruciales. Si las tolerancias se eligen demasiado pequeñas, la iteración podría continuar sin fin. Una buena elección es tomarlas unas 100 veces mayores que 10^{-M} , donde M es el número de cifras decimales de la representación en coma flotante que usa el computador. La proximidad entre las abscisas puede comprobarse con uno de los siguientes criterios

$$|p_n - p_{n-1}| < \delta \quad (\text{estimación del error absoluto})$$

o bien

$$\frac{2|p_n - p_{n-1}|}{|p_n| + |p_{n-1}|} < \delta \quad (\text{estimación del error relativo}).$$

La proximidad de la ordenada se comprueba normalmente mediante $|f(p_n)| < \varepsilon$.

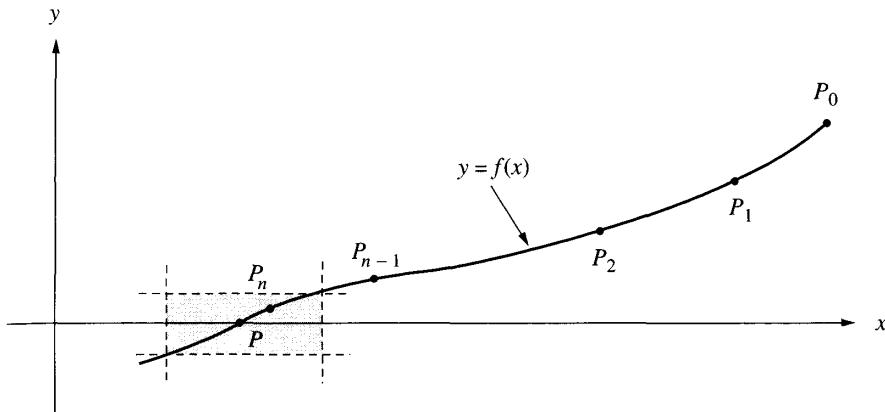


Figura 2.12 (a) Región rectangular definida por $|x - p| < \delta$ y $|y| < \varepsilon$.

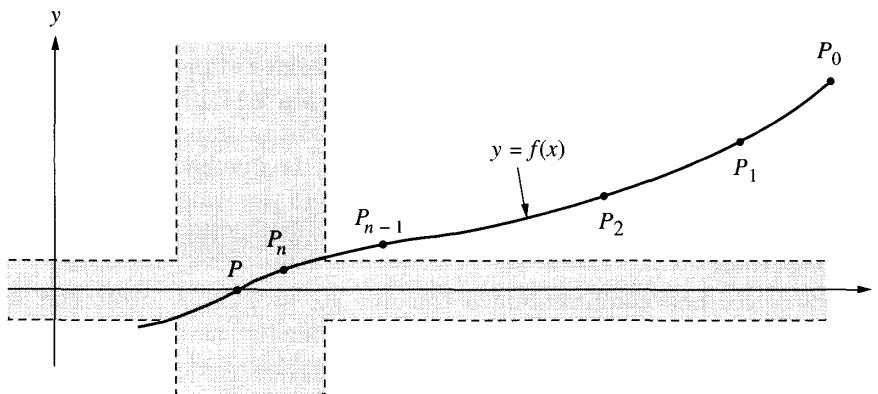


Figura 2.12 (b) Región no acotada definida por $|x - p| < \delta$ o bien $|y| < \varepsilon$.

Funciones problemáticas

Una solución aproximada de $f(x) = 0$ calculada con un computador estará siempre sometida a errores debidos a los redondeos o a la inestabilidad de los algoritmos. Si la curva $y = f(x)$ es muy inclinada cerca de la raíz $(p, 0)$, entonces el problema de hallar la raíz está bien condicionado (o sea, es fácil obtener la solución con varias cifras significativas). Si la curva $y = f(x)$ es casi horizontal cerca de $(p, 0)$, entonces el problema de hallar la raíz está mal condicionado (o sea, la solución calculada podría tener sólo unas pocas cifras significativas de precisión); esto ocurre cuando $f(x)$ tiene una raíz múltiple en p . Discutiremos este problema más a fondo en la siguiente sección.

Programa 2.4 (Localización aproximada de raíces). Estimación aproximada de la localización de una raíz de la ecuación $f(x) = 0$ en el intervalo $[a, b]$ mediante el uso de puntos de muestra equiespaciados $(x_k, f(x_k))$ y de acuerdo con los siguientes criterios:

- (i) $(y_{k-1})(y_k) < 0$, o
- (ii) $|y_k| < \varepsilon$ e $(y_k - y_{k-1})(y_{k+1} - y_k) < 0$.

Esto es, o bien $f(x_{k-1})$ y $f(x_k)$ tienen distinto signo, o bien $|f(x_k)|$ es pequeño y la pendiente de la curva $y = f(x)$ cambia de signo cerca de $(x_k, f(x_k))$.

```
function R = approot(X,epsilon)
% Datos
%     - f es la función, almacenada en el archivo f.m
%     - X es el vector de abscisas
%     - epsilon es la tolerancia
% Resultado
%         - R es el vector de raíces aproximadas
Y=f(X);
yrange = max(Y)-min(Y);
epsilon2 = yrange*epsilon;
n=length(X);
m=0;
X(n+1)=X(n);
Y(n+1)=Y(n);
for k=2:n,
    if Y(k-1)*Y(k)<=0,
        m=m+1;
        R(m)=(X(k-1)+X(k))/2;
    end
    s=(Y(k)-Y(k-1))*(Y(k+1)-Y(k));
    if (abs(Y(k)) < epsilon2) & (s<=0),
        m=m+1;
        R(m)=X(k);
    end
end
```

Ejemplo 2.10. Usemos approot para hallar la localización aproximada de las raíces de $f(x) = \operatorname{sen}(\cos(x^3))$ en el intervalo $[-2, 2]$. Primero hay que almacenar f como un archivo llamado f.m. Puesto que los resultados los usaremos como aproximaciones iniciales en un algoritmo de cálculo de raíces, construiremos X de manera que las aproximaciones tengan cuatro cifras decimales de precisión

```
>>X=-2:.001:2;
>>approot(X,0.00001)
```

ans=

-1.9875 -1.6765 -1.1625 1.1625 1.6765 1.9875

Comparando los resultados con la gráfica de f , puede verse que ahora disponemos de buenas aproximaciones iniciales para cualquiera de nuestros algoritmos de cálculo de raíces.

Ejercicios

En los Ejercicios 1 a 6, utilice un computador o una calculadora con pantalla gráfica para determinar gráficamente la localización aproximada de las raíces de $f(x) = 0$ en el intervalo dado. En cada caso, determine un intervalo $[a, b]$ en el que los Programas 2.2 y 2.3 puedan ser usados para hallar las raíces (o sea, $f(a)f(b) < 0$).

1. $f(x) = x^2 - e^x$ para $-2 \leq x \leq 2$
2. $f(x) = x - \cos(x)$ para $-2 \leq x \leq 2$
3. $f(x) = \operatorname{sen}(x) - 2 \cos(x)$ para $-2 \leq x \leq 2$
4. $f(x) = \cos(x) + (1 + x^2)^{-1}$ para $-2 \leq x \leq 2$
5. $f(x) = (x - 2)^2 - \ln(x)$ para $0.5 \leq x \leq 4.5$
6. $f(x) = 2x - \tan(x)$ para $-1.4 \leq x \leq 1.4$

Algoritmos y programas

En los Problemas 1 y 2, utilice un computador o una calculadora con pantalla gráfica y el Programa 2.4 para aproximar, con una precisión de cuatro cifras decimales, las raíces reales de cada una de las funciones dadas en el intervalo correspondiente. Luego utilice el Programa 2.2 o el Programa 2.3 para aproximar cada raíz con una precisión de 12 cifras decimales.

1. $f(x) = 1\,000\,000x^3 - 111\,000x^2 + 1110x - 1$ para $-2 \leq x \leq 2$.
2. $f(x) = 5x^{10} - 38x^9 + 21x^8 - 5\pi x^6 - 3\pi x^5 - 5x^2 + 8x - 3$ para $-15 \leq x \leq 15$.

3. Los programas de computador para dibujar curvas $y = f(x)$ (con x en un intervalo $[a, b]$) y mediante los puntos $(x_0, y_0), (x_1, y_1), \dots$, y (x_N, y_N)) suelen realizar un escalado de la altura de la gráfica y deben incluir un procedimiento que determine los valores máximo y mínimo de f en el intervalo.
- Construya un algoritmo que encuentre los valores $Y_{\max} = \max_k \{y_k\}$ e $Y_{\min} = \min_k \{y_k\}$.
 - Escriba un programa con el paquete MATLAB que determine la localización aproximada y los valores extremos de una función $f(x)$ en un intervalo $[a, b]$.
 - Utilice su programa del apartado (b) para hallar la localización aproximada y los valores extremos de las funciones de los Problemas 1 y 2. Compare sus resultados con los valores exactos.

2.4 Los métodos de Newton-Raphson y de la secante

Métodos tangenciales para el cálculo de raíces

Si $f(x)$, $f'(x)$ y $f''(x)$ son continuas cerca de una raíz p , esta información adicional sobre la naturaleza de $f(x)$ puede usarse para desarrollar algoritmos que produzcan sucesiones $\{p_k\}$ que converjan a p más rápidamente que en el método de bisección o en el de la *régula falsi*. El método de Newton-Raphson (o, simplemente, de Newton), que descansa en la continuidad de $f'(x)$ y $f''(x)$, es uno de los algoritmos más útiles y mejor conocidos. Lo introduciremos gráficamente y luego daremos un tratamiento más riguroso basado en el teorema de Taylor.

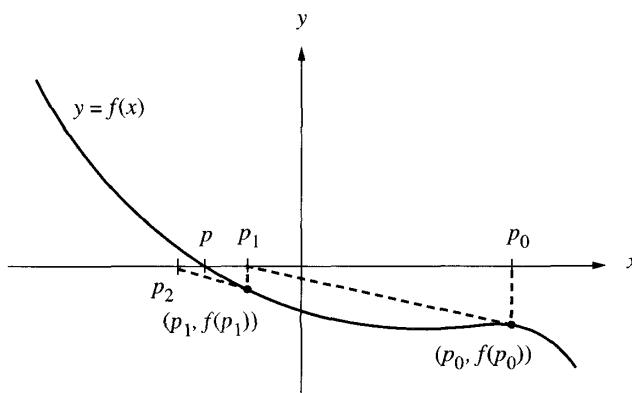


Figura 2.13 Construcción geométrica de p_1 y p_2 en el método de Newton-Raphson.

Supongamos que la aproximación inicial p_0 está cerca de la raíz p . Entonces la curva $y = f(x)$ y el eje de abscisas se cortan en el punto $(p, 0)$ y, además, el punto $(p_0, f(p_0))$ está situado en la curva cerca de $(p, 0)$ (véase la Figura 2.13). Definimos p_1 como el punto de intersección del eje de abscisas con la recta tangente a la curva en el punto $(p_0, f(p_0))$. La Figura 2.13 muestra que p_1 estará, en este caso, más cerca de p que p_0 . Podemos encontrar la ecuación que relaciona p_1 con p_0 igualando dos fórmulas distintas para la pendiente m de la recta tangente. Por un lado,

$$(1) \quad m = \frac{0 - f(p_0)}{p_1 - p_0},$$

que es la pendiente de la línea recta que pasa por $(p_1, 0)$ y $(p_0, f(p_0))$; por otro lado,

$$(2) \quad m = f'(p_0),$$

que es la pendiente de la recta tangente a la curva en el punto $(p_0, f(p_0))$. Igualando los valores de la pendiente m en las fórmulas (1) y (2) y despejando p_1 obtenemos

$$(3) \quad p_1 = p_0 - \frac{f(p_0)}{f'(p_0)}.$$

Este proceso puede repetirse para obtener una sucesión $\{p_k\}$ que converge a p . Hagamos más precisas estas ideas.

Teorema 2.5 (Teorema de Newton-Raphson). Supongamos que la función $f \in C^2[a, b]$ y que existe un número $p \in [a, b]$ tal que $f(p) = 0$. Si $f'(p) \neq 0$, entonces existe $\delta > 0$ tal que la sucesión $\{p_k\}_{k=0}^{\infty}$ definida por el proceso iterativo

$$(4) \quad p_k = g(p_{k-1}) = p_{k-1} - \frac{f(p_{k-1})}{f'(p_{k-1})} \quad \text{para } k = 1, 2, \dots$$

converge a p cualquiera que sea la aproximación inicial $p_0 \in [p - \delta, p + \delta]$.

Observación. La función $g(x)$ definida por la relación

$$(5) \quad g(x) = x - \frac{f(x)}{f'(x)}$$

se llama **función de iteración de Newton-Raphson**. Puesto que $f(p) = 0$, es fácil ver que $g(p) = p$, lo que nos dice que la iteración de Newton-Raphson para hallar una raíz de la ecuación $f(x) = 0$ consiste en hallar un punto fijo de $g(x)$.

Demostración. La construcción geométrica de p_1 que se muestra en la Figura 2.13 no nos ayuda a entender por qué p_0 debe estar cerca de p ni por qué la continuidad de $f''(x)$ es esencial. Nuestro análisis comienza con el polinomio de Taylor de grado $n = 1$ de f alrededor de p_0 y su correspondiente resto:

$$(6) \quad f(x) = f(p_0) + f'(p_0)(x - p_0) + \frac{f''(c)(x - p_0)^2}{2!},$$

donde c es un punto intermedio entre p_0 y x . Poniendo $x = p$ en la relación (6) y usando que $f(p) = 0$ obtenemos

$$(7) \quad 0 = f(p_0) + f'(p_0)(p - p_0) + \frac{f''(c)(p - p_0)^2}{2!}.$$

Si p_0 está suficientemente cerca de p , entonces el último sumando del miembro derecho de (7) será pequeño, comparado con la suma de los dos primeros, así que podemos despreciarlo y usar la aproximación

$$(8) \quad 0 \approx f(p_0) + f'(p_0)(p - p_0).$$

Despejando p en la relación (8), obtenemos $p \approx p_0 - f(p_0)/f'(p_0)$, expresión que usamos para definir p_1 , la siguiente aproximación a la raíz

$$(9) \quad p_1 = p_0 - \frac{f(p_0)}{f'(p_0)}.$$

Poniendo p_{k-1} en lugar de p_0 en la relación (9), la regla general (4) queda establecida. En muchas aplicaciones, esto es todo lo que hace falta entender y saber usar; sin embargo, para comprender totalmente lo que ocurre, necesitamos considerar la iteración de Newton-Raphson como una iteración de punto fijo y aplicar el Teorema 2.2 en nuestra situación. La clave nos la da el análisis de $g'(x)$:

$$g'(x) = 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2}.$$

Por hipótesis, sabemos que $f(p) = 0$; luego $g'(p) = 0$. Como $g(x)$ es continua y $g'(p) = 0$, podemos encontrar $\delta > 0$ tal que la hipótesis $|g'(x)| < 1$ del Teorema 2.2 se cumple en el intervalo $(p - \delta, p + \delta)$. Por consiguiente, que $p_0 \in (p - \delta, p + \delta)$ es una condición suficiente para que p_0 sea el punto de partida de una sucesión $\{p_k\}_{k=0}^{\infty}$ que converge a la única raíz de $f(x) = 0$ en dicho intervalo, siempre que δ sea elegido tal que

$$(10) \quad \frac{|f(x)f''(x)|}{|f'(x)|^2} < 1 \quad \text{para todo } x \in (p - \delta, p + \delta). \quad \bullet$$

Corolario 2.2 (Iteración de Newton-Raphson para el cálculo de raíces cuadradas). Supongamos que $A > 0$ es un número real y sea $p_0 > 0$ una aproximación inicial a \sqrt{A} . Definimos una sucesión $\{p_k\}_{k=0}^{\infty}$ mediante el proceso recursivo

$$(11) \quad p_k = \frac{1}{2} \left(p_{k-1} + \frac{A}{p_{k-1}} \right) \quad \text{para } k = 1, 2, \dots$$

Entonces la sucesión $\{p_k\}_{k=0}^{\infty}$ converge a \sqrt{A} ; es decir, $\lim_{n \rightarrow \infty} p_k = \sqrt{A}$.

Esbozo de la demostración. Empezamos con la función $f(x) = x^2 - A$. Notemos que las raíces de la ecuación $x^2 - A = 0$ son $\pm\sqrt{A}$. Ahora usamos la función $f(x)$ y su derivada $f'(x)$ en la fórmula (5) y escribimos la correspondiente fórmula de iteración de Newton-Raphson

$$(12) \quad g(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^2 - A}{2x},$$

que podemos simplificar para obtener

$$(13) \quad g(x) = \frac{1}{2} \left(x + \frac{A}{x} \right).$$

Cuando usamos la función de iteración $g(x)$ dada en (13) para definir el proceso recursivo (4), el resultado es la fórmula (11). Puede probarse que la sucesión generada por (11) converge cualquiera que sea el valor inicial $p_0 > 0$; los detalles se dejan como ejercicio.

Un aspecto importante del Corolario 2.2 es que la función de iteración $g(x)$ sólo necesita las operaciones aritméticas $+$, $-$, \times y $/$; si $g(x)$ hubiera necesitado el cálculo de una raíz cuadrada, estaríamos atrapados en un razonamiento circular: el uso del cálculo de raíces cuadradas para calcular raíces cuadradas (de forma más complicada). Por esta razón se eligió $f(x) = x^2 - A$, porque sólo involucra las operaciones aritméticas elementales.

Ejemplo 2.11. Vamos a usar el algoritmo de Newton-Raphson dado inmediatamente antes para calcular $\sqrt{5}$.

Empezando con $p_0 = 2$ y usando la fórmula (11), calculamos:

$$p_1 = \frac{2 + 5/2}{2} = 2.25$$

$$p_2 = \frac{2.25 + 5/2.25}{2} = 2.236111111$$

$$p_3 = \frac{2.236111111 + 5/2.236111111}{2} = 2.236067978$$

$$p_4 = \frac{2.36067978 + 5/2.36067978}{2} = 2.236067978.$$

Si siguiéramos iterando llegaríamos a $p_k \approx 2.236067978$ para $k > 4$, de manera que hemos conseguido una exactitud de nueve cifras decimales. ■

Vamos a prestar un poco de atención ahora a un conocido problema de física elemental, en el que la determinación de una raíz de una cierta ecuación es fundamental. Supongamos que disparamos un proyectil desde el origen con un ángulo de elevación b_0 y velocidad inicial v_0 . Cuando se analiza este problema en los primeros cursos, se desprecia la resistencia del aire y aprendemos que la altura $y = y(t)$ y el alcance horizontal $x = x(t)$, medidos en metros, obedecen las reglas

$$(14) \quad y = v_y t - 4.9t^2 \quad y \quad x = v_x t,$$

donde las componentes horizontal y vertical de la velocidad inicial son, respectivamente, $v_x = v_0 \cos(b_0)$ y $v_y = v_0 \sin(b_0)$. Es fácil trabajar con el modelo matemático expresado por las fórmulas de (14), pero tiende a dar valores demasiado altos de la altura y del alcance horizontal de la trayectoria del proyectil. Si hacemos la hipótesis adicional de que la resistencia del aire es proporcional a la velocidad, entonces las ecuaciones del movimiento pasan a ser

$$(15) \quad y = f(t) = (Cv_y + 9.8C^2) \left(1 - e^{-t/C}\right) - 9.8Ct$$

y

$$(16) \quad x = r(t) = Cv_x \left(1 - e^{-t/C}\right),$$

siendo $C = m/\kappa$, donde κ es el coeficiente de resistencia del aire y m es la masa del proyectil. Un valor grande de C corresponderá a una altura máxima y un alcance mayores del proyectil. La gráfica de la trayectoria de vuelo de un proyectil cuando tenemos en cuenta la resistencia del aire se muestra en la Figura 2.14. Este modelo mejorado es más realista, pero requiere el uso de un algoritmo de cálculo de raíces para resolver $f(t) = 0$ y poder así determinar el tiempo transcurrido hasta que el proyectil cae al suelo. El modelo elemental dado en (14) no requiere ningún procedimiento sofisticado para hallar el tiempo transcurrido.

Ejemplo 2.12. Se dispara un proyectil con un ángulo de elevación $b_0 = 45^\circ$, velocidades iniciales $v_y = v_x = 100 \text{ m/s}$ y $C = 10$. Vamos a determinar el tiempo transcurrido hasta el impacto en el suelo.

Usando las fórmulas (15) y (16), las ecuaciones del movimiento del proyectil son $y = f(t) = 1980(1 - e^{-t/10}) - 98t$ y $x = r(t) = 1000(1 - e^{-t/10})$. Puesto que $f(16) = 12.24489437$ y $f(17) = -47.71337762$, usaremos como aproximación inicial $p_0 = 16$. La derivada es $f'(t) = 198e^{-t/10} - 98$ y, en el punto inicial, vale $f'(p_0) = f'(16) = -58.02448937$ que, en la fórmula (4), nos proporciona

$$p_1 = 16 - \frac{12.24489437}{-58.02448937} = 16.21102977.$$

Los cálculos se muestran en la Tabla 2.4.

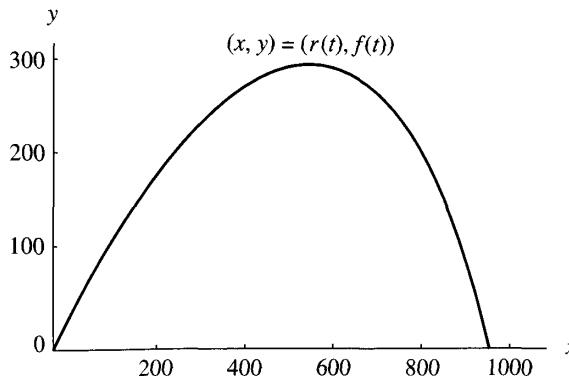


Figura 2.14 Trayectoria de un proyectil cuando tenemos en cuenta la resistencia del aire.

Tabla 2.4 Cálculo del instante en el que la altura $f(t)$ vale cero.

k	Tiempo, p_k	$p_{k+1} - p_k$	Altura, $f(p_k)$
0	16.00000000	0.21102977	12.24489437
1	16.21102977	-0.00150171	-0.08838974
2	16.20952805	-0.00000007	-0.000000441
3	16.20952798	0.00000000	0.00000000
4	16.20957798	0.00000000	0.00000000

El valor p_4 tiene ocho cifras decimales de precisión y el instante del impacto en el suelo es $t \approx 16.20957798$ segundos. Podemos calcular el alcance usando $r(t)$ y obtenemos

$$r(16.20957798) = 1000 (1 - e^{-1.620957798}) = 802.28976853 \text{ m.}$$

Errores por la división entre cero

Uno de los inconvenientes obvios del método de Newton-Raphson es la posibilidad de que se divida entre cero en la fórmula (4), lo que ocurriría si $f'(p_{k-1}) = 0$. El Programa 2.5 contiene instrucciones que permiten detectar esta situación, pero ¿de qué nos sirve en ese caso la última aproximación calculada p_{k-1} ? Es posible que $|f(p_{k-1})|$ sea suficientemente pequeño y que p_{k-1} sea una aproximación aceptable a la raíz. Investigaremos ahora esta situación y, de paso, descubriremos un aspecto interesante: la velocidad de convergencia del método.

Definición 2.4 (Orden de una raíz). Supongamos que $f(x)$ y sus derivadas $f'(x), \dots, f^{(M)}(x)$ están definidas y son continuas en un intervalo centrado

en el punto p . Diremos que $f(x) = 0$ tiene una raíz de orden M en $x = p$ si

$$(17) \quad f(p) = 0, f'(p) = 0, \dots, f^{(M-1)}(p) = 0 \text{ y } f^{(M)}(p) \neq 0.$$

Las raíces de orden $M = 1$ se suelen llamar **raíces simples**, mientras que si $M > 1$, entonces se llaman **raíces múltiples**. En particular, las raíces de orden $M = 2$ se conocen como **raíces dobles**, y así sucesivamente. El siguiente resultado nos servirá para aclarar estos conceptos. ▲

Lema 2.1. Si una ecuación $f(x) = 0$ tiene una raíz de orden M en $x = p$, entonces existe una función continua $h(x)$ tal que $f(x)$ puede expresarse como el producto

$$(18) \quad f(x) = (x - p)^M h(x), \quad \text{siendo } h(p) \neq 0.$$

Ejemplo 2.13. La función $f(x) = x^3 - 3x + 2$ tiene una raíz simple en $p = -2$ y una raíz doble en $p = 1$. Para verificar este hecho basta considerar las derivadas $f'(x) = 3x^2 - 3$ y $f''(x) = 6x$ y evaluar. Para $p = -2$, tenemos $f(-2) = 0$ y $f'(-2) = 9$, de manera que $M = 1$ en la Definición 2.4 y, por tanto, $p = -2$ es una raíz simple. Para $p = 1$, tenemos $f(1) = 0$, $f'(1) = 0$ y $f''(1) = 6$, de manera que $M = 2$ en la Definición 2.4 y, por tanto, $p = 1$ es una raíz doble. Hagamos notar, por otro lado, que $f(x)$ puede factorizarse como $f(x) = (x + 2)(x - 1)^2$. ■

Velocidad de convergencia

Como vamos a mostrar, la propiedad distintiva que caracteriza cada caso es la siguiente: Si p es una raíz simple de $f(x) = 0$, entonces el método de Newton-Raphson converge rápidamente, de forma que en cada iteración doblamos (aproximadamente) el número de cifras decimales exactas. Si, por el contrario, p es una raíz múltiple, entonces el error en cada paso es una fracción del error en el paso anterior. Para describir estos hechos de manera precisa, definimos a continuación la noción de **orden de convergencia**, que es una medida de la velocidad de convergencia de una sucesión.

Definición 2.5 (Orden de convergencia). Supongamos que $\{p_n\}_{n=0}^{\infty}$ converge a p y sea $E_n = p - p_n$ para cada $n \geq 0$. Si existen dos constantes positivas $A > 0$ y $R > 0$ tales que

$$(19) \quad \lim_{n \rightarrow \infty} \frac{|p - p_{n+1}|}{|p - p_n|^R} = \lim_{n \rightarrow \infty} \frac{|E_{n+1}|}{|E_n|^R} = A,$$

entonces se dice que la sucesión converge a p con orden de convergencia R y el número A se llama constante asintótica del error. Los casos $R = 1, 2$ merecen una consideración especial:

(20) Si $R = 1$, la convergencia de $\{p_n\}_{n=0}^{\infty}$ se llama **lineal**.

(21) Si $R = 2$, la convergencia de $\{p_n\}_{n=0}^{\infty}$ se llama **cuadrática**. ▲

Tabla 2.5 El método de Newton-Raphson converge cuadráticamente en una raíz simple.

k	p_k	$p_{k+1} - p_k$	$E_k = p - p_k$	$\frac{ E_{k+1} }{ E_k ^2}$
0	-2.400000000	0.323809524	0.400000000	0.476190475
1	-2.076190476	0.072594465	0.076190476	0.619469086
2	-2.003596011	0.003587422	0.003596011	0.664202613
3	-2.000008589	0.000008589	0.000008589	
4	-2.000000000	0.000000000	0.000000000	

Si R es grande, entonces la sucesión $\{p_n\}$ converge rápidamente a p ; esto es, la relación (19) implica que para valores grandes de n tenemos la aproximación $|E_{n+1}| \approx A|E_n|^R$. Por ejemplo, supongamos que $R = 2$ y que $|E_n| \approx 10^{-2}$, entonces cabe esperar que $|E_{n+1}| \approx A \times 10^{-4}$.

Algunas sucesiones convergen con un orden que no es un número natural; veremos, por ejemplo, que el orden de convergencia del método de la secante es $R = (1 + \sqrt{5})/2 \approx 1.618033989$.

Ejemplo 2.14 (Convergencia cuadrática cuando la raíz es simple). Partiendo de $p_0 = -2.4$ y usando la iteración de Newton-Raphson, vamos a aproximarnos a la raíz simple $p = -2$ del polinomio $f(x) = x^3 - 3x + 2$.

La fórmula de iteración para calcular $\{p_k\}$ es

$$(22) \quad p_{k+1} = g(p_k) = \frac{2p_k^3 - 2}{3p_k^2 - 3}.$$

Usando la fórmula (19) con $R = 2$ para comprobar si la convergencia es cuadrática, obtenemos los valores que se muestran en la Tabla 2.5.

Analizando con atención la velocidad de convergencia en el Ejemplo 2.14 vemos que el error en una iteración es proporcional al cuadrado del error en la iteración previa; concretamente,

$$|p - p_{k+1}| \approx A|p - p_k|^2,$$

donde $A \approx 2/3$. Para comprobar este hecho, usamos, por ejemplo, que

$$|p - p_3| = 0.000008589 \quad \text{y} \quad |p - p_2|^2 = |0.003596011|^2 = 0.000012931,$$

con lo cual

$$|p - p_3| = 0.000008589 \approx 0.000008621 = \frac{2}{3}|p - p_2|^2.$$

Tabla 2.6 El método de Newton-Raphson converge linealmente en una raíz doble.

k	p_k	$p_{k+1} - p_k$	$E_k = p - p_k$	$\frac{ E_{k+1} }{ E_k }$
0	1.200000000	-0.096969697	-0.200000000	0.515151515
1	1.103030303	-0.050673883	-0.103030303	0.508165253
2	1.052356420	-0.025955609	-0.052356420	0.496751115
3	1.026400811	-0.013143081	-0.026400811	0.509753688
4	1.013257730	-0.006614311	-0.013257730	0.501097775
5	1.006643419	-0.003318055	-0.006643419	0.500550093
:	:	:	:	:

Ejemplo 2.15 (Convergencia lineal cuando la raíz es doble). Partiendo de $p_0 = 1.2$ y usando la iteración de Newton-Raphson, vamos a aproximarnos a la raíz doble $p = 1$ del polinomio $f(x) = x^3 - 3x + 2$.

Usando la fórmula (19) con $R = 1$ para comprobar si la convergencia es lineal, obtenemos los valores que se muestran en la Tabla 2.6. ■

Hagamos notar que el método de Newton-Raphson converge a la raíz doble, pero con una velocidad bastante baja. Los valores de $f(p_k)$ en el Ejemplo 2.15 convergen a cero más rápidamente que los valores de $f'(p_k)$, de manera que el cociente $f(p_k)/f'(p_k)$ que aparece en la fórmula (4) puede calcularse cuando $p_k \neq p$. La sucesión converge linealmente y el error decrece en cada paso hasta, aproximadamente, la mitad de su tamaño. El siguiente teorema recoge la eficacia del método de Newton-Raphson, en términos de su orden de convergencia, para raíces simples y dobles.

Teorema 2.6 (Orden de convergencia del método de Newton-Raphson). Supongamos que el método de Newton-Raphson genera una sucesión $\{p_n\}_{n=0}^{\infty}$ que converge a un cero p de la función $f(x)$. Si p es una raíz simple, entonces la convergencia es cuadrática:

$$(23) \quad |E_{n+1}| \approx \frac{|f''(p)|}{2|f'(p)|} |E_n|^2 \quad \text{para } n \text{ suficientemente grande.}$$

Si p es una raíz múltiple de orden $M > 1$, entonces la convergencia es lineal:

$$(24) \quad |E_{n+1}| \approx \frac{M-1}{M} |E_n| \quad \text{para } n \text{ suficientemente grande.}$$

Inconvenientes y dificultades

Es fácil predecir el error que puede aparecer al dividir entre cero, pero hay otras dificultades que no son tan fáciles de advertir. Supongamos, por ejemplo, que

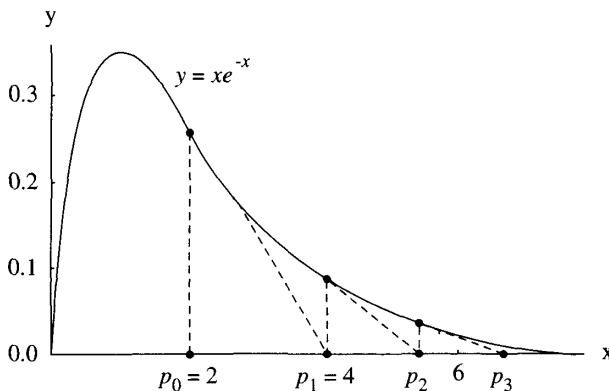


Figura 2.15 (a) La iteración de Newton-Raphson para $f(x) = xe^{-x}$ puede producir una sucesión divergente.

la función es $f(x) = x^2 - 4x + 5$, entonces la sucesión $\{p_k\}$ generada por la fórmula (4) oscilará de izquierda a derecha sin converger; un análisis somero de la situación revela que $f(x) > 0$ luego f no tiene ceros reales.

Algunas veces la aproximación inicial p_0 está demasiado lejos de la raíz deseada y la sucesión $\{p_k\}$ converge a otra raíz; esto suele ocurrir cuando la pendiente $f'(p_0)$ es pequeña y la recta tangente a la curva $y = f(x)$ es casi horizontal. Por ejemplo, si $f(x) = \cos(x)$, si la raíz que buscamos es $p = \pi/2$ y si empezamos con $p_0 = 3$, entonces los cálculos producen $p_1 = -4.01525255$, $p_2 = -4.85265757, \dots$, de forma que la sucesión $\{p_k\}$ converge a una raíz distinta: $-3\pi/2 \approx -4.71238898$.

Supongamos que $f(x)$ es positiva y monótona decreciente en un intervalo no acotado $[a, \infty)$ y que $p_0 > a$, entonces la sucesión $\{p_k\}$ podría diverger a $+\infty$. Por ejemplo, con $f(x) = xe^{-x}$ y $p_0 = 2.0$, tenemos

$$p_1 = 4.0, p_2 = 5.333333333, \dots, p_{15} = 19.723549434, \dots,$$

y $\{p_k\}$ diverge lentamente a $+\infty$ (véase la Figura 2.15(a)). Esta función concreta presenta otro problema inesperado: el valor de $f(x)$ tiende a cero rápidamente conforme x crece; por ejemplo, $f(p_{15}) = 0.0000000536$ y es posible que p_{15} pudiera tomarse erróneamente como una aproximación a una raíz. Por esta razón hemos diseñado un criterio de parada en el Programa 2.5 que involucra el error relativo $2|p_{k+1} - p_k|/(|p_k| + 10^{-6})$, de manera que para $k = 15$ este valor es 0.106817 y la tolerancia $\delta = 10^{-6}$ evita que el computador nos dé una raíz falsa.

Otro fenómeno, **la periodicidad**, ocurre cuando los términos de la sucesión $\{p_k\}$ tienden a repetirse o casi repetirse. Por ejemplo, si $f(x) = x^3 - x - 3$ y la

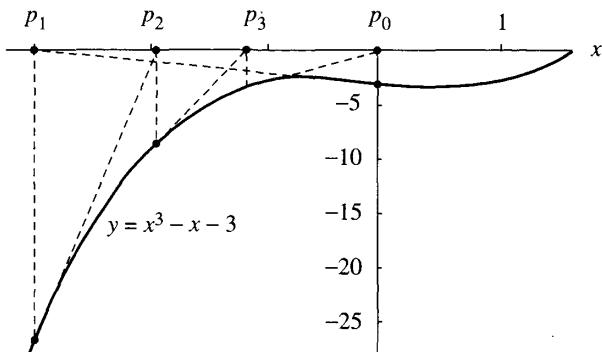


Figura 2.15 (b) La iteración de Newton-Raphson para $f(x) = x^3 - x - 3$ puede producir una sucesión periódica.

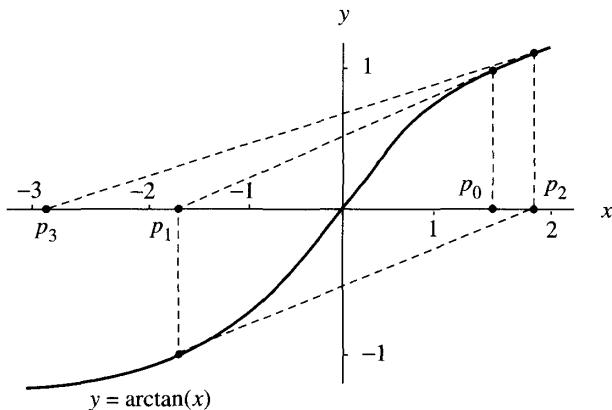


Figura 2.15 (c) La iteración de Newton-Raphson para $f(x) = \arctan(x)$ puede producir una sucesión oscilante y divergente.

aproximación inicial es $p_0 = 0$, entonces la sucesión que se obtiene es

$$\begin{aligned} p_1 &= -3.000000, & p_2 &= -1.961538, & p_3 &= -1.147176, & p_4 &= -0.006579, \\ p_5 &= -3.000389, & p_6 &= -1.961818, & p_7 &= -1.147430, & \dots \end{aligned}$$

y quedamos atrapados en un ciclo en el que $p_{k+4} \approx p_k$ para $k = 0, 1, \dots$ (véase la Figura 2.15(b)). Sin embargo, si el valor inicial p_0 está suficientemente cerca de la raíz $p \approx 1.671699881$, entonces $\{p_k\}$ converge; por ejemplo, si $p_0 = 2$, entonces la sucesión converge: $p_1 = 1.72727272$, $p_2 = 1.67369173$, $p_3 = 1.671702570$ y $p_4 = 1.671699881$.

Cuando $|g'(x)| \geq 1$ en un intervalo que contiene la raíz p , entonces hay una posibilidad de que aparezca una oscilación divergente: por ejemplo, sea $f(x) = \arctan(x)$; entonces la función de iteración para el método de Newton-Raphson es $g(x) = x - (1 + x^2) \arctan(x)$ y se tiene que $g'(x) = -2x \arctan(x)$. Si el valor de partida elegido es $p_0 = 1.45$, entonces

$$p_1 = -1.550263297, p_2 = 1.845931751, p_3 = -2.889109054, \text{ etc.}$$

(véase la Figura 2.15(c)). Pero si el valor de partida está suficientemente cerca de la raíz $p = 0$, entonces la sucesión converge; así, para $p_0 = 0.5$, tenemos

$$p_1 = -0.079559511, p_2 = 0.000335302, p_3 = 0.000000000.$$

Los ejemplos y situaciones que acabamos de describir inciden en el hecho de que debemos tener cuidado a la hora de proporcionar una respuesta. Algunas veces la sucesión no converge; no siempre se da el caso de que encontraremos una solución después de N iteraciones. Los programas de cálculo de raíces deben ser capaces de informar de esta situación. Si disponemos de información adicional sobre el contexto del problema, entonces disminuirá la posibilidad de que hallemos una solución errónea. A veces, es posible determinar un intervalo preciso en el que $f(x)$ tiene una raíz; el conocimiento del comportamiento de la función o una gráfica aceptable pueden ser de gran ayuda para elegir p_0 .

El método de la secante

En el algoritmo de Newton-Raphson hay que evaluar dos funciones en cada iteración, $f(p_{k-1})$ y $f'(p_{k-1})$. Tradicionalmente, el cálculo de la derivada de una función elemental puede llegar a suponer un esfuerzo considerable. Sin embargo, con los modernos paquetes de cálculo simbólico, esto no es un problema serio hoy en día. Aun así, hay muchas funciones dadas de forma no elemental (como integrales, o sumas de series, etc.) para las que sería deseable disponer de un método que converja casi tan rápidamente como el de Newton-Raphson y que necesite únicamente evaluaciones de $f(x)$ y no de $f'(x)$. El método de la secante necesita sólo una evaluación de $f(x)$ por paso y en una raíz simple tiene un orden de convergencia $R \approx 1.618033989$; es casi tan veloz como el de Newton, cuyo orden de convergencia es 2.

La fórmula de iteración del método de la secante es la misma que la que aparece en el método de la *régula falsi*, la diferencia entre ambos estriba en la estructura lógica de la forma de decidir cómo se elige el siguiente término. Partimos de dos puntos iniciales $(p_0, f(p_0))$ y $(p_1, f(p_1))$ cercanos al punto $(p, 0)$, como se muestra en la Figura 2.16, y se define p_2 como la abscisa del punto de intersección de la recta que pasa por estos dos puntos con el eje OX . La Figura 2.16 muestra que p_2 estará más cerca de p que p_0 y que p_1 . La

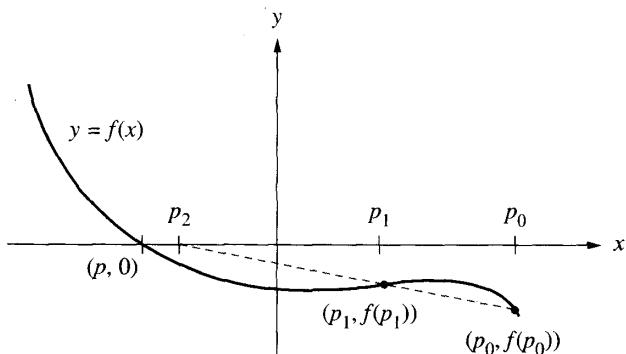


Figura 2.16 Construcción geométrica de p_2 en el método de la secante.

fórmula que relaciona p_2 , p_1 y p_0 se halla escribiendo la pendiente de la recta en cuestión:

$$(25) \quad m = \frac{f(p_1) - f(p_0)}{p_1 - p_0} \quad \text{y} \quad m = \frac{0 - f(p_1)}{p_2 - p_1}.$$

El primer valor de m en (25) es la pendiente de la recta secante que pasa por los dos puntos iniciales y el segundo valor es la pendiente de la recta que pasa por $(p_1, f(p_1))$ y $(p_2, 0)$. Igualando los miembros derechos de las dos fórmulas de (20) y despejando $p_2 = g(p_1, p_0)$ obtenemos

$$(26) \quad p_2 = g(p_1, p_0) = p_1 - \frac{f(p_1)(p_1 - p_0)}{f(p_1) - f(p_0)}.$$

El término general de la sucesión generada por este método viene dado por la fórmula de iteración de dos puntos

$$(27) \quad p_{k+1} = g(p_k, p_{k-1}) = p_k - \frac{f(p_k)(p_k - p_{k-1})}{f(p_k) - f(p_{k-1})}.$$

Ejemplo 2.16 (Método de la secante en una raíz simple). Empezando con $p_0 = -2.6$ y $p_1 = -2.4$, vamos a usar el método de la secante para aproximarnos a la raíz $p = -2$ de la función polinomial $f(x) = x^3 - 3x + 2$.

En este caso, la fórmula de iteración (27) es

$$(28) \quad p_{k+1} = g(p_k, p_{k-1}) = p_k - \frac{(p_k^3 - 3p_k + 2)(p_k - p_{k-1})}{p_k^3 - p_{k-1}^3 - 3p_k + 3p_{k-1}},$$

Tabla 2.7 Convergencia del método de la secante en una raíz simple.

k	p_k	$p_{k+1} - p_k$	$E_k = p - p_k$	$\frac{ E_{k+1} }{ E_k ^{1.618}}$
0	-2.600000000	0.200000000	0.600000000	0.914152831
1	-2.400000000	0.293401015	0.400000000	0.469497765
2	-2.106598985	0.083957573	0.106598985	0.847290012
3	-2.022641412	0.021130314	0.022641412	0.693608922
4	-2.001511098	0.001488561	0.001511098	0.825841116
5	-2.000022537	0.000022515	0.000022537	0.727100987
6	-2.000000022	0.000000022	0.000000022	
7	-2.000000000	0.000000000	0.000000000	

que podemos manipular algebraicamente y simplificar para obtener

$$(29) \quad p_{k+1} = g(p_k, p_{k-1}) = \frac{p_k^2 p_{k-1} + p_k p_{k-1}^2 - 2}{p_k^2 + p_k p_{k-1} + p_{k-1}^2 - 3}.$$

La sucesión generada aparece en la Tabla 2.7.

El método de la secante y el método de Newton-Raphson están relacionados de la siguiente manera: Para una función polinomial $f(x)$, la fórmula de dos puntos del método de la secante $p_{k+1} = g(p_k, p_{k-1})$ se reduce, después de simplificar, a la fórmula de un sólo punto de Newton-Raphson $p_{k+1} = g(p_k)$ cuando p_k sustituye a p_{k-1} ; de hecho, si reemplazamos p_{k-1} por p_k en (29), entonces el miembro derecho de esta fórmula se convierte en el miembro derecho de la fórmula (22) dada en el Ejemplo 2.14.

Los términos de la sucesión de los errores verifican, cuando la raíz es simple, la relación

$$(30) \quad |E_{k+1}| \approx |E_k|^{1.618} \left| \frac{f''(p)}{2f'(p)} \right|^{0.618},$$

siendo el orden de convergencia $R = (1 + \sqrt{5})/2 \approx 1.618$. Pueden encontrarse demostraciones de este hecho en los libros de análisis numérico avanzado. Vamos, no obstante, a hacer una comprobación usando el Ejemplo 2.16 y los valores concretos

$$|p - p_5| = 0.000022537$$

$$|p - p_4|^{1.618} = 0.001511098^{1.618} = 0.000027296,$$

y

$$A = |f''(-2)/2f'(-2)|^{0.618} = (2/3)^{0.618} = 0.778351205.$$

Tabla 2.8 Aceleración de la convergencia en una raíz doble.

k	p_k	$p_{k+1} - p_k$	$E_k = p - p_k$	$\frac{ E_{k+1} }{ E_k ^2}$
0	1.200000000	-0.193939394	-0.200000000	0.151515150
1	1.006060606	-0.006054519	-0.006060606	0.165718578
2	1.000006087	-0.000006087	-0.000006087	
3	1.000000000	0.000000000	0.000000000	

Combinando estos valores adecuadamente vemos que

$$|p - p_5| = 0.000022537 \approx 0.000021246 = A|p - p_4|^{1.618}.$$

Aceleración de la convergencia

Cabe esperar que existan métodos de cálculo de raíces que tengan un orden de convergencia mejor que el lineal cuando p sea una raíz múltiple de orden M . El último resultado que presentamos en esta sección muestra una modificación del método de Newton-Raphson cuya convergencia pasa a ser cuadrática en una raíz múltiple.

Teorema 2.7 (Iteración de Newton-Raphson acelerada). Supongamos que el algoritmo de Newton-Raphson produce una sucesión que converge linealmente a una raíz $x = p$ de orden $M > 1$. Entonces la fórmula de iteración de Newton-Raphson acelerada

$$(31) \quad p_k = p_{k-1} - \frac{Mf(p_{k-1})}{f'(p_{k-1})}$$

genera una sucesión $\{p_k\}_{k=0}^{\infty}$ que converge cuadráticamente a p .

Ejemplo 2.17 (Aceleración de la convergencia en una raíz doble). Partiendo de $p_0 = 1.2$ y usando la iteración de Newton-Raphson acelerada, vamos a aproximarnos a la raíz doble $p = 1$ de $f(x) = x^3 - 3x + 2$.

Puesto que $M = 2$, la fórmula de aceleración (31) es, en este caso,

$$p_k = p_{k-1} - 2 \frac{f(p_{k-1})}{f'(p_{k-1})} = \frac{p_{k-1}^3 + 3p_{k-1} - 4}{3p_{k-1}^2 - 3},$$

con la que obtenemos los valores que se muestran en la Tabla 2.8. ■

En la Tabla 2.9 se comparan las velocidades de convergencia de los métodos de cálculo de raíces que hemos visto hasta ahora. (El valor de la constante A es diferente para cada método.)

Tabla 2.9 Comparación de las velocidades de convergencia.

Método	Consideraciones especiales	Relación entre los sucesivos términos de error
Bisección		$E_{k+1} \approx \frac{1}{2} E_k $
<i>Régula falsi</i>		$E_{k+1} \approx A E_k $
Secante	Raíz múltiple	$E_{k+1} \approx A E_k $
Newton-Raphson	Raíz múltiple	$E_{k+1} \approx A E_k $
Secante	Raíz simple	$E_{k+1} \approx A E_k ^{1.618}$
Newton-Raphson	Raíz simple	$E_{k+1} \approx A E_k ^2$
Newton-Raphson acelerado	Raíz múltiple	$E_{k+1} \approx A E_k ^2$

Programa 2.5 (Iteración de Newton-Raphson). Aproximación a una raíz de $f(x) = 0$ a partir de un valor inicial p_0 mediante la iteración

$$p_k = p_{k-1} - \frac{f(p_{k-1})}{f'(p_{k-1})} \quad \text{para } k = 1, 2, \dots$$

```

function [p0,err,k,y]=newton(f,df,p0,delta,epsilon,max1)
% Datos
%   - f es la función,
%     introducida como una cadena de caracteres 'f'
%   - df es la derivada de f, introducida como una cadena 'df'
%   - p0 es la aproximación inicial a un cero de f
%   - delta es la tolerancia para p0
%   - epsilon es la tolerancia para los valores de la función
%   - max1 es el número máximo de iteraciones
% Resultados
%   - p0 es la aproximación al cero,
%     obtenida con el método de Newton-Raphson
%   - err es una estimación del error de p0
%   - k es el número de iteraciones realizadas
%   - y es el valor de la función f(p0)
for k=1:max1
    p1=p0-feval(f,p0)/feval(df,p0);
    err=abs(p1-p0);
    relerr=2*err/(abs(p1)+delta);
    p0=p1;
    y=feval(f,p0);
    if (err<delta)|(relerr<delta)|(abs(y)<epsilon),break,end
end

```

Programa 2.6 (Método de la secante). Aproximación a una raíz de $f(x) = 0$ a partir de unos valores iniciales p_0 y p_1 mediante la iteración

$$p_{k+1} = p_k - \frac{f(p_k)(p_k - p_{k-1})}{f(p_k) - f(p_{k-1})} \quad \text{para } k = 1, 2, \dots$$

```
function [p1,err,k,y]=secant(f,p0,p1,delta,epsilon,max1)
% Datos
%     - f la función,
%     introducida como una cadena de caracteres 'f'
%     - p0 y p1 son las aproximaciones iniciales a un cero de f
%     - delta es la tolerancia para p1
%     - epsilon es la tolerancia para los valores de la función
%     - max1 es el número máximo de iteraciones
% Resultados
%     - p1 es la aproximación al cero,
%     obtenida con el método de la secante
%     - err es una estimación del error de p1
%     - k es el número de iteraciones realizadas
%     - y es el valor de la función f(p1)

for k=1:max1
    p2=p1-feval(f,p1)*(p1-p0)/(feval(f,p1)-feval(f,p0));
    err=abs(p2-p1);
    relerr=2*err/(abs(p2)+delta);
    p0=p1;
    p1=p2;
    y=feval(f,p1);
    if (err<delta)|(relerr<delta)|(abs(y)<epsilon),break,end
end
```

Ejercicios

Las operaciones aritméticas pueden hacerse con una calculadora o con un computador.

1. Sea $f(x) = x^2 - x + 2$.
 - Determine la fórmula de Newton-Raphson $p_k = g(p_{k-1})$.
 - Empiece con $p_0 = -1.5$ y calcule p_1, p_2 y p_3 .
2. Sea $f(x) = x^2 - x - 3$.
 - Determine la fórmula de Newton-Raphson $p_k = g(p_{k-1})$.
 - Empiece con $p_0 = 1.6$ y calcule p_1, p_2 y p_3 .

- (c) Empiece con $p_0 = 0.0$ y calcule p_1, p_2, p_3 y p_4 . ¿Qué puede conjeturarse sobre esta sucesión?
3. Sea $f(x) = (x - 2)^4$.
- Determine la fórmula de Newton-Raphson $p_k = g(p_{k-1})$.
 - Empiece con $p_0 = 2.1$ y calcule p_1, p_2, p_3 y p_4 .
 - Esta sucesión, ¿converge linealmente o cuadráticamente?
4. Sea $f(x) = x^3 - 3x - 2$.
- Determine la fórmula de Newton-Raphson $p_k = g(p_{k-1})$.
 - Empiece con $p_0 = 2.1$ y calcule p_1, p_2, p_3 y p_4 .
 - Esta sucesión, ¿converge linealmente o cuadráticamente?
5. Considere la función $f(x) = \cos(x)$.
- Determine la fórmula de Newton-Raphson $p_k = g(p_{k-1})$.
 - Deseamos aproximarnos a la raíz $p = 3\pi/2$. ¿Podemos usar $p_0 = 3$? ¿Por qué?
 - Deseamos aproximarnos a la raíz $p = 3\pi/2$. ¿Podemos usar $p_0 = 5$? ¿Por qué?
6. Considere la función $f(x) = \arctan(x)$.
- Determine la fórmula de Newton-Raphson $p_k = g(p_{k-1})$.
 - Para $p_0 = 1.0$, calcule p_1, p_2, p_3 y p_4 . ¿Cuánto vale $\lim_{k \rightarrow \infty} p_k$?
 - Para $p_0 = 2.0$, calcule p_1, p_2, p_3 y p_4 . ¿Cuánto vale $\lim_{k \rightarrow \infty} p_k$?
7. Considere la función $f(x) = xe^{-x}$.
- Determine la fórmula de Newton-Raphson $p_k = g(p_{k-1})$.
 - Para $p_0 = 0.2$, calcule p_1, p_2, p_3 y p_4 . ¿Cuánto vale $\lim_{k \rightarrow \infty} p_k$?
 - Para $p_0 = 20$ calcule p_1, p_2, p_3 y p_4 . ¿Cuánto vale $\lim_{k \rightarrow \infty} p_k$?
 - ¿Cuánto vale $f(p_4)$ en el apartado (c)?

En los Ejercicios 8 a 10, use el método de la secante, con la fórmula (27), y calcule los puntos siguientes p_2 y p_3 .

- Sea $f(x) = x^2 - 2x - 1$. Empiece con $p_0 = 2.6$ y $p_1 = 2.5$.
- Sea $f(x) = x^2 - x - 3$. Empiece con $p_0 = 1.7$ y $p_1 = 1.67$.
- Sea $f(x) = x^3 - x + 2$. Empiece con $p_0 = -1.5$ y $p_1 = -1.52$.
- Algoritmo para la raíz cúbica.** A partir de $f(x) = x^3 - A$, donde A es un número real cualquiera, deduzca la fórmula de recursión

$$p_k = \frac{2p_{k-1} + A/p_{k-1}^2}{3} \quad \text{para } k = 1, 2, \dots$$

12. Considere $f(x) = x^N - A$, siendo N un número natural.

- ¿Cuáles son las soluciones reales de $f(x) = 0$ según las diversas elecciones de N y A que pueden hacerse?

(b) Deduzca la fórmula de recursión

$$p_k = \frac{(N-1)p_{k-1} + A/p_{k-1}^{N-1}}{N} \quad \text{para } k = 1, 2, \dots$$

para hallar la raíz N -ésima de A .

13. ¿Podemos usar el método de Newton-Raphson para resolver $f(x) = 0$ siendo $f(x) = x^2 - 14x + 50$? ¿Por qué?
14. ¿Podemos usar el método de Newton-Raphson para resolver $f(x) = 0$ siendo $f(x) = x^{1/3}$? ¿Por qué?
15. ¿Podemos usar el método de Newton-Raphson para resolver $f(x) = 0$ siendo $f(x) = (x - 3)^{1/2}$ tomando $p_0 = 4$ como valor inicial? ¿Por qué?
16. Establezca el límite de la sucesión definida en (11).
17. Pruebe que la sucesión $\{p_k\}$ generada con la fórmula (4) del Teorema 2.5 converge a p siguiendo los pasos que se relacionan a continuación.
 - (a) Pruebe que si p es un punto fijo de la función $g(x)$ dada en (5) entonces p es un cero de $f(x)$.
 - (b) Pruebe que si p es un cero de $f(x)$ y $f'(p) \neq 0$, entonces $g'(p) = 0$. Utilice el apartado (a) y el Teorema 2.3 para probar que la sucesión $\{p_k\}$ generada con la fórmula (4) converge a p .
18. Pruebe la estimación dada en la fórmula (23) del Teorema 2.6 siguiendo los pasos que se relacionan a continuación. De acuerdo con el Teorema 1.11, podemos desarrollar $f(x)$ alrededor de $x = p_k$ de manera que

$$f(x) = f(p_k) + f'(p_k)(x - p_k) + \frac{1}{2}f''(c_k)(x - p_k)^2.$$

Puesto que p es un cero de $f(x)$, ponemos $x = p$ y nos queda

$$0 = f(p_k) + f'(p_k)(p - p_k) + \frac{1}{2}f''(c_k)(p - p_k)^2.$$

- (a) Ahora supongamos que $f'(x) \neq 0$ para todo x cerca de la raíz p . Use la expresión anterior junto con $f'(p_k) \neq 0$ para deducir que

$$p - p_k + \frac{f(p_k)}{f'(p_k)} = \frac{-f''(c_k)}{2f'(p_k)}(p - p_k)^2.$$

- (b) Supongamos que $f'(x)$ y $f''(x)$ no cambian demasiado rápidamente de manera que podemos usar las aproximaciones $f'(p_k) \approx f'(p)$ y $f''(c_k) \approx f''(p)$. Utilice entonces el apartado (a) para obtener

$$E_{k+1} \approx \frac{-f''(p)}{2f'(p)} E_k^2.$$

19. Supongamos que A es un número real y positivo.

- (a) Pruebe que A puede escribirse como $A = q \times 2^{2m}$, siendo $1/4 \leq q < 1$ y m un entero.
- (b) Use el apartado (a) para probar que la raíz cuadrada de A es $A^{1/2} = q^{1/2} \times 2^m$.

Observación. Si tomamos $p_0 = (2q + 1)/3$, siendo $1/4 \leq q < 1$, y usamos la fórmula de Newton (11), entonces después de tres iteraciones, p_3 será una aproximación a $q^{1/2}$ con 24 cifras binarias de precisión. Este es el algoritmo que los computadores usan habitualmente para hallar raíces cuadradas.

20. (a) Pruebe que la fórmula (27) del método de la secante es algebraicamente equivalente a

$$p_{k+1} = \frac{p_{k-1}f(p_k) - p_kf(p_{k-1})}{f(p_k) - f(p_{k-1})}.$$

- (b) Explique por qué la pérdida de cifras significativas en la resta hace que esta fórmula sea peor, para propósitos computacionales, que la fórmula dada en (27).

21. Supongamos que p es una raíz de orden $M = 2$ de la ecuación $f(x) = 0$. Pruebe que la iteración de Newton-Raphson acelerada

$$p_k = p_{k-1} - \frac{2f(p_{k-1})}{f'(p_{k-1})}$$

converge cuadráticamente (véase el Ejercicio 18).

22. El **método de Halley** es otra forma de acelerar la convergencia del método de Newton-Raphson. La fórmula de iteración de Halley es

$$g(x) = x - \frac{f(x)}{f'(x)} \left(1 - \frac{f(x)f''(x)}{2(f'(x))^2} \right)^{-1};$$

el término entre paréntesis es la modificación introducida en la fórmula del método de Newton-Raphson. El método de Halley proporciona un orden de convergencia triple ($R = 3$) en los ceros simples de $f(x)$.

- (a) A partir de $f(x) = x^2 - A$, determine la función de iteración de Halley $g(x)$ para hallar \sqrt{A} . Empiece con $p_0 = 2$ para aproximar $\sqrt{5}$ y calcule p_1 , p_2 y p_3 .
- (b) A partir de $f(x) = x^3 - 3x + 2$, determine la función de iteración de Halley $g(x)$. Empiece con $p_0 = -2.4$ y calcule p_1 , p_2 y p_3 .

23. Un **método de Newton-Raphson modificado para raíces múltiples**. Si p es una raíz de multiplicidad M , entonces $f(x) = (x - p)^M q(x)$, donde $q(p) \neq 0$.

- (a) Pruebe que $h(x) = f(x)/f'(x)$ tiene una raíz simple en p .

- (b) Pruebe que si aplicamos el método de Newton-Raphson para hallar la raíz simple p de $h(x)$, entonces se obtiene $g(x) = x - h(x)/h'(x)$ que, en este caso, es

$$g(x) = x - \frac{f(x)f'(x)}{(f'(x))^2 - f(x)f''(x)}.$$

- (c) La iteración que usa la función $g(x)$ dada en el apartado (b) converge cuadráticamente a p . Explique por qué ocurre esto.
 (d) La función $f(x) = \operatorname{sen}(x^3)$ tiene en $p = 0$ un cero triple. Partiendo de $p_0 = 1$, calcule p_1 , p_2 y p_3 usando el método de Newton-Raphson modificado.
24. Supongamos que un método iterativo para resolver una ecuación $f(x) = 0$ produce los siguientes términos de error (véase el Ejemplo 2.11): $E_0 = 0.400000$, $E_1 = 0.043797$, $E_2 = 0.000062$ y $E_3 = 0.000000$. Estime la constante asintótica del error A y el orden de convergencia R de la sucesión generada por este método iterativo.

Algoritmos y programas

- Modifique los Programas 2.5 y 2.6 para que generen mensajes de error apropiados cuando (i) se divida entre cero en (4) o (27), respectivamente, o bien (ii) se exceda el número máximo de iteraciones `max1`.
- A menudo resulta instructivo disponer de todos los términos de las sucesiones generadas por las fórmulas que aparecen en (4) y (27) (o sea, la segunda columna de la Tabla 2.4). Modifique los Programas 2.5 y 2.6 para que den como resultado las sucesiones generadas por (4) y (27), respectivamente.
- Modifique el Programa 2.5 de manera que pueda usar el método de Newton para calcular raíces cuadradas, y aproxime cada una de las siguientes raíces cuadradas con una precisión de 10 cifras decimales.
 - Empiece con $p_0 = 3$ y aproxime $\sqrt{8}$.
 - Empiece con $p_0 = 10$ y aproxime $\sqrt{91}$.
 - Empiece con $p_0 = -3$ y aproxime $-\sqrt{8}$.
- Modifique el Programa 2.5 de manera que pueda usar el algoritmo para calcular raíces cúbicas, descrito en el Ejercicio 11, para aproximar las siguientes raíces cúbicas hasta la décima cifra decimal.
 - Empiece con $p_0 = 2$ y aproxime $7^{1/3}$.
 - Empiece con $p_0 = 6$ y aproxime $200^{1/3}$.
 - Empiece con $p_0 = -2$ y aproxime $(-7)^{1/3}$.

5. Modifique el Programa 2.5 de manera que pueda usar el algoritmo de Newton-Raphson acelerado, descrito en el Teorema 2.7, para hallar la raíz p de orden M de cada una de las siguientes funciones.
- $f(x) = (x - 2)^5$, $M = 5$, $p = 2$; empiece con $p_0 = 1$.
 - $f(x) = \operatorname{sen}(x^3)$, $M = 3$, $p = 0$; empiece con $p_0 = 1$.
 - $f(x) = (x - 1) \ln(x)$, $M = 2$, $p = 1$; empiece con $p_0 = 2$.
6. Modifique el Programa 2.5 para que pueda usar el método de Halley, descrito en el Ejercicio 22, para hallar el cero simple de $f(x) = x^3 - 3x + 2$ partiendo de $p_0 = -2.4$.
7. Supongamos que las ecuaciones del movimiento de un proyectil son

$$y = f(t) = 4605(1 - e^{-t/15}) - 147t,$$

$$x = r(t) = 2400(1 - e^{-t/15}).$$

- Determine el tiempo transcurrido hasta el impacto con diez cifras decimales de precisión.
 - Determine el alcance del disparo con diez cifras decimales de precisión.
8. (a) Halle el punto de la parábola $y = x^2$ que está más cerca del punto $(3, 1)$ con diez cifras decimales de precisión.
- (b) Halle el punto de la curva $y = \operatorname{sen}(x - \operatorname{sen}(x))$ que está más cerca del punto $(2.1, 0.5)$ con diez cifras decimales de precisión.
- (c) Halle, con una precisión de diez cifras decimales, el valor de x para el que es mínima la distancia vertical entre las gráficas de las funciones $f(x) = x^2 + 2$ y $g(x) = (x/5) - \operatorname{sen}(x)$.
9. Se construye una caja sin tapadera a partir de una hoja metálica rectangular que mide 10 por 16 centímetros. ¿Cuál debe ser el lado de los cuadrados que hay que recortar en cada esquina para que el volumen de la caja sea 100 centímetros cúbicos? Precisión: 0.000000001 centímetros.
10. La curva formada por un cable colgante se llama catenaria. Supongamos que el punto más bajo de una catenaria es el origen $(0, 0)$, entonces la ecuación de la catenaria es $y = C \cosh(x/C) - C$. Si queremos determinar la catenaria que pasa por los puntos $(\pm a, b)$, entonces debemos resolver la ecuación $b = C \cosh(a/C) - C$ donde la incógnita es C .
- Pruebe que la catenaria que pasa por los puntos $(\pm 10, 6)$ es

$$y = 9.1889 \cosh(x/9.1889) - 9.1889.$$

- Halle la catenaria que pasa por los puntos $(\pm 12, 5)$.

2.5 Los métodos de Aitken, Steffensen y Muller (opcional)

En la Sección 2.4 vimos que el método de Newton-Raphson converge lentamente en una raíz múltiple y que la sucesión de aproximaciones $\{p_k\}$ presenta convergencia lineal. El Teorema 2.7 proporciona un método de aceleración de la convergencia, pero este método depende del conocimiento previo del orden de la raíz.

El método de Aitken

Una técnica conocida como **método Δ^2 de Aitken** permite acelerar la convergencia de cualquier sucesión que sea linealmente convergente. Para presentar esta técnica necesitamos una definición previa.

Definición 2.6. Dada una sucesión $\{p_n\}_{n=0}^{\infty}$, se define la diferencia progresiva Δp_n mediante

$$(1) \quad \Delta p_n = p_{n+1} - p_n \quad \text{para } n \geq 0.$$

Las potencias superiores $\Delta^k p_n$ se definen de manera recursiva mediante

$$(2) \quad \Delta^k p_n = \Delta^{k-1}(\Delta p_n) \quad \text{para } k \geq 2.$$

Teorema 2.8 (Aceleración de Aitken). Sea $\{p_n\}_{n=0}^{\infty}$ una sucesión que converge linealmente a su límite p y tal que $p - p_n \neq 0$ para todo $n \geq 0$. Si existe un número real A con $|A| < 1$ y tal que

$$(3) \quad \lim_{n \rightarrow \infty} \frac{p - p_{n+1}}{p - p_n} = A,$$

entonces la sucesión $\{q_n\}_{n=0}^{\infty}$ definida por

$$(4) \quad q_n = p_n - \frac{(\Delta p_n)^2}{\Delta^2 p_n} = p_n - \frac{(p_{n+1} - p_n)^2}{p_{n+2} - 2p_{n+1} + p_n}$$

converge a p más rápidamente que $\{p_n\}_{n=0}^{\infty}$, en el sentido de que

$$(5) \quad \lim_{n \rightarrow \infty} \left| \frac{p - q_n}{p - p_n} \right| = 0.$$

Demostración. Mostraremos cómo se deduce la fórmula (4) y dejaremos la demostración de (5) como un ejercicio. Usando la relación (3), podemos escribir

$$(6) \quad \frac{p - p_{n+1}}{p - p_n} \approx A \quad \text{y} \quad \frac{p - p_{n+2}}{p - p_{n+1}} \approx A \quad \text{cuando } n \text{ es grande.}$$

Tabla 2.10 Sucesión linealmente convergente $\{p_n\}$.

n	p_n	$E_n = p_n - p$	$A_n = \frac{E_n}{E_{n-1}}$
1	0.606530660	0.039387369	-0.586616609
2	0.545239212	-0.021904079	-0.556119357
3	0.579703095	0.012559805	-0.573400269
4	0.560064628	-0.007078663	-0.563596551
5	0.571172149	0.004028859	-0.569155345
6	0.564862947	-0.002280343	-0.566002341

Tabla 2.11 Sucesión $\{q_n\}$ obtenida con el método de Aitken.

n	q_n	$q_n - p$
1	0.567298989	0.000155699
2	0.567193142	0.000049852
3	0.567159364	0.000016074
4	0.567148453	0.000005163
5	0.567144952	0.000001662
6	0.567143825	0.000000534

Las aproximaciones dadas en (6) implican que

$$(7) \quad (p - p_{n+1})^2 \approx (p - p_{n+2})(p - p_n).$$

Haciendo las operaciones en ambos miembros de (7) y simplificando los términos p^2 , obtenemos

$$(8) \quad p \approx \frac{p_{n+2}p_n - p_{n+1}^2}{p_{n+2} - 2p_{n+1} + p_n} = q_n \quad \text{para } n = 0, 1, \dots$$

La fórmula (8), que define el término q_n , puede ser manipulada algebraicamente para convertirla en la fórmula (4), que tiene mejor comportamiento frente a la propagación de los errores cuando se usan con un computador.

Ejemplo 2.18. Veamos que la sucesión $\{p_n\}$ del Ejemplo 2.2 presenta convergencia lineal y que la sucesión $\{q_n\}$ obtenida con el método Δ^2 de Aitken converge más rápidamente.

La sucesión $\{p_n\}$ se obtuvo aplicando iteración de punto fijo a la función dada por $g(x) = e^{-x}$ a partir del valor inicial $p_0 = 0.5$. Una vez detenidas las iteraciones,

el límite calculado fue $p \approx 0.567143290$. Los valores p_n y q_n se muestran en las Tablas 2.10 y 2.11. Para ilustrar lo que decimos, observemos que el valor q_1 se obtiene con el cálculo

$$\begin{aligned} q_1 &= p_1 - \frac{(p_2 - p_1)^2}{p_3 - 2p_2 + p_1} \\ &= 0.606530660 - \frac{(-0.061291448)^2}{0.095755331} = 0.567298989. \end{aligned}$$

■

Aunque la sucesión $\{q_n\}$ de la Tabla 2.11 converge linealmente, su velocidad de convergencia es mayor que la de $\{p_n\}$ en el sentido dado en el Teorema 2.8; de hecho, el método de Aitken produce, usualmente, mejoras más sustanciales que la de este caso. Cuando el método de Aitken se combina con una iteración de punto fijo, el resultado se conoce como **método de aceleración de Steffensen**. Los detalles de este método se dan en el Programa 2.7 y en los ejercicios.

El método de Muller

El método de Muller es una generalización del método de la secante, en el sentido de que no necesita el cálculo de la derivada de la función. Es un método iterativo que necesita tres puntos iniciales $(p_0, f(p_0))$, $(p_1, f(p_1))$ y $(p_2, f(p_2))$. Entonces se construye la parábola que pasa por estos puntos y se usa la fórmula de resolución de las ecuaciones de segundo grado para determinar el punto de corte de dicha parábola con el eje OX ; la abscisa de este punto se usa para construir la siguiente aproximación. Puede probarse que, cerca de una raíz simple, el método de Muller converge más rápidamente que el método de la secante y es casi tan veloz como el método de Newton-Raphson. Este método puede emplearse tanto para hallar ceros reales como para hallar ceros complejos y puede programarse para que trabaje con aritmética compleja. Sin pérdida de generalidad, supongamos que p_2 es la mejor aproximación a la raíz y consideremos la parábola, que se muestra en la Figura 2.17, que pasa por los tres puntos de partida. Hagamos el cambio de variable

$$(9) \quad t = x - p_2,$$

y usemos las diferencias

$$(10) \quad h_0 = p_0 - p_2 \quad \text{y} \quad h_1 = p_1 - p_2.$$

Consideremos el polinomio cuadrático en la variable t :

$$(11) \quad y = at^2 + bt + c.$$

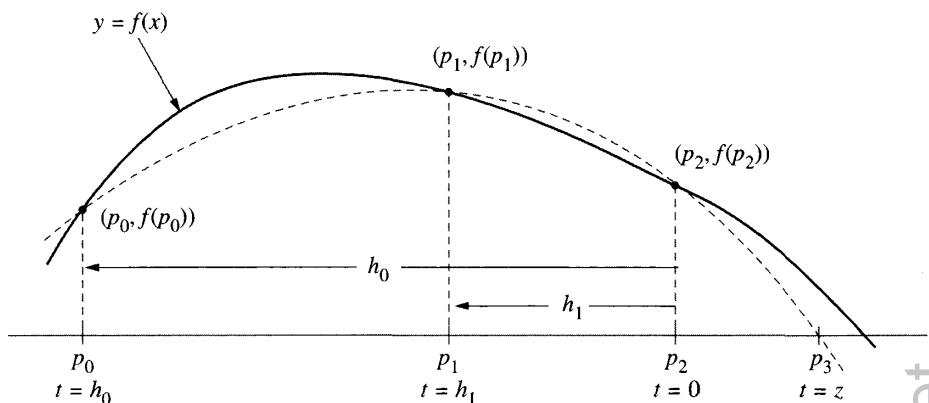


Figura 2.17 Los valores iniciales p_0 , p_1 y p_2 para el método de Muller y las diferencias h_0 y h_1 .

Cada punto inicial nos proporciona una ecuación para a , b y c :

$$(12) \quad \begin{aligned} \text{En } t = h_0: \quad & ah_0^2 + bh_0 + c = f(p_0), \\ \text{En } t = h_1: \quad & ah_1^2 + bh_1 + c = f(p_1), \\ \text{En } t = 0: \quad & a0^2 + b0 + c = f(p_2). \end{aligned}$$

La tercera ecuación de (12) nos dice que

$$(13) \quad c = f(p_2).$$

Sustituyendo (13) en las primeras dos ecuaciones de (12) y poniendo $e_0 = f(p_0) - c$ y $e_1 = f(p_1) - c$ nos queda

$$(14) \quad \begin{aligned} ah_0^2 + bh_0 &= f(p_0) - c = e_0, \\ ah_1^2 + bh_1 &= f(p_1) - c = e_1. \end{aligned}$$

Ahora resolvemos este sistema y obtenemos a y b :

$$(15) \quad \begin{aligned} a &= \frac{e_0 h_1 - e_1 h_0}{h_1 h_0^2 - h_0 h_1^2}, \\ b &= \frac{e_1 h_0^2 - e_0 h_1^2}{h_1 h_0^2 - h_0 h_1^2}. \end{aligned}$$

Las raíces $t = z_1, z_2$ de (11) se obtienen usando la fórmula

$$(16) \quad z = \frac{-2c}{b \pm \sqrt{b^2 - 4ac}}.$$

Esta fórmula (16) es equivalente a la fórmula habitual para hallar las raíces de una ecuación de segundo grado; pero, en este caso, (16) es mejor porque ya sabemos que $c = f(p_2)$.

Para asegurar la estabilidad del método hay que elegir la raíz de (16) que tenga menor valor absoluto; así, si $b > 0$, entonces usamos el signo positivo de la raíz cuadrada, mientras que si $b < 0$, entonces usamos el signo negativo. El nuevo punto p_3 , que se muestra en la Figura 2.17, viene dado por

$$(17) \quad p_3 = p_2 + z.$$

Para actualizar los valores y dar el siguiente paso, elegimos los nuevos p_0 y p_1 como los dos puntos más cercanos a p_3 de entre $\{p_0, p_1, p_2\}$ (o sea, rechazamos el más lejano) y p_3 pasa a ser el nuevo p_2 . Aunque hace falta realizar muchos cálculos adicionales en el método de Muller, sólo hace falta una evaluación de la función f en cada iteración.

Si usamos el método de Muller para hallar las raíces reales de $f(x) = 0$, es posible que nos encontremos con aproximaciones complejas, ya que las raíces de (16) podrían ser números complejos (o sea, tener parte imaginaria no nula). En estos casos las partes imaginarias serán de pequeña magnitud y podremos despreciarlas, de manera que nuestros cálculos se hagan con números reales.

Comparación entre los métodos

El método de Steffensen puede usarse junto con la función de iteración del método de Newton-Raphson $g(x) = x - f(x)/f'(x)$. En los dos ejemplos siguientes nos vamos a plantear el cálculo de las raíces del polinomio $f(x) = x^3 - 3x + 2$, para el cual la función de iteración del método de Newton-Raphson es $g(x) = (2x^3 - 2)/(3x^2 - 3)$. Cuando usamos esta función en el Programa 2.7, obtenemos los cálculos que se muestran bajo la cabecera Steffensen con Newton en las Tablas 2.12 y 2.13. Por ejemplo, empezando con $p_0 = -2.4$, calculamos

$$(18) \quad p_1 = g(p_0) = -2.076190476,$$

y

$$(19) \quad p_2 = g(p_1) = -2.003596011.$$

Entonces la fórmula de Aitken nos proporciona $p_3 = -1.982618143$.

Ejemplo 2.19 (Convergencia cerca de una raíz simple.). Vamos a comparar los métodos dados usando la función $f(x) = x^3 - 3x + 2$ cerca de su raíz simple $p = -2$.

Los resultados producidos por los métodos de Newton-Raphson y de la secante para este caso ya se mostraron en los Ejemplos 2.14 y 2.16, respectivamente. La Tabla 2.12 proporciona un resumen de los resultados obtenidos con todos los métodos. ■

Tabla 2.12 Comparación de las convergencias cerca de una raíz simple.

k	Método de la secante	Método de Muller	Método de Newton-Raphson	Steffensen con Newton-Raphson
0	-2.600000000	-2.600000000	-2.400000000	-2.400000000
1	-2.400000000	-2.500000000	-2.076190476	-2.076190476
2	-2.106598985	-2.400000000	-2.003596011	-2.003596011
3	-2.022641412	-1.985275287	-2.000008589	-1.982618143
4	-2.001511098	-2.000334062	-2.000000000	-2.000204982
5	-2.000022537	-2.000000218		-2.000000028
6	-2.000000022	-2.000000000		-2.0000002389
7	-2.000000000			-2.000000000

Ejemplo 2.20 (Convergencia cerca de una raíz doble). Vamos a comparar ahora los métodos dados usando la función $f(x) = x^3 - 3x + 2$ cerca de su raíz doble $p = 1$. La Tabla 2.13 proporciona un resumen de los resultados obtenidos con todos los métodos.

El método de Newton-Raphson es la mejor elección para hallar una raíz simple como se ve en la Tabla 2.12. En una raíz doble, tanto el método de Muller como el método de Steffensen aplicado a la fórmula de iteración de Newton-Raphson resultan ser una buena elección, como se ve en la Tabla 2.13. Hagamos notar que, en la fórmula del método de aceleración de Aitken (4), podríamos tener una división entre cero conforme la sucesión $\{p_k\}$ converge. En este caso, la última aproximación calculada es la que usamos como aproximación al cero.

En el programa que damos a continuación, la sucesión $\{p_k\}$, generada con el método de Steffensen aplicado a la fórmula de iteración de Newton-Raphson, se almacena en una matriz Q que tiene `max1` filas y tres columnas: La primera columna de Q contiene la aproximación inicial p_0 y los términos $p_3, p_6, \dots, p_{3k}, \dots$ generados por el método de aceleración de Aitken (4), mientras que las columnas segunda y tercera de Q contienen los términos generados por el método de Newton. El criterio de parada del programa se basa en la diferencia entre dos términos consecutivos de la primera columna de Q .

Programa 2.7 (Aceleración de Steffensen). Aceleración de la convergencia de la iteración de punto fijo para resolver la ecuación $f(x) = 0$ a partir de p_0 ; donde se supone que f y f' son continuas y que el método de Newton-Raphson converge lentamente (linealmente) a p .

Tabla 2.13 Comparación de la convergencia cerca de una raíz doble.

<i>k</i>	Método de la secante	Método de Muller	Método de Newton-Raphson	Steffensen con Newton-Raphson
0	1.400000000	1.400000000	1.200000000	1.200000000
1	1.200000000	1.300000000	1.103030303	1.103030303
2	1.138461538	1.200000000	1.052356417	1.052356417
3	1.083873738	1.003076923	1.026400814	0.996890433
4	1.053093854	1.003838922	1.013257734	0.998446023
5	1.032853156	1.000027140	1.006643418	0.999223213
6	1.020429426	0.999997914	1.003325375	0.999999193
7	1.012648627	0.999999747	1.001663607	0.999999597
8	1.007832124	1.000000000	1.000832034	0.999999798
9	1.004844757		1.000416075	0.999999999
	:		:	

```

function [p,Q]=steff(f,df,p0,delta,epsilon,max1)

% Datos
%     - f la función, introducida como
%       una cadena de caracteres 'f'
%     - df es la derivada de f, introducida como una cadena 'df'
%     - p0 es la aproximación inicial a un cero de f
%     - delta es la tolerancia para p0
%     - epsilon es la tolerancia para los valores de la función
%     - max1 es el número máximo de iteraciones
%
% Resultados
%     - p es la aproximación al cero,
%       obtenida con el método de Steffensen
%     - Q es la matriz que contiene la sucesión de Steffensen

% Inicializamos la matriz R
R=zeros(max1,3);
R(1,1)=p0;
for k=1:max1
    for j=2:3
        % Denominador del método de Newton-Raphson
        nrdenom=feval(df,R(k,j-1));
        % Apoximación de Newton-Raphson
        if nrdenom==0

```

Tabla 2.12 Comparación de las convergencias cerca de una raíz simple.

k	Método de la secante	Método de Muller	Método de Newton-Raphson	Steffensen con Newton-Raphson
0	-2.600000000	-2.600000000	-2.400000000	-2.400000000
1	-2.400000000	-2.500000000	-2.076190476	-2.076190476
2	-2.106598985	-2.400000000	-2.003596011	-2.003596011
3	-2.022641412	-1.985275287	-2.000008589	-1.982618143
4	-2.001511098	-2.000334062	-2.000000000	-2.000204982
5	-2.000022537	-2.000000218		-2.000000028
6	-2.000000022	-2.000000000		-2.000002389
7	-2.000000000			-2.000000000

Ejemplo 2.20 (Convergencia cerca de una raíz doble). Vamos a comparar ahora los métodos dados usando la función $f(x) = x^3 - 3x + 2$ cerca de su raíz doble $p = 1$. La Tabla 2.13 proporciona un resumen de los resultados obtenidos con todos los métodos.

El método de Newton-Raphson es la mejor elección para hallar una raíz simple como se ve en la Tabla 2.12. En una raíz doble, tanto el método de Muller como el método de Steffensen aplicado a la fórmula de iteración de Newton-Raphson resultan ser una buena elección, como se ve en la Tabla 2.13. Hagamos notar que, en la fórmula del método de aceleración de Aitken (4), podríamos tener una división entre cero conforme la sucesión $\{p_k\}$ converge. En este caso, la última aproximación calculada es la que usamos como aproximación al cero.

En el programa que damos a continuación, la sucesión $\{p_k\}$, generada con el método de Steffensen aplicado a la fórmula de iteración de Newton-Raphson, se almacena en una matriz Q que tiene max1 filas y tres columnas: La primera columna de Q contiene la aproximación inicial p_0 y los términos $p_3, p_6, \dots, p_{3k}, \dots$ generados por el método de aceleración de Aitken (4), mientras que las columnas segunda y tercera de Q contienen los términos generados por el método de Newton. El criterio de parada del programa se basa en la diferencia entre dos términos consecutivos de la primera columna de Q .

Programa 2.7 (Aceleración de Steffensen). Aceleración de la convergencia de la iteración de punto fijo para resolver la ecuación $f(x) = 0$ a partir de p_0 ; donde se supone que f y f' son continuas y que el método de Newton-Raphson converge lentamente (linealmente) a p .

Tabla 2.13 Comparación de la convergencia cerca de una raíz doble.

<i>k</i>	Método de la secante	Método de Muller	Método de Newton-Raphson	Steffensen con Newton-Raphson
0	1.400000000	1.400000000	1.200000000	1.200000000
1	1.200000000	1.300000000	1.103030303	1.103030303
2	1.138461538	1.200000000	1.052356417	1.052356417
3	1.083873738	1.003076923	1.026400814	0.996890433
4	1.053093854	1.003838922	1.013257734	0.998446023
5	1.032853156	1.000027140	1.006643418	0.999223213
6	1.020429426	0.999997914	1.003325375	0.999999193
7	1.012648627	0.999999747	1.001663607	0.999999597
8	1.007832124	1.000000000	1.000832034	0.999999798
9	1.004844757		1.000416075	0.999999999
	:		:	:

```

function [p,Q]=steff(f,df,p0,delta,epsilon,max1)
% Datos
% - f la función, introducida como
%   una cadena de caracteres 'f'
% - df es la derivada de f, introducida como una cadena 'df'
% - p0 es la aproximación inicial a un cero de f
% - delta es la tolerancia para p0
% - epsilon es la tolerancia para los valores de la función
% - max1 es el número máximo de iteraciones
% Resultados
% - p es la aproximación al cero,
%   obtenida con el método de Steffensen
% - Q es la matriz que contiene la sucesión de Steffensen
% Inicializamos la matriz R
R=zeros(max1,3);
R(1,1)=p0;
for k=1:max1
    for j=2:3
        % Denominador del método de Newton-Raphson
        nrdenom=feval(df,R(k,j-1));
        % Apoximación de Newton-Raphson
        if nrdenom==0

```

```

'división entre cero en el método de Newton-Raphson'
break
else
    R(k,j)=R(k,j-1)-feval(f,R(k,j-1))/nrdenom;
end

% Denominador del método de aceleración de Aitken
aadenom=R(k,3)-2*R(k,2)+R(k,1);

% Cálculo de las aproximaciones de Aitken
if aadenom==0
    'división entre cero en el método de Aitken'
    break
else
    R(k+1,1)=R(k,1)-(R(k,2)-R(k,1))^2/aadenom;
end

end

% Final del programa si ocurre división entre cero
if (nrdenom==0)|(aadenom==0)
    break
end

% Criterios de parada
err=abs(R(k,1)-R(k+1,1));
relerr=err/(abs(R(k+1,1))+delta);
y=feval(f,R(k+1,1));
if (err<delta)|(relerr<delta)|(y<epsilon)
    % determinación de p y de la matriz Q
    p=R(k+1,1);
    Q=R(1:k+1,:);
    break
end

end

```

Programa 2.8 (Método de Muller). Cálculo de una raíz de $f(x) = 0$ a partir de tres aproximaciones iniciales p_0 , p_1 y p_2 .

```

function [p,y,err]=muller(f,p0,p1,p2,delta epsilon,max1)

% Datos
% - f la función, introducida como
%   una cadena de caracteres 'f'
% - p0, p1 y p2 son las aproximaciones iniciales
% - delta es la tolerancia para p0, p1 y p2
% - epsilon es la tolerancia para los valores de la función
% - max1 es el número máximo de iteraciones

```

```
% Resultados
%      - p es la aproximación al cero,
%      obtenida con el método de Muller
%      - err es el error en la aproximación p
%
% Inicializamos las matrices P e Y
P=[p0 p1 p2];
Y=feval(f,P);
%
% Cálculo de a y b en la fórmula (15)
for k=1:max1
    h0=P(1)-P(3);h1=P(2)-P(3);e0=Y(1)-Y(3);e1=Y(2)-Y(3);c=Y(3);
    denom=h1*h0^2-h0*h1^2;
    a=(e0*h1-e1*h0)/denom;
    b=(e1*h0^2-e0*h1^2)/denom;
%
    % Supresión de la parte imaginaria de las raíces complejas
    if b^2-4*a*c > 0
        disc=sqrt(b^2-4*a*c);
    else
        disc=0;
    end
%
    % Cálculo de la menor raíz de (17)
    if b < 0
        disc=-disc;
    end
    z=-2*c/(b+disc);
    p=P(3)+z;
%
    % Ordenamos los elementos de P para hallar el más próximo a p
    if abs(p-P(2))<abs(p-P(1))
        Q=[P(2) P(1) P(3)];
        P=Q;
        Y=feval(f,P);
    end
    if abs(p-P(3))<abs(p-P(2))
        R=[P(1) P(3) P(2)];
        P=R;
        Y=feval(f,P);
    end
%
    % Reemplazamos el elemento de P más lejano de p por el propio p
    P(3)=p;
    Y(3) = feval(f,P(3));
    y=Y(3);
%
    % Criterio de parada
```

```

err=abs(z);
relerr=err/(abs(p)+delta);
if (err<delta)|(relerr<delta)|(abs(y)<epsilon)
    break
end
end

```

Ejercicios

1. Halle Δp_n , siendo
 - (a) $p_n = 5$
 - (b) $p_n = 6n + 2$
 - (c) $p_n = n(n + 1)$
2. Sea $p_n = 2n^2 + 1$. Halle $\Delta^k p_n$ para
 - (a) $k = 2$
 - (b) $k = 3$
 - (c) $k = 4$
3. Sea $p_n = 1/2^n$. Pruebe que $q_n = 0$ para todo n , siendo q_n el término generado por la fórmula (4).
4. Sea $p_n = 1/n$. Pruebe que $q_n = 1/(2n + 2)$ para todo n ; lo que significa que la aceleración de la convergencia conseguida en este caso es muy pequeña. ¿Converge $\{p_n\}$ a 0 linealmente? ¿Por qué?
5. Sea $p_n = 1/(2^n - 1)$. Pruebe que $q_n = 1/(4^{n+1} - 1)$ para todo n .
6. La sucesión $p_n = 1/(4^n + 4^{-n})$ converge linealmente a 0. Use la fórmula del método de Aitken (4) para hallar q_1, q_2 y q_3 y, así, acelerar la convergencia.

n	p_n	q_n
0	0.5	-0.26437542
1	0.23529412	
2	0.06225681	
3	0.01562119	
4	0.00390619	
5	0.00097656	

7. La sucesión $\{p_n\}$, generada por iteración de punto fijo con la función dada por $g(x) = (6 + x)^{1/2}$ y a partir del valor inicial $p_0 = 2.5$, converge linealmente a $p = 3$. Use la fórmula del método de Aitken (4) para hallar q_1, q_2 y q_3 y, así, acelerar la convergencia.
8. La sucesión $\{p_n\}$, generada por iteración de punto fijo con la función dada por $g(x) = \ln(x) + 2$ y a partir del valor inicial $p_0 = 3.14$, converge linealmente a $p \approx 3.1419322$. Use la fórmula del método de Aitken (4) para hallar q_1, q_2 y q_3 y, así, acelerar la convergencia.

9. La función del método de iteración de Newton-Raphson para resolver la ecuación $\cos(x) - 1 = 0$ es $g(x) = x - (1 - \cos(x))/\operatorname{sen}(x) = x - \tan(x/2)$. Use el método de Steffensen con esta función $g(x)$, empezando en $p_0 = 0.5$, y calcule p_1 , p_2 y p_3 ; después calcule p_4 , p_5 y p_6 .
10. *Convergencia de series.* El método de Aitken puede usarse para acelerar la convergencia de una serie. Si la suma parcial n -ésima de la serie es

$$S_n = \sum_{k=1}^n a_k,$$

demuestre que la sucesión obtenida al aplicar el método de Aitken a la sucesión $\{S_n\}$ es

$$T_n = S_n + \frac{a_{n+1}^2}{a_{n+1} - a_{n+2}}.$$

En los Ejercicios 11 a 14, aplique el método de Aitken y el resultado del Ejercicio 10 para acelerar la convergencia de la correspondiente serie.

11. $S_n = \sum_{k=1}^n (0.99)^k$

12. $S_n = \sum_{k=1}^n \frac{1}{4^k + 4^{-k}}$

13. $S_n = \sum_{k=1}^n \frac{k}{2^{k-1}}$

14. $S_n = \sum_{k=1}^n \frac{1}{2^k k}$

15. Use el método de Muller para aproximar una raíz de $f(x) = x^3 - x - 2$. Empiece con $p_0 = 1.0$, $p_1 = 1.2$ y $p_2 = 1.4$ y calcule p_3 , p_4 y p_5 .

16. Use el método de Muller para hallar una raíz de $f(x) = 4x^2 - e^x$. Empiece con $p_0 = 4.0$, $p_1 = 4.1$ y $p_2 = 4.2$ y calcule p_3 , p_4 y p_5 .

17. Sean $\{p_n\}$ y $\{q_n\}$ dos sucesiones cualesquiera de números reales. Demuestre

(a) $\Delta(p_n + q_n) = \Delta p_n + \Delta q_n$

(b) $\Delta(p_n q_n) = p_{n+1} \Delta q_n + q_n \Delta p_n$

18. En la fórmula (8), sume los términos p_{n+2} y $-p_{n+2}$ al miembro derecho para obtener la fórmula equivalente

$$p \approx p_{n+2} - \frac{(p_{n+2} - p_{n+1})^2}{p_{n+2} - 2p_{n+1} + p_n} = q_n.$$

19. Suponga que el error en un proceso iterativo verifica la relación $E_{n+1} = KE_n$ para alguna constante K tal que $|K| < 1$.

(a) Halle una expresión de E_n en función de E_0 , K y n .

(b) Halle el menor número natural N tal que $|E_N| < 10^{-8}$.

Algoritmos y programas

1. Use el método de Steffensen con el valor inicial $p_0 = 0.5$ para hallar una aproximación a un cero de $f(x) = x - \operatorname{sen}(x)$ con una precisión de diez cifras decimales.
2. Use el método de Steffensen con el valor inicial $p_0 = 0.5$ para hallar una aproximación al cero de $f(x) = \operatorname{sen}(x^3)$ más próximo a 0.5 con una precisión de diez cifras decimales.
3. Use el método de Muller con valores iniciales $p_0 = 1.5$, $p_1 = 1.4$ y $p_2 = 1.3$ para hallar una aproximación a un cero de $f(x) = 1 + 2x - \tan(x)$ con una precisión de doce cifras decimales.
4. En el Programa 2.8 (el método de Muller) se inicializa una matriz P de dimensiones 1×3 mediante p_0 , p_1 y p_2 . Después, al final del bucle, uno de los tres valores p_0 , p_1 o p_2 se reemplaza con la nueva aproximación al cero y el proceso continúa hasta que se satisfaga el criterio de parada, por ejemplo, en la iteración $k = K$. Modifique el Programa 2.8 de manera que, además de la solución y del error, se obtenga una matriz Q de dimensiones $(K+1) \times 3$ cuya primera fila sea $(p_0 \ p_1 \ p_2)$ y cuya fila k -ésima contenga las tres aproximaciones al cero del paso k -ésimo. Use esta modificación del Programa 2.8 con valores iniciales $p_0 = 2.4$, $p_1 = 2.3$ y $p_2 = 2.2$ para hallar una aproximación a un cero de $f(x) = 3 \cos(x) + 2 \operatorname{sen}(x)$ con una precisión de ocho cifras decimales.

Resolución de sistemas lineales

En la Figura 3.1 se muestra un sólido limitado por seis caras planas, tres correspondientes a los planos coordenados y otras tres a los planos de ecuaciones

$$\begin{aligned}5x + y + z &= 5 \\x + 4y + z &= 4 \\x + y + 3z &= 3.\end{aligned}$$

¿Cuáles son las coordenadas del punto de intersección de estos tres planos? Podemos usar el método de eliminación de Gauss para hallar la solución del sistema lineal:

$$x = 0.76, \quad y = 0.68 \quad y \quad z = 0.52.$$

En este capítulo desarrollaremos métodos para resolver sistemas de ecuaciones lineales.

1.1 Vectores y matrices

Un vector real N -dimensional \mathbf{X} es un conjunto ordenado de N números reales que normalmente se escribe como

$$(1) \quad \mathbf{X} = (x_1, x_2, \dots, x_N).$$

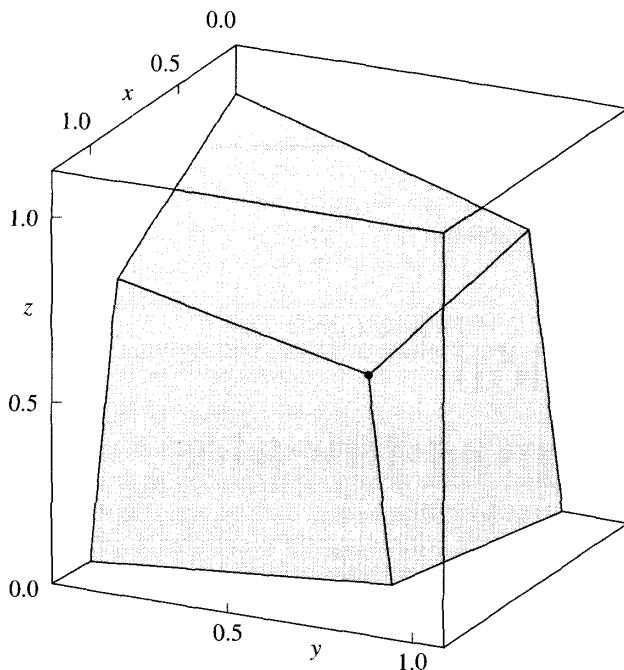


Figura 3.1 La intersección de tres planos.

Los números x_1, x_2, \dots y x_N se llaman **componentes** o **coordenadas** de \mathbf{X} . El conjunto formado por todos los vectores N -dimensionales se llama **espacio N -dimensional** y se denota por \mathbb{R}^N . Cuando un vector se utiliza para denotar un punto o una posición en el espacio, se suele llamar **vector de posición**, mientras que cuando se usa para indicar un movimiento entre dos puntos del espacio, se suele llamar **vector de desplazamiento**.

Sea $\mathbf{Y} = (y_1, y_2, \dots, y_N)$ otro vector. Se dice que los dos vectores \mathbf{X} e \mathbf{Y} son iguales cuando sus correspondientes coordenadas son iguales; o sea,

$$(2) \quad \mathbf{X} = \mathbf{Y} \quad \text{si, y sólo si, } \quad x_j = y_j \quad \text{para } j = 1, 2, \dots, N.$$

La suma de los vectores \mathbf{X} e \mathbf{Y} se calcula componente a componente; es decir,

$$(3) \quad \mathbf{X} + \mathbf{Y} = (x_1 + y_1, x_2 + y_2, \dots, x_N + y_N).$$

El opuesto del vector \mathbf{X} es el vector que se obtiene reemplazando cada coordenada por su opuesta:

$$(4) \quad -\mathbf{X} = (-x_1, -x_2, \dots, -x_N).$$

La diferencia $\mathbf{Y} - \mathbf{X}$ se obtiene restando coordenada a coordenada:

$$(5) \quad \mathbf{Y} - \mathbf{X} = (y_1 - x_1, y_2 - x_2, \dots, y_N - x_N).$$

Los vectores del espacio N -dimensional cumplen la siguiente propiedad algebraica

$$(6) \quad \mathbf{Y} - \mathbf{X} = \mathbf{Y} + (-\mathbf{X}).$$

Si c es un número real (un escalar), se define el **producto** de c y \mathbf{X} como

$$(7) \quad c\mathbf{X} = (cx_1, cx_2, \dots, cx_N),$$

también se dice que $c\mathbf{X}$ es un **múltiplo escalar** de \mathbf{X} .

Si c y d son escalares, entonces la suma ponderada $c\mathbf{X} + d\mathbf{Y}$ se llama **combinación lineal** de \mathbf{X} e \mathbf{Y} , y se tiene

$$(8) \quad c\mathbf{X} + d\mathbf{Y} = (cx_1 + dy_1, cx_2 + dy_2, \dots, cx_N + dy_N).$$

El **producto escalar** de dos vectores \mathbf{X} e \mathbf{Y} es un escalar (un número real) definido por la relación

$$(9) \quad \mathbf{X} \cdot \mathbf{Y} = x_1y_1 + x_2y_2 + \cdots + x_Ny_N.$$

La **norma** (o **módulo**) del vector \mathbf{X} se define como

$$(10) \quad \|\mathbf{X}\| = (x_1^2 + x_2^2 + \cdots + x_N^2)^{1/2},$$

a la que también nos referiremos como la **norma euclídea** del vector \mathbf{X} .

El producto $c\mathbf{X}$ dilata el vector \mathbf{X} cuando $|c| > 1$ y lo contrae cuando $|c| < 1$. Para ver esto usamos la relación (10):

$$(11) \quad \begin{aligned} \|c\mathbf{X}\| &= (c^2x_1^2 + c^2x_2^2 + \cdots + c^2x_N^2)^{1/2} \\ &= |c|(x_1^2 + x_2^2 + \cdots + x_N^2)^{1/2} = |c|\|\mathbf{X}\|. \end{aligned}$$

Existe una relación importante entre el producto escalar y la norma de un vector: si elevamos al cuadrado en (10) y usamos (9) para el caso en que $\mathbf{X} = \mathbf{Y}$, tenemos

$$(12) \quad \|\mathbf{X}\|^2 = x_1^2 + x_2^2 + \cdots + x_N^2 = \mathbf{X} \cdot \mathbf{X}.$$

Si \mathbf{X} e \mathbf{Y} son los vectores de posición de dos puntos (x_1, x_2, \dots, x_N) e (y_1, y_2, \dots, y_N) en el espacio N -dimensional, entonces el **vector de desplazamiento** desde \mathbf{X} hasta \mathbf{Y} es el dado por la diferencia

$$(13) \quad \mathbf{Y} - \mathbf{X} \quad (\text{desplazamiento desde la posición } \mathbf{X} \text{ hasta la posición } \mathbf{Y}).$$

Nótese que si una partícula parte de la posición \mathbf{X} y su desplazamiento es $\mathbf{Y} - \mathbf{X}$, entonces su posición de llegada es \mathbf{Y} , para ver esto basta sumar:

$$(14) \quad \mathbf{Y} = \mathbf{X} + (\mathbf{Y} - \mathbf{X}).$$

Usando las relaciones (10) y (13), podemos escribir la fórmula de la **distan-
cia entre dos puntos** en el espacio N -dimensional:

$$(15) \quad \|\mathbf{Y} - \mathbf{X}\| = ((y_1 - x_1)^2 + (y_2 - x_2)^2 + \cdots + (y_N - x_N)^2)^{1/2}.$$

Cuando la distancia entre dos puntos se calcula usando la fórmula (15), se dice que los puntos están en el **espacio euclídeo N -dimensional**.

Ejemplo 3.1. Sean $\mathbf{X} = (2, -3, 5, -1)$ e $\mathbf{Y} = (6, 1, 2, -4)$. Vamos a ilustrar los conceptos anteriores con estos dos vectores en el espacio 4-dimensional.

Suma	$\mathbf{X} + \mathbf{Y} = (8, -2, 7, -5)$
Diferencia	$\mathbf{X} - \mathbf{Y} = (-4, -4, 3, 3)$
Múltiplo escalar	$3\mathbf{X} = (6, -9, 15, -3)$
Módulo	$\ \mathbf{X}\ = (4 + 9 + 25 + 1)^{1/2} = 39^{1/2}$
Producto escalar	$\mathbf{X} \cdot \mathbf{Y} = 12 - 3 + 10 + 4 = 23$
Desplazamiento desde \mathbf{X} hasta \mathbf{Y}	$\mathbf{Y} - \mathbf{X} = (4, 4, -3, -3)$
Distancia entre \mathbf{X} e \mathbf{Y}	$\ \mathbf{Y} - \mathbf{X}\ = (16 + 16 + 9 + 9)^{1/2} = 50^{1/2}$

A veces es conveniente escribir los vectores como columnas en vez de escribirlos como filas. Por ejemplo, si

$$(16) \quad \mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \quad \text{e} \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix},$$

entonces la combinación lineal $c\mathbf{X} + d\mathbf{Y}$ es

$$(17) \quad c\mathbf{X} + d\mathbf{Y} = \begin{bmatrix} cx_1 + dy_1 \\ cx_2 + dy_2 \\ \vdots \\ cx_N + dy_N \end{bmatrix}.$$

Eligiendo c y d adecuadamente en la ecuación (17), podemos obtener la suma $1\mathbf{X} + 1\mathbf{Y}$, la diferencia $1\mathbf{X} - 1\mathbf{Y}$, y el múltiplo escalar $c\mathbf{X} + 0\mathbf{Y}$. Usaremos la virgulilla “ $'$ ” para indicar la trasposición de un vector fila en un vector columna y viceversa. Así,

$$(18) \quad (x_1, x_2, \dots, x_N)' = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \quad \text{y} \quad \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}' = (x_1, x_2, \dots, x_N).$$

El espacio N -dimensional tiene un elemento neutro o nulo: el vector $\mathbf{0}$ definido por

$$(19) \quad \mathbf{0} = (0, 0, \dots, 0).$$

Teorema 3.1 (Álgebra de vectores). Sean \mathbf{X}, \mathbf{Y} y \mathbf{Z} vectores N -dimensionales y sean a y b escalares (números reales). Entonces se verifican las siguientes propiedades de la suma de vectores y del producto por escalares;

- | | |
|----------------------------------------------------------------------------------------|---------------------------------------|
| (20) $\mathbf{Y} + \mathbf{X} = \mathbf{X} + \mathbf{Y}$ | propiedad conmutativa de la suma |
| (21) $\mathbf{0} + \mathbf{X} = \mathbf{X} + \mathbf{0}$ | elemento neutro de la suma |
| (22) $\mathbf{X} - \mathbf{X} = \mathbf{X} + (-\mathbf{X}) = \mathbf{0}$ | elemento opuesto de la suma |
| (23) $(\mathbf{X} + \mathbf{Y}) + \mathbf{Z} = \mathbf{X} + (\mathbf{Y} + \mathbf{Z})$ | propiedad asociativa de la suma |
| (24) $(a + b)\mathbf{X} = a\mathbf{X} + b\mathbf{X}$ | propiedad distributiva para escalares |
| (25) $a(\mathbf{X} + \mathbf{Y}) = a\mathbf{X} + a\mathbf{Y}$ | propiedad distributiva para vectores |
| (26) $a(b\mathbf{X}) = (ab)\mathbf{X}$ | propiedad asociativa de escalares |

Matrices

Una matriz es una colección de números reales dispuestos de forma rectangular en filas y columnas. Una matriz con M filas y N columnas se dice que es de orden, o dimensiones, $M \times N$ (léase “ M por N ”). En general, una letra mayúscula \mathbf{A} denota una matriz, mientras que las correspondientes minúsculas a_{ij} indican uno de los números que forman la matriz, de manera que escribimos

$$(27) \quad \mathbf{A} = [a_{ij}]_{M \times N} \quad \text{para } 1 \leq i \leq M, 1 \leq j \leq N,$$

siendo a_{ij} el número que ocupa la posición (i, j) (o sea, que está en la i -ésima fila y la j -ésima columna de la matriz), y nos referimos a a_{ij} como el elemento que ocupa la posición (i, j) . Podemos escribir lo anterior de forma desarrollada:

$$(28) \quad \text{fila } i \rightarrow \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{iN} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{Mj} & \cdots & a_{MN} \end{bmatrix} = \mathbf{A}.$$

↑
columna j

Las filas de una matriz \mathbf{A} de orden $M \times N$ son vectores N -dimensionales:

$$(29) \quad \mathbf{V}_i = (a_{i1}, a_{i2}, \dots, a_{iN}) \quad \text{para } i = 1, 2, \dots, M.$$

Los vectores filas de (29) pueden verse también como matrices de orden $1 \times N$ (matrices-fila) y, entonces, lo que hemos hecho es rebanar la matriz \mathbf{A} de orden $M \times N$ en M trozos (submatrices) que son matrices de orden $1 \times N$.

En este caso podríamos expresar \mathbf{A} como una matriz de orden $M \times 1$ cuyos elementos son las matrices-fila \mathbf{V}_i de orden $1 \times N$; esto es,

$$(30) \quad \mathbf{A} = \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \\ \vdots \\ \mathbf{V}_i \\ \vdots \\ \mathbf{V}_M \end{bmatrix} = [\mathbf{V}_1 \quad \mathbf{V}_2 \quad \cdots \quad \mathbf{V}_i \quad \cdots \quad \mathbf{V}_M]'.$$

De manera similar, las columnas de una matriz \mathbf{A} de orden $M \times N$ son matrices de orden $M \times 1$ (matrices-columna):

$$(31) \quad \mathbf{C}_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{i1} \\ \vdots \\ a_{M1} \end{bmatrix}, \quad \dots, \quad \mathbf{C}_j = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{ij} \\ \vdots \\ a_{Mj} \end{bmatrix}, \quad \dots, \quad \mathbf{C}_N = \begin{bmatrix} a_{1N} \\ a_{2N} \\ \vdots \\ a_{iN} \\ \vdots \\ a_{MN} \end{bmatrix}.$$

En cuyo caso podríamos expresar la matriz \mathbf{A} como una matriz de orden $1 \times N$ cuyos elementos son las matrices-columna \mathbf{C}_j de orden $M \times 1$:

$$(32) \quad \mathbf{A} = [\mathbf{C}_1 \quad \mathbf{C}_2 \quad \cdots \quad \mathbf{C}_j \quad \cdots \quad \mathbf{C}_N].$$

Ejemplo 3.2. Vamos a identificar las matrices-fila y columna asociadas a la siguiente matriz de orden 4×3

$$\mathbf{A} = \begin{bmatrix} -2 & 4 & 9 \\ 5 & -7 & 1 \\ 0 & -3 & 8 \\ -4 & 6 & -5 \end{bmatrix}.$$

Las cuatro matrices-fila son $\mathbf{V}_1 = [-2 \quad 4 \quad 9]$, $\mathbf{V}_2 = [5 \quad -7 \quad 1]$, $\mathbf{V}_3 = [0 \quad -3 \quad 8]$ y $\mathbf{V}_4 = [-4 \quad 6 \quad -5]$. Las tres matrices-columna son

$$\mathbf{C}_1 = \begin{bmatrix} -2 \\ 5 \\ 0 \\ -4 \end{bmatrix}, \quad \mathbf{C}_2 = \begin{bmatrix} 4 \\ -7 \\ -3 \\ 6 \end{bmatrix} \quad \text{y} \quad \mathbf{C}_3 = \begin{bmatrix} 9 \\ 1 \\ 8 \\ -5 \end{bmatrix}.$$

Hagamos notar ahora cómo podemos representar \mathbf{A} mediante estas matrices:

$$\mathbf{A} = \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \\ \mathbf{V}_3 \\ \mathbf{V}_4 \end{bmatrix} = [\mathbf{C}_1 \quad \mathbf{C}_2 \quad \mathbf{C}_3]. \quad \blacksquare$$

Sean $\mathbf{A} = [a_{ij}]_{M \times N}$ y $\mathbf{B} = [b_{ij}]_{M \times N}$ dos matrices del mismo orden. Se dice que las dos matrices \mathbf{A} y \mathbf{B} son iguales cuando sus elementos correspondientes son iguales; es decir,

$$(33) \quad \mathbf{A} = \mathbf{B} \quad \text{si, y sólo si, } a_{ij} = b_{ij} \quad \text{para } 1 \leq i \leq M, 1 \leq j \leq N.$$

La suma de dos matrices \mathbf{A} y \mathbf{B} del mismo orden $M \times N$ se calcula elemento a elemento; es decir, se define

$$(34) \quad \mathbf{A} + \mathbf{B} = [a_{ij} + b_{ij}]_{M \times N} \quad \text{para } 1 \leq i \leq M, 1 \leq j \leq N.$$

La opuesta de una matriz \mathbf{A} se obtiene sustituyendo cada elemento por su opuesto:

$$(35) \quad -\mathbf{A} = [-a_{ij}]_{M \times N} \quad \text{para } 1 \leq i \leq M, 1 \leq j \leq N.$$

La diferencia $\mathbf{A} - \mathbf{B}$ se forma restando elemento a elemento:

$$(36) \quad \mathbf{A} - \mathbf{B} = [a_{ij} - b_{ij}]_{M \times N} \quad \text{para } 1 \leq i \leq M, 1 \leq j \leq N.$$

Si c es un número real (un escalar), definimos el producto (múltiplo escalar) $c\mathbf{A}$ como

$$(37) \quad c\mathbf{A} = [ca_{ij}]_{M \times N} \quad \text{para } 1 \leq i \leq M, 1 \leq j \leq N.$$

Si p y q son escalares, la suma ponderada $p\mathbf{A} + q\mathbf{B}$ se llama combinación lineal de las matrices \mathbf{A} y \mathbf{B} , y se tiene

$$(38) \quad p\mathbf{A} + q\mathbf{B} = [pa_{ij} + qb_{ij}]_{M \times N} \quad \text{para } 1 \leq i \leq M, 1 \leq j \leq N.$$

La matriz cero de orden $M \times N$ es aquella cuyos elementos son todos cero:

$$(39) \quad \mathbf{0} = [0]_{M \times N}.$$

Ejemplo 3.3. Calculemos los múltiplos escalares $2\mathbf{A}$ y $3\mathbf{B}$ y la combinación lineal $2\mathbf{A} - 3\mathbf{B}$ de las matrices

$$\mathbf{A} = \begin{bmatrix} -1 & 2 \\ 7 & 5 \\ 3 & -4 \end{bmatrix} \quad \text{y} \quad \mathbf{B} = \begin{bmatrix} -2 & 3 \\ 1 & -4 \\ -9 & 7 \end{bmatrix}.$$

Usando la fórmula (37), tenemos

$$2\mathbf{A} = \begin{bmatrix} -2 & 4 \\ 14 & 10 \\ 6 & -8 \end{bmatrix} \quad \text{y} \quad 3\mathbf{B} = \begin{bmatrix} -6 & 9 \\ 3 & -12 \\ -27 & 21 \end{bmatrix}.$$

Con lo que podemos calcular la combinación lineal $2\mathbf{A} - 3\mathbf{B}$:

$$2\mathbf{A} - 3\mathbf{B} = \begin{bmatrix} -2 + 6 & 4 - 9 \\ 14 - 3 & 10 + 12 \\ 6 + 27 & -8 - 21 \end{bmatrix} = \begin{bmatrix} 4 & -5 \\ 11 & 22 \\ 33 & -29 \end{bmatrix}.$$

Teorema 3.2 (Suma de matrices). Supongamos que \mathbf{A} , \mathbf{B} y \mathbf{C} son matrices de orden $M \times N$ y que p y q son escalares. Entonces se verifican las siguientes propiedades de la suma de matrices y del producto de una matriz por un escalar:

- | | |
|----------------------------------------------------------------------------------------|---------------------------------------|
| (40) $\mathbf{B} + \mathbf{A} = \mathbf{A} + \mathbf{B}$ | propiedad commutativa de la suma |
| (41) $\mathbf{0} + \mathbf{A} = \mathbf{A} + \mathbf{0}$ | elemento neutro de la suma |
| (42) $\mathbf{A} - \mathbf{A} = \mathbf{A} + (-\mathbf{A}) = \mathbf{0}$ | elemento opuesto de la suma |
| (43) $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$ | propiedad asociativa de la suma |
| (44) $(p + q)\mathbf{A} = p\mathbf{A} + q\mathbf{A}$ | propiedad distributiva para escalares |
| (45) $p(\mathbf{A} + \mathbf{B}) = p\mathbf{A} + p\mathbf{B}$ | propiedad distributiva para matrices |
| (46) $p(q\mathbf{A}) = (pq)\mathbf{A}$ | propiedad asociativa para escalares |

Ejercicios

Le animamos a que realice estos ejercicios tanto a mano como con el paquete de programas MATLAB.

- Dados los vectores \mathbf{X} e \mathbf{Y} , calcule (a) $\mathbf{X} + \mathbf{Y}$, (b) $\mathbf{X} - \mathbf{Y}$, (c) $3\mathbf{X}$, (d) $\|\mathbf{X}\|$, (e) $7\mathbf{Y} - 4\mathbf{X}$, (f) $\mathbf{X} \cdot \mathbf{Y}$ y (g) $\|7\mathbf{Y} - 4\mathbf{X}\|$.
 - $\mathbf{X} = (3, -4)$ e $\mathbf{Y} = (-2, 8)$
 - $\mathbf{X} = (-6, 3, 2)$ e $\mathbf{Y} = (-8, 5, 1)$
 - $\mathbf{X} = (4, -8, 1)$ e $\mathbf{Y} = (1, -12, -11)$
 - $\mathbf{X} = (1, -2, 4, 2)$ e $\mathbf{Y} = (3, -5, -4, 0)$
- Usando la ley de los cosenos, puede probarse que el ángulo θ que forman dos vectores no nulos \mathbf{X} e \mathbf{Y} viene dado por la relación

$$\cos(\theta) = \frac{\mathbf{X} \cdot \mathbf{Y}}{\|\mathbf{X}\| \|\mathbf{Y}\|}.$$

Encuentre el ángulo, en radianes, formado por los siguientes vectores

- $\mathbf{X} = (-6, 3, 2)$ e $\mathbf{Y} = (2, -2, 1)$
- $\mathbf{X} = (4, -8, 1)$ e $\mathbf{Y} = (3, 4, 12)$

3. Se dice que dos vectores \mathbf{X} e \mathbf{Y} son ortogonales (o perpendiculares) si forman un ángulo de $\pi/2$.

(a) Pruebe que \mathbf{X} e \mathbf{Y} son ortogonales si, y sólo si, $\mathbf{X} \cdot \mathbf{Y} = 0$.

Use el apartado (a) para determinar cuáles de las siguientes parejas de vectores son ortogonales.

(b) $\mathbf{X} = (-6, 4, 2)$ e $\mathbf{Y} = (6, 5, 8)$

(c) $\mathbf{X} = (-4, 8, 3)$ e $\mathbf{Y} = (2, 5, 16)$

(d) $\mathbf{X} = (-5, 7, 2)$ e $\mathbf{Y} = (4, 1, 6)$

(e) Determine dos vectores diferentes que sean ambos ortogonales al vector $\mathbf{X} = (1, 2, -5)$.

4. Calcule (a) $\mathbf{A} + \mathbf{B}$, (b) $\mathbf{A} - \mathbf{B}$ y (c) $3\mathbf{A} - 2\mathbf{B}$ para las matrices

$$\mathbf{A} = \begin{bmatrix} -1 & 9 & 4 \\ 2 & -3 & -6 \\ 0 & 5 & 7 \end{bmatrix} \quad \text{y} \quad \mathbf{B} = \begin{bmatrix} -4 & 9 & 2 \\ 3 & -5 & 7 \\ 8 & 1 & -6 \end{bmatrix}.$$

5. La **traspuesta** o **adjunta** de una matriz \mathbf{A} de orden $M \times N$ es la matriz \mathbf{A}' de orden $N \times M$ que se obtiene de \mathbf{A} convirtiendo las filas de \mathbf{A} en las columnas de \mathbf{A}' ; es decir, si $\mathbf{A} = [a_{ij}]_{M \times N}$ y $\mathbf{A}' = [b_{ij}]_{N \times M}$, entonces sus elementos satisfacen la relación

$$b_{ji} = a_{ij} \quad \text{para} \quad 1 \leq i \leq M, 1 \leq j \leq N.$$

Halle la traspuesta de las siguientes matrices:

$$(a) \begin{bmatrix} -2 & 5 & 12 \\ 1 & 4 & -1 \\ 7 & 0 & 6 \\ 11 & -3 & 8 \end{bmatrix} \quad (b) \begin{bmatrix} 4 & 9 & 2 \\ 3 & 5 & 7 \\ 8 & 1 & 6 \end{bmatrix}$$

6. Se dice que una matriz cuadrada \mathbf{A} de orden $N \times N$ es simétrica si $\mathbf{A} = \mathbf{A}'$ (véase la definición de \mathbf{A}' en el Ejercicio 5). Determine si las siguientes matrices son simétricas:

$$(a) \begin{bmatrix} 1 & -7 & 4 \\ -7 & 2 & 0 \\ 4 & 0 & 3 \end{bmatrix} \quad (b) \begin{bmatrix} 4 & -7 & 1 \\ 0 & 2 & -7 \\ 3 & 0 & 4 \end{bmatrix}$$

$$(c) \mathbf{A} = [a_{ij}]_{N \times N}, \text{ siendo } a_{ij} = \begin{cases} ij & i = j \\ i - ij + j & i \neq j \end{cases}$$

$$(d) \mathbf{A} = [a_{ij}]_{N \times N}, \text{ siendo } a_{ij} = \begin{cases} \cos(ij) & i = j \\ i - ij - j & i \neq j \end{cases}$$

7. Demuestre las propiedades (20), (24) y (25) del Teorema 3.1.

3.2 Multiplicación de matrices

Una combinación lineal de las variables x_1, x_2, \dots, x_N es una suma

$$(1) \quad a_1x_1 + a_2x_2 + \cdots + a_Nx_N$$

en la que a_k se llama coeficiente de x_k para $k = 1, 2, \dots, N$. Si exigimos que la combinación lineal (1) sea igual a un valor prefijado b , obtenemos una ecuación lineal para las incógnitas x_1, x_2, \dots, x_N :

$$(2) \quad a_1x_1 + a_2x_2 + \cdots + a_Nx_N = b.$$

En la práctica es frecuente encontrarse con sistemas de ecuaciones lineales que, si lo que tenemos son M ecuaciones con N incógnitas, escribimos como

$$(3) \quad \begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1N}x_N &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2N}x_N &= b_2 \\ \vdots &\quad \vdots &\quad \vdots &\quad \vdots \\ a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kN}x_N &= b_k \\ \vdots &\quad \vdots &\quad \vdots &\quad \vdots \\ a_{M1}x_1 + a_{M2}x_2 + \cdots + a_{MN}x_N &= b_M. \end{aligned}$$

Para localizar los diferentes coeficientes de cada ecuación es necesario usar los dos subíndices (k, j) ; el primero nos indica que es un coeficiente de la ecuación k -ésima y el segundo nos indica que es el coeficiente de la variable j -ésima x_j .

Una solución de (3) es un conjunto de números reales x_1, x_2, \dots, x_N que verifica simultáneamente todas las ecuaciones de (3). Por tanto una solución puede ser considerada como un vector N -dimensional:

$$(4) \quad \mathbf{X} = (x_1, x_2, \dots, x_N).$$

Ejemplo 3.4. La mezcla que se emplea para construir aceras se compone de cemento, arena y grava en distintas proporciones. Un distribuidor tiene ya preparados sacos con tres tipos de mezclas diferentes: El primer tipo contiene cemento, arena y grava mezclados según las proporciones $1/8, 3/8, 4/8$, las proporciones en el segundo tipo son $2/10, 5/10, 3/10$ y las proporciones en el tercero son $2/5, 3/5, 0/5$.

Sean x_1, x_2 y x_3 las cantidades (en metros cúbicos) de cada uno de los tipos anteriores que hay que usar para formar una cantidad total de 10 metros cúbicos de mezcla que contenga $b_1 = 2.3$, $b_2 = 4.8$ y $b_3 = 2.9$ metros cúbicos de cemento, arena y grava, respectivamente. Entonces el sistema de ecuaciones lineales para los ingredientes es:

$$(5) \quad \begin{aligned} 0.125x_1 + 0.200x_2 + 0.400x_3 &= 2.3 && \text{(cemento),} \\ 0.375x_1 + 0.500x_2 + 0.600x_3 &= 4.8 && \text{(arena),} \\ 0.500x_1 + 0.300x_2 + 0.000x_3 &= 2.9 && \text{(grava).} \end{aligned}$$

La solución del sistema lineal (5) es $x_1 = 4$, $x_2 = 3$ y $x_3 = 3$, como puede comprobarse sustituyendo directamente en las ecuaciones:

$$(0.125)(4) + (0.200)(3) + (0.400)(3) = 2.3,$$

$$(0.375)(4) + (0.500)(3) + (0.600)(3) = 4.8,$$

$$(0.500)(4) + (0.300)(3) + (0.000)(3) = 2.9.$$
■

Multiplicación de matrices

Definición 3.1. Si $\mathbf{A} = [a_{ik}]_{M \times N}$ y $\mathbf{B} = [b_{kj}]_{N \times P}$ son dos matrices con la propiedad de que \mathbf{A} tiene tantas columnas como \mathbf{B} filas, entonces la matriz producto \mathbf{AB} se define como la matriz \mathbf{C} de orden $M \times P$

$$(6) \quad \mathbf{AB} = \mathbf{C} = [c_{ij}]_{M \times P}$$

cuyo elemento c_{ij} es el producto escalar de la i -ésima fila de \mathbf{A} por la j -ésima columna de \mathbf{B} :

$$(7) \quad c_{ij} = \sum_{k=1}^N a_{ik} b_{kj} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{iN}b_{Nj}$$

para $i = 1, 2, \dots, M$ y $j = 1, 2, \dots, P$.

■

Ejemplo 3.5. Vamos a calcular el producto $\mathbf{C} = \mathbf{AB}$ de las siguientes matrices \mathbf{A} y \mathbf{B} ; también explicaremos por qué \mathbf{BA} no está definido. Las matrices son

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ -1 & 4 \end{bmatrix} \quad \text{y} \quad \mathbf{B} = \begin{bmatrix} 5 & -2 & 1 \\ 3 & 8 & -6 \end{bmatrix}.$$

La matriz \mathbf{A} tiene dos columnas y la matriz \mathbf{B} tiene dos filas, así que el producto matricial \mathbf{AB} puede hacerse. El producto de una matriz de orden 2×2 por una matriz de orden 2×3 es una matriz de orden 2×3 . Si hacemos los cálculos en nuestro caso, resulta

$$\begin{aligned} \mathbf{AB} &= \begin{bmatrix} 2 & 3 \\ -1 & 4 \end{bmatrix} \begin{bmatrix} 5 & -2 & 1 \\ 3 & 8 & -6 \end{bmatrix} \\ &= \begin{bmatrix} 10 + 9 & -4 + 24 & 2 - 18 \\ -5 + 12 & 2 + 32 & -1 - 24 \end{bmatrix} = \begin{bmatrix} 19 & 20 & -16 \\ 7 & 34 & -25 \end{bmatrix} = \mathbf{C}. \end{aligned}$$

Cuando intentamos formar el producto \mathbf{BA} , descubrimos que no puede hacerse porque las filas de \mathbf{B} son vectores tridimensionales y las columnas de \mathbf{A} son vectores bidimensionales; de manera que el producto escalar de una fila de \mathbf{B} por una columna de \mathbf{A} no está definido.

■

Cuando se tiene que $\mathbf{AB} = \mathbf{BA}$, entonces se dice que \mathbf{A} y \mathbf{B} comutan. En la mayoría de las ocasiones, incluso cuando \mathbf{AB} y \mathbf{BA} pueden ambos calcularse, los productos resultantes no tienen por qué ser iguales.

A continuación vamos a mostrar cómo podemos usar matrices para representar un sistema de ecuaciones lineales. Las ecuaciones lineales de (3) pueden escribirse como un producto matricial: Los coeficientes a_{kj} constituyen una matriz \mathbf{A} (llamada matriz de los coeficientes) de orden $M \times N$, las incógnitas x_j constituyen una matriz-columna \mathbf{X} de orden $N \times 1$ y las constantes b_k constituyen otra matriz-columna \mathbf{B} de orden $M \times 1$, de manera que podemos escribir

$$(8) \quad \mathbf{AX} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kj} & \cdots & a_{kN} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{Mj} & \cdots & a_{MN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_j \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_j \\ \vdots \\ b_M \end{bmatrix} = \mathbf{B}.$$

La multiplicación matricial $\mathbf{AX} = \mathbf{B}$ que aparece en (8) nos recuerda el producto escalar de vectores normales y corrientes, ya que cada elemento b_k de \mathbf{B} es el resultado que se obtiene al hacer el producto escalar de la fila k -ésima de la matriz \mathbf{A} por la matriz-columna \mathbf{X} .

Ejemplo 3.6. Vamos a expresar el sistema de ecuaciones lineales (5) del Ejemplo 3.4 en forma matricial; luego usaremos la multiplicación matricial para comprobar que $[4 \ 3 \ 3]'$ es la solución:

$$(9) \quad \begin{bmatrix} 0.125 & 0.200 & 0.400 \\ 0.375 & 0.500 & 0.600 \\ 0.500 & 0.300 & 0.000 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2.3 \\ 4.8 \\ 2.9 \end{bmatrix}.$$

Para verificar que $[4 \ 3 \ 3]'$ es la solución del sistema (5), debemos probar que se tiene $\mathbf{A} [4 \ 3 \ 3]' = [2.3 \ 4.8 \ 2.9]'$:

$$\begin{bmatrix} 0.125 & 0.200 & 0.400 \\ 0.375 & 0.500 & 0.600 \\ 0.500 & 0.300 & 0.000 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 0.5 + 0.6 + 1.2 \\ 1.5 + 1.5 + 1.8 \\ 2.0 + 0.9 + 0.0 \end{bmatrix} = \begin{bmatrix} 2.3 \\ 4.8 \\ 2.9 \end{bmatrix}.$$

Algunas matrices especiales

Como ya hemos visto antes, la matriz de orden $M \times N$ cuyos elementos son todos cero se llama **matriz cero, o matriz nula, de orden $M \times N$** y se denota por

$$(10) \quad \mathbf{0} = [0]_{M \times N}.$$

Cuando las dimensiones estén claras en el contexto, escribiremos simplemente **0** para denotar la matriz cero.

La **matriz identidad, o matriz unidad, de orden N** es la matriz cuadrada dada por

$$(11) \quad \mathbf{I}_N = [\delta_{ij}]_{N \times N} \quad \text{siendo} \quad \delta_{ij} = \begin{cases} 1 & \text{si } i = j, \\ 0 & \text{si } i \neq j. \end{cases}$$

Esta matriz es el elemento neutro de la multiplicación matricial, lo que se ilustra en el siguiente ejemplo.

Ejemplo 3.7. Sea \mathbf{A} una matriz de orden 2×3 . Entonces $\mathbf{I}_2\mathbf{A} = \mathbf{A}\mathbf{I}_3 = \mathbf{A}$, ya que multiplicando a la izquierda de \mathbf{A} por \mathbf{I}_2 se obtiene

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} = \begin{bmatrix} a_{11} + 0 & a_{12} + 0 & a_{13} + 0 \\ a_{21} + 0 & a_{22} + 0 & a_{23} + 0 \end{bmatrix} = \mathbf{A}$$

y multiplicando a la derecha de \mathbf{A} por \mathbf{I}_3 se obtiene

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} + 0 + 0 & 0 + a_{12} + 0 & 0 + 0 + a_{13} \\ a_{21} + 0 + 0 & 0 + a_{22} + 0 & 0 + 0 + a_{23} \end{bmatrix} = \mathbf{A}. \blacksquare$$

Algunas propiedades de la multiplicación matricial se recogen en el siguiente teorema.

Teorema 3.3 (Multiplicación de matrices). Supongamos que c es un escalar y que \mathbf{A} , \mathbf{B} y \mathbf{C} son matrices tales que las sumas y productos que se indican a continuación están definidos. Entonces se verifica:

- | | |
|---------------------------------------------------------------------------|-----------------------------------------------|
| (12) $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$ | propiedad asociativa del producto de matrices |
| (13) $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$ | elemento neutro del producto de matrices |
| (14) $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$ | propiedad distributiva por la izquierda |
| (15) $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$ | propiedad distributiva por la derecha |
| (16) $c(\mathbf{AB}) = (c\mathbf{A})\mathbf{B} = \mathbf{A}(c\mathbf{B})$ | propiedad asociativa de escalares |

Matrices Invertibles

El concepto de inversa puede aplicarse a matrices, pero hay que prestarle una especial atención. Se dice que una matriz \mathbf{A} cuadrada de orden $N \times N$ es **invertible, o no singular**, si existe una matriz \mathbf{B} también de orden $N \times N$ tal que

$$(17) \quad \mathbf{AB} = \mathbf{BA} = \mathbf{I}.$$

Si no es posible encontrar una matriz B que verifique lo anterior, entonces se dice que A es *singular*, o *no invertible*. Cuando A es invertible, la matriz B que verifica las igualdades de (17) se llama inversa de A y se representa por $B = A^{-1}$, lo que nos permite usar la relación

$$(18) \quad AA^{-1} = A^{-1}A = I \quad \text{si } A \text{ es invertible.}$$

Es fácil probar que, como mucho, sólo hay una matriz B que verifique las igualdades de (17): Supongamos que C es también una inversa de A (o sea, que $AC = CA = I$); entonces, usando las propiedades (12) y (13), obtenemos

$$C = IC = (BA)C = B(AC) = BI = B.$$

Determinantes

El determinante de una matriz cuadrada A es una cantidad escalar (un número real) que se denota por $\det(A)$ o bien $|A|$. Si A es una matriz de orden $N \times N$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix},$$

entonces se suele escribir

$$\det(A) = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{vmatrix}.$$

Aunque la notación para el determinante es muy parecida a la notación para una matriz, sus propiedades son muy distintas; para empezar, el determinante es un número real. La definición de $\det(A)$ puede encontrarse en la mayoría de los libros de texto de álgebra lineal, pero esta definición no es adecuada para propósitos computacionales si $N > 3$. Recordaremos la definición recursiva de determinante usando el desarrollo por una fila o por una columna mediante cofactores. El cálculo de los determinantes de orden alto se realiza mediante el método de eliminación de Gauss y se menciona en el cuerpo del Programa 3.3.

Si $A = [a_{ij}]$ es una matriz de orden 1×1 , se define $\det(A) = a_{11}$. Si $A = [a_{ij}]_{N \times N}$, siendo $N \geq 2$, entonces sea M_{ij} el determinante de la submatriz de orden $N-1 \times N-1$ extraída de A borrando la fila i -ésima y la columna j -ésima de A ; este determinante M_{ij} se llama el *menor* de a_{ij} . El *cofactor* A_{ij} de

a_{ij} se define, entonces, como $A_{ij} = (-1)^{i+j} M_{ij}$ y, finalmente, el determinante de la matriz \mathbf{A} de orden $N \times N$ viene dado por

$$(19) \quad \det(\mathbf{A}) = \sum_{j=1}^N a_{ij} A_{ij} \quad (\text{desarrollo por la } i\text{-ésima fila})$$

o bien por

$$(20) \quad \det(\mathbf{A}) = \sum_{i=1}^N a_{ij} A_{ij} \quad (\text{desarrollo por la } j\text{-ésima columna}).$$

Aplicando la fórmula (19), con $i = 1$, a la matriz de orden 2×2

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

vemos que $\det(\mathbf{A}) = a_{11}a_{22} - a_{12}a_{21}$. El siguiente ejemplo muestra cómo se utilizan las fórmulas (19) y (20) para reducir el cálculo del determinante de una matriz de orden $N \times N$ al cálculo de un cierto número de determinantes de orden 2×2 .

Ejemplo 3.8. Vamos a usar la fórmula (19) con $i = 1$ y la fórmula (20) con $j = 2$ para calcular el determinante de la matriz

$$\mathbf{A} = \begin{bmatrix} 2 & 3 & 8 \\ -4 & 5 & -1 \\ 7 & -6 & 9 \end{bmatrix}.$$

Usando la fórmula (19) con $i = 1$, obtenemos

$$\begin{aligned} \det(\mathbf{A}) &= (2) \begin{vmatrix} 5 & -1 \\ -6 & 9 \end{vmatrix} - (3) \begin{vmatrix} -4 & -1 \\ 7 & 9 \end{vmatrix} + (8) \begin{vmatrix} -4 & 5 \\ 7 & -6 \end{vmatrix} \\ &= (2)(45 - 6) - (3)(-36 + 7) + (8)(24 - 35) \\ &= 77. \end{aligned}$$

Usando la fórmula (20) con $j = 2$, obtenemos

$$\begin{aligned} \det(\mathbf{A}) &= -(3) \begin{vmatrix} -4 & -1 \\ 7 & 9 \end{vmatrix} + (5) \begin{vmatrix} 2 & 8 \\ 7 & 9 \end{vmatrix} - (-6) \begin{vmatrix} 2 & 8 \\ -4 & -1 \end{vmatrix} \\ &= 77. \end{aligned}$$

El siguiente teorema nos proporciona condiciones necesarias y suficientes para la existencia y unicidad de la solución de un sistema de ecuaciones lineales $\mathbf{AX} = \mathbf{B}$ cuando la matriz de los coeficientes es cuadrada; o sea, cuando hay tantas ecuaciones como incógnitas.

Teorema 3.4. Supongamos que \mathbf{A} es una matriz cuadrada de orden $N \times N$. Entonces las siguientes condiciones son equivalentes:

- (21) Dada cualquier matriz \mathbf{B} de orden $N \times 1$, el sistema de ecuaciones lineales $\mathbf{AX} = \mathbf{B}$ tiene solución única.
- (22) La matriz \mathbf{A} es invertible (es decir, existe \mathbf{A}^{-1}).
- (23) El sistema de ecuaciones $\mathbf{AX} = \mathbf{0}$ tiene como única solución $\mathbf{X} = \mathbf{0}$.
- (24) $\det(\mathbf{A}) \neq 0$.

Los Teoremas 3.3 y 3.4 nos permiten relacionar el álgebra de matrices con el álgebra ordinaria de números. Si cualquiera de las condiciones (21)–(24) se verifica, entonces, junto con las propiedades (12) y (13), podemos deducir

$$(25) \quad \mathbf{AX} = \mathbf{B} \text{ implica } \mathbf{A}^{-1}\mathbf{AX} = \mathbf{A}^{-1}\mathbf{B}; \text{ luego, } \mathbf{X} = \mathbf{A}^{-1}\mathbf{B}.$$

Ejemplo 3.9. Vamos a usar la matriz inversa

$$\mathbf{A}^{-1} = \frac{1}{5} \begin{bmatrix} 4 & -1 \\ -7 & 3 \end{bmatrix}$$

y el razonamiento expuesto en (25) para resolver el sistema de ecuaciones lineales $\mathbf{AX} = \mathbf{B}$:

$$\mathbf{AX} = \begin{bmatrix} 3 & 1 \\ 7 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \end{bmatrix} = \mathbf{B}.$$

Usando (25), obtenemos

$$\mathbf{X} = \mathbf{A}^{-1}\mathbf{B} = \frac{1}{5} \begin{bmatrix} 4 & -1 \\ -7 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \end{bmatrix} = \frac{1}{5} \begin{bmatrix} 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.6 \\ 0.2 \end{bmatrix}.$$

Observación. En la práctica nunca se calculan la inversa ni el determinante de una matriz cuadrada. Estos conceptos se utilizan como herramientas teóricas para establecer la existencia y unicidad de soluciones o como medios para expresar algebraicamente la solución de un sistema de ecuaciones lineales (como en el Ejemplo 3.9).

Rotaciones planas

Supongamos que \mathbf{A} es una matriz de orden 3×3 y que $\mathbf{U} = [x \ y \ z]'$ es una matriz de orden 3×1 , entonces el producto $\mathbf{V} = \mathbf{AU}$ es otra matriz de orden 3×1 . Este es un ejemplo de aplicación lineal, un concepto que se utiliza en el campo de la generación de gráficos por computador. La matriz \mathbf{U} equivale al

Tabla 3.1 Coordenadas de los vértices de un cubo tras sucesivas rotaciones.

U	$V = R_z(\frac{\pi}{4})U$	$W = R_y(\frac{\pi}{6})R_z(\frac{\pi}{4})U$
$(0, 0, 0)'$	$(0.000000, 0.000000, 0)'$	$(0.000000, 0.000000, 0.000000)'$
$(1, 0, 0)'$	$(0.707107, 0.707107, 0)'$	$(0.612372, 0.707107, -0.353553)'$
$(0, 1, 0)'$	$(-0.707107, 0.707107, 0)'$	$(-0.612372, 0.707107, 0.353553)'$
$(0, 0, 1)'$	$(0.000000, 0.000000, 1)'$	$(0.500000, 0.000000, 0.866025)'$
$(1, 1, 0)'$	$(0.000000, 1.414214, 0)'$	$(0.000000, 1.414214, 0.000000)'$
$(1, 0, 1)'$	$(0.707107, 0.707107, 1)'$	$(1.112372, 0.707107, 0.512472)'$
$(0, 1, 1)'$	$(-0.707107, 0.707107, 1)'$	$(-0.112372, 0.707107, 1.219579)'$
$(1, 1, 1)'$	$(0.000000, 1.414214, 1)'$	$(0.500000, 1.414214, 0.866025)'$

vector de posición $\mathbf{U} = (x, y, z)$ que representa las coordenadas de un punto en el espacio tridimensional. Consideremos tres matrices especiales:

$$(26) \quad \mathbf{R}_x(\alpha) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{bmatrix},$$

$$(27) \quad \mathbf{R}_y(\beta) = \begin{bmatrix} \cos(\beta) & 0 & \sin(\beta) \\ 0 & 1 & 0 \\ -\sin(\beta) & 0 & \cos(\beta) \end{bmatrix},$$

$$(28) \quad \mathbf{R}_z(\gamma) = \begin{bmatrix} \cos(\gamma) & -\sin(\gamma) & 0 \\ \sin(\gamma) & \cos(\gamma) & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Estas matrices $\mathbf{R}_x(\alpha)$, $\mathbf{R}_y(\beta)$ y $\mathbf{R}_z(\gamma)$ se usan, respectivamente, para hacer girar los puntos alrededor del eje OX , OY y OZ un ángulo α , β y γ . Las inversas de estas matrices son $\mathbf{R}_x(-\alpha)$, $\mathbf{R}_y(-\beta)$ y $\mathbf{R}_z(-\gamma)$, que giran los puntos alrededor del eje OX , OY y OZ un ángulo $-\alpha$, $-\beta$ y $-\gamma$, respectivamente. El siguiente ejemplo ilustra esta situación; dejamos como ejercicio la realización de investigaciones posteriores.

Ejemplo 3.10. Un cubo unidad se sitúa en el primer octante teniendo el origen como uno de sus vértices. Primero, hacemos girar el cubo un ángulo de $\pi/4$ alrededor del eje OZ ; luego, giramos su imagen un ángulo de $\pi/6$ alrededor del eje OY . Vamos a determinar las imágenes de los ocho vértices del cubo.

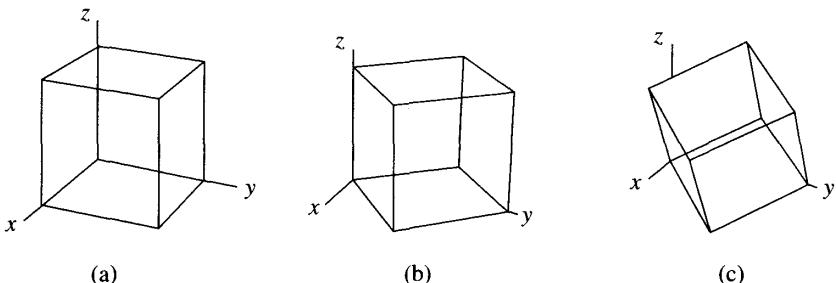


Figura 3.2 (a) El cubo de partida. (b) $\mathbf{V} = \mathbf{R}_z(\pi/4)\mathbf{U}$, rotación alrededor del eje OZ . (c) $\mathbf{W} = \mathbf{R}_y(\pi/6)\mathbf{V}$, rotación alrededor del eje OY .

La primera rotación viene dada por la transformación

$$\mathbf{V} = \mathbf{R}_z\left(\frac{\pi}{4}\right)\mathbf{U} = \begin{bmatrix} \cos\left(\frac{\pi}{4}\right) & -\sin\left(\frac{\pi}{4}\right) & 0 \\ \sin\left(\frac{\pi}{4}\right) & \cos\left(\frac{\pi}{4}\right) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$= \begin{bmatrix} 0.707107 & -0.707107 & 0.000000 \\ 0.707107 & 0.707107 & 0.000000 \\ 0.000000 & 0.000000 & 1.000000 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}.$$

La segunda rotación viene dada por

$$\begin{aligned} \mathbf{W} &= \mathbf{R}_y\left(\frac{\pi}{6}\right) \mathbf{V} = \begin{bmatrix} \cos\left(\frac{\pi}{6}\right) & 0 & \sin\left(\frac{\pi}{6}\right) \\ 0 & 1 & 0 \\ -\sin\left(\frac{\pi}{6}\right) & 0 & \cos\left(\frac{\pi}{6}\right) \end{bmatrix} \mathbf{V} \\ &= \begin{bmatrix} 0.866025 & 0.000000 & 0.500000 \\ 0.000000 & 1.000000 & 0.000000 \\ -0.500000 & 0.000000 & 0.866025 \end{bmatrix} \mathbf{V}. \end{aligned}$$

La composición de las dos rotaciones es

$$W = R_y\left(\frac{\pi}{6}\right) R_z\left(\frac{\pi}{4}\right) U = \begin{bmatrix} 0.612372 & -0.612372 & 0.500000 \\ 0.707107 & 0.707107 & 0.000000 \\ -0.353553 & 0.353553 & 0.866025 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

En la Tabla 3.1 se muestran los cálculos de las coordenadas de los vértices al comienzo y tras cada una de las rotaciones (como vectores de posición); las imágenes de estos cubos se muestran en la Figura 3.2(a)–(c). ■

MATLAB

Las funciones `det(A)` e `inv(A)` del paquete de programas MATLAB proporcionan, respectivamente, el determinante y la inversa (si A es invertible) de una matriz cuadrada A .

Ejemplo 3.11. Vamos a usar el paquete MATLAB, con el método de la matriz inversa descrito en (25), para resolver el sistema de ecuaciones lineales del Ejemplo 3.6.

Primero comprobamos que \mathbf{A} es invertible mostrando que $\det(\mathbf{A}) \neq 0$ (Teorema 3.4).

```
>>A=[0.125 0.200 0.400;0.375 0.500 0.600;0.500 0.300 0.000];
>>det(A)
ans=
-0.0175
```

Siguiendo el razonamiento hecho en (25), la solución de $\mathbf{AX} = \mathbf{B}$ es $\mathbf{X} = \mathbf{A}^{-1}\mathbf{B}$.

```
>>X=inv(A)*[2.3 4.8 2.9]'
```

```
X=
4.0000
3.0000
3.0000
```

Podemos comprobar nuestra solución verificando que $\mathbf{AX} = \mathbf{B}$.

```
>>B=A*X
B=
2.3000
4.8000
2.9000
```

Ejercicios

Le animamos a que realice estos ejercicios tanto a mano como con el paquete de programas MATLAB.

- Calcule \mathbf{AB} y \mathbf{BA} para las siguientes matrices:

$$\mathbf{A} = \begin{bmatrix} -3 & 2 \\ 1 & 4 \end{bmatrix} \quad \text{y} \quad \mathbf{B} = \begin{bmatrix} 5 & 0 \\ 2 & -6 \end{bmatrix}.$$

- Calcule \mathbf{AB} y \mathbf{BA} para las siguientes matrices:

$$\mathbf{A} = \begin{bmatrix} 1 & -2 & 3 \\ 2 & 0 & 5 \end{bmatrix} \quad \text{y} \quad \mathbf{B} = \begin{bmatrix} 3 & 0 \\ -1 & 5 \\ 3 & -2 \end{bmatrix}.$$

- Sean \mathbf{A} , \mathbf{B} y \mathbf{C} dadas por

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 0 & 4 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & 2 \\ -2 & -6 \end{bmatrix} \quad \text{y} \quad \mathbf{C} = \begin{bmatrix} 2 & -5 \\ 3 & 4 \end{bmatrix}.$$

- (a) Calcule $(\mathbf{AB})\mathbf{C}$ y $\mathbf{A}(\mathbf{BC})$.

- (b) Calcule $\mathbf{A}(\mathbf{B} + \mathbf{C})$ y $\mathbf{AB} + \mathbf{AC}$.
 (c) Calcule $(\mathbf{A} + \mathbf{B})\mathbf{C}$ y $\mathbf{AC} + \mathbf{BC}$.
 (d) Calcule $(\mathbf{AB})'$ y $\mathbf{B}'\mathbf{A}'$.
4. Calcule \mathbf{A}^2 y \mathbf{B}^2 , donde usamos la notación $\mathbf{A}^2 = \mathbf{AA}$, para las matrices

$$\mathbf{A} = \begin{bmatrix} -1 & -7 \\ 5 & 2 \end{bmatrix} \quad \text{y} \quad \mathbf{B} = \begin{bmatrix} 2 & 0 & 6 \\ -1 & 5 & -4 \\ 3 & -5 & 2 \end{bmatrix}$$

5. Calcule, si existe, el determinante de las siguientes matrices
- (a) $\begin{bmatrix} -1 & -7 \\ 5 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 0 & 6 \\ -1 & 5 & -4 \\ 3 & -5 & 2 \end{bmatrix}$
 (c) $\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 0 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 4 & 6 \\ 0 & 0 & 5 & 4 \\ 0 & 0 & 0 & 7 \end{bmatrix}$
6. Demuestre que $\mathbf{R}_x(\alpha)\mathbf{R}_x(-\alpha) = \mathbf{I}$ multiplicando directamente las matrices $\mathbf{R}_x(\alpha)$ y $\mathbf{R}_x(-\alpha)$, (vea la fórmula (26)).
7. (a) Demuestre que
- $$\mathbf{R}_x(\alpha)\mathbf{R}_y(\beta) = \begin{bmatrix} \cos(\beta) & 0 & \sin(\beta) \\ \sin(\beta)\sin(\alpha) & \cos(\alpha) & -\cos(\beta)\sin(\alpha) \\ -\cos(\alpha)\sin(\beta) & \sin(\alpha) & \cos(\beta)\cos(\alpha) \end{bmatrix}$$
- (vea las fórmulas (26) y (27)).
- (b) Demuestre que
- $$\mathbf{R}_y(\beta)\mathbf{R}_x(\alpha) = \begin{bmatrix} \cos(\beta) & \sin(\beta)\sin(\alpha) & \cos(\alpha)\sin(\beta) \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ -\sin(\alpha) & \cos(\beta)\sin(\alpha) & \cos(\beta)\cos(\alpha) \end{bmatrix}.$$
8. Sean \mathbf{A} y \mathbf{B} matrices invertibles de orden $N \times N$ y sea $\mathbf{C} = \mathbf{AB}$. Demuestre que $\mathbf{C}^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$. *Indicación.* Use la propiedad asociativa del producto de matrices.
9. Demuestre las propiedades (13) y (16) del Teorema 3.3.
10. Sea \mathbf{A} una matriz de orden $M \times N$ y \mathbf{X} una matriz de orden $N \times 1$.
- (a) ¿Cuántas multiplicaciones hacen falta para calcular \mathbf{AX} ?
 (b) ¿Cuántas sumas hacen falta para calcular \mathbf{AX} ?
11. Sea \mathbf{A} una matriz de orden $M \times N$ y sean \mathbf{B} y \mathbf{C} dos matrices de orden $N \times P$. Pruebe la propiedad distributiva de la multiplicación matricial por la izquierda: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$.

12. Sean \mathbf{A} y \mathbf{B} dos matrices de orden $M \times N$ y sea \mathbf{C} una matriz de orden $N \times P$. Pruebe la propiedad distributiva de la multiplicación matricial por la derecha: $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$.
13. Calcule \mathbf{XX}' y $\mathbf{X}'\mathbf{X}$, donde $\mathbf{X} = [1 \ -1 \ 2]$. Nota. \mathbf{X}' es el traspuesto del vector \mathbf{X} .
14. Sea \mathbf{A} una matriz de orden $M \times N$ y sea \mathbf{B} una matriz de orden $N \times P$. Pruebe que $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$. *Indicación.* Siendo $\mathbf{C} = \mathbf{AB}$ pruebe, usando la definición del producto de matrices, que el elemento (i, j) de \mathbf{C}' es igual al elemento (i, j) de $\mathbf{B}'\mathbf{A}'$.
15. Use el resultado del Ejercicio 14 y la propiedad asociativa del producto de matrices para probar que $(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'$.

Algoritmos y programas

La primera columna de la Tabla 3.1 contiene las coordenadas de los vértices de un cubo unidad situado en el primer octante y uno de cuyos vértices es el origen. Los ocho vértices pueden almacenarse en una matriz \mathbf{U} de orden 8×3 , en la que cada fila representa las coordenadas de uno de los vértices. De acuerdo con el Ejercicio 14, el producto de \mathbf{U} por la traspuesta de $\mathbf{R}_z(\pi/4)$ será una matriz de orden 8×3 (que representa la segunda columna de la Tabla 3.1, en la que cada fila representa las coordenadas del correspondiente vértice de \mathbf{U} después del primer giro). Combinando esta idea con el resultado del Ejercicio 15, obtenemos que las coordenadas de los vértices del cubo después de un número cualquiera de rotaciones consecutivas se puede representar mediante el producto por otras tantas matrices.

1. Un cubo unidad está situado en el primer octante y uno de sus vértices es el origen. Primero rotamos el cubo un ángulo de $\pi/6$ alrededor del eje OY , luego rotamos su imagen un ángulo de $\pi/4$ alrededor del eje OZ . Calcule las posiciones finales de los ocho vértices del cubo y compare su resultado con lo obtenido en el Ejemplo 3.10.

¿Cuál es la diferencia? Explique su respuesta usando que, en general, el producto de matrices no es comutativo (vea la Figura 3.3(a)–(c)). Use la instrucción `plot3` del paquete MATLAB para dibujar cada uno de los tres cubos.

2. Un cubo unidad está situado en el primer octante y uno de sus vértices es el origen. Primero rotamos el cubo un ángulo de $\pi/12$ alrededor del eje OX , luego rotamos su imagen un ángulo de $\pi/6$ alrededor del eje OZ . Calcule las posiciones finales de los ocho vértices del cubo y use la instrucción `plot3` para dibujar cada uno de los tres cubos.

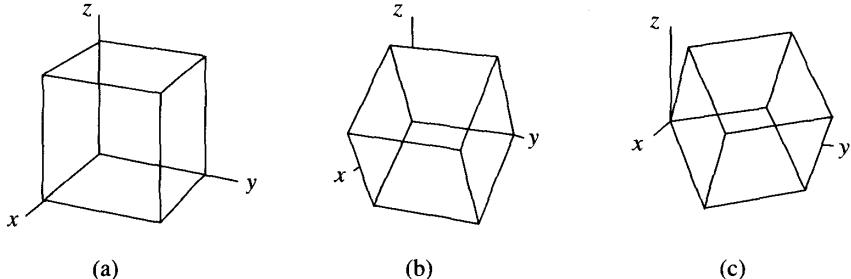


Figura 3.3 (a) El cubo de partida. (b) $V = R_y(\pi/6)U$, rotación alrededor de OY . (c) $W = R_z(\pi/4)V$, rotación alrededor de OZ .

3. Considere el tetraedro cuyos vértices son $(0, 0, 0)$, $(1, 0, 0)$, $(0, 1, 0)$ y $(0, 0, 1)$. Primero rotamos el tetraedro un ángulo de 0.15 radianes alrededor del eje OY , luego un ángulo de -1.5 radianes alrededor del eje OZ y finalmente un ángulo de 2.7 radianes alrededor del eje OX . Calcule las imágenes de los cuatro vértices y use la instrucción `plot3` para dibujar las cuatro imágenes.

3.3 Sistemas lineales triangulares

Desarrollaremos ahora el **algoritmo de sustitución regresiva**, con el que podremos resolver un sistema de ecuaciones lineales cuya matriz de coeficientes sea triangular superior. Este algoritmo será luego incorporado al algoritmo de resolución de un sistema de ecuaciones lineales general en la Sección 3.4.

Definición 3.2. Se dice que una matriz $A = [a_{ij}]$ de orden $N \times N$ es **triangular superior** cuando sus elementos verifican $a_{ij} = 0$ siempre que $i > j$. Se dice que una matriz $A = [a_{ij}]$ de orden $N \times N$ es **triangular inferior** si $a_{ij} = 0$ siempre que $i < j$.

Vamos a desarrollar un método para hallar la solución de un sistema de ecuaciones lineales triangular superior y dejamos como ejercicio el caso de los sistemas triangulares inferiores. Si A es una matriz triangular superior, entonces se dice que el sistema de ecuaciones $AX = B$ es un **un sistema triangular superior** de ecuaciones lineales, sistema que tiene la siguiente forma:

$$\begin{aligned}
 (1) \quad & a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1N-1}x_{N-1} + a_{1N}x_N = b_1 \\
 & a_{22}x_2 + a_{23}x_3 + \cdots + a_{2N-1}x_{N-1} + a_{2N}x_N = b_2 \\
 & a_{33}x_3 + \cdots + a_{3N-1}x_{N-1} + a_{3N}x_N = b_3 \\
 & \vdots \qquad \qquad \qquad \vdots \\
 & a_{N-1N-1}x_{N-1} + a_{NN}x_N = b_{N-1} \\
 & a_{NN}x_N = b_N.
 \end{aligned}$$

Teorema 3.5 (Sustitución regresiva). Supongamos que $AX = B$ es un sistema triangular superior como el dado en (1). Si

$$(2) \quad a_{kk} \neq 0 \quad \text{para } k = 1, 2, \dots, N,$$

entonces existe una solución única de (1).

Demostración constructiva. La solución es fácil de hallar. La última ecuación sólo contiene la incógnita x_N , así que empezamos por ésta:

$$(3) \quad x_N = \frac{b_N}{a_{NN}}.$$

Ahora ya conocemos x_N así que podemos usarla en la penúltima ecuación:

$$(4) \quad x_{N-1} = \frac{b_{N-1} - a_{N-1N}x_N}{a_{N-1N-1}}.$$

Ahora usamos x_N y x_{N-1} para hallar x_{N-2} :

$$(5) \quad x_{N-2} = \frac{b_{N-2} - a_{N-2N-1}x_{N-1} - a_{N-2N}x_N}{a_{N-2N-2}}.$$

Una vez calculados los valores $x_N, x_{N-1}, \dots, x_{k+1}$, el paso general es

$$(6) \quad x_k = \frac{b_k - \sum_{j=k+1}^N a_{kj}x_j}{a_{kk}} \quad \text{para } k = N-1, N-2, \dots, 1.$$

La unicidad de la solución es fácil de ver. La última ecuación implica que b_N/a_{NN} es el único posible valor de x_N y, por inducción finita, los valores de $x_{N-1}, x_{N-2}, \dots, x_1$ también son únicos.

Ejemplo 3.12. Vamos a usar el método de sustitución regresiva para resolver el sistema lineal

$$\begin{aligned} 4x_1 - x_2 + 2x_3 + 3x_4 &= 20 \\ -2x_2 + 7x_3 - 4x_4 &= -7 \\ 6x_3 + 5x_4 &= 4 \\ 3x_4 &= 6. \end{aligned}$$

Despejando x_4 en la última ecuación obtenemos

$$x_4 = \frac{6}{3} = 2.$$

Usando que $x_4 = 2$ en la tercera ecuación, obtenemos

$$x_3 = \frac{4 - 5(2)}{6} = -1.$$

Ahora usamos los valores $x_3 = -1$ y $x_4 = 2$ para despejar x_2 en la segunda ecuación:

$$x_2 = \frac{-7 - 7(-1) + 4(2)}{-2} = -4.$$

Finalmente, x_1 se obtiene de la primera ecuación:

$$x_1 = \frac{20 + 1(-4) - 2(-1) - 3(2)}{4} = 3.$$

La condición $a_{kk} \neq 0$ es esencial porque en la fórmula (6) hay que dividir entre a_{kk} . Si este requisito no se cumple, entonces o bien no hay solución o bien hay infinitas soluciones.

Ejemplo 3.13. Vamos a probar que el siguiente sistema no tiene solución:

$$(7) \quad \begin{aligned} 4x_1 - x_2 + 2x_3 + 3x_4 &= 20 \\ 0x_2 + 7x_3 - 4x_4 &= -7 \\ 6x_3 + 5x_4 &= 4 \\ 3x_4 &= 6. \end{aligned}$$

Usando la última ecuación de (7), obtenemos $x_4 = 2$, que se sustituye en las ecuaciones segunda y tercera para obtener

$$(8) \quad \begin{aligned} 7x_3 - 8 &= -7 \\ 6x_3 + 10 &= 4. \end{aligned}$$

La primera ecuación de (8) implica que $x_3 = 1/7$ y la segunda implica que $x_3 = -1$. Esta contradicción nos lleva a la conclusión de que el sistema lineal (7) no tiene solución.

Ejemplo 3.14. Veamos ahora que el siguiente sistema tiene infinitas soluciones:

$$(9) \quad \begin{aligned} 4x_1 - x_2 + 2x_3 + 3x_4 &= 20 \\ 0x_2 + 7x_3 + 0x_4 &= -7 \\ 6x_3 + 5x_4 &= 4 \\ 3x_4 &= 6. \end{aligned}$$

Usando la última ecuación de (9), tenemos que $x_4 = 2$, que se sustituye en las ecuaciones segunda y tercera para obtener $x_3 = -1$, valor que verifica ambas. Pero ahora resulta que, de las tres últimas ecuaciones, sólo hemos obtenido valores para x_3 y x_4 y cuando sustituimos estos valores en la primera ecuación de (9), el resultado es la ecuación

$$(10) \quad x_2 = 4x_1 - 16,$$

que tiene infinitas soluciones. En resumen, el sistema (9) tiene infinitas soluciones. Para cada valor de x_1 que elijamos en (10), el valor de x_2 estará únicamente determinado. Por ejemplo, si incluimos la ecuación $x_1 = 2$ en el sistema (9), entonces obtenemos $x_2 = -8$ en (10). ■

El Teorema 3.4 establece que un sistema lineal $\mathbf{AX} = \mathbf{B}$, siendo \mathbf{A} una matriz de orden $N \times N$, tiene solución única si, y sólo si, $\det(\mathbf{A}) \neq 0$. El siguiente teorema establece que si un elemento de la diagonal principal de una matriz triangular, superior o inferior, es cero, entonces $\det(\mathbf{A}) = 0$. Por tanto, viendo cómo son los coeficientes de los tres ejemplos anteriores, queda claro que el sistema del Ejemplo 3.12 tiene solución única y que los sistemas de los Ejemplos 3.13 y 3.14 no tienen solución única. La prueba del Teorema 3.6 puede hallarse en la mayoría de los textos de introducción al álgebra lineal.

Teorema 3.6. Si una matriz $\mathbf{A} = [a_{ij}]$ de orden $N \times N$ es triangular superior o inferior, entonces

$$(11) \quad \det(\mathbf{A}) = a_{11}a_{22} \cdots a_{NN} = \prod_{i=1}^N a_{ii}.$$

El valor del determinante de la matriz de los coeficientes del Ejemplo 3.12 es $\det \mathbf{A} = 4(-2)(6)(3) = -144$. Los valores de los determinantes de las matrices de los coeficientes de los Ejemplos 3.13 y 3.14 son ambos $4(0)(6)(3) = 0$.

El siguiente programa sirve para resolver un sistema triangular superior como el dado en (1) por el método de sustitución regresiva, supuesto que $a_{kk} \neq 0$ para $k = 1, 2, \dots, N$.

Programa 3.1 (Sustitución regresiva). Resolución de un sistema triangular superior $\mathbf{AX} = \mathbf{B}$ por el método de sustitución regresiva. El método funciona sólo si todos los elementos diagonales son distintos de cero. Primero se calcula $x_N = b_N/a_{NN}$ y luego se usa la regla

$$x_k = \frac{b_k - \sum_{j=k+1}^N a_{kj}x_j}{a_{kk}} \quad \text{para } k = N-1, N-2, \dots, 1.$$

```
function X=backsub(A,B)
% Datos
%     - A es una matriz triangular superior
%         invertible de orden n x n
%     - B es una matriz de orden n x 1
% Resultado
%     - X es la solución del sistema lineal AX = B
% Cálculo de la dimensión de B e inicialización de X
n=length(B);
X=zeros(n,1); X(n)=B(n)/A(n,n);
for k=n-1:-1:1
    X(k)=(B(k)-A(k,k+1:n)*X(k+1:n))/A(k,k);
end
```

Ejercicios

En los Ejercicios 1 a 3, resuelva el sistema triangular superior y halle el valor del determinante de la matriz de los coeficientes.

$$1. \begin{aligned} 3x_1 - 2x_2 + x_3 - x_4 &= 8 \\ 4x_2 - x_3 + 2x_4 &= -3 \\ 2x_3 + 3x_4 &= 11 \\ 5x_4 &= 15 \end{aligned}$$

$$2. \begin{aligned} 5x_1 - 3x_2 - 7x_3 + x_4 &= -14 \\ 11x_2 + 9x_3 + 5x_4 &= 22 \\ 3x_3 - 13x_4 &= -11 \\ 7x_4 &= 14 \end{aligned}$$

$$\begin{aligned} 4x_1 - x_2 + 2x_3 + 2x_4 - x_5 &= 4 \\ -2x_2 + 6x_3 + 2x_4 + 7x_5 &= 0 \\ x_3 - x_4 - 2x_5 &= 3 \\ -2x_4 - x_5 &= 10 \\ 3x_5 &= 6 \end{aligned}$$

$$3. \begin{aligned} 4x_1 - x_2 + 2x_3 + 2x_4 - x_5 &= 4 \\ -2x_2 + 6x_3 + 2x_4 + 7x_5 &= 0 \\ x_3 - x_4 - 2x_5 &= 3 \\ -2x_4 - x_5 &= 10 \\ 3x_5 &= 6 \end{aligned}$$

4. (a) Consideremos las dos matrices triangulares superiores

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix} \quad \text{y} \quad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ 0 & b_{22} & b_{23} \\ 0 & 0 & b_{33} \end{bmatrix}.$$

Pruebe que su producto $\mathbf{C} = \mathbf{AB}$ es también triangular superior.

- (b) Sean \mathbf{A} y \mathbf{B} dos matrices triangulares superiores de orden $N \times N$. Pruebe que su producto también es triangular superior.

5. Resuelva el sistema triangular inferior $\mathbf{AX} = \mathbf{B}$ siguiente y calcule $\det(A)$.

$$\begin{aligned} 2x_1 &= 6 \\ -x_1 + 4x_2 &= 5 \\ 3x_1 - 2x_2 - x_3 &= 4 \\ x_1 - 2x_2 + 6x_3 + 3x_4 &= 2 \end{aligned}$$

6. Resuelva el sistema triangular inferior $\mathbf{AX} = \mathbf{B}$ siguiente y calcule $\det(A)$.

$$\begin{aligned} 5x_1 &= -10 \\ x_1 + 3x_2 &= 4 \\ 3x_1 + 4x_2 + 2x_3 &= 2 \\ -x_1 + 3x_2 - 6x_3 - x_4 &= 5 \end{aligned}$$

7. Pruebe que en el método de sustitución regresiva se realizan N divisiones, $(N^2 - N)/2$ multiplicaciones y $(N^2 - N)/2$ sumas y restas. *Indicación.* Use la fórmula

$$\sum_{k=1}^M k = M(M+1)/2.$$

Algoritmos y programas

1. Use el Programa 3.1 para resolver el sistema $\mathbf{U}\mathbf{X} = \mathbf{B}$ siendo

$$\mathbf{U} = [u_{ij}]_{10 \times 10} \quad \text{con} \quad u_{ij} = \begin{cases} \cos(ij) & i \leq j, \\ 0 & i > j \end{cases}$$

y $\mathbf{B} = [b_{i1}]_{10 \times 1}$ siendo $b_{i1} = \tan(i)$.

2. *Algoritmo de sustitución progresiva.* Se dice que un sistema lineal $\mathbf{AX} = \mathbf{B}$ es triangular inferior cuando $a_{ij} = 0$ siempre que $i < j$. Construya un programa `forsub`, análogo al Programa 3.1, para resolver el siguiente sistema triangular inferior. *Observación.* Este programa lo usaremos en la Sección 3.5.

$$\begin{aligned} a_{11}x_1 &= b_1 \\ a_{21}x_1 + a_{22}x_2 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \\ \vdots &\vdots \vdots \vdots \\ a_{N-1,1}x_1 + a_{N-1,2}x_2 + a_{N-1,3}x_3 + \cdots + a_{N-1,N-1}x_{N-1} &= b_{N-1} \\ a_{N,1}x_1 + a_{N,2}x_2 + a_{N,3}x_3 + \cdots + a_{N,N-1}x_{N-1} + a_{NN}x_N &= b_N \end{aligned}$$

3. Use el programa `forsub` para resolver el sistema $\mathbf{L}\mathbf{X} = \mathbf{B}$, siendo

$$\mathbf{L} = [l_{ij}]_{20 \times 20} \quad \text{con} \quad l_{ij} = \begin{cases} i+j & i \geq j, \\ 0 & i < j, \end{cases} \quad \text{y} \quad \mathbf{B} = [b_{i1}]_{20 \times 1} \quad \text{con} \quad b_{i1} = i.$$

3.4 Eliminación gaussiana y pivoteo

En esta sección desarrollamos un método para resolver un sistema de ecuaciones lineales general $\mathbf{AX} = \mathbf{B}$ de N ecuaciones con N incógnitas. El objetivo es construir un sistema triangular superior equivalente $\mathbf{UX} = \mathbf{Y}$ que podamos resolver usando el método de la Sección 3.3.

Se dice que dos sistemas de orden $N \times N$ son *equivalentes* cuando tienen el mismo conjunto de soluciones. Los teoremas del álgebra lineal prueban que hay ciertas transformaciones que no cambian el conjunto de soluciones de un sistema de ecuaciones lineales.

Teorema 3.7 (Transformaciones elementales). Cualquiera de las siguientes operaciones aplicadas a un sistema de ecuaciones lineales produce un sistema equivalente.

- (1) Intercambio: El orden de las ecuaciones puede cambiarse.
- (2) Escalado: Multiplicar una ecuación por una constante no nula.
- (3) Sustitución: Una ecuación puede ser reemplazada por la suma de ella misma más un múltiplo de otra ecuación.

La forma habitual de usar (3) es reemplazar una ecuación por la diferencia de esa ecuación y un múltiplo de otra. Estos conceptos se ilustran en el siguiente ejemplo.

Ejemplo 3.15. Vamos a determinar la parábola de ecuación $y = A + Bx + Cx^2$ que pasa por los puntos $(1, 1)$, $(2, -1)$ y $(3, 1)$.

Para cada punto obtenemos una ecuación que relaciona el valor de la abscisa x con el de la ordenada y . El resultado es el sistema lineal

$$(4) \quad \begin{aligned} A + B + C &= 1 && \text{en } (1, 1) \\ A + 2B + 4C &= -1 && \text{en } (2, -1) \\ A + 3B + 9C &= 1 && \text{en } (3, 1). \end{aligned}$$

Aplicando la transformación de sustitución (3), la incógnita A es eliminada de las ecuaciones segunda y tercera sustrayendo la primera de ambas. El sistema lineal equivalente es

$$(5) \quad \begin{aligned} A + B + C &= 1 \\ B + 3C &= -2 \\ 2B + 8C &= 0. \end{aligned}$$

Ahora, la variable B se elimina de la tercera ecuación de (5) restándole a dicha ecuación el doble de la segunda, lo que nos lleva a un sistema equivalente que es triangular superior:

$$(6) \quad \begin{aligned} A + B + C &= 1 \\ B + 3C &= -2 \\ 2C &= 4. \end{aligned}$$

Finalmente, usamos el algoritmo de sustitución regresiva para hallar las incógnitas, los coeficientes de la ecuación de la parábola, $C = 4/2 = 2$, $B = -2 - 3(2) = -8$, y $A = 1 - (-8) - 2 = 7$, con lo que dicha ecuación es $y = 7 - 8x + 2x^2$. ■

Una forma eficaz de trabajar es almacenar todas las constantes del sistema lineal $\mathbf{A}\mathbf{x} = \mathbf{b}$ en una matriz de orden $N \times (N + 1)$ que se obtiene añadiendo a la matriz \mathbf{A} una columna, la columna $(N + 1)$ -ésima, en la que se almacenan los términos de \mathbf{b} (es decir, $a_{kN+1} = b_k$). Cada fila de esta matriz, que se llama **matriz ampliada** del sistema y se denota por $[\mathbf{A}|\mathbf{b}]$, contiene toda la

información necesaria para representar la correspondiente ecuación del sistema lineal:

$$(7) \quad [\mathbf{A}|\mathbf{B}] = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1N} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2N} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} & b_N \end{array} \right].$$

Un sistema $\mathbf{AX} = \mathbf{B}$, cuya matriz ampliada viene dada en (7), puede resolverse realizando las operaciones elementales con las filas de la matriz ampliada $[\mathbf{A}|\mathbf{B}]$. Las variables x_k no sirven más que para marcar el sitio de los coeficientes y pueden ser omitidas hasta el final de los cálculos.

Teorema 3.8 (Operaciones elementales con las filas). Cualquiera de las siguientes operaciones aplicada a la matriz ampliada (7) produce un sistema lineal equivalente.

- (8) Intercambio: El orden de las filas puede cambiarse.
- (9) Escalado: Multiplicar una fila por una constante no nula.
- (10) Sustitución: Una fila puede ser reemplazada por la suma de esa fila más un múltiplo de cualquier otra fila; o sea,
 $\text{fila}_r = \text{fila}_r - m_{rq} \times \text{fila}_q$.

Como se conoce de los cursos previos de álgebra lineal, la forma habitual de usar (10) es reemplazar una fila por la diferencia entre esa fila y un múltiplo de otra.

Definición 3.3 (Pivotes y multiplicadores). El elemento a_{qq} de la matriz de los coeficientes en el paso $q+1$ que se usará en la eliminación de a_{rq} , para $r = q+1, q+2, \dots, N$, se llama q -ésimo **pivote** y la fila q -ésima se llama **fila pivote**. Los números $m_{rq} = a_{rq}/a_{qq}$ ($r = q+1, q+2, \dots, N$) por los que se multiplica la fila pivote para restarla de las correspondientes filas posteriores se llaman **multiplicadores** de la eliminación. ▲

El siguiente ejemplo muestra cómo se usan las operaciones descritas en el Teorema 3.8 para obtener un sistema triangular superior $\mathbf{UX} = \mathbf{Y}$ que sea equivalente a un sistema lineal $\mathbf{AX} = \mathbf{B}$ en el que \mathbf{A} es una matriz de orden $N \times N$.

Ejemplo 3.16. Vamos a expresar el siguiente sistema en forma de matriz ampliada, luego hallaremos un sistema triangular superior que sea equivalente y, final-

mente, la solución.

$$\begin{aligned}x_1 + 2x_2 + x_3 + 4x_4 &= 13 \\2x_1 + 0x_2 + 4x_3 + 3x_4 &= 28 \\4x_1 + 2x_2 + 2x_3 + x_4 &= 20 \\-3x_1 + x_2 + 3x_3 + 2x_4 &= 6.\end{aligned}$$

La matriz ampliada es

$$\text{pivot } \rightarrow \left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 2 & 0 & 4 & 3 & 28 \\ 4 & 2 & 2 & 1 & 20 \\ -3 & 1 & 3 & 2 & 6 \end{array} \right].$$

$$\begin{aligned}m_{21} &= 2 \\m_{31} &= 4 \\m_{41} &= -3\end{aligned}$$

La primera fila se usa para eliminar los elementos de la primera columna que están por debajo de la diagonal principal. En este paso, esta primera fila es la fila pivote y su elemento $a_{11} = 1$ es el elemento pivote. Los valores $m_{k1} = a_{k1}/a_{11}$ (para $k = 2, 3, 4$) son los multiplicadores, o sea, los escalares por los que hay que multiplicar la primera fila para, restando de la fila k -ésima el correspondiente múltiplo de la primera fila, hacer cero el elemento a_{k1} . El resultado de la eliminación es

$$\text{pivot } \rightarrow \left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 0 & -4 & 2 & -5 & 2 \\ 0 & -6 & -2 & -15 & -32 \\ 0 & 7 & 6 & 14 & 45 \end{array} \right].$$

$$\begin{aligned}m_{32} &= 1.5 \\m_{42} &= -1.75\end{aligned}$$

Ahora, usamos la segunda fila para eliminar los elementos de la segunda columna que están por debajo de la diagonal principal. Esta segunda fila es la fila pivote y los valores $m_{k2} = a_{k2}/a_{22}$ (para $k = 3, 4$) son los multiplicadores. El resultado de la eliminación es

$$\text{pivot } \rightarrow \left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 0 & -4 & 2 & -5 & 2 \\ 0 & 0 & -5 & -7.5 & -35 \\ 0 & 0 & 9.5 & 5.25 & 48.5 \end{array} \right].$$

$$m_{43} = -1.9$$

Finalmente, restamos de la cuarta fila la tercera multiplicada por $m_{43} = a_{43}/a_{33} = -1.9$ y el resultado es el sistema triangular superior

$$(11) \quad \left[\begin{array}{cccc|c} 1 & 2 & 1 & 4 & 13 \\ 0 & -4 & 2 & -5 & 2 \\ 0 & 0 & -5 & -7.5 & -35 \\ 0 & 0 & 0 & -9 & -18 \end{array} \right].$$

Usando el algoritmo de sustitución regresiva para resolver (11) obtenemos

$$x_4 = 2, \quad x_3 = 4, \quad x_2 = -1, \quad x_1 = 3.$$

El proceso que acabamos de describir se llama *eliminación gaussiana* o *método de eliminación de Gauss* pero debemos modificarlo si queremos que funcione en casi todas las situaciones. El problema que puede aparecer es el siguiente: Si $a_{kk} = 0$, entonces no podemos usar la fila k -ésima para eliminar los elementos de la columna k -ésima que están por debajo de la diagonal principal. Lo que hacemos es intercambiar la fila k -ésima con alguna fila posterior para conseguir un elemento pivote que no sea cero; si esto no puede hacerse, entonces la matriz de los coeficientes del sistema es singular y el sistema no tiene solución única.

Teorema 3.9 (Eliminación gaussiana con sustitución regresiva). Si \mathbf{A} es una matriz invertible de orden $N \times N$, entonces existe un sistema lineal $\mathbf{U}\mathbf{X} = \mathbf{Y}$, equivalente al sistema $\mathbf{AX} = \mathbf{B}$, en el que \mathbf{U} es una matriz triangular superior con elementos diagonales $u_{kk} \neq 0$. Una vez construidos \mathbf{U} e \mathbf{Y} , se usa el algoritmo de sustitución regresiva para resolver $\mathbf{UX} = \mathbf{Y}$ y, así, calcular la solución \mathbf{X} .

Demuestra. Usaremos la matriz ampliada del sistema, en la que \mathbf{B} se almacena como su columna $(N+1)$ -ésima:

$$\mathbf{AX} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2N}^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3N}^{(1)} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{N1}^{(1)} & a_{N2}^{(1)} & a_{N3}^{(1)} & \cdots & a_{NN}^{(1)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} a_{1N+1}^{(1)} \\ a_{2N+1}^{(1)} \\ a_{3N+1}^{(1)} \\ \vdots \\ a_{NN+1}^{(1)} \end{bmatrix} = \mathbf{B}$$

y construiremos un sistema triangular superior equivalente $\mathbf{UX} = \mathbf{Y}$:

$$\mathbf{UX} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & a_{NN}^{(N)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} a_{1N+1}^{(1)} \\ a_{2N+1}^{(2)} \\ a_{3N+1}^{(3)} \\ \vdots \\ a_{NN+1}^{(N)} \end{bmatrix} = \mathbf{Y}.$$

Paso 1. Almacenamos todos los coeficientes en la matriz ampliada. El superíndice (1) indica que ésta es la primera vez que se almacena un número en

la posición (r, c) :

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2N}^{(1)} & a_{2N+1}^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3N}^{(1)} & a_{3N+1}^{(1)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ a_{N1}^{(1)} & a_{N2}^{(1)} & a_{N3}^{(1)} & \cdots & a_{NN}^{(1)} & a_{NN+1}^{(1)} \end{array} \right].$$

Paso 2. Si es necesario, intercambiamos filas de manera que $a_{11}^{(1)} \neq 0$; entonces se elimina la incógnita x_1 en todas las filas desde la segunda hasta la última. En este proceso, m_{r1} es el número por el que hay que multiplicar la primera fila para restarla de la fila r -ésima. Describiremos este paso como si fueran instrucciones del paquete de programas MATLAB

```
for r = 2 : N
    m_r1 = a_r1^(1) / a_11^(1);
    a_r1^(2) = 0;
    for c = 2 : N + 1
        a_rc^(2) = a_rc^(1) - m_r1 * a_1c^(1);
    end
end
```

Los nuevos elementos $a_{rc}^{(2)}$ se superindizan con un (2) para señalar que ésta es la segunda vez que se almacena un número en la posición (r, c) de la matriz. El resultado tras el paso 2 es

$$\left[\begin{array}{ccccc|c} a_{12}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2N+1}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & \cdots & a_{3N}^{(2)} & a_{3N+1}^{(2)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & a_{N2}^{(2)} & a_{N3}^{(2)} & \cdots & a_{NN}^{(2)} & a_{NN+1}^{(2)} \end{array} \right].$$

Paso 3. Si es necesario, intercambiamos la segunda fila con alguna posterior para que $a_{22}^{(2)} \neq 0$; luego, eliminamos la incógnita x_2 en todas las filas desde la tercera hasta la última. En este proceso, m_{r2} es el número por el que hay que multiplicar la segunda fila para restarla de la fila r -ésima.

```

for r = 3 : N
     $m_{r2} = a_{r2}^{(2)} / a_{22}^{(2)}$ ;
     $a_{r2}^{(3)} = 0$ ;
    for c = 3 : N + 1
         $a_{rc}^{(3)} = a_{rc}^{(2)} - m_{r2} * a_{2c}^{(2)}$ ;
    end
end

```

Los nuevos elementos $a_{rc}^{(3)}$ se superindizan con un (3) para señalar que ésta es la tercera vez que se almacena un número en la posición (r, c) de la matriz. El resultado tras el paso 3 es

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2N+1}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} & a_{3N+1}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & a_{N3}^{(3)} & \cdots & a_{NN}^{(3)} & a_{NN+1}^{(3)} \end{array} \right].$$

Paso q + 1. Este es el paso general. Si es necesario, intercambiamos la fila que ocupa el lugar q -ésimo con alguna posterior para que $a_{qq}^{(q)} \neq 0$; luego, eliminamos la incógnita x_q en todas las filas desde la $(q + 1)$ -ésima hasta la última. Ahora, m_{rq} es el número por el que hay que multiplicar la q -ésima para restarla de la fila r -ésima.

```

for r = q + 1 : N
     $m_{rq} = a_{rq}^{(q)} / a_{qq}^{(q)}$ ;
     $a_{rq}^{(q+1)} = 0$ ;
    for c = q + 1 : N + 1
         $a_{rc}^{(q+1)} = a_{rc}^{(q)} - m_{rq} * a_{qc}^{(q)}$ ;
    end
end

```

El resultado final, una vez que hemos eliminado la incógnita x_{N-1} en la

última fila es

$$\left[\begin{array}{cccc|c|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1\ N+1}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2\ N+1}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} & a_{3\ N+1}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{NN}^{(N)} & a_{N\ N+1}^{(N)} \end{array} \right].$$

Y el proceso de triangularización está ya terminado.

Puesto que A es invertible, cuando se realizan las operaciones con las filas, las matrices que se van obteniendo sucesivamente son también invertibles. Esto garantiza que $a_{kk}^{(k)} \neq 0$ para todo k a lo largo del proceso. Por tanto, podemos usar el algoritmo de sustitución regresiva para resolver $UX = Y$, lo que finaliza la prueba del teorema

Pivoteo para evitar $a_{qq}^{(q)} = 0$

Cuando $a_{qq}^{(q)} = 0$, la fila q -ésima no puede usarse para eliminar los elementos de la columna q -ésima que están por debajo de la diagonal principal. Se hace necesario entonces hallar una fila posterior, digamos la k -ésima con $k > q$, en la que $a_{kq}^{(q)} \neq 0$, e intercambiarlas para obtener un pivote no nulo. Este proceso se llama pivoteo y el criterio para decidir qué fila escoger se llama estrategia de pivoteo. La estrategia de **pivoteo trivial** es la siguiente: Si $a_{qq}^{(q)} \neq 0$, entonces no se hace intercambio de fila; mientras que si $a_{qq}^{(q)} = 0$, entonces se localiza la primera fila por debajo de la q -ésima en la cual se tenga $a_{kq}^{(q)} \neq 0$ y se intercambia la fila q -ésima con la k -ésima. Esto proporciona el pivote no nulo deseado.

Estrategias de pivoteo para reducir los errores

Como los computadores usan una aritmética cuya precisión está fijada de antemano, es posible que cada vez que se realice una operación aritmética se introduzca un pequeño error. El siguiente ejemplo pone de manifiesto que el uso de la estrategia de pivoteo trivial en la eliminación gaussiana puede llevar aparejado un error apreciable en la solución de un sistema de ecuaciones lineales calculada con un computador.

Ejemplo 3.17. Los valores $x_1 = x_2 = 1.000$ son la solución del sistema

$$(12) \quad \begin{aligned} 1.133x_1 + 5.281x_2 &= 6.414 \\ 24.14x_1 - 1.210x_2 &= 22.93. \end{aligned}$$

Haciendo las operaciones con una precisión de cuatro cifras decimales significativas (véanse los Ejercicios 6 y 7 de la Sección 1.3), vamos a usar el método de eliminación de Gauss con la estrategia de pivoteo trivial para hallar una solución aproximada del sistema.

El multiplicador que usamos para obtener un sistema triangular superior es $m_{21} = 24.14/1.133 = 21.31$. Haciendo las operaciones con cuatro cifras de precisión, los nuevos coeficientes son:

$$a_{22}^{(2)} = -1.210 - 21.31(5.281) = -1.210 - 112.5 = -113.7$$

$$a_{23}^{(2)} = 22.93 - 21.31(6.414) = 22.93 - 136.7 = -113.8,$$

y el sistema triangular superior resultante es

$$1.133x_1 + 5.281x_2 = 6.414$$

$$-113.7x_2 = -113.8.$$

Usando ahora el algoritmo de sustitución regresiva, obtenemos como solución

$$x_2 = \frac{-113.8}{-113.7} = 1.001 \text{ y } x_1 = \frac{6.414 - 5.281(1.001)}{1.133} = \frac{6.414 - 5.286}{1.133} = 0.9956. \blacksquare$$

El error de la solución del sistema lineal (12) se debe a la magnitud del multiplicador $m_{21} = 21.31$. En el siguiente ejemplo reducimos el tamaño del multiplicador intercambiando previamente la primera ecuación con la segunda y, a continuación, resolvemos el sistema usando el método de eliminación de Gauss con la estrategia de pivoteo trivial.

Ejemplo 3.18. Haciendo las operaciones con una precisión de cuatro cifras, vamos a usar el método de eliminación de Gauss con la estrategia de pivoteo trivial para hallar una solución aproximada del sistema

$$24.14x_1 - 1.210x_2 = 22.93$$

$$1.133x_1 + 5.281x_2 = 6.414.$$

Esta vez, el multiplicador es $m_{21} = 1.133/24.14 = 0.04693$ y los nuevos coeficientes son

$$a_{22}^{(2)} = 5.281 - 0.04693(-1.210) = 5.281 + 0.05679 = 5.338$$

$$a_{23}^{(2)} = 6.414 - 0.04693(22.93) = 6.414 - 1.076 = 5.338.$$

El sistema triangular superior resultante es

$$24.14x_1 - 1.210x_2 = 22.93$$

$$5.338x_2 = 5.338,$$

que resolvemos por sustitución regresiva, obteniendo $x_2 = 5.338/5.338 = 1.000$ y $x_1 = (22.93 + 1.210(1.000))/(24.14) = 1.000$. \blacksquare

El propósito de las estrategias de pivoteo es usar como pivote el elemento de mayor magnitud y, una vez colocado en la diagonal principal, usarlo para eliminar los restantes elementos de su columna que están por debajo de él. Si en la columna q hay más de un elemento no nulo en la diagonal principal o por debajo de ésta, entonces hay varias formas de elegir qué filas se intercambian. La estrategia de **pivoteo parcial**, que hemos mostrado en el Ejemplo 3.18, es la más habitual y la usaremos en el Programa 3.2. Consiste en lo siguiente: Para reducir la propagación de los errores de redondeo, se sugiere que se compare el tamaño de todos los elementos de la columna q desde el que está en la diagonal hasta el de la última fila. Una vez localizada la fila, digamos la k -ésima, en la que se encuentra el elemento de mayor valor absoluto; o sea, si

$$|a_{kq}| = \max\{|a_{qq}|, |a_{q+1q}|, \dots, |a_{N-1q}|, |a_{Nq}|\},$$

entonces intercambiaremos la fila q -ésima con la fila k -ésima, salvo que $k = q$; de esa manera, los multiplicadores m_{rq} , para $r = q+1, \dots, N$, serán todos menores que 1 en valor absoluto. Este proceso suele conservar las magnitudes relativas de los elementos de la matriz \mathbf{U} del Teorema 3.9 del mismo orden que las de los coeficientes de la matriz original. Normalmente, la elección como pivote del mayor elemento también ayuda a que se propague un error más pequeño.

En la sección 3.5 veremos que hace falta un total de $(4N^3 + 9N^2 - 7N)/6$ operaciones aritméticas para resolver un sistema de orden $N \times N$. Si $N = 20$, el número total de operaciones que hay que efectuar es 5 910 y la propagación de los errores en los cálculos podría dar lugar a una respuesta incorrecta. La técnica de **pivoteo parcial escalado** puede usarse para reducir aún más los efectos de la propagación de los errores. En el pivoteo parcial escalado se elige el elemento de la columna q -ésima, en o por debajo de la diagonal principal, que tiene mayor tamaño relativo con respecto al resto de los elementos de su fila. Es decir, primero se busca en cada fila, desde la q -ésima hasta la última, el elemento de mayor tamaño, digamos s_r :

$$(13) \quad s_r = \max\{|a_{rq}|, |a_{r,q+1}|, \dots, |a_{rN}|\} \quad \text{para } r = q, q+1, \dots, N.$$

La fila pivote será la que ocupa el lugar k para el cual

$$(14) \quad \frac{|a_{kq}|}{s_k} = \max \left\{ \frac{|a_{qq}|}{s_q}, \frac{|a_{q+1q}|}{s_{q+1}}, \dots, \frac{|a_{Nq}|}{s_N} \right\}.$$

Ahora se intercambian las filas q -ésima y k -ésima, salvo que $q = k$. De nuevo, el propósito de esta estrategia de pivoteo es mantener las magnitudes relativas de los elementos de la matriz \mathbf{U} del Teorema 3.9 del mismo orden que las de los coeficientes de la matriz original.

Matrices mal condicionadas

Se dice que una matriz \mathbf{A} está **mal condicionada** si existe una matriz \mathbf{B} de manera que cambios pequeños en los elementos de \mathbf{A} o \mathbf{B} provocan cambios

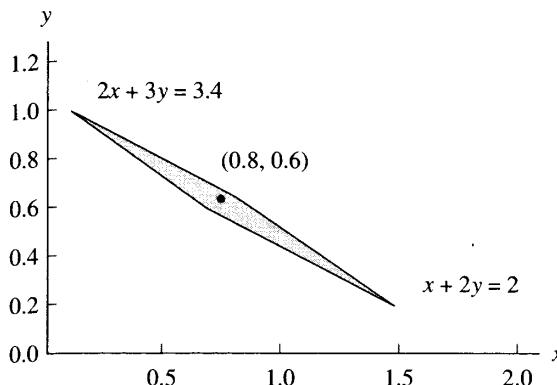


Figura 3.4 Una región donde dos ecuaciones “casi se verifican”.

grandes en $\mathbf{X} = \mathbf{A}^{-1}\mathbf{B}$. Se dice que el sistema $\mathbf{AX} = \mathbf{B}$ está mal condicionado si \mathbf{A} está mal condicionada y, cuando esto ocurre, los métodos numéricos para obtener una solución aproximada son proclives a tener más errores.

Una circunstancia que suele llevar aparejada la mala condición es que la matriz sea “casi singular” y su determinante sea casi cero. Otra posible causa de una mala condición es, por ejemplo, que un sistema de dos ecuaciones corresponda a dos líneas rectas casi paralelas (o que un sistema de tres ecuaciones corresponda a tres planos casi paralelos). Una consecuencia de la mala condición de un sistema es que algunos valores erróneos pueden parecer soluciones aceptables al sustituirlos en las ecuaciones. Por ejemplo, consideremos las dos ecuaciones

$$(15) \quad \begin{aligned} x + 2y - 2.00 &= 0 \\ 2x + 3y - 3.40 &= 0. \end{aligned}$$

Si sustituimos $x_0 = 1.00$ y $y_0 = 0.48$ en estas ecuaciones, obtenemos “casi ceros”:

$$\begin{aligned} 1 + 2(0.48) - 2.00 &= 1.96 - 2.00 = -0.04 \approx 0 \\ 2 + 3(0.48) - 3.40 &= 3.44 - 3.40 = 0.04 \approx 0. \end{aligned}$$

La discrepancia del 0 es sólo de ± 0.04 . Sin embargo, la solución correcta de este sistema lineal es $x = 0.8$ e $y = 0.6$, de manera que los errores en la solución aproximada son $x - x_0 = 0.80 - 1.00 = -0.20$ e $y - y_0 = 0.60 - 0.48 = 0.12$. Es decir, la mera sustitución de valores en un sistema de ecuaciones no es una garantía de exactitud. La región romboidal R que se muestra en la Figura 3.4 representa un conjunto en el que las dos ecuaciones (15) “casi se verifican”:

$$R = \{(x, y) : |x + 2y - 2.00| < 0.1 \quad \text{y} \quad |2x + 3y - 3.40| < 0.2\}.$$

Hay puntos en R que están lejos del punto solución $(0.8, 0.6)$ y, aún así, producen valores muy pequeños cuando se sustituyen en las ecuaciones (15).

Si se sospecha que un sistema lineal está mal condicionado, entonces hay que realizar las operaciones con una aritmética de precisión múltiple. Para obtener más información sobre este fenómeno, la persona interesada puede consultar los estudios sobre el número de condición de una matriz.

La mala condición puede tener consecuencias más drásticas cuando hay varias ecuaciones involucradas. Consideremos el problema de hallar el polinomio cúbico $y = c_1x^3 + c_2x^2 + c_3x + c_4$ que pasa por los cuatro puntos $(2, 8)$, $(3, 27)$, $(4, 64)$ y $(5, 125)$ (claramente, $y = x^3$ es el polinomio cúbico deseado). En el Capítulo 5 veremos cómo se puede resolver un problema de este tipo que, en este caso, nos llevaría a resolver el siguiente sistema lineal

$$\begin{bmatrix} 20514 & 4424 & 978 & 224 \\ 4424 & 978 & 224 & 54 \\ 978 & 224 & 54 & 14 \\ 224 & 54 & 14 & 4 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} 20514 \\ 4424 \\ 978 \\ 224 \end{bmatrix}.$$

Un computador que trabaja con una precisión de nueve cifras ofreció como solución:

$$c_1 = 1.000004, \quad c_2 = -0.000038, \quad c_3 = 0.000126, \quad \text{y} \quad c_4 = -0.000131.$$

Aunque este resultado está cerca de la solución correcta, $c_1 = 1$ y $c_2 = c_3 = c_4 = 0$, se observa lo fácil que es el que aparezcan errores en la solución. Es más, supongamos que el coeficiente $a_{11} = 20514$ de la esquina superior izquierda cambia su valor por 20515 y que resolvemos el sistema perturbado. El mismo computador nos da como solución:

$$c_1 = 0.642857, \quad c_2 = 3.75000, \quad c_3 = -12.3928, \quad \text{y} \quad c_4 = 12.7500,$$

que es una respuesta muy incorrecta. La mala condición no es fácilmente detectable. Un síntoma de mala condición es que, volviendo a resolver el sistema perturbando ligeramente los coeficientes, la nueva respuesta difiera sustancialmente de la anterior. El análisis de la sensibilidad de un sistema a las perturbaciones es un tema que se estudia normalmente en cursos de análisis numérico avanzado.

MATLAB

En el Programa 3.2 que daremos a continuación, la instrucción `[A B]` es la que permite construir la matriz ampliada del sistema lineal $\mathbf{AX} = \mathbf{B}$ y la instrucción `max` se usa para determinar el pivote en la estrategia de pivoteo parcial. Una vez que tenemos la matriz ampliada $[\mathbf{U}|\mathbf{Y}]$ del sistema triangular equivalente, ésta se separa en \mathbf{U} e \mathbf{Y} con las que podemos usar el Programa 3.1 para llevar a cabo el algoritmo de sustitución regresiva (`backsub(U, Y)`).

El siguiente ejemplo ilustra el uso de estas instrucciones y procedimientos.

Ejemplo 3.19. En (a) vamos a usar el paquete MATLAB para construir la matriz ampliada del sistema lineal del Ejemplo 3.16; luego, en (b), usaremos la instrucción `max` para hallar el elemento de mayor valor absoluto de la primera columna de la matriz de los coeficientes \mathbf{A} y finalmente, en (c), separaremos la matriz ampliada dada en (11) en la matriz de coeficientes \mathbf{U} y la matriz de términos independientes \mathbf{Y} del sistema triangular superior $\mathbf{U}\mathbf{X} = \mathbf{Y}$.

(a)

```
>> A=[1 2 1 4;2 0 4 3;4 2 2 1;-3 1 3 2];
>> B=[13 28 20 6]';
>> Aug=[A B]
Aug=
 1 2 1 4 13
 2 0 4 3 28
 4 2 2 1 20
 -3 1 3 2 6
```

(b) En la siguiente reproducción de una sesión de trabajo con el paquete MATLAB, a es el elemento de mayor valor absoluto de la primera columna de \mathbf{A} y j es el número de la fila donde dicho máximo se alcanza.

```
>> [a,j]=max{abs(A(1:4,1))}

a=
 4
j=
 3
```

(c) Sea $\text{Augup} = [\mathbf{U}|\mathbf{Y}]$ la matriz triangular superior que aparece en (11). Entonces

```
>> Augup=[1 2 1 4 13;0 -4 2 -5 2;0 0 -5 -7.5 -35;0 0 0 -9 -18];
>> U=Augup(1:4,1:4)
U=
 1.0000 2.0000 1.0000 4.0000
 0 -4.0000 2.0000 -5.0000
 0 0 -5.0000 -7.5000
 0 0 0 -9.0000
>> Y=Augup(1:4,5)
Y=
 13
 2
 -35
 -18
```

Programa 3.2 (Triangularización superior seguida de sustitución regresiva). Cálculo de la solución del sistema lineal $\mathbf{AX} = \mathbf{B}$ mediante la reducción a forma triangular superior de la matriz ampliada $[\mathbf{A}| \mathbf{B}]$ seguida de la sustitución regresiva.

```

function X = uptrbk(A,B)

% Datos
%     - A es una matriz invertible de orden N x N
%     - B es una matriz de orden N x 1
% Resultados
%     - X es una matriz de orden N x 1 que contiene
%       la solución de AX=B.

% Inicializamos X y una matriz C que sirve de almácen temporal
[N N]=size(A);
X=zeros(N,1);
C=zeros(1,N+1);

% Cálculo de la matriz ampliada Aug=[A|B]
Aug=[A B];

for q=1:N-1
    % Pivoteo parcial en la columna q-ésima
    [Y,j]=max(abs(Aug(q:N,q)));
    % Intercambiamos las filas q-ésima y (j+q-1)-ésima
    C=Aug(q,:);
    Aug(q,:)=Aug(j+q-1,:);
    Aug(j+q-1,:)=C;
    if Aug(q,q)==0
        'A es singular. No hay solución o no es única.'
        break
    end

    % Proceso de eliminación en la columna q-ésima
    for k=q+1:N
        m=Aug(k,q)/Aug(q,q);
        Aug(k,q:N+1)=Aug(k,q:N+1)-m*Aug(q,q:N+1);
    end
end

% Sustitución regresiva en [U|Y] usando el Programa 3.1
X=backsub(Aug(1:N,1:N),Aug(1:N,N+1));

```

Ejercicios

En los Ejercicios 1 a 4 pruebe que $\mathbf{AX} = \mathbf{B}$ es equivalente al sistema triangular superior $\mathbf{UX} = \mathbf{Y}$ que se da y halle la solución.

$$1. \quad \begin{array}{rcl} 2x_1 + 4x_2 - 6x_3 & = & -4 \\ x_1 + 5x_2 + 3x_3 & = & 10 \\ x_1 + 3x_2 + 2x_3 & = & 5 \end{array} \quad \begin{array}{rcl} 2x_1 + 4x_2 - 6x_3 & = & -4 \\ 3x_2 + 6x_3 & = & 12 \\ 3x_3 & = & 3 \end{array}$$

$$2. \quad \begin{array}{rcl} x_1 + x_2 + 6x_3 & = & 7 \\ -x_1 + 2x_2 + 9x_3 & = & 2 \\ x_1 - 2x_2 + 3x_3 & = & 10 \end{array} \quad \begin{array}{rcl} x_1 + x_2 + 6x_3 & = & 7 \\ 3x_2 + 15x_3 & = & 9 \\ 12x_3 & = & 12 \end{array}$$

$$3. \quad \begin{array}{rcl} 2x_1 - 2x_2 + 5x_3 & = & 6 \\ 2x_1 + 3x_2 + x_3 & = & 13 \\ -x_1 + 4x_2 - 4x_3 & = & 3 \end{array} \quad \begin{array}{rcl} 2x_1 - 2x_2 + 5x_3 & = & 6 \\ 5x_2 - 4x_3 & = & 7 \\ 0.9x_3 & = & 1.8 \end{array}$$

$$4. \quad \begin{array}{rcl} -5x_1 + 2x_2 - x_3 & = & -1 \\ x_1 + 0x_2 + 3x_3 & = & 5 \\ 3x_1 + x_2 + 6x_3 & = & 17 \end{array} \quad \begin{array}{rcl} -5x_1 + 2x_2 - x_3 & = & -1 \\ 0.4x_2 + 2.8x_3 & = & 4.8 \\ -10x_3 & = & -10 \end{array}$$

5. Halle la parábola $y = A + Bx + Cx^2$ que pasa por los puntos $(1, 4)$, $(2, 7)$ y $(3, 14)$.

6. Halle la parábola $y = A + Bx + Cx^2$ que pasa por los puntos $(1, 6)$, $(2, 5)$ y $(3, 2)$.

7. Halle la cúbica $y = A + Bx + Cx^2 + Dx^3$ que pasa por los puntos $(0, 0)$, $(1, 1)$, $(2, 2)$ y $(3, 2)$.

En los Ejercicios 8 a 10 pruebe que $\mathbf{AX} = \mathbf{B}$ es equivalente al sistema triangular superior $\mathbf{UX} = \mathbf{Y}$ que se da y halle la solución.

$$8. \quad \begin{array}{rcl} 4x_1 + 8x_2 + 4x_3 + 0x_4 & = & 8 \\ x_1 + 5x_2 + 4x_3 - 3x_4 & = & -4 \\ x_1 + 4x_2 + 7x_3 + 2x_4 & = & 10 \\ x_1 + 3x_2 + 0x_3 - 2x_4 & = & -4 \end{array} \quad \begin{array}{rcl} 4x_1 + 8x_2 + 4x_3 + 0x_4 & = & 8 \\ 3x_2 + 3x_3 - 3x_4 & = & -6 \\ 4x_3 + 4x_4 & = & 12 \\ x_4 & = & 2 \end{array}$$

$$9. \quad \begin{array}{rcl} 2x_1 + 4x_2 - 4x_3 + 0x_4 & = & 12 \\ x_1 + 5x_2 - 5x_3 - 3x_4 & = & 18 \\ 2x_1 + 3x_2 + x_3 + 3x_4 & = & 8 \\ x_1 + 4x_2 - 2x_3 + 2x_4 & = & 8 \end{array} \quad \begin{array}{rcl} 2x_1 + 4x_2 - 4x_3 + 0x_4 & = & 12 \\ 3x_2 - 3x_3 - 3x_4 & = & 12 \\ 4x_3 + 2x_4 & = & 0 \\ 3x_4 & = & -6 \end{array}$$

$$10. \quad \begin{array}{rcl} x_1 + 2x_2 + 0x_3 - x_4 & = & 9 \\ 2x_1 + 3x_2 - x_3 + 0x_4 & = & 9 \\ 0x_1 + 4x_2 + 2x_3 - 5x_4 & = & 26 \\ 5x_1 + 5x_2 + 2x_3 - 4x_4 & = & 32 \end{array} \quad \begin{array}{rcl} x_1 + 2x_2 + 0x_3 - x_4 & = & 9 \\ -x_2 - x_3 + 2x_4 & = & -9 \\ -2x_3 + 3x_4 & = & -10 \\ 1.5x_4 & = & -3 \end{array}$$

- 11.** Halle la solución del siguiente sistema lineal

$$\begin{aligned}x_1 + 2x_2 &= 7 \\2x_1 + 3x_2 - x_3 &= 9 \\4x_2 + 2x_3 + 3x_4 &= 10 \\2x_3 - 4x_4 &= 12\end{aligned}$$

- 12.** Halle la solución del siguiente sistema lineal

$$\begin{aligned}x_1 + x_2 &= 5 \\2x_1 - x_2 + 5x_3 &= -9 \\3x_2 - 4x_3 + 2x_4 &= 19 \\2x_3 + 6x_4 &= 2\end{aligned}$$

- 13.** Para decidir qué computador comprar, si el ENC 174 o el MGR 11, una compañía ha decidido evaluar la precisión con la que cada uno de estos modelos resuelve el sistema

$$\begin{aligned}34x + 55y - 21 &= 0 \\55x + 89y - 34 &= 0.\end{aligned}$$

El computador ENC 174 da como solución $x = -0.11$ e $y = 0.45$ y, para comprobar su exactitud se sustituye en el sistema y se obtiene

$$\begin{aligned}34(-0.11) + 55(0.45) - 21 &= 0.01 \\55(-0.11) + 89(0.45) - 34 &= 0.00.\end{aligned}$$

El computador MGR 11 da como solución $x = -0.99$ e $y = 1.01$ y, para comprobar su exactitud se sustituye en el sistema y se obtiene

$$\begin{aligned}34(-0.99) + 55(1.01) - 21 &= 0.89 \\55(-0.99) + 89(1.01) - 34 &= 1.44.\end{aligned}$$

¿Qué computador da mejor respuesta? ¿Por qué?

- 14.** Resuelva los siguientes sistemas lineales usando (i) el método de eliminación de Gauss con pivote parcial y (ii) el método de eliminación de Gauss con pivote parcial escalado.

(a) $2x_1 - 3x_2 + 100x_3 = 1$ $x_1 + 10x_2 - 0.001x_3 = 0$ $3x_1 - 100x_2 + 0.01x_3 = 0$	(b) $x_1 + 20x_2 - x_3 + 0.001x_4 = 0$ $2x_1 - 5x_2 + 30x_3 - 0.1x_4 = 1$ $5x_1 + x_2 - 100x_3 - 10x_4 = 0$ $2x_1 - 100x_2 - x_3 + x_4 = 0$
--------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------

15. La matriz de Hilbert es un ejemplo clásico de matriz mal condicionada: cambios pequeños en sus coeficientes provocan cambios grandes en la solución del sistema perturbado.

- (a) Calcule la solución exacta de $\mathbf{AX} = \mathbf{B}$ (deje todos los números como fracciones y utilice aritmética exacta) siendo la matriz de los coeficientes la matriz de Hilbert de orden 4×4 dada por:

$$\mathbf{A} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{bmatrix} \quad \text{y} \quad \mathbf{B} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

- (b) Ahora resuelva $\mathbf{AX} = \mathbf{B}$ usando aritmética en coma flotante con una precisión de cuatro cifras decimales y redondeo:

$$\mathbf{A} = \begin{bmatrix} 1.0000 & 0.5000 & 0.3333 & 0.2500 \\ 0.5000 & 0.3333 & 0.2500 & 0.2000 \\ 0.3333 & 0.2500 & 0.2000 & 0.1667 \\ 0.2500 & 0.2000 & 0.1667 & 0.1429 \end{bmatrix} \quad \text{y} \quad \mathbf{B} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Nota. La matriz de coeficientes del apartado (b) es la correspondiente aproximación a la matriz de coeficientes del apartado (a).

Algoritmos y programas

1. En muchas aplicaciones nos encontramos con matrices que tienen muchos ceros. Son particularmente importantes los *sistemas tridiagonales* (véanse los Ejercicios 11 y 12), que son los de la forma

$$d_1x_1 + c_1x_2 = b_1$$

$$a_1x_1 + d_2x_2 + c_2x_3 = b_2$$

$$a_2x_2 + d_3x_3 + c_3x_4 = b_3$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$a_{N-2}x_{N-2} + d_{N-1}x_{N-1} + c_{N-1}x_N = b_{N-1}$$

$$a_{N-1}x_{N-1} + d_Nx_N = b_N.$$

Construya un programa que resuelva un sistema tridiagonal. Puede suponer que no hace falta intercambiar filas, de manera que la fila k -ésima puede usarse para eliminar la incógnita x_k en la fila siguiente.

2. Use el Programa 3.2 para hallar el polinomio de grado seis $y = a_1 + a_2x + a_3x^2 + a_4x^3 + a_5x^4 + a_6x^5 + a_7x^6$ que pasa por los puntos $(0, 1), (1, 3), (2, 2), (3, 1), (4, 3), (5, 2)$ y $(6, 1)$. Use la instrucción `plot` para dibujar el polinomio obtenido y los puntos dados sobre la misma gráfica. Explique las discrepancias que puedan aparecer en su dibujo.
3. Use el Programa 3.2 para resolver el sistema lineal $\mathbf{AX} = \mathbf{B}$, siendo $\mathbf{A} = [a_{ij}]_{N \times N}$ con $a_{ij} = i^{j-1}$ y $\mathbf{B} = [b_{ij}]_{N \times 1}$ con $b_{11} = N$ y $b_{i1} = i^{N-2}/(i-1)$ para $i \geq 2$, en los casos $N = 3, 7$ y 11 . Sabiendo que la solución exacta es $\mathbf{X} = [1 \ 1 \ \dots \ 1 \ 1]'$, explique las desviaciones de la solución calculada.
4. Construya un programa que cambie la estrategia de pivoteo parcial del Programa 3.2 por la estrategia de pivoteo parcial escalado.
5. Use su programa con la estrategia de pivoteo parcial escalado del Problema 4 para resolver el sistema dado en el Problema 3 con $N = 11$. Explique las mejoras que obtenga en las soluciones.
6. Modifique el Programa 3.2 de manera que resuelva eficientemente M sistemas lineales

$$\mathbf{AX}_1 = \mathbf{B}_1, \quad \mathbf{AX}_2 = \mathbf{B}_2, \quad \dots \quad \text{y} \quad \mathbf{AX}_M = \mathbf{B}_M.$$

que tengan la misma matriz de coeficientes \mathbf{A} pero diferentes matrices \mathbf{B} .

7. La discusión que sigue se plantea con una matriz \mathbf{A} de orden 3×3 , pero se puede aplicar a matrices de orden $N \times N$. Si \mathbf{A} es invertible, entonces \mathbf{A}^{-1} existe y cumple $\mathbf{AA}^{-1} = \mathbf{I}$. Sean $\mathbf{C}_1, \mathbf{C}_2$ y \mathbf{C}_3 las columnas de \mathbf{A}^{-1} y sean $\mathbf{E}_1, \mathbf{E}_2$ y \mathbf{E}_3 las columnas de \mathbf{I} , entonces la relación $\mathbf{AA}^{-1} = \mathbf{I}$ puede representarse como

$$\mathbf{A} [\mathbf{C}_1 \ \mathbf{C}_2 \ \mathbf{C}_3] = [\mathbf{E}_1 \ \mathbf{E}_2 \ \mathbf{E}_3].$$

Este producto matricial es equivalente a los tres sistemas lineales

$$\mathbf{AC}_1 = \mathbf{E}_1, \quad \mathbf{AC}_2 = \mathbf{E}_2 \quad \text{y} \quad \mathbf{AC}_3 = \mathbf{E}_3,$$

de manera que hallar \mathbf{A}^{-1} es equivalente a resolver estos tres sistemas.

Usando el Programa 3.2 o su programa del Problema 6, halle la inversa de cada una de las siguientes matrices y compruebe su respuesta calculando los correspondientes productos \mathbf{AA}^{-1} así como usando la instrucción `inv(A)`. Explique las diferencias que aparezcan.

$$(a) \begin{bmatrix} 2 & 0 & 1 \\ 3 & 2 & 5 \\ 1 & -1 & 0 \end{bmatrix} \qquad (b) \begin{bmatrix} 16 & -120 & 240 & -140 \\ -120 & 1200 & -2700 & 1680 \\ 240 & -2700 & 6480 & -4200 \\ -140 & 1680 & -4200 & 2800 \end{bmatrix}$$

3.5 Factorización triangular

En la Sección 3.3 vimos lo fácil que resulta resolver un sistema triangular superior. Ahora introduciremos el concepto de factorización triangular de una matriz: la posibilidad de escribir una matriz dada A como el producto de una matriz triangular inferior L , cuyos elementos en la diagonal principal son todos iguales a 1, por una matriz triangular superior U , cuyos elementos diagonales son distintos de cero. Para facilitar la notación, ilustraremos los conceptos con matrices de orden 4×4 , pero se aplican a matrices de orden arbitrario $N \times N$.

Definición 3.4. Diremos que una matriz invertible A admite una **factorización triangular** o **factorización LU** si puede expresarse como el producto de una matriz triangular inferior L , cuyos elementos diagonales son todos iguales a 1, por una matriz triangular superior U :

$$(1) \quad A = LU,$$

o, escrito de manera desarrollada,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ m_{21} & 1 & 0 & 0 \\ m_{31} & m_{32} & 1 & 0 \\ m_{41} & m_{42} & m_{43} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}. \quad \blacktriangle$$

La condición de que A sea invertible implica que $u_{kk} \neq 0$ para todo k . La notación para los elementos de L es m_{ij} ; la razón para elegir m_{ij} en vez de l_{ij} la daremos enseguida.

Solución de un sistema lineal

Supongamos que la matriz de los coeficientes A de un sistema lineal $AX = B$ admite una factorización triangular como la de (1), entonces la solución de

$$(2) \quad LU\mathbf{X} = \mathbf{B}$$

puede obtenerse definiendo $\mathbf{Y} = U\mathbf{X}$ y resolviendo dos sistemas lineales:

$$(3) \quad \text{primero se halla } \mathbf{Y} \text{ en } LY = \mathbf{B} \quad \text{y luego } \mathbf{X} \text{ en } UX = \mathbf{Y}.$$

En forma desarrollada, primero debemos resolver el sistema triangular inferior

$$(4) \quad \begin{array}{rcl} y_1 & & = b_1 \\ m_{21}y_1 + y_2 & & = b_2 \\ m_{31}y_1 + m_{32}y_2 + y_3 & & = b_3 \\ m_{41}y_1 + m_{42}y_2 + m_{43}y_3 + y_4 & & = b_4 \end{array}$$

para obtener y_1 , y_2 , y_3 e y_4 y, una vez que los tenemos, resolver el sistema triangular superior

$$(5) \quad \begin{aligned} u_{11}x_1 + u_{12}x_2 + u_{13}x_3 + u_{14}x_4 &= y_1 \\ u_{22}x_2 + u_{23}x_3 + u_{24}x_4 &= y_2 \\ u_{33}x_3 + u_{34}x_4 &= y_3 \\ u_{44}x_4 &= y_4. \end{aligned}$$

Ejemplo 3.20. Vamos a resolver

$$\begin{aligned} x_1 + 2x_2 + 4x_3 + x_4 &= 21 \\ 2x_1 + 8x_2 + 6x_3 + 4x_4 &= 52 \\ 3x_1 + 10x_2 + 8x_3 + 8x_4 &= 79 \\ 4x_1 + 12x_2 + 10x_3 + 6x_4 &= 82, \end{aligned}$$

usando el método descrito antes y sabiendo que la matriz de los coeficientes admite la factorización triangular

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 4 & 1 \\ 2 & 8 & 6 & 4 \\ 3 & 10 & 8 & 8 \\ 4 & 12 & 10 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 1 & 1 & 0 \\ 4 & 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 & 1 \\ 0 & 4 & -2 & 2 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & 0 & -6 \end{bmatrix} = \mathbf{L}\mathbf{U}.$$

Usando el algoritmo de sustitución progresiva resolvemos $\mathbf{LY} = \mathbf{B}$:

$$(6) \quad \begin{aligned} y_1 &= 21 \\ 2y_1 + y_2 &= 52 \\ 3y_1 + y_2 + y_3 &= 79 \\ 4y_1 + y_2 + 2y_3 + y_4 &= 82, \end{aligned}$$

obteniendo $y_1 = 21$, $y_2 = 52 - 2(21) = 10$, $y_3 = 79 - 3(21) - 10 = 6$ e $y_4 = 82 - 4(21) - 10 - 2(6) = -24$, o sea, $\mathbf{Y} = [21 \ 10 \ 6 \ -24]'$. Ahora escribimos el sistema $\mathbf{UX} = \mathbf{Y}$:

$$(7) \quad \begin{aligned} x_1 + 2x_2 + 4x_3 + x_4 &= 21 \\ 4x_2 - 2x_3 + 2x_4 &= 10 \\ -2x_3 + 3x_4 &= 6 \\ -6x_4 &= -24 \end{aligned}$$

y, con el método de sustitución regresiva, calculamos la solución $x_4 = -24/(-6) = 4$, $x_3 = (6 - 3(4))/(-2) = 3$, $x_2 = (10 - 2(4) + 2(3))/4 = 2$ y $x_1 = 21 - 4 - 4(3) - 2(2) = 1$, o sea $\mathbf{X} = [1 \ 2 \ 3 \ 4]'$.

Factorización triangular

Ahora discutiremos la manera de obtener factorizaciones triangulares. Si no hace falta realizar intercambios de filas cuando usamos eliminación gaussiana, entonces los multiplicadores m_{ij} son los elementos subdiagonales de \mathbf{L} .

Ejemplo 3.21. Vamos a usar el método de eliminación de Gauss para construir la factorización triangular de la matriz

$$\mathbf{A} = \begin{bmatrix} 4 & 3 & -1 \\ -2 & -4 & 5 \\ 1 & 2 & 6 \end{bmatrix}.$$

La matriz \mathbf{L} se construye sobre una matriz identidad que ponemos a la izquierda: cada vez que realicemos una operación con las filas en la construcción de la matriz triangular superior, colocamos el multiplicador m_{ij} utilizado en la posición que le corresponde en la matriz de la izquierda. Empezamos con

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & -1 \\ -2 & -4 & 5 \\ 1 & 2 & 6 \end{bmatrix}.$$

Usando la primera fila para eliminar los elementos de la primera columna de \mathbf{A} que están por debajo del elemento diagonal a_{11} , los multiplicadores son, respectivamente, $m_{21} = -0.5$ y $m_{31} = 0.25$. Estos multiplicadores se ponen en la matriz de la izquierda y el resultado es:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.25 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & -1 \\ 0 & -2.5 & 4.5 \\ 0 & 1.25 & 6.25 \end{bmatrix}.$$

Ahora usamos la segunda fila para eliminar el elemento de la segunda columna de la matriz de la derecha que está debajo del elemento diagonal. El multiplicador es $m_{32} = -0.5$ que, al ponerlo en su posición en la matriz de la izquierda, nos da la factorización triangular deseada:

$$(8) \quad \mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.25 & -0.5 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & -1 \\ 0 & -2.5 & 4.5 \\ 0 & 0 & 8.5 \end{bmatrix}.$$

Teorema 3.10 (Factorización directa $\mathbf{A} = \mathbf{LU}$ sin intercambios de filas). Supongamos que podemos llevar a cabo hasta el final el proceso de eliminación gaussiana, sin intercambios de filas, para resolver un sistema de ecuaciones lineales cualquiera $\mathbf{AX} = \mathbf{B}$. Entonces la matriz \mathbf{A} puede factorizarse como el producto de una matriz triangular inferior \mathbf{L} por una matriz triangular superior \mathbf{U} ; es decir, $\mathbf{A} = \mathbf{LU}$.

Es más, \mathbf{L} puede ser construida de manera que sus elementos diagonales son todos iguales a 1 y \mathbf{U} tiene todos sus elementos diagonales distintos de cero. Una vez halladas \mathbf{L} y \mathbf{U} , la solución \mathbf{X} puede calcularse en dos pasos:

1. Hallar \mathbf{Y} resolviendo $\mathbf{LY} = \mathbf{B}$ con el método de sustitución progresiva.
2. Hallar \mathbf{X} resolviendo $\mathbf{UX} = \mathbf{Y}$ con el método de sustitución regresiva.

Demostración. Probaremos que, cuando el proceso de eliminación gaussiana puede llevarse a cabo hasta el final en la matriz ampliada que se obtiene añadiendo la columna \mathbf{B} a la matriz \mathbf{A} , entonces los elementos subdiagonales de \mathbf{L} coinciden con los multiplicadores correspondientes que se usan en la eliminación, la matriz triangular superior \mathbf{U} es la matriz de los coeficientes del sistema triangular superior $\mathbf{UX} = \mathbf{Y}$ obtenido al final del proceso de eliminación y el segundo miembro de este sistema es, precisamente, el vector \mathbf{Y} solución del sistema que hay que resolver en el paso 1 mencionado al final del enunciado. Las matrices \mathbf{L} , \mathbf{U} , \mathbf{B} e \mathbf{Y} serán, entonces,

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ m_{21} & 1 & 0 & \cdots & 0 \\ m_{31} & m_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ m_{N1} & m_{N2} & m_{N3} & \cdots & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} a_{1N+1}^{(1)} \\ a_{2N+1}^{(1)} \\ a_{3N+1}^{(1)} \\ \vdots \\ a_{NN+1}^{(1)} \end{bmatrix}$$

$$\mathbf{U} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & a_{NN}^{(N)} \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} a_{1N+1}^{(1)} \\ a_{2N+1}^{(2)} \\ a_{3N+1}^{(3)} \\ \vdots \\ a_{NN+1}^{(N)} \end{bmatrix}.$$

Observación. Si sólo queremos calcular las matrices \mathbf{L} y \mathbf{U} de la factorización, entonces la columna $(N + 1)$ -ésima no es necesaria.

Paso 1. Almacenamos los coeficientes en la matriz ampliada; el superíndice (1) de $a_{rc}^{(1)}$ señala que ésta es la primera vez que se almacena un número en la

posición (r, c) .

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2N}^{(1)} & a_{2N+1}^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3N}^{(1)} & a_{3N+1}^{(1)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ a_{N1}^{(1)} & a_{N2}^{(1)} & a_{N3}^{(1)} & \cdots & a_{NN}^{(1)} & a_{NN+1}^{(1)} \end{array} \right].$$

Paso 2. Eliminamos la incógnita x_1 en todas las filas desde la segunda hasta la última y, en la posición $(r, 1)$ de la matriz, almacenamos el multiplicador m_{r1} usado para eliminar x_1 en la fila r -ésima. Describiremos este paso como si fueran instrucciones del paquete MATLAB:

```
for r = 2 : N
     $m_{r1} = a_{r1}^{(1)} / a_{11}^{(1)}$ ;
     $a_{r1} = m_{r1}$ ;
    for c = 2 : N + 1
         $a_{rc}^{(2)} = a_{rc}^{(1)} - m_{r1} * a_{1c}^{(1)}$ ;
    end
end
```

El superíndice (2) de los elementos nuevos $a_{rc}^{(2)}$ señala que ésta es la segunda vez que se almacena un número en la posición (r, c) de la matriz. Después del paso 2, la matriz queda:

$$\left[\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ m_{21} & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2N+1}^{(2)} \\ m_{31} & a_{32}^{(2)} & a_{33}^{(2)} & \cdots & a_{3N}^{(2)} & a_{3N+1}^{(2)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ m_{N1} & a_{N2}^{(2)} & a_{N3}^{(2)} & \cdots & a_{NN}^{(2)} & a_{NN+1}^{(2)} \end{array} \right].$$

Paso 3. Eliminamos la incógnita x_2 en todas las filas desde la tercera hasta la última y, en la posición $(r, 2)$ de la matriz, almacenamos el multiplicador m_{r2} usado para eliminar x_2 en la fila r -ésima:

```
for r = 3 : N
     $m_{r2} = a_{r2}^{(2)} / a_{22}^{(2)}$ ;
     $a_{r2} = m_{r2}$ ;
```

```

        for c = 3 : N + 1
             $a_{rc}^{(3)} = a_{rc}^{(2)} - m_{r2} * a_{2c}^{(2)}$ ;
        end
    end

```

El superíndice (3) de los elementos nuevos $a_{rc}^{(3)}$ señala que ésta es la tercera vez que se almacena un número en la posición (r, c) de la matriz. Después del paso 3, la matriz queda:

$$\left[\begin{array}{cccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ m_{21} & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2N+1}^{(2)} \\ m_{31} & m_{32} & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} & a_{3N+1}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ m_{N1} & m_{N2} & a_{N3}^{(3)} & \cdots & a_{NN}^{(3)} & a_{NN+1}^{(3)} \end{array} \right].$$

Paso q + 1. Este es el paso general: Eliminamos la incógnita x_q en todas las filas desde la q -ésima hasta la última y, en la posición (r, q) de la matriz, almacenamos el multiplicador m_{rq} usado para eliminar x_q en la fila r -ésima.

```

        for r = q + 1 : N
             $m_{rq} = a_{rq}^{(q)} / a_{qq}^{(q)}$ ;
             $a_{rq} = m_{rq};$ 
            for c = q + 1 : N + 1
                 $a_{rc}^{(q+1)} = a_{rc}^{(q)} - m_{rq} * a_{qc}^{(q)}$ ;
            end
        end

```

El resultado final, tras haber eliminado x_{N-1} de la última fila es

$$\left[\begin{array}{cccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1N}^{(1)} & a_{1N+1}^{(1)} \\ m_{21} & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2N}^{(2)} & a_{2N+1}^{(2)} \\ m_{31} & m_{32} & a_{33}^{(3)} & \cdots & a_{3N}^{(3)} & a_{3N+1}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ m_{N1} & m_{N2} & m_{N3} & \cdots & a_{NN}^{(N)} & a_{NN+1}^{(N)} \end{array} \right].$$

El proceso de triangularización ya está completo. Hagamos notar que sólo hemos necesitado una matriz para almacenar todos los elementos de \mathbf{L} y \mathbf{U} : no se guardan los unos de la diagonal de \mathbf{L} ni los ceros que hay en \mathbf{L} y \mathbf{U} por encima

y por debajo de la diagonal principal, respectivamente; ¡sólo se almacenan los coeficientes esenciales para reconstruir \mathbf{L} y \mathbf{U} !

Debemos comprobar que $\mathbf{LU} = \mathbf{A}$. Sea $\mathbf{D} = \mathbf{LU}$ y consideremos el caso en que $r \leq c$, entonces d_{rc} es

$$(9) \quad d_{rc} = m_{r1}a_{1c}^{(1)} + m_{r2}a_{2c}^{(2)} + \cdots + m_{rr-1}a_{r-1c}^{(r-1)} + a_{rc}^{(r)}.$$

Las ecuaciones de reemplazamiento de los pasos 1 hasta $q+1=r$ han sido:

$$(10) \quad \begin{aligned} m_{r1}a_{1c}^{(1)} &= a_{rc}^{(1)} - a_{rc}^{(2)}, \\ m_{r2}a_{2c}^{(2)} &= a_{rc}^{(2)} - a_{rc}^{(3)}, \\ &\vdots \\ m_{rr-1}a_{r-1c}^{(r-1)} &= a_{rc}^{(r-1)} - a_{rc}^{(r)}. \end{aligned}$$

Sustituyendo los reemplazamientos de (10) en la relación (9), obtenemos:

$$d_{rc} = a_{rc}^{(1)} - a_{rc}^{(2)} + a_{rc}^{(2)} - a_{rc}^{(3)} + \cdots + a_{rc}^{(r-1)} - a_{rc}^{(r)} + a_{rc}^{(r)} = a_{rc}^{(1)}.$$

El caso restante, cuando $r > c$, se prueba de manera parecida.

La equivalencia de los sistemas $\mathbf{AX} = \mathbf{B}$ y $\mathbf{UX} = \mathbf{Y}$ y la factorización $\mathbf{A} = \mathbf{UL}$ prueban que la última columna de la matriz ampliada del final del proceso es la solución \mathbf{Y} del sistema $\mathbf{LY} = \mathbf{B}$. En otras palabras, el efecto del proceso de eliminación gaussiana en la última columna de la matriz ampliada es, precisamente, el algoritmo de sustitución progresiva del sistema $\mathbf{LY} = \mathbf{B}$. •

Complejidad computacional

El proceso de triangularización es el mismo para la eliminación gaussiana que para el método de factorización triangular. Contaremos el número de operaciones necesarias para realizar el proceso descrito en el Teorema 3.10, fijándonos sólo en las N primeras columnas de la matriz ampliada. El bucle exterior del paso $p+1$ conlleva $N-q = N-(q+1)+1$ divisiones para calcular los multiplicadores m_{rq} ; en el bucle interior, pero sólo para las primeras N columnas, el cálculo de los nuevos elementos $a_{rc}^{(q+1)}$ conlleva un total de $(N-q)(N-q)$ multiplicaciones y otras tantas substracciones. Puesto que este proceso se realiza para $q = 1, 2, \dots, N-1$, tenemos que la factorización triangular $\mathbf{A} = \mathbf{LU}$ conlleva

$$(11) \quad \sum_{q=1}^{N-1} (N-q)(N-q+1) = \frac{N^3 - N}{3} \quad \text{multiplicaciones y divisiones}$$

y

$$(12) \quad \sum_{q=1}^{N-1} (N - pq)(N - q) = \frac{2N^3 - 3N^2 + N}{6} \quad \text{substracciones.}$$

Para establecer (11) usamos las siguientes fórmulas de sumación:

$$\sum_{k=1}^M k = \frac{M(M+1)}{2} \quad \text{y} \quad \sum_{k=1}^M k^2 = \frac{M(M+1)(2M+1)}{6}.$$

Haciendo el cambio de variables $k = N - q$, podemos escribir (11) como

$$\begin{aligned} \sum_{q=1}^{N-1} (N - q)(N - q + 1) &= \sum_{q=1}^{N-1} (N - q) + \sum_{q=1}^{N-1} (N - q)^2 \\ &= \sum_{k=1}^{N-1} k + \sum_{k=1}^{N-1} k^2 \\ &= \frac{(N-1)N}{2} + \frac{(N-1)(N)(2N-1)}{6} \\ &= \frac{N^3 - N}{3}. \end{aligned}$$

Una prueba similar permite establecer (12).

Una vez obtenida la factorización triangular $\mathbf{A} = \mathbf{LU}$, la solución del sistema triangular inferior $\mathbf{LY} = \mathbf{B}$ conlleva $0 + 1 + \dots + N - 1 = (N^2 - N)/2$ multiplicaciones y substracciones; no hacen falta divisiones porque los elementos diagonales de \mathbf{L} son todos iguales a 1. Finalmente, la solución del sistema triangular superior $\mathbf{UX} = \mathbf{Y}$ conlleva $1 + 2 + \dots + N = (N^2 + N)/2$ multiplicaciones y divisiones y $(N^2 - N)/2$ substracciones. En consecuencia, el cálculo de la solución de $\mathbf{LUX} = \mathbf{B}$, una vez que se tienen \mathbf{L} y \mathbf{U} conlleva

N^2 multiplicaciones y divisiones y $N^2 - N$ substracciones.

Vemos, entonces, que la mayor parte del coste computacional recae en el cálculo de la factorización triangular. Si debemos resolver varios sistemas que tienen la misma matriz de coeficientes \mathbf{A} pero diferentes columnas \mathbf{B} de términos independientes, no es necesario hacer la factorización cada vez; basta con hacerla la primera vez y almacenar los factores. Esta es la razón por la que se suele elegir el método de la factorización triangular antes que el método de eliminación de Gauss; sin embargo, si sólo hay que resolver un sistema de ecuaciones, entonces los dos métodos son iguales, salvo que en la factorización triangular se guardan los multiplicadores.

Matrices de permutación

Para llevar a cabo el proceso de factorización $\mathbf{A} = \mathbf{LU}$ descrito en el Teorema 3.10 hemos supuesto que no se hacen intercambios de filas. Puede ocurrir que una matriz invertible \mathbf{A} no admita factorización $\mathbf{A} = \mathbf{LU}$.

Ejemplo 3.22. Vamos a probar que la siguiente matriz \mathbf{A} no admite factorización $\mathbf{A} = \mathbf{LU}$,

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 6 \\ 4 & 8 & -1 \\ -2 & 3 & 5 \end{bmatrix}.$$

Supongamos que \mathbf{A} sí admite factorización $\mathbf{A} = \mathbf{LU}$, o sea,

$$(13) \quad \begin{bmatrix} 1 & 2 & 6 \\ 4 & 8 & -1 \\ -2 & 3 & 5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}.$$

Si hacemos el producto de las matrices \mathbf{L} y \mathbf{U} del miembro derecho de (13) y comparamos los elementos del producto con los elementos correspondientes de \mathbf{A} , nos queda lo siguiente: En la primera columna, $1 = 1u_{11}$, después $4 = m_{21}u_{11} = m_{21}$ y, finalmente, $-2 = m_{31}u_{11} = m_{31}$. En la segunda columna, $2 = 1u_{12}$, luego $8 = m_{21}u_{12} = (4)(2) + u_{22}$, lo cual implica que $u_{22} = 0$ y, finalmente, $3 = m_{31}u_{12} + m_{32}u_{22} = (-2)(2) + m_{32}(0) = -4$, que es una contradicción. En consecuencia, \mathbf{A} no admite factorización triangular. ■

Una permutación de los N primeros números naturales $1, 2, \dots, N$ es un cambio de orden k_1, k_2, \dots, k_N de éstos. Por ejemplo, $1, 4, 2, 3, 5$ es una permutación de $1, 2, 3, 4, 5$. En la siguiente definición usaremos también la base canónica del espacio \mathbb{R}^N formada por los vectores $\mathbf{E}_i = [0 \ 0 \ \cdots \ 0 \ 1_i \ 0 \ \cdots \ 0]$, donde el subíndice i señala la posición, para $i = 1, 2, \dots, N$.

Definición 3.5. Una *matriz de permutación* \mathbf{P} es una matriz de orden $N \times N$ tal que en cada fila y en cada columna sólo tiene un elemento igual a 1 siendo todos los demás iguales a cero. Las filas de \mathbf{P} son, entonces, una permutación de las filas de la matriz identidad y \mathbf{P} puede escribirse como

$$(14) \quad \mathbf{P} = [\mathbf{E}'_{k_1} \ \mathbf{E}'_{k_2} \ \dots \ \mathbf{E}'_{k_N}]',$$

siendo k_1, k_2, \dots, k_N la permutación de $1, 2, \dots, N$, en cuyo caso, los elementos de $\mathbf{P} = [p_{ij}]$ son de la forma

$$p_{ij} = \begin{cases} 1 & \text{si } j = k_i, \\ 0 & \text{en otro caso.} \end{cases}$$

Por ejemplo, la siguiente matriz de orden 4×4 es una matriz de permutación,

$$(15) \quad \mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} = [\mathbf{E}'_2 \quad \mathbf{E}'_1 \quad \mathbf{E}'_4 \quad \mathbf{E}'_3]' . \quad \blacktriangle$$

Teorema 3.11. Supongamos que $\mathbf{P} = [\mathbf{E}'_{k_1} \quad \mathbf{E}'_{k_2} \quad \dots \quad \mathbf{E}'_{k_N}]'$ es una matriz de permutación. Entonces \mathbf{PA} es la matriz que se obtiene permutando las filas de \mathbf{A} en el mismo orden: fila _{k_1} \mathbf{A} , fila _{k_2} \mathbf{A} , ..., fila _{k_N} \mathbf{A} .

Ejemplo 3.23. Sea \mathbf{A} una matriz de orden 4×4 y sea \mathbf{P} la matriz de permutación dada en (15), entonces \mathbf{PA} es la matriz que se obtiene permutando las filas de \mathbf{A} en el mismo orden: fila₂ \mathbf{A} , fila₁ \mathbf{A} , fila₄ \mathbf{A} , fila₃ \mathbf{A} .

Calculando el producto tenemos

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} a_{21} & a_{22} & a_{23} & a_{24} \\ a_{11} & a_{12} & a_{13} & a_{14} \\ a_{41} & a_{42} & a_{43} & a_{44} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix} .$$

Teorema 3.12. Si \mathbf{P} es una matriz de permutación, entonces es invertible y se tiene que $\mathbf{P}^{-1} = \mathbf{P}'$.

Teorema 3.13. Si \mathbf{A} es una matriz invertible, entonces existe una matriz de permutación \mathbf{P} tal que \mathbf{PA} admite una factorización triangular

$$(16) \quad \mathbf{PA} = \mathbf{LU}.$$

Las demostraciones de estos teorema pueden hallarse en los libros de álgebra lineal de carácter avanzado.

Ejemplo 3.24. Si intercambiamos las filas segunda y tercera de la matriz del Ejemplo 3.22, entonces la matriz resultante \mathbf{PA} sí admite factorización triangular.

La matriz de permutación \mathbf{P} que intercambia las filas segunda y tercera es $\mathbf{P} = [\mathbf{E}'_1 \quad \mathbf{E}'_3 \quad \mathbf{E}'_2]'$. Calculando el producto \mathbf{PA} , obtenemos

$$\mathbf{PA} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 6 \\ 4 & 8 & -1 \\ -2 & 3 & 5 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 6 \\ -2 & 3 & 5 \\ 4 & 8 & -1 \end{bmatrix} .$$

Ahora podemos usar el proceso de eliminación gaussiana sin intercambio de filas:

$$\begin{aligned} \text{pivote} &\rightarrow \begin{bmatrix} 1 & 2 & 6 \\ -2 & 3 & 5 \\ 4 & 8 & -1 \end{bmatrix} \\ m_{21} &= -2 \begin{bmatrix} 1 & 2 & 6 \\ -2 & 3 & 5 \\ 4 & 8 & -1 \end{bmatrix} \\ m_{31} &= 4 \begin{bmatrix} 1 & 2 & 6 \\ -2 & 3 & 5 \\ 4 & 8 & -1 \end{bmatrix} \end{aligned}$$

Al final de este paso nos encontramos con que x_2 ya está eliminado de la tercera ecuación, así que no hay que hacer el siguiente (el multiplicador m_{32} es cero)

$$\text{pivote } \rightarrow \begin{bmatrix} 1 & 2 & 6 \\ 0 & -7 & 17 \\ 0 & 0 & -25 \end{bmatrix} = \mathbf{U}.$$

■

Extensión del proceso de eliminación gaussiana

El siguiente teorema es una extensión del Teorema 3.10 que incluye el caso en que es necesario, o conveniente, realizar intercambios de filas. El teorema nos dice, en particular, que podemos usar el método de factorización triangular para resolver un sistema lineal cualquiera $\mathbf{AX} = \mathbf{B}$ en el que \mathbf{A} sea invertible.

Teorema 3.14 (Factorización indirecta: $\mathbf{PA} = \mathbf{LU}$). Sea \mathbf{A} una matriz de orden $N \times N$. Supongamos que el proceso de eliminación gaussiana puede llevarse a cabo hasta el final para resolver un sistema cualquiera $\mathbf{AX} = \mathbf{B}$, pero que hemos realizado intercambios de filas. Entonces existe una matriz de permutación \mathbf{P} (la que recoge todos los intercambios de filas realizados) tal que el producto \mathbf{PA} puede factorizarse como el producto de una matriz triangular inferior \mathbf{L} por una matriz triangular superior \mathbf{U} :

$$\mathbf{PA} = \mathbf{LU}.$$

Es más, \mathbf{L} puede ser construida de manera que sus elementos diagonales son todos iguales a 1 y \mathbf{U} tiene elementos diagonales distintos de cero. La solución \mathbf{X} puede hallarse en cuatro pasos:

1. Construir las matrices \mathbf{L} , \mathbf{U} y \mathbf{P} .
2. Calcular el vector columna \mathbf{PB} .
3. Hallar \mathbf{Y} resolviendo $\mathbf{LY} = \mathbf{PB}$ con el algoritmo de sustitución progresiva.
4. Hallar \mathbf{X} resolviendo $\mathbf{UX} = \mathbf{Y}$ con el algoritmo de sustitución regresiva.

Observación. Supongamos que debemos resolver $\mathbf{AX} = \mathbf{B}$ para una matriz fija \mathbf{A} y varias matrices-columna \mathbf{B} . Entonces el primer paso sólo se lleva a cabo una vez y los pasos segundo a cuarto se usan para hallar la solución \mathbf{X} que corresponde a cada \mathbf{B} . Una vez dado el primer paso, los pasos dos a cuatro constituyen un método muy eficaz desde el punto de vista computacional porque sólo requieren del orden de $\mathcal{O}(N^2)$ operaciones, en vez de las $\mathcal{O}(N^3)$ operaciones que se requieren en el proceso de eliminación gaussiana del sistema completo.

MATLAB

La instrucción `[L,U,P]=lu(A)` del paquete MATLAB proporciona, usando el método de eliminación de Gauss con pivote parcial, la matriz triangular inferior \mathbf{L} , la matriz triangular superior \mathbf{U} y la matriz de permutación \mathbf{P} de la factorización $\mathbf{PA} = \mathbf{LU}$ dada en el Teorema 3.14.

Ejemplo 3.25. Vamos a aplicar la instrucción $[L, U, P] = \text{lu}(A)$ a la matriz A del Ejemplo 3.22; comprobaremos que $A = P^{-1}LU$ (lo que es equivalente a probar que $PA = LU$).

```
>>A=[1 2 6 ;4 8 -1;-2 3 -5];
>>[L,U,P]=lu(A)
L=
    1.0000   0         0
   -0.5000   1.0000   0
    0.2500   0         1.0000
U=
    4.0000   8.0000  -1.0000
     0        7.0000  4.5000
     0        0         6.2500
P=
    0  1  0
    0  0  1
    1  0  0
>>inv(P)*L*U
  1   2   6
  4   8  -1
 -2   3   5
```

Como indicamos antes, se suele preferir el método de factorización triangular indirecta antes que el método de eliminación. La factorización triangular indirecta también se usa en las instrucciones `inv(A)` y `det(A)` del paquete MATLAB. Por ejemplo, sabemos que el determinante de una matriz no singular A es igual a $(-1)^p \det(U)$, siendo U la matriz triangular superior de la factorización triangular indirecta y p el número de intercambios de filas necesarios que hay que realizar en la matriz identidad I para obtener P . Puesto que U es una matriz triangular, el determinante de U es simplemente el producto de los elementos de su diagonal principal (Teorema 3.6). Dejamos como ejercicio la verificación de que $\det(A) = 175 = (-1)^2(175) = (-1)^2 \det(U)$ en el Ejemplo 3.25.

En el siguiente programa se construye el proceso descrito en la prueba del Teorema 3.10 pero incluyendo la estrategia de pivoteo parcial, por lo que constituye una extensión del Programa 3.2. El intercambio de filas debido al pivoteo parcial se almacena en la matriz R que luego se usa en la sustitución progresiva para hallar la matriz Y .

Programa 3.3 ($PA = LU$: factorización con pivoteo). Cálculo de la solución de un sistema lineal $AX = B$ cuando la matriz A es invertible.

```
function X = lufact(A,B) .
% Datos
%      - A es una matriz de orden N x N
```

```

% - B es una matriz de orden N x 1
% Resultado
% - X es la matriz de orden N x 1 solución de AX =B.
% Inicializamos X, Y, la matriz de almacenamiento temporal C y
% la matriz fila R donde se registran los intercambios de filas
[N,N]=size(A);
X=zeros(N,1);
Y=zeros(N,1);
C=zeros(1,N);
R=1:N;
for q=1:N-1
% Determinación de la fila pivote para la columna q-ésima
[max1,j]=max(abs(A(q:N,q)));
% Intercambio de las filas q-ésima y j-ésima
C=A(q,:);
A(q,:)=A(j+q-1,:);
A(j+q-1,:)=C;
d=R(q);
R(q)=R(j+q-1);
R(j+q-1)=d;
if A(q,q)==0
'A es singular, no hay solución o no es única'
break
end
% Cálculo del multiplicador,
% que se guarda en la parte subdiagonal de A
for k=q+1:N
mult=A(k,q)/A(q,q);
A(k,q) = mult;
A(k,q+1:N)=A(k,q+1:N)-mult*A(q,q+1:N);
end
end
% Resolución para hallar Y
Y(1) = B(R(1));
for k=2:N
Y(k)= B(R(k))-A(k,1:k-1)*Y(1:k-1);
end
% Resolución para hallar X
X(N)=Y(N)/A(N,N);
for k=N-1:-1:1
X(k)=(Y(k)-A(k,k+1:N)*X(k+1:N))/A(k,k);
end

```

Ejercicios

1. Resuelva $\mathbf{LY} = \mathbf{B}$, $\mathbf{UX} = \mathbf{Y}$ y compruebe que se verifica $\mathbf{B} = \mathbf{AX}$ para
 (a) $\mathbf{B} = [-4 \quad 10 \quad 5]'$ y (b) $\mathbf{B} = [20 \quad 49 \quad 32]'$, donde $\mathbf{A} = \mathbf{LU}$ es

$$\begin{bmatrix} 2 & 4 & -6 \\ 1 & 5 & 3 \\ 1 & 3 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/2 & 1/3 & 1 \end{bmatrix} \begin{bmatrix} 2 & 4 & -6 \\ 0 & 3 & 6 \\ 0 & 0 & 3 \end{bmatrix}.$$

2. Resuelva $\mathbf{LY} = \mathbf{B}$, $\mathbf{UX} = \mathbf{Y}$ y compruebe que se verifica $\mathbf{B} = \mathbf{AX}$ para
 (a) $\mathbf{B} = [7 \quad 2 \quad 10]'$ y (b) $\mathbf{B} = [23 \quad 35 \quad 7]'$, donde $\mathbf{A} = \mathbf{LU}$ es

$$\begin{bmatrix} 1 & 1 & 6 \\ -1 & 2 & 9 \\ 1 & -2 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 6 \\ 0 & 3 & 15 \\ 0 & 0 & 12 \end{bmatrix}.$$

3. Halle la factorización triangular $\mathbf{A} = \mathbf{LU}$ de las siguientes matrices

$$(a) \begin{bmatrix} -5 & 2 & -1 \\ 1 & 0 & 3 \\ 3 & 1 & 6 \end{bmatrix}$$

$$(b) \begin{bmatrix} 1 & 0 & 3 \\ 3 & 1 & 6 \\ -5 & 2 & -1 \end{bmatrix}$$

4. Halle la factorización triangular $\mathbf{A} = \mathbf{LU}$ de las siguientes matrices

$$(a) \begin{bmatrix} 4 & 2 & 1 \\ 2 & 5 & -2 \\ 1 & -2 & 7 \end{bmatrix}$$

$$(b) \begin{bmatrix} 1 & -2 & 7 \\ 4 & 2 & 1 \\ 2 & 5 & -2 \end{bmatrix}$$

5. Resuelva $\mathbf{LY} = \mathbf{B}$, $\mathbf{UX} = \mathbf{Y}$ y compruebe que se verifica $\mathbf{B} = \mathbf{AX}$ para

(a) $\mathbf{B} = [8 \quad -4 \quad 10 \quad -4]'$ y (b) $\mathbf{B} = [28 \quad 13 \quad 23 \quad 4]'$, donde $\mathbf{A} = \mathbf{LU}$ es

$$\begin{bmatrix} 4 & 8 & 4 & 0 \\ 1 & 5 & 4 & -3 \\ 1 & 4 & 7 & 2 \\ 1 & 3 & 0 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{4} & 1 & 0 & 0 \\ \frac{1}{4} & \frac{2}{3} & 1 & 0 \\ \frac{1}{4} & \frac{1}{3} & -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 4 & 8 & 4 & 0 \\ 0 & 3 & 3 & -3 \\ 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

6. Halle la factorización triangular $\mathbf{A} = \mathbf{LU}$ de la matriz

$$\begin{bmatrix} 1 & 1 & 0 & 4 \\ 2 & -1 & 5 & 0 \\ 5 & 2 & 1 & 2 \\ -3 & 0 & 2 & 6 \end{bmatrix}.$$

7. Establezca la fórmula que aparece en (12).

8. Pruebe que una factorización triangular es única en el siguiente sentido: Si \mathbf{A} es invertible y $\mathbf{L}_1\mathbf{U}_1 = \mathbf{A} = \mathbf{L}_2\mathbf{U}_2$, entonces $\mathbf{L}_1 = \mathbf{L}_2$ y $\mathbf{U}_1 = \mathbf{U}_2$.

9. Demuestre el caso $r > c$ que quedó pendiente en el Teorema 3.10.

10. (a) Compruebe que se verifica el Teorema 3.12 probando que se verifica $\mathbf{P}\mathbf{P}' = \mathbf{I} = \mathbf{P}'\mathbf{P}$ para la matriz de permutación

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

- (b) Pruebe el Teorema 3.12. *Indicación.* Use la definición de producto de matrices y el hecho de que cada fila y cada columna de \mathbf{P} y \mathbf{P}' contiene exactamente un elemento igual a 1.
11. Pruebe que la inversa de una matriz triangular superior e invertible de orden $N \times N$ también es triangular superior.

Algoritmos y programas

1. Utilice el Programa 3.3 para resolver el sistema $\mathbf{AX} = \mathbf{B}$, siendo

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 5 & 7 \\ 2 & -1 & 3 & 5 \\ 0 & 0 & 2 & 5 \\ -2 & -6 & -3 & 1 \end{bmatrix} \quad \text{y} \quad \mathbf{B} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}.$$

Luego utilice la instrucción $[L, U, P] = \text{lu}(\mathbf{A})$ para comprobar su respuesta.

2. Use el Programa 3.3 para resolver el sistema lineal $\mathbf{AX} = \mathbf{B}$, siendo $\mathbf{A} = [a_{ij}]_{N \times N}$ con $a_{ij} = i^{j-1}$ y $\mathbf{B} = [b_{ij}]_{N \times 1}$ con $b_{11} = N$ y $b_{i1} = i^{N-2}/(i-1)$ para $i \geq 2$. Hágalo en los casos $N = 3, 7$ y 11 . Sabiendo que la solución exacta es $\mathbf{X} = [1 \ 1 \ \dots \ 1 \ 1]^T$, explique cualquier desviación que aparezca entre la solución exacta y la calculada.
3. Modifique el Programa 3.3 de manera que calcule \mathbf{A}^{-1} resolviendo N sistemas de ecuaciones lineales

$$\mathbf{AC}_J = \mathbf{E}_J \quad \text{para } J = 1, 2, \dots, N.$$

Con lo cual

$$\mathbf{A} [\mathbf{C}_1 \ \mathbf{C}_2 \ \dots \ \mathbf{C}_N] = [\mathbf{E}_1 \ \mathbf{E}_2 \ \dots \ \mathbf{E}_N]$$

y, por tanto,

$$\mathbf{A}^{-1} = [\mathbf{C}_1 \ \mathbf{C}_2 \ \dots \ \mathbf{C}_N].$$

¡Debe calcular la factorización $\mathbf{A} = \mathbf{LU}$ sólo una vez!

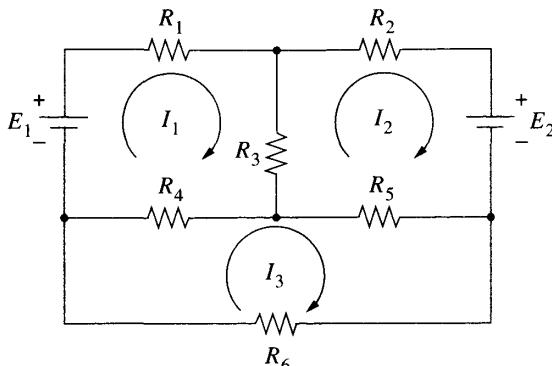


Figura 3.5 La red eléctrica del Ejercicio 4.

4. La ley de Kirchhoff para el voltaje aplicada al circuito que se muestra en la Figura 3.5 produce el siguiente sistema de ecuaciones:

$$(17) \quad \begin{aligned} (R_1 + R_3 + R_4)I_1 + & R_3 I_2 + & R_4 I_3 = E_1 \\ R_3 I_1 + (R_2 + R_3 + R_5)I_2 - & R_5 I_3 = E_2 \\ R_4 I_1 - & R_5 I_2 + (R_4 + R_5 + R_6)I_3 = 0. \end{aligned}$$

Use el Programa 3.3 para hallar las intensidades de corriente I_1 , I_2 e I_3 en los siguientes casos:

- (a) $R_1 = 1$, $R_2 = 1$, $R_3 = 2$, $R_4 = 1$, $R_5 = 2$, $R_6 = 4$, $E_1 = 23$ y $E_2 = 29$
 (b) $R_1 = 1$, $R_2 = 0.75$, $R_3 = 1$, $R_4 = 2$, $R_5 = 1$, $R_6 = 4$, $E_1 = 12$ y $E_2 = 21.5$
 (c) $R_1 = 1$, $R_2 = 2$, $R_3 = 4$, $R_4 = 3$, $R_5 = 1$, $R_6 = 5$ $E_1 = 41$ y $E_2 = 38$.

5. Las técnicas de cálculo infinitesimal nos dicen que la siguiente integral se determina descomponiendo el denominador en fracciones simples

$$\int \frac{x^2 + x + 1}{(x - 1)(x - 2)(x - 3)^2(x^2 + 1)} dx.$$

Esto requeriría hallar los coeficientes A_i , para $i = 1, 2, \dots, 6$, en la expresión

$$\begin{aligned} \frac{x^2 + x + 1}{(x - 1)(x - 2)(x - 3)^2(x^2 + 1)} &= \frac{A_1}{(x - 1)} + \frac{A_2}{(x - 2)} + \frac{A_3}{(x - 3)^2} + \frac{A_4}{(x - 3)} + \frac{A_5x + A_6}{(x^2 + 1)}. \end{aligned}$$

Use el Programa 3.3 para hallar los coeficientes de las fracciones simples.

6. Use el Programa 3.3 para resolver el sistema lineal $\mathbf{AX} = \mathbf{B}$, donde \mathbf{A} se genera mediante la instrucción $\mathbf{A}=\text{rand}(10,10)$ y $\mathbf{B}=[1 \ 2 \ 3 \ \dots \ 10]'$ del paquete MATLAB. Recuerde que debe verificar que \mathbf{A} es invertible ($\det(\mathbf{A}) \neq 0$) antes de usar el Programa 3.3. Compruebe la precisión de su respuesta formando

la diferencia de matrices $\mathbf{AX} - \mathbf{B}$ y examinando cuánto de cerca están sus elementos de cero (una respuesta exacta daría $\mathbf{AX} - \mathbf{B} = \mathbf{0}$). Repita el ejercicio usando una matriz de coeficientes \mathbf{A} generada mediante la instrucción $\mathbf{A}=\text{rand}(20,20)$ y $\mathbf{B}=[1 \ 2 \ 3 \ \dots \ 20]'$. Explique las diferencias que puedan aparecer en la precisión de las soluciones obtenidas con el Programa 3.3 en estos dos sistemas.

7. En la relación (8) de la Sección 3.1 definimos el concepto de combinación lineal en el espacio N -dimensional. Por ejemplo, el vector $(4, -3)$, que es equivalente a la matriz $[4 \ -3]'$, puede ser escrito como combinación lineal de los vectores $[1 \ 0]'$ y $[0 \ 1]'$:

$$\begin{bmatrix} 4 \\ -3 \end{bmatrix} = 4 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + (-3) \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Use el Programa 3.3 para probar que la matriz-columna $[1 \ 3 \ 5 \ 7 \ 9]'$ puede ser escrita como combinación lineal de las matrices-columna

$$\begin{bmatrix} 0 \\ 4 \\ -2 \\ 3 \\ -1 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 0 \\ 4 \\ 4 \end{bmatrix}, \begin{bmatrix} 3 \\ 2 \\ 0 \\ 5 \\ 1 \end{bmatrix}, \begin{bmatrix} 5 \\ 6 \\ -3 \\ 0 \\ 2 \end{bmatrix} \quad \text{y} \quad \begin{bmatrix} 1 \\ 4 \\ -2 \\ 7 \\ 0 \end{bmatrix}.$$

Explique por qué cualquier matriz-columna $[x_1 \ x_2 \ x_3 \ x_4 \ x_5]'$ puede ser escrita como combinación lineal de las matrices anteriores.

3.6 Métodos iterativos para sistemas lineales

El objetivo de esta sección es el extender a espacios de dimensión mayor que uno algunos de los métodos iterativos introducidos en el Capítulo 2. Consideraremos extensiones del método de iteración de punto fijo que se aplican a sistemas de ecuaciones lineales.

Método de iteración de Jacobi

Ejemplo 3.26. Consideremos el sistema de ecuaciones

$$(1) \quad \begin{aligned} 4x - y + z &= 7 \\ 4x - 8y + z &= -21 \\ -2x + y + 5z &= 15. \end{aligned}$$

Estas ecuaciones las podemos escribir como

$$(2) \quad \begin{aligned} x &= \frac{7 + y - z}{4} \\ y &= \frac{21 + 4x + z}{8} \\ z &= \frac{15 + 2x - y}{5}, \end{aligned}$$

lo que sugiere el siguiente proceso iterativo:

$$(3) \quad \begin{aligned} x_{k+1} &= \frac{7 + y_k - z_k}{4} \\ y_{k+1} &= \frac{21 + 4x_k + z_k}{8} \\ z_{k+1} &= \frac{15 + 2x_k - y_k}{5}. \end{aligned}$$

Vamos a comprobar que si empezamos con $\mathbf{P}_0 = (x_0, y_0, z_0) = (1, 2, 2)$, entonces la iteración (3) parece converger a la solución $(2, 4, 3)$.

Sustituyendo $x_0 = 1$, $y_0 = 2$ y $z_0 = 2$ en el miembro derecho de la relación (3), obtenemos

$$\begin{aligned} x_1 &= \frac{7 + 2 - 2}{4} = 1.75 \\ y_1 &= \frac{21 + 4 + 2}{8} = 3.375 \\ z_1 &= \frac{15 + 2 - 2}{5} = 3.00. \end{aligned}$$

El nuevo punto $\mathbf{P}_1 = (1.75, 3.375, 3.00)$ está más cerca de $(2, 4, 3)$ que \mathbf{P}_0 . En la Tabla 3.2 se muestra cómo los puntos $\{\mathbf{P}_k\}$ generados por la iteración (3) convergen a $(2, 4, 3)$.

Este proceso se conoce como **método de iteración de Jacobi** y puede usarse para resolver algunas clases de sistemas de ecuaciones lineales. Tras 19 pasos, vemos que, al aplicarlo al sistema (3), conseguimos 9 cifras decimales de aproximación $(2.00000000, 4.00000000, 3.00000000)$.

En la resolución numérica de ecuaciones en derivadas parciales suelen aparecer sistemas de ecuaciones lineales con incluso 100 000 incógnitas; en estos sistemas la matriz de los coeficientes es dispersa; es decir, un alto porcentaje de los elementos de la matriz son iguales a cero. Si hay algún tipo de patrón en la distribución de los elementos distintos de cero (ejemplo: los sistemas tridiagonales), entonces un método iterativo puede resultar muy eficaz en la resolución de estos sistemas tan enormes.

Algunas veces el método iterativo de Jacobi no funciona. Vamos a realizar un experimento para comprobar que una reordenación de las ecuaciones del sistema original puede tener como consecuencia que el método iterativo de Jacobi aplicado al nuevo sistema produzca un sucesión de puntos divergente.

Tabla 3.2 Convergencia del método iterativo de Jacobi para el sistema (1).

k	x_k	y_k	z_k
0	1.0	2.0	2.0
1	1.75	3.375	3.0
2	1.84375	3.875	3.025
3	1.9625	3.925	2.9625
4	1.99062500	3.97656250	3.00000000
5	1.99414063	3.99531250	3.00093750
⋮	⋮	⋮	⋮
15	1.99999993	3.99999985	2.99999993
⋮	⋮	⋮	⋮
19	2.00000000	4.00000000	3.00000000

Ejemplo 3.27. Reordenemos el sistema (1) como sigue:

$$(4) \quad \begin{aligned} -2x + y + 5z &= 15 \\ 4x - 8y + z &= -21 \\ 4x - y + z &= 7. \end{aligned}$$

Si escribimos estas ecuaciones como

$$(5) \quad \begin{aligned} x &= \frac{-15 + y + 5z}{3} \\ y &= \frac{21 + 4x + z}{8} \\ z &= 7 - 4x + y, \end{aligned}$$

entonces el método iterativo de Jacobi es, en este caso,

$$(6) \quad \begin{aligned} x_{k+1} &= \frac{-15 + y_k + 5z_k}{3} \\ y_{k+1} &= \frac{21 + 4x_k + z_k}{8} \\ z_{k+1} &= 7 - 4x_k + y_k. \end{aligned}$$

Veamos que si empezamos con el punto $P_0 = (x_0, y_0, z_0) = (1, 2, 2)$, entonces el proceso iterativo (6) diverge.

Tabla 3.3 Divergencia del método iterativo de Jacobi para el sistema (4).

k	x_k	y_k	z_k
0	1.0	2.0	2.0
1	-1.5	3.375	5.0
2	6.6875	2.5	16.375
3	34.6875	8.015625	-17.25
4	-46.617188	17.8125	-123.73438
5	-307.929688	-36.150391	211.28125
6	502.62793	-124.929688	1202.56836
:	:	:	:

Sustituimos $x_0 = 1$, $y_0 = 2$ y $z_0 = 2$ en el miembro derecho de (6) y obtenemos los nuevos valores x_1 , y_1 y z_1 siguientes:

$$\begin{aligned}x_1 &= \frac{-15 + 2 + 10}{2} = -1.5 \\y_1 &= \frac{21 + 4 + 2}{8} = 3.375 \\z_1 &= 7 - 4 + 2 = 5.00.\end{aligned}$$

El nuevo punto $\mathbf{P}_1 = (-1.5, 3.375, 5.00)$ está más lejos de la solución $(2, 4, 3)$ que \mathbf{P}_0 . De hecho, el proceso iterativo (6) diverge como se muestra en la Tabla 3.3.

Método iterativo de Gauss-Seidel

Algunas veces podemos acelerar la convergencia. Observemos que en el método iterativo de Jacobi (3) produce tres sucesiones $\{x_k\}$, $\{y_k\}$ y $\{z_k\}$ que convergen, respectivamente a 2, 4 y 3 (véase la Tabla 3.2). Puesto que x_{k+1} es, probablemente, mejor aproximación al límite que x_k , sería razonable usar x_{k+1} en vez de x_k a la hora de calcular y_{k+1} y, de forma semejante, sería mejor usar x_{k+1} e y_{k+1} en el cálculo de z_{k+1} . El siguiente ejemplo muestra lo que ocurre cuando se aplica este razonamiento al sistema de ecuaciones del Ejemplo 3.26.

Ejemplo 3.28. Consideremos el sistema de ecuaciones dado en (1) y el proceso iterativo, llamado **método de Gauss-Seidel**, sugerido por (2):

$$(7) \quad \begin{aligned}x_{k+1} &= \frac{7 + y_k - z_k}{4} \\y_{k+1} &= \frac{21 + 4x_{k+1} + z_k}{8} \\z_{k+1} &= \frac{15 + 2x_{k+1} - y_{k+1}}{5}.\end{aligned}$$

Tabla 3.4 Convergencia del método iterativo de Gauss-Seidel para el sistema (1).

k	x_k	y_k	z_k
0	1.0	2.0	2.0
1	1.75	3.75	2.95
2	1.95	3.96875	2.98625
3	1.995625	3.99609375	2.99903125
\vdots	\vdots	\vdots	\vdots
8	1.99999983	3.99999988	2.99999996
9	1.99999998	3.99999999	3.00000000
10	2.00000000	4.00000000	3.00000000

Veamos que si empezamos con $\mathbf{P}_0 = (x_0, y_0, z_0) = (1, 2, 2)$, entonces el proceso iterativo (7) converge a la solución $(2, 4, 3)$.

Sustituyendo $y_0 = 2$ y $z_0 = 2$ en la primera ecuación de (7) obtenemos

$$x_1 = \frac{7 + 2 - 2}{4} = 1.75.$$

Sustituyendo ahora $x_1 = 1.75$ y $z_0 = 2$ en la segunda ecuación de (7) obtenemos

$$y_1 = \frac{21 + 4(1.75) + 2}{8} = 3.75.$$

Finalmente, sustituyendo $x_1 = 1.75$ e $y_1 = 3.75$ en la tercera ecuación de (7) obtenemos

$$z_1 = \frac{15 + 2(1.75) - 3.75}{5} = 2.95.$$

El nuevo punto $\mathbf{P}_1 = (1.75, 3.75, 2.95)$ está más cerca de $(2, 4, 3)$ que \mathbf{P}_0 y es mejor que el punto obtenido en el Ejemplo 3.26. En la Tabla 3.4 se muestra cómo los puntos $\{\mathbf{P}_k\}$ generados por la iteración (7) convergen a $(2, 4, 3)$. ■

A la vista de los Ejemplos 3.26 y 3.27, se hace necesario disponer de algún criterio que determine si el método iterativo de Jacobi converge. Para ello damos la siguiente definición.

Definición 3.6. Se dice que una matriz \mathbf{A} de orden $N \times N$ es de *diagonal estrictamente dominante* cuando

$$(8) \quad |a_{kk}| > \sum_{\substack{j=1 \\ j \neq k}}^N |a_{kj}| \quad \text{para } k = 1, 2, \dots, N.$$

Esto significa que en cada fila de la matriz, el tamaño del elemento que está en la diagonal principal debe ser mayor que la suma de los tamaños de todos los demás elementos de la fila. La matriz de los coeficientes del sistema lineal (1), en el Ejemplo 3.26, es de diagonal estrictamente dominante porque

$$\text{En la primera fila: } |4| > |-1| + |1|$$

$$\text{En la segunda fila: } |-8| > |4| + |1|$$

$$\text{En la tercera fila: } |5| > |-2| + |1|.$$

Todas las filas verifican la relación (8) de la Definición 3.6; por tanto, la matriz de los coeficientes \mathbf{A} del sistema lineal (1) es de diagonal estrictamente dominante.

La matriz de los coeficientes \mathbf{A} del sistema lineal (4), en el Ejemplo 3.27, no es de diagonal estrictamente dominante porque

$$\text{En la primera fila: } |-2| < |1| + |5|$$

$$\text{En la segunda fila: } |-8| > |4| + |1|$$

$$\text{En la tercera fila: } |1| < |4| + |-1|.$$

Las filas primera y tercera no cumplen la relación (8) de la Definición 3.6; por tanto, la matriz de los coeficientes \mathbf{A} del sistema lineal (4) no es de diagonal estrictamente dominante.

Vamos a considerar ahora los procesos iterativos de Jacobi y Gauss-Seidel con mayor generalidad. Supongamos que tenemos un sistema de ecuaciones lineales

$$(9) \quad \begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1j}x_j + \cdots + a_{1N}x_N &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2j}x_j + \cdots + a_{2N}x_N &= b_2 \\ \vdots &\quad \vdots & \vdots &\quad \vdots &\quad \vdots \\ a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jj}x_j + \cdots + a_{jN}x_N &= b_j \\ \vdots &\quad \vdots & \vdots &\quad \vdots &\quad \vdots \\ a_{N1}x_1 + a_{N2}x_2 + \cdots + a_{Nj}x_j + \cdots + a_{NN}x_N &= b_N. \end{aligned}$$

Sea $\mathbf{P}_k = (x_1^{(k)}, x_2^{(k)}, \dots, x_j^{(k)}, \dots, x_N^{(k)})$ el k -ésimo punto obtenido, de manera que el siguiente punto es $\mathbf{P}_{k+1} = (x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_j^{(k+1)}, \dots, x_N^{(k+1)})$. El superíndice (k) de las coordenadas de \mathbf{P}_k nos permite identificar las coordenadas que pertenecen a dicho punto. Las fórmulas de iteración usan la fila j -ésima de (9) para despejar $x_j^{(k+1)}$ como una combinación lineal de los valores previamente obtenidos:

Método iterativo de Jacobi:

$$(10) \quad x_j^{(k+1)} = \frac{b_j - a_{j1}x_1^{(k)} - \cdots - a_{jj-1}x_{j-1}^{(k)} - a_{jj+1}x_{j+1}^{(k)} - \cdots - a_{jN}x_N^{(k)}}{a_{jj}}$$

para $j = 1, 2, \dots, N$.

En el método iterativo de Jacobi se usan todas las coordenadas del punto anterior en la obtención de las coordenadas del punto nuevo, mientras que en el método iterativo de Gauss-Seidel se emplean las coordenadas nuevas conforme se van generando:

Método iterativo de Gauss-Seidel:

$$(11) \quad x_j^{(k+1)} = \frac{b_j - a_{j1}x_1^{(k+1)} - \cdots - a_{jj-1}x_{j-1}^{(k+1)} - a_{jj+1}x_{j+1}^{(k)} - \cdots - a_{jN}x_N^{(k)}}{a_{jj}}$$

para $j = 1, 2, \dots, N$.

Es importante darse cuenta de la pequeña modificación de la fórmula (10) que nos conduce a la fórmula (11).

Teorema 3.15 (Método iterativo de Jacobi). Supongamos que \mathbf{A} es una matriz de diagonal estrictamente dominante. Entonces el sistema de ecuaciones lineales $\mathbf{AX} = \mathbf{B}$ tiene solución única $\mathbf{X} = \mathbf{P}$. Además, el proceso iterativo dado por la fórmula (10) produce una sucesión de vectores $\{\mathbf{P}_k\}$ que converge a \mathbf{P} cualquiera que sea el vector de partida \mathbf{P}_0 .

Demostración. La demostración puede encontrarse en textos de análisis numérico de carácter avanzado. •

Puede probarse que el método iterativo de Gauss-Seidel también converge cuando la matriz \mathbf{A} es de diagonal estrictamente dominante así como para matrices simétricas definidas positivas (véase la referencia [66]). Normalmente, el método de Gauss-Seidel converge más rápidamente que el de Jacobi, por lo que es el que se suele preferir (compárense los Ejemplos 3.26 y 3.28). Se dan casos, sin embargo, en los que el método de Jacobi converge pero el de Gauss-Seidel no.

Convergencia

Para determinar si una sucesión $\{\mathbf{P}_k\}$ converge a \mathbf{P} , es necesario tener una medida de la cercanía entre vectores. La distancia euclídea (véase la Sección 3.1) entre $\mathbf{P} = (x_1, x_2, \dots, x_N)$ y $\mathbf{Q} = (y_1, y_2, \dots, y_N)$, dada por

$$(12) \quad \|\mathbf{P} - \mathbf{Q}\| = \left(\sum_{j=1}^N (x_j - y_j)^2 \right)^{1/2},$$

tiene la desventaja de que requiere un esfuerzo computacional considerable. Por esa razón se introducen otras normas, como la norma $\|\mathbf{X}\|_1$:

$$(13) \quad \|\mathbf{X}\|_1 = \sum_{j=1}^N |x_j|.$$

El siguiente resultado asegura que $\|\mathbf{X}\|_1$ tiene las propiedades matemáticas de una métrica, lo que significa, en otras palabras, que es una fórmula adecuada para medir “distancias”. Sabemos, por la teoría del álgebra lineal, que todas las normas en un espacio vectorial de dimensión finita son equivalentes; es decir, si dos vectores están próximos según la norma $\|\cdot\|_1$, entonces también están próximos según la norma euclídea $\|\cdot\|$ y recíprocamente.

Teorema 3.16. Sean \mathbf{X} e \mathbf{Y} dos vectores N -dimensionales y sea c un escalar. Entonces la función $\|\mathbf{X}\|_1$ tiene las siguientes propiedades:

$$(14) \quad \|\mathbf{X}\|_1 \geq 0,$$

$$(15) \quad \|\mathbf{X}\|_1 = 0 \quad \text{si, y sólo si,} \quad \mathbf{X} = \mathbf{0},$$

$$(16) \quad \|c\mathbf{X}\|_1 = |c| \|\mathbf{X}\|_1,$$

$$(17) \quad \|\mathbf{X} + \mathbf{Y}\|_1 \leq \|\mathbf{X}\|_1 + \|\mathbf{Y}\|_1.$$

Demostración. Probamos (17) y dejamos las demás como ejercicio. Para cada j , la desigualdad triangular para el valor absoluto de números reales nos dice que $|x_j + y_j| \leq |x_j| + |y_j|$; la suma de todas estas desigualdades nos da la desigualdad (17):

$$\|\mathbf{X} + \mathbf{Y}\|_1 = \sum_{j=1}^N |x_j + y_j| \leq \sum_{j=1}^N |x_j| + \sum_{j=1}^N |y_j| = \|\mathbf{X}\|_1 + \|\mathbf{Y}\|_1.$$

La norma dada en (13) puede ser usada, entonces, para definir una nueva noción de distancia entre puntos.

Definición 3.7. Supongamos que \mathbf{X} e \mathbf{Y} son dos puntos en \mathbb{R}^N . Definimos la distancia entre \mathbf{X} e \mathbf{Y} según la norma $\|\cdot\|_1$ como

$$\|\mathbf{X} - \mathbf{Y}\|_1 = \sum_{j=1}^N |x_j - y_j|. \quad \blacktriangle$$

Ejemplo 3.29. Vamos a determinar la distancia euclídea y la distancia según la norma $\|*\|_1$ entre los puntos $P = (2, 4, 3)$ y $Q = (1.75, 3.75, 2.95)$.

La distancia euclídea es

$$\|P - Q\| = ((2 - 1.75)^2 + (4 - 3.75)^2 + (3 - 2.95)^2)^{1/2} = 0.3570$$

y la distancia según la norma $\|*\|_1$ es

$$\|P - Q\|_1 = |2 - 1.75| + |4 - 3.75| + |3 - 2.95| = 0.55.$$

La norma $\|*\|_1$ es más fácil de calcular y, por eso, se suele usar para determinar la convergencia en el espacio N -dimensional space. ■

MATLAB

En el Programa 3.4 se usa la instrucción $A(j, [1:j-1, j+1:N])$ del paquete MATLAB; esta instrucción permite seleccionar todos los elementos de la fila j -ésima de A excepto el elemento de la diagonal principal $A(j, j)$, o sea, el que está en la columna j -ésima. Esta notación permite simplificar el paso general de la iteración de Jacobi (10) en el Programa 3.4.

Tanto en el Programa 3.4 como en el 3.5 hemos usado la instrucción del paquete MATLAB `norm`, que calcula la norma euclídea. La norma $\|*\|_1$ también puede usarse; le animamos a que consulte la información sobre la instrucción `norm` que hay en el menú de ayuda de MATLAB, o cualquiera de las obras de referencia.

Programa 3.4 (Método iterativo de Jacobi). Resolución de un sistema lineal $AX = B$ mediante la generación de una sucesión $\{P_k\}$ que converge a la solución, a partir de un punto inicial P_0 . Una condición suficiente para que el método sea aplicable es que A sea de diagonal estrictamente dominante.

```
function X=jacobi(A,B,P,delta,max1)
% Datos
%     - A es una matriz invertible de orden N x N
%     - B es una matriz de orden N x 1
%     - P es una matriz de orden N x 1: el punto inicial
%     - delta es la tolerancia para P
%     - max1 es el número máximo de iteraciones
% Resultados
%     - X es una matriz de orden N x 1:
%         la aproximación a la solución de AX=B
%         generada por el método iterativo de Jacobi
N = length(B);
for k=1:max1
```

```

for j=1:N
    X(j)=(B(j)-A(j,[1:j-1,j+1:N])*P([1:j-1,j+1:N]))/A(j,j);
end
err=abs(norm(X'-P));
relerr=err/(norm(X)+eps);
P=X';
if(err<delta)|(relerr<delta)
    break
end
end
X=X';

```

Programa 3.5 (Método iterativo de Gauss-Seidel). Resolución de un sistema lineal $\mathbf{AX} = \mathbf{B}$ mediante la generación de una sucesión $\{\mathbf{P}_k\}$ que converge a la solución, a partir de un punto inicial \mathbf{P}_0 . Una condición suficiente para que el método sea aplicable es que \mathbf{A} sea de diagonal estrictamente dominante.

```

function X=gseid(A,B,P,delta,max1)
% Datos
%
% - A es una matriz invertible de orden N x N
% - B es una matriz de orden N x 1
% - P es una matriz de orden N x 1: el punto inicial
% - delta es la tolerancia para P
% - max1 es el número máximo de iteraciones
% Resultados
%
% - X es una matriz de orden N x 1:
%   la aproximación a la solución de AX=B
%   generada por el método iterativo de Gauss-Seidel
N = length(B);
for k=1:max1
    for j=1:N
        if j==1
            X(1)=(B(1)-A(1,2:N)*P(2:N))/A(1,1);
        elseif j==N
            X(N)=(B(N)-A(N,1:N-1)*(X(1:N-1))')/A(N,N);
        else
            % X contiene la aproximación k-ésima
            % y P la (k-1)-ésima
            X(j)=(B(j)-A(j,1:j-1)*X(1:j-1)
                  -A(j,j+1:N)*P(j+1:N))/A(j,j);
        end
    end
end

```

```

err=abs(norm(X'-P));
relerr=err/(norm(X)+eps);
P=X';
if(err<delta)|(relerr<delta)
    break
end
X=X';

```

Ejercicios

En los Ejercicios 1 a 8:

(a) Empiece con $\mathbf{P}_0 = \mathbf{0}$, use el método iterativo de Jacobi y calcule \mathbf{P}_k para $k = 1, 2, 3$. ¿Converge la iteración de Jacobi a la solución?

(b) Empiece con $\mathbf{P}_0 = \mathbf{0}$, use el método iterativo de Gauss-Seidel y calcule \mathbf{P}_k para $k = 1, 2, 3$. ¿Converge la iteración de Gauss-Seidel a la solución?

$$\begin{aligned} 1. \quad 4x - y &= 15 \\ &x + 5y = 9 \end{aligned}$$

$$\begin{aligned} 2. \quad 8x - 3y &= 10 \\ &-x + 4y = 6 \end{aligned}$$

$$\begin{aligned} 3. \quad -x + 3y &= 1 \\ &6x - 2y = 2 \end{aligned}$$

$$\begin{aligned} 4. \quad 2x + 3y &= 1 \\ &7x - 2y = 1 \end{aligned}$$

$$\begin{aligned} 5. \quad 5x - y + z &= 10 \\ &2x + 8y - z = 11 \\ &-x + y + 4z = 3 \end{aligned}$$

$$\begin{aligned} 6. \quad 2x + 8y - z &= 11 \\ &5x - y + z = 10 \\ &-x + y + 4z = 3 \end{aligned}$$

$$\begin{aligned} 7. \quad x - 5y - z &= -8 \\ &4x + y - z = 13 \\ &2x - y - 6z = -2 \end{aligned}$$

$$\begin{aligned} 8. \quad 4x + y - z &= 13 \\ &x - 5y - z = -8 \\ &2x - y - 6z = -2 \end{aligned}$$

9. Sea $\mathbf{X} = (x_1, x_2, \dots, x_N)$. Pruebe que la norma $\|*\|_1$

$$\|\mathbf{X}\|_1 = \sum_{k=1}^N |x_k|$$

verifica las tres propiedades (14)–(16).

10. Sea $\mathbf{X} = (x_1, x_2, \dots, x_N)$. Pruebe que la norma euclídea

$$\|\mathbf{X}\| = \left(\sum_{k=1}^N (x_k)^2 \right)^{1/2}$$

verifica las cuatro propiedades (14)–(17).

11. Sea $\mathbf{X} = (x_1, x_2, \dots, x_N)$. Pruebe que la norma $\|*\|_\infty$ definida por

$$\|\mathbf{X}\|_\infty = \max\{|x_k| : 1 \leq k \leq N\}$$

verifica las cuatro propiedades (14)–(17).

Algoritmos y programas

- Use los dos Programas 3.4 y 3.5 para resolver los Ejercicios 1 a 8 empleando la instrucción `format long` del paquete MATLAB con `delta = 10^-9`.
- En el Teorema 3.14 la condición de que \mathbf{A} sea de diagonal estrictamente dominante es suficiente pero no necesaria. Use los dos Programas 3.4 y 3.5, tomando varios puntos iniciales \mathbf{P}_0 , para resolver el siguiente sistema lineal. Nota. La iteración de Jacobi parece converger, mientras que la de Gauss-Seidel diverge.

$$\begin{array}{rcl} x & + & z = 2 \\ -x & + & y = 0 \\ x & + 2y & - 3z = 0 \end{array}$$

- Consideremos el siguiente sistema lineal tridiagonal y supongamos que la matriz de los coeficientes es de diagonal estrictamente dominante.

$$\begin{array}{lll} d_1 x_1 + c_1 x_2 & & = b_1 \\ a_1 x_1 + d_2 x_2 + c_2 x_3 & & = b_2 \\ a_2 x_2 + d_3 x_3 + c_3 x_4 & & = b_3 \\ & \ddots & \vdots \\ & & \\ a_{N-2} x_{N-2} + d_{N-1} x_{N-1} + c_{N-1} x_N & & = b_{N-1} \\ a_{N-1} x_{N-1} + d_N x_N & & = b_N \end{array}$$

(i) Escriba un algoritmo iterativo, siguiendo (9)–(11), que resuelva este sistema. Su algoritmo debería hacer un uso eficiente del hecho de que la matriz es dispersa.

(ii) Construya un programa con el paquete MATLAB basado en su algoritmo y resuelva los sistemas tridiagonales que se relacionan a continuación.

<p>(a) $4m_1 + m_2 = 3$</p> $m_1 + 4m_2 + m_3 = 3$ $m_2 + 4m_3 + m_4 = 3$ $m_3 + 4m_4 + m_5 = 3$ $\vdots \quad \vdots \quad \vdots \quad \vdots$ $m_{48} + 4m_{49} + m_{50} = 3$ $m_{49} + 4m_{50} = 3$	<p>(b) $4m_1 + m_2 = 1$</p> $m_1 + 4m_2 + m_3 = 2$ $m_2 + 4m_3 + m_4 = 1$ $m_3 + 4m_4 + m_5 = 2$ $\vdots \quad \vdots \quad \vdots \quad \vdots$ $m_{48} + 4m_{49} + m_{50} = 1$ $m_{49} + 4m_{50} = 2$
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

4. Use el método iterativo de Gauss-Seidel para resolver el siguiente sistema lineal con estructura de banda:

$$\begin{aligned}
 12x_1 - 2x_2 + x_3 &= 5 \\
 -2x_1 + 12x_2 - 2x_3 + x_4 &= 5 \\
 x_1 - 2x_2 + 12x_3 - 2x_4 + x_5 &= 5 \\
 x_2 - 2x_3 + 12x_4 - 2x_5 + x_6 &= 5 \\
 \vdots &\quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
 x_{46} - 2x_{47} + 12x_{48} - 2x_{49} + x_{50} &= 5 \\
 x_{47} - 2x_{48} + 12x_{49} - 2x_{50} &= 5 \\
 x_{48} - 2x_{49} + 12x_{50} &= 5
 \end{aligned}$$

5. En los Programas 3.4 y 3.5 se emplea el error relativo entre dos puntos consecutivos de la iteración como criterio de parada. Los problemas que puede acarrear el uso exclusivo de este criterio ya fueron discutidos extensamente en la Sección 2.3. Si escribimos el sistema lineal $\mathbf{AX} = \mathbf{B}$ como $\mathbf{AX} - \mathbf{B} = \mathbf{0}$ y si \mathbf{X}_k es el vector obtenido en el paso k -ésimo del método iterativo de Jacobi, o del método de Gauss-Seidel, entonces la norma del **residuo** $\|\mathbf{AX}_k - \mathbf{B}\|$ es, en general, un criterio de parada más adecuado.

Modifique los Programas 3.4 y 3.5 para emplear el tamaño del residuo como criterio de parada y use los programas modificados para resolver el sistema con estructura de banda del Problema 4.

3.7 Métodos iterativos para sistemas no lineales (opcional)

En esta sección discutiremos técnicas iterativas que extienden los métodos del Capítulo 2 y de la Sección 3.6 al caso de los sistemas de ecuaciones no lineales. Consideraremos las funciones

$$\begin{aligned}
 (1) \qquad f_1(x, y) &= x^2 - 2x - y + 0.5 \\
 f_2(x, y) &= x^2 + 4y^2 - 4.
 \end{aligned}$$

Queremos hallar un método para resolver el sistema de ecuaciones no lineales:

$$(2) \qquad f_1(x, y) = 0 \quad y \quad f_2(x, y) = 0.$$

Las ecuaciones $f_1(x, y) = 0$ y $f_2(x, y) = 0$ definen implícitamente sendas curvas en el plano XOY ; por tanto, una solución del sistema (2) es un punto (p, q) en el que ambas curvas se cruzan (o sea, $f_1(p, q) = 0$ y $f_2(p, q) = 0$). Las curvas del sistema (1) son bien conocidas:

$$\begin{aligned}
 (3) \qquad x^2 - 2x - y + 0.5 = 0 &\quad \text{es una parábola,} \\
 x^2 + 4y^2 - 4 = 0 &\quad \text{es una ellipse.}
 \end{aligned}$$

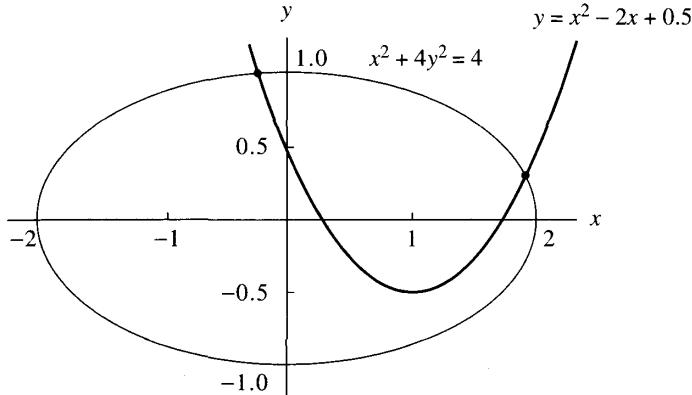


Figura 3.6 Representación gráfica del sistema no lineal $y = x^2 - 2x + 0.5$ y $x^2 + 4y^2 = 4$.

En la Figura 3.6 podemos ver que hay dos puntos de corte entre ambas curvas y que estos puntos están, respectivamente, cerca de $(-0.2, 1.0)$ y de $(1.9, 0.3)$.

La primera técnica es iteración de punto fijo. Debemos desarrollar un método que genere una sucesión $\{(p_k, q_k)\}$ convergente a una de las soluciones. En la primera ecuación de (3) podemos despejar directamente x ; sin embargo, para despejar y en la segunda, añadimos un múltiplo adecuado de y a cada miembro, en este caso $-8y$, para obtener $x^2 + 4y^2 - 8y - 4 = -8y$. La elección de $-8y$ es crucial y será explicada más adelante. Tenemos entonces el sistema equivalente de ecuaciones:

$$(4) \quad \begin{aligned} x &= \frac{x^2 - y + 0.5}{2} \\ y &= \frac{-x^2 - 4y^2 + 8y + 4}{8}. \end{aligned}$$

Ahora usamos estas dos ecuaciones para escribir fórmulas recursivas: Empezando con un punto inicial (p_0, q_0) , construimos la sucesión $\{(p_{k+1}, q_{k+1})\}$ mediante el esquema iterativo:

$$(5) \quad \begin{aligned} p_{k+1} &= g_1(p_k, q_k) = \frac{p_k^2 - q_k + 0.5}{2} \\ q_{k+1} &= g_2(p_k, q_k) = \frac{-p_k^2 - 4q_k^2 + 8q_k + 4}{8}. \end{aligned}$$

Caso (i): Si usamos como punto inicial $(p_0, q_0) = (0, 1)$, entonces

$$p_1 = \frac{0^2 - 1 + 0.5}{2} = -0.25 \quad \text{y} \quad q_1 = \frac{-0^2 - 4(1)^2 + 8(1) + 4}{8} = 1.0.$$

Tabla 3.5 Iteración de punto fijo con las fórmulas de (5).

Caso (i): Punto inicial (0, 1)			Caso (ii): Punto inicial (2, 0)		
k	p_k	q_k	k	p_k	q_k
0	0.00	1.00	0	2.00	0.00
1	-0.25	1.00	1	2.25	0.00
2	-0.21875	0.9921875	2	2.78125	-0.1328125
3	-0.2221680	0.9939880	3	4.184082	-0.6085510
4	-0.2223147	0.9938121	4	9.307547	-2.4820360
5	-0.2221941	0.9938029	5	44.80623	-15.891091
6	-0.2222163	0.9938095	6	1,011.995	-392.60426
7	-0.2222147	0.9938083	7	512,263.2	-205,477.82
8	-0.2222145	0.9938084			Esta sucesión diverge.
9	-0.2222146	0.9938084			

La iteración que genera la sucesión del caso (i) se muestra en la Tabla 3.5. En este caso la sucesión converge a la solución que está cerca del punto inicial (0, 1).

Caso (ii): Si usamos como punto inicial $(p_0, q_0) = (2, 0)$, entonces

$$p_1 = \frac{2^2 - 0 + 0.5}{2} = 2.25 \quad \text{y} \quad q_1 = \frac{-2^2 - 4(0)^2 + 8(0) + 4}{8} = 0.0.$$

La iteración que genera la sucesión del caso (ii) se muestra en la Tabla 3.5. En este caso la sucesión diverge.

El esquema de iteración dado en (5) no puede usarse para dar una aproximación a la segunda solución $(1.900677, 0.3112186)$. Para hallar este punto necesitamos un par de fórmulas de iteración distintas de las de (5). Si añadimos $-2x$ a la primera ecuación de (3) y $-11y$ a la segunda, obtenemos

$$x^2 - 4x - y + 0.5 = -2x \quad \text{y} \quad x^2 + 4y^2 - 11y - 4 = -11y,$$

a partir de las cuales generamos las fórmulas de iteración

$$(6) \quad p_{k+1} = g_1(p_k, q_k) = \frac{-p_k^2 + 4p_k + q_k - 0.5}{2}$$

$$q_{k+1} = g_2(p_k, q_k) = \frac{-p_k^2 - 4q_k^2 + 11q_k + 4}{11}.$$

En la Tabla 3.6 se muestran los resultados obtenidos al usar (6) para hallar la segunda solución.

Teoría

¿Por qué resultaron adecuadas las fórmulas de (6) para hallar la solución cerca de $(1.9, 0.3)$, mientras que las de (5) no? En la Sección 2.1 vimos que el aspecto

Tabla 3.6 Iteración de punto fijo con las fórmulas de (6).

k	p_k	q_k
0	2.00	0.00
1	1.75	0.0
2	1.71875	0.0852273
3	1.753063	0.1776676
4	1.808345	0.2504410
8	1.903595	0.3160782
12	1.900924	0.3112267
16	1.900652	0.3111994
20	1.900677	0.3112196
24	1.900677	0.3112186

crucial era el tamaño de la derivada en el punto fijo. Cuando se trabaja con funciones de varias variables, hay que usar las derivadas parciales. La generalización de la noción de derivada para sistemas de funciones de varias variables es la matriz jacobiana, sobre la que recordaremos algunas ideas introductorias. Los detalles adicionales se pueden encontrar en los libros de cálculo infinitesimal para funciones de varias variables.

Definición 3.8 (Matriz jacobiana). Sean $f_1(x, y)$ y $f_2(x, y)$ funciones de dos variables independientes x e y . Entonces su matriz jacobiana $\mathbf{J}(x, y)$ es

$$(7) \quad \mathbf{J}(x, y) = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{bmatrix}.$$

Análogamente, si $f_1(x, y, z)$, $f_2(x, y, z)$ y $f_3(x, y, z)$ son funciones de tres variables independientes x , y y z , entonces su matriz jacobiana es la matriz $\mathbf{J}(x, y, z)$ de orden 3×3 definida por:

$$(8) \quad \mathbf{J}(x, y, z) = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} & \frac{\partial f_1}{\partial z} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} & \frac{\partial f_2}{\partial z} \\ \frac{\partial f_3}{\partial x} & \frac{\partial f_3}{\partial y} & \frac{\partial f_3}{\partial z} \end{bmatrix}.$$

Ejemplo 3.30. Vamos a calcular la matriz jacobiana $\mathbf{J}(x, y, z)$ de orden 3×3 en el punto $(1, 3, 2)$ para las funciones

$$f_1(x, y, z) = x^3 - y^2 + y - z^4 + z^2$$

$$f_2(x, y, z) = xy + yz + xz$$

$$f_3(x, y, z) = \frac{y}{xz}.$$

La matriz jacobiana es

$$\mathbf{J}(x, y, z) = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} & \frac{\partial f_1}{\partial z} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} & \frac{\partial f_2}{\partial z} \\ \frac{\partial f_3}{\partial x} & \frac{\partial f_3}{\partial y} & \frac{\partial f_3}{\partial z} \end{bmatrix} = \begin{bmatrix} 3x^2 & -2y+1 & -4z^3+2z \\ y+z & x+z & y+x \\ -\frac{y}{x^2 z} & \frac{1}{xz} & \frac{-y}{xz^2} \end{bmatrix}.$$

Luego la matriz jacobiana evaluada en el punto $(1, 3, 2)$ es la matriz de orden 3×3

$$\mathbf{J}(1, 3, 2) = \begin{bmatrix} 3 & -5 & -28 \\ 5 & 3 & 4 \\ -\frac{3}{2} & \frac{1}{2} & -\frac{3}{4} \end{bmatrix}.$$

La diferencial

Cuando tenemos una función de varias variables, la diferencial es el instrumento que se usa para mostrar el efecto de los cambios de las variables independientes en los cambios de las variables dependientes. Consideremos las funciones

$$(9) \quad u = f_1(x, y, z), \quad v = f_2(x, y, z) \quad \text{y} \quad w = f_3(x, y, z).$$

Supongamos que los valores de las funciones de (9) se conocen en el punto (x_0, y_0, z_0) y que queremos estimar sus valores en un punto cercano (x, y, z) . Si denotamos por du , dv y dw los cambios diferenciales en las variables dependientes y por dx , dy y dz los cambios diferenciales en las variables independientes, entonces estos cambios obedecen las relaciones

$$(10) \quad \begin{aligned} du &= \frac{\partial f_1}{\partial x}(x_0, y_0, z_0) dx + \frac{\partial f_1}{\partial y}(x_0, y_0, z_0) dy + \frac{\partial f_1}{\partial z}(x_0, y_0, z_0) dz, \\ dv &= \frac{\partial f_2}{\partial x}(x_0, y_0, z_0) dx + \frac{\partial f_2}{\partial y}(x_0, y_0, z_0) dy + \frac{\partial f_2}{\partial z}(x_0, y_0, z_0) dz, \\ dw &= \frac{\partial f_3}{\partial x}(x_0, y_0, z_0) dx + \frac{\partial f_3}{\partial y}(x_0, y_0, z_0) dy + \frac{\partial f_3}{\partial z}(x_0, y_0, z_0) dz. \end{aligned}$$

Usando notación vectorial, (10) puede escribirse de forma más compacta usando la matriz jacobiana. Si denotamos los cambios en la función vectorial por $d\mathbf{F}$ y los cambios en las variables por $d\mathbf{X}$, entonces

$$(11) \quad d\mathbf{F} = \begin{bmatrix} du \\ dv \\ dw \end{bmatrix} = \mathbf{J}(x_0, y_0, z_0) \begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} = \mathbf{J}(x_0, y_0, z_0) d\mathbf{X}.$$

Ejemplo 3.31. Vamos a usar la matriz jacobiana para estimar los cambios diferenciales (du, dv, dw) cuando las variables independientes cambian de $(1, 3, 2)$ a $(1.02, 2.97, 2.01)$ en el sistema de funciones

$$\begin{aligned} u &= f_1(x, y, z) = x^3 - y^2 + y - z^4 + z^2 \\ v &= f_2(x, y, z) = xy + yz + xz \\ w &= f_3(x, y, z) = \frac{y}{xz}. \end{aligned}$$

Usando la relación (11) junto con la matriz $\mathbf{J}(1, 3, 2)$ del Ejemplo 3.30 y los incrementos $(dx, dy, dz) = (0.02, -0.03, 0.01)$, obtenemos

$$\begin{bmatrix} du \\ dv \\ dw \end{bmatrix} = \begin{bmatrix} 3 & -5 & -28 \\ 5 & 3 & 4 \\ -\frac{3}{2} & \frac{1}{2} & -\frac{3}{4} \end{bmatrix} \begin{bmatrix} 0.02 \\ -0.03 \\ 0.01 \end{bmatrix} = \begin{bmatrix} -0.07 \\ 0.05 \\ -0.0525 \end{bmatrix}.$$

Hagamos notar que las aproximaciones lineales que se obtienen al sumar los incrementos $du = -0.07$, $dv = 0.05$ y $dw = -0.0525$ a los valores de las correspondientes funciones $f_1(1, 3, 2) = -17$, $f_2(1, 3, 2) = 11$ y $f_3(1, 3, 2) = 1.5$ están cerca de los valores de las funciones en el punto $(1.02, 2.97, 2.01)$:

$$\begin{aligned} f_1(1.02, 2.97, 2.01) &= -17.072 \approx -17.01 = f_1(1, 3, 2) + du \\ f_2(1.02, 2.97, 2.01) &= 11.0493 \approx 11.05 = f_2(1, 3, 2) + dv \\ f_3(1.02, 2.97, 2.01) &= 1.44864 \approx 1.4475 = f_3(1, 3, 2) + dw. \end{aligned}$$

Convergencia cerca de los puntos fijos

Damos ahora las extensiones de las definiciones y los teoremas de la Sección 2.1 a los casos bidimensional y tridimensional. No mencionaremos las extensiones al caso general N -dimensional que pueden hallarse en los textos de análisis numérico.

Definición 3.9. Un *punto fijo* del sistema de dos ecuaciones

$$(12) \quad x = g_1(x, y) \quad \text{e} \quad y = g_2(x, y)$$

es un punto (p, q) tal que $p = g_1(p, q)$ y $q = g_2(p, q)$. Análogamente, en el caso tridimensional, un punto fijo del sistema

$$(13) \quad x = g_1(x, y, z), \quad y = g_2(x, y, z) \quad \text{y} \quad z = g_3(x, y, z)$$

es un punto (p, q, r) tal que $p = g_1(p, q, r)$, $q = g_2(p, q, r)$ y $r = g_3(p, q, r)$. ▲

Definición 3.10. Para las funciones de (12), el **método de iteración de punto fijo** es

$$(14) \quad p_{k+1} = g_1(p_k, q_k) \quad \text{y} \quad q_{k+1} = g_2(p_k, q_k)$$

para $k = 0, 1, \dots$

Análogamente, para las funciones de (13), el **método de iteración de punto fijo** es

$$(15) \quad \begin{aligned} p_{k+1} &= g_1(p_k, q_k, r_k) \\ q_{k+1} &= g_2(p_k, q_k, r_k) \\ r_{k+1} &= g_3(p_k, q_k, r_k) \end{aligned}$$

para $k = 0, 1, \dots$

Teorema 3.17 (Iteración de punto fijo). Supongamos que las funciones de (12) y (13) así como sus derivadas parciales son continuas en una región que contiene un punto fijo (p, q) o (p, q, r) , respectivamente. Entonces tenemos los siguientes casos:

Caso (i): Bidimensional. Si (p_0, q_0) está suficientemente cerca de (p, q) y

$$(16) \quad \begin{aligned} \left| \frac{\partial g_1}{\partial x}(p, q) \right| + \left| \frac{\partial g_1}{\partial y}(p, q) \right| &< 1, \\ \left| \frac{\partial g_2}{\partial x}(p, q) \right| + \left| \frac{\partial g_2}{\partial y}(p, q) \right| &< 1, \end{aligned}$$

entonces la iteración de punto fijo descrita en (14) genera una sucesión convergente al punto fijo (p, q) .

Caso (ii): Tridimensional. Si (p_0, q_0, r_0) está suficientemente cerca de (p, q, r) y

$$(17) \quad \begin{aligned} \left| \frac{\partial g_1}{\partial x}(p, q, r) \right| + \left| \frac{\partial g_1}{\partial y}(p, q, r) \right| + \left| \frac{\partial g_1}{\partial z}(p, q, r) \right| &< 1, \\ \left| \frac{\partial g_2}{\partial x}(p, q, r) \right| + \left| \frac{\partial g_2}{\partial y}(p, q, r) \right| + \left| \frac{\partial g_2}{\partial z}(p, q, r) \right| &< 1, \\ \left| \frac{\partial g_3}{\partial x}(p, q, r) \right| + \left| \frac{\partial g_3}{\partial y}(p, q, r) \right| + \left| \frac{\partial g_3}{\partial z}(p, q, r) \right| &< 1, \end{aligned}$$

entonces la iteración de punto fijo descrita en (15) genera una sucesión convergente al punto fijo (p, q, r) .

Si las condiciones (16) o (17) no se cumplen, entonces la iteración podría ser divergente. Esto es lo que suele ocurrir cuando la suma de los tamaños de las derivadas parciales es mucho mayor que 1. Podemos usar el Teorema 3.17

para ver por qué la iteración (5) converge al punto fijo cercano a $(-0.2, 1.0)$. Las derivadas parciales son

$$\begin{aligned}\frac{\partial}{\partial x}g_1(x, y) &= x, & \frac{\partial}{\partial y}g_1(x, y) &= -\frac{1}{2}, \\ \frac{\partial}{\partial x}g_2(x, y) &= -\frac{x}{4}, & \frac{\partial}{\partial y}g_2(x, y) &= -y + 1.\end{aligned}$$

Luego, para todo (x, y) tal que $-0.5 < x < 0.5$ y $0.5 < y < 1.5$, se tiene que las derivadas parciales cumplen

$$\begin{aligned}\left|\frac{\partial}{\partial x}g_1(x, y)\right| + \left|\frac{\partial}{\partial y}g_1(x, y)\right| &= |x| + |-0.5| < 1, \\ \left|\frac{\partial}{\partial x}g_2(x, y)\right| + \left|\frac{\partial}{\partial y}g_2(x, y)\right| &= \frac{|-x|}{4} + |-y + 1| < 0.625 < 1.\end{aligned}$$

Por tanto, las condiciones sobre las derivadas parciales dadas en (16) se cumplen y el Teorema 3.17 garantiza que la iteración de punto fijo converge a $(p, q) \approx (-0.2222146, 0.9938084)$. En las cercanías del otro punto fijo $(1.90068, 0.31122)$, las derivadas parciales no cumplen las condiciones de (16), así que la convergencia no se asegura, ya que

$$\begin{aligned}\left|\frac{\partial}{\partial x}g_1(1.90068, 0.31122)\right| + \left|\frac{\partial}{\partial y}g_1(1.90068, 0.31122)\right| &= 2.40068 > 1, \\ \left|\frac{\partial}{\partial x}g_2(1.90068, 0.31122)\right| + \left|\frac{\partial}{\partial y}g_2(1.90068, 0.31122)\right| &= 1.16395 > 1.\end{aligned}$$

El método iterativo de Seidel

Podemos llevar a cabo una mejora, análoga al método iterativo de Gauss-Seidel para sistemas lineales, del método de iteración de punto fijo. Consiste en usar p_{k+1} en el cálculo de q_{k+1} (y, en el caso tridimensional, usar p_{k+1} y q_{k+1} en el cálculo de r_{k+1}). La incorporación de estas modificaciones a las fórmulas (14) y (15) se conoce como **método iterativo de Seidel**:

$$(18) \quad p_{k+1} = g_1(p_k, q_k) \quad \text{y} \quad q_{k+1} = g_2(p_{k+1}, q_k),$$

y

$$(19) \quad \begin{aligned}p_{k+1} &= g_1(p_k, q_k, r_k) \\ q_{k+1} &= g_2(p_{k+1}, q_k, r_k) \\ r_{k+1} &= g_3(p_{k+1}, q_{k+1}, r_k).\end{aligned}$$

El Programa 3.6 utiliza el método iterativo de Seidel para sistemas no lineales; dejamos como ejercicio la construcción de un programa similar que emplee iteración de punto fijo.

El método de Newton-Raphson para sistemas no lineales

Vamos a construir el método de Newton-Raphson en el caso bidimensional; construcción que se generaliza fácilmente a dimensiones mayores.

Consideremos el sistema

$$(20) \quad \begin{aligned} u &= f_1(x, y) \\ v &= f_2(x, y), \end{aligned}$$

que puede verse como una transformación del plano XOY en el plano UOV . Si estamos interesados en el comportamiento de esta transformación cerca del punto (x_0, y_0) , cuya imagen es el punto (u_0, v_0) , y si las dos funciones tienen derivadas parciales continuas, entonces podemos usar la diferencial del sistema para escribir un sistema de aproximaciones incrementales lineales válidas cerca del punto (x_0, y_0) en cuestión:

$$(21) \quad \begin{aligned} u - u_0 &\approx \frac{\partial}{\partial x} f_1(x_0, y_0)(x - x_0) + \frac{\partial}{\partial y} f_1(x_0, y_0)(y - y_0), \\ v - v_0 &\approx \frac{\partial}{\partial x} f_2(x_0, y_0)(x - x_0) + \frac{\partial}{\partial y} f_2(x_0, y_0)(y - y_0). \end{aligned}$$

El sistema (21) es una aproximación lineal local que nos da una idea del efecto que pequeños cambios en las variables independientes producen en las variables dependientes. Si usamos la matriz jacobiana $\mathbf{J}(x_0, y_0)$, esta relación se escribe de forma más cómoda como

$$(22) \quad \begin{bmatrix} u - u_0 \\ v - v_0 \end{bmatrix} \approx \begin{bmatrix} \frac{\partial}{\partial x} f_1(x_0, y_0) & \frac{\partial}{\partial y} f_1(x_0, y_0) \\ \frac{\partial}{\partial x} f_2(x_0, y_0) & \frac{\partial}{\partial y} f_2(x_0, y_0) \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}.$$

Si escribimos el sistema de (20) como una función vectorial $\mathbf{V} = \mathbf{F}(\mathbf{X})$, entonces la matriz jacobiana $\mathbf{J}(x, y)$ es al análogo bidimensional de la derivada, porque la relación (22) queda

$$(23) \quad \Delta \mathbf{F} \approx \mathbf{J}(x_0, y_0) \Delta \mathbf{X}.$$

Usaremos la aproximación (23) para desarrollar el método de Newton bidimensional. Consideremos el sistema de ecuaciones que resulta de igualar u y v a cero en (20):

$$(24) \quad \begin{aligned} 0 &= f_1(x, y) \\ 0 &= f_2(x, y). \end{aligned}$$

Supongamos que (p, q) es una solución de (24); es decir,

$$(25) \quad \begin{aligned} 0 &= f_1(p, q) \\ 0 &= f_2(p, q). \end{aligned}$$

Si consideramos pequeños cambios de las funciones cerca de un punto inicial (p_0, q_0) próximo a la solución (p, q) :

$$(26) \quad \begin{aligned} \Delta u &= u - u_0, & \Delta p &= x - p_0, \\ \Delta v &= v - v_0, & \Delta q &= y - q_0, \end{aligned}$$

ponemos $(x, y) = (p, q)$ en (20) y usamos (25), de manera que $(u, v) = (0, 0)$, entonces los cambios en las variables dependientes son

$$(27) \quad \begin{aligned} u - u_0 &= f_1(p, q) - f_1(p_0, q_0) = 0 - f_1(p_0, q_0) \\ v - v_0 &= f_2(p, q) - f_2(p_0, q_0) = 0 - f_2(p_0, q_0). \end{aligned}$$

Ahora usamos los resultados de (27) en la aproximación lineal (22) y obtenemos

$$(28) \quad \begin{bmatrix} \frac{\partial}{\partial x} f_1(p_0, q_0) & \frac{\partial}{\partial y} f_1(p_0, q_0) \\ \frac{\partial}{\partial x} f_2(p_0, q_0) & \frac{\partial}{\partial y} f_2(p_0, q_0) \end{bmatrix} \begin{bmatrix} \Delta p \\ \Delta q \end{bmatrix} \approx - \begin{bmatrix} f_1(p_0, q_0) \\ f_2(p_0, q_0) \end{bmatrix}.$$

Si la matriz jacobiana $\mathbf{J}(p_0, q_0)$ que aparece en (28) es invertible, entonces podemos despejar $\Delta \mathbf{P} = [\Delta p \quad \Delta q]' = [p \quad q]' - [p_0 \quad q_0]'$ de manera que

$$(29) \quad \Delta \mathbf{P} \approx -\mathbf{J}(p_0, q_0)^{-1} \mathbf{F}(p_0, q_0).$$

Esto nos proporciona la siguiente aproximación \mathbf{P}_1 a la solución $\mathbf{P} = [p \quad q]$:

$$(30) \quad \mathbf{P}_1 = \mathbf{P}_0 + \Delta \mathbf{P} = \mathbf{P}_0 - \mathbf{J}(p_0, q_0)^{-1} \mathbf{F}(p_0, q_0).$$

Hagamos notar que la fórmula (30) es la generalización de la fórmula de iteración del método de Newton-Raphson para funciones de una variable que, como vimos, es $p_1 = p_0 - f(p_0)/f'(p_0)$.

Esquema del método de Netwon-Raphson

Supongamos que hemos obtenido \mathbf{P}_k .

Paso 1. Evaluamos la función

$$\mathbf{F}(\mathbf{P}_k) = \begin{bmatrix} f_1(p_k, q_k) \\ f_2(p_k, q_k) \end{bmatrix}.$$

Paso 2. Evaluamos la matriz jacobiana

$$\mathbf{J}(\mathbf{P}_k) = \begin{bmatrix} \frac{\partial}{\partial x} f_1(p_k, q_k) & \frac{\partial}{\partial y} f_1(p_k, q_k) \\ \frac{\partial}{\partial x} f_2(p_k, q_k) & \frac{\partial}{\partial y} f_2(p_k, q_k) \end{bmatrix}.$$

Paso 3. Calculamos $\Delta\mathbf{P}$ resolviendo el sistema lineal

$$\mathbf{J}(\mathbf{P}_k)\Delta\mathbf{P} = -\mathbf{F}(\mathbf{P}_k).$$

Paso 4. Calculamos el siguiente punto

$$\mathbf{P}_{k+1} = \mathbf{P}_k + \Delta\mathbf{P}.$$

Y se repite el proceso.

Ejemplo 3.32. Consideremos el sistema no lineal

$$0 = x^2 - 2x - y + 0.5$$

$$0 = x^2 + 4y^2 - 4.$$

Vamos a usar el método de Newton-Raphson tomando $(p_0, q_0) = (2.00, 0.25)$ como punto inicial y calculando (p_1, q_1) , (p_2, q_2) y (p_3, q_3) .

La función vectorial y la matriz jacobiana son

$$\mathbf{F}(x, y) = \begin{bmatrix} x^2 - 2x - y + 0.5 \\ x^2 + 4y^2 - 4 \end{bmatrix}, \quad \mathbf{J}(x, y) = \begin{bmatrix} 2x - 2 & -1 \\ 2x & 8y \end{bmatrix}$$

que, en el punto $(2.00, 0.25)$, valen

$$\mathbf{F}(2.00, 0.25) = \begin{bmatrix} 0.25 \\ 0.25 \end{bmatrix}, \quad \mathbf{J}(2.00, 0.25) = \begin{bmatrix} 2.0 & -1.0 \\ 4.0 & 2.0 \end{bmatrix}.$$

Los incrementos Δp y Δq son las soluciones del sistema lineal

$$\begin{bmatrix} 2.0 & -1.0 \\ 4.0 & 2.0 \end{bmatrix} \begin{bmatrix} \Delta p \\ \Delta q \end{bmatrix} = -\begin{bmatrix} 0.25 \\ 0.25 \end{bmatrix}.$$

Haciendo los cálculos, obtenemos

$$\Delta\mathbf{P} = \begin{bmatrix} \Delta p \\ \Delta q \end{bmatrix} = \begin{bmatrix} -0.09375 \\ 0.0625 \end{bmatrix},$$

así que el siguiente punto de la iteración es

$$\mathbf{P}_1 = \mathbf{P}_0 + \Delta\mathbf{P} = \begin{bmatrix} 2.00 \\ 0.25 \end{bmatrix} + \begin{bmatrix} -0.09375 \\ 0.0625 \end{bmatrix} = \begin{bmatrix} 1.90625 \\ 0.3125 \end{bmatrix}.$$

De manera similar se calculan los siguientes puntos, que son

$$\mathbf{P}_2 = \begin{bmatrix} 1.900691 \\ 0.311213 \end{bmatrix} \quad \text{y} \quad \mathbf{P}_3 = \begin{bmatrix} 1.900677 \\ 0.311219 \end{bmatrix}.$$

Las coordenadas de \mathbf{P}_3 tienen una precisión de seis cifras decimales. Los cálculos necesarios para hallar \mathbf{P}_2 y \mathbf{P}_3 se resumen en la Tabla 3.7. ■

Tabla 3.7 Valores funcionales, matrices jacobianas e incrementos necesarios en cada iteración del método de Newton-Raphson en el Ejemplo 3.32.

\mathbf{P}_k	Solución del sistema lineal $J(\mathbf{P}_k)\Delta\mathbf{P} = -\mathbf{F}(\mathbf{P}_k)$	$\mathbf{P}_k + \Delta\mathbf{P}$
$\begin{bmatrix} 2.00 \\ 0.25 \end{bmatrix}$	$\begin{bmatrix} 2.0 & -1.0 \\ 4.0 & 2.0 \end{bmatrix} \begin{bmatrix} -0.09375 \\ 0.0625 \end{bmatrix} = - \begin{bmatrix} 0.25 \\ 0.25 \end{bmatrix}$	$\begin{bmatrix} 1.90625 \\ 0.3125 \end{bmatrix}$
$\begin{bmatrix} 1.90625 \\ 0.3125 \end{bmatrix}$	$\begin{bmatrix} 1.8125 & -1.0 \\ 3.8125 & 2.5 \end{bmatrix} \begin{bmatrix} -0.005559 \\ -0.001287 \end{bmatrix} = - \begin{bmatrix} 0.008789 \\ 0.024414 \end{bmatrix}$	$\begin{bmatrix} 1.900691 \\ 0.311213 \end{bmatrix}$
$\begin{bmatrix} 1.900691 \\ 0.311213 \end{bmatrix}$	$\begin{bmatrix} 1.801381 & -1.000000 \\ 3.801381 & 2.489700 \end{bmatrix} \begin{bmatrix} -0.000014 \\ 0.000006 \end{bmatrix} = - \begin{bmatrix} 0.000031 \\ 0.000038 \end{bmatrix}$	$\begin{bmatrix} 1.900677 \\ 0.311219 \end{bmatrix}$

Para usar el método de Newton-Raphson es necesario calcular varias derivadas parciales. Esto puede hacerse mediante aproximaciones numéricas (veremos las técnicas en el Capítulo 6), pero hay que tener cuidado a la hora de determinar el tamaño de paso adecuado. Otra observación es que, para resolver el sistema $J(\mathbf{P}_k)\Delta\mathbf{P} = -\mathbf{F}(\mathbf{P}_k)$, habrá que utilizar las técnicas introducidas en las secciones anteriores de este capítulo; lo que nunca debe hacerse, porque supone una pérdida de tiempo enorme, es calcular la inversa de $J(\mathbf{P}_k)$ y luego despejar $\Delta\mathbf{P} = -(J(\mathbf{P}_k))^{-1}\mathbf{F}(\mathbf{P}_k)$.

MATLAB

Los Programas 3.6 (Iteración de Seidel no lineal) y 3.7 (Método de Newton-Raphson) necesitan que las funciones de los sistemas no lineales $\mathbf{X} = \mathbf{G}(\mathbf{X})$ y, respectivamente, $\mathbf{F}(\mathbf{X}) = \mathbf{0}$ así como la matriz jacobiana \mathbf{JF} de este último, se almacenen previamente como archivos con la extensión .m; por ejemplo, el sistema no lineal del Ejemplo 3.32 y la matriz jacobiana correspondiente pueden almacenarse, respectivamente, en los archivos F.m y JF.m siguientes:

```
function Z=F(X)
x=X(1);y=X(2);
Z=zeros(1,2);
Z(1)=x^2-2*x-y+0.5;
Z(2)=x^2+4*y^2-4;
```

```
function W=JF(X)
x=X(1);y=X(2);
W=[2*x-2 -1;2*x 8*y];
```

Estas funciones pueden evaluarse con las instrucciones habituales del paquete MATLAB.

```
>>A=feval('F',[2.00 0.25])
A=
0.2500 0.2500
```

```
>>V=JF([2.00 0.25])
```

```
B=
```

```
2 -1  
4 2
```

Programa 3.6 (Iteración de Seidel no lineal). Resolución del problema de punto fijo no lineal $\mathbf{X} = \mathbf{G}(\mathbf{X})$ generando, a partir de una aproximación inicial \mathbf{P}_0 , una sucesión $\{\mathbf{P}_k\}$ que converge a la solución \mathbf{P} .

```
function [P,iter] = seidel(G,P,delta, max1)  
  
% Datos  
% - G es el sistema no lineal, archivado como G.m  
% - P es el punto inicial  
% - delta es la tolerancia  
% - max1 es el número máximo de iteraciones  
% Resultados  
% - P es la aproximación a la solución que se obtiene  
% - iter es el número de iteraciones realizadas  
  
N=length(P);  
  
for k=1:max1  
    X=P;  
    % X es la k-ésima aproximación a la solución  
    for j=1:N  
        A=feval('G',X);  
        % Las coordenadas de X se actualizan conforme se calculan  
        X(j)=A(j);  
    end  
    err=abs(norm(X-P));  
    relerr=err/(norm(X)+eps);  
    P=X;  
    iter=k;  
    if(err<delta)|(relerr<delta)  
        break  
    end  
end
```

En el siguiente programa se utiliza la instrucción $\mathbf{A}\backslash\mathbf{B}$ del paquete MATLAB para resolver el sistema lineal $\mathbf{AX} = \mathbf{B}$ (como se ve en la línea $\mathbf{Q}=\mathbf{P}-(\mathbf{J}\backslash\mathbf{Y}')'$). Otra posibilidad sería emplear los programas desarrollados en las secciones anteriores de este capítulo. La elección del programa adecuado para resolver el sistema lineal depende del tamaño y características de la matriz jacobiana.

Programa 3.7 (Método de Newton-Raphson). Resolución del sistema de ecuaciones no lineales $\mathbf{F}(\mathbf{X}) = \mathbf{0}$, generando, a partir de una aproximación inicial \mathbf{P}_0 , una sucesión $\{\mathbf{P}_k\}$ que converge a la solución \mathbf{P} .

```

function [P,iter,err]=newdim(F,JF,P,delta,epsilon,max1)
% Datos
%   - F es el sistema no lineal, archivado como F.m
%   - JF es la matriz jacobiana de F, archivada como JF.m
%   - P es el punto inicial
%   - delta es la tolerancia para P
%   - epsilon es la tolerancia para F(P)
%   - maxi es el número máximo de iteraciones
% Resultados
%   - P es la aproximación a la solución que se obtiene
%   - iter es el número de iteraciones realizadas
%   - err es la estimación del error de P.
Y=feval(F,P);
for k=1:max1
    J=feval(JF,P);
    Q=P-(J\Y)';
    Z=feval(F,Q);
    err=norm(Q-P);
    relerr=err/(norm(Q)+eps);
    P=Q;
    Y=Z;
    iter=k;
    if (err<delta)|(relerr<delta)|(abs(Y)<epsilon)
        break
    end
end

```

Ejercicios

- Determine (analíticamente) los puntos fijos de cada uno de los sistemas que se relacionan a continuación.
 - $x = g_1(x, y) = x - y^2$
 $y = g_2(x, y) = -x + 6y$
 - $x = g_1(x, y) = (x^2 - y^2 - x - 3)/3$
 $y = g_2(x, y) = (-x + y - 1)/3$
 - $x = g_1(x, y) = \operatorname{sen}(y)$
 $y = g_2(x, y) = -6x + y$

- (d) $x = g_1(x, y, z) = 9 - 3y - 2z$
 $y = g_2(x, y, z) = 2 - x + z$
 $z = g_3(x, y, z) = -9 + 3x + 4y - z$
2. Determine (analíticamente) los ceros de cada uno de los sistemas que se relacionan a continuación y evalúe la matriz jacobiana de cada sistema en el cero correspondiente
- (a) $0 = f_1(x, y) = 2x + y - 6$
 $0 = f_2(x, y) = x + 2y$
- (b) $0 = f_1(x, y) = 3x^2 + 2y - 4$
 $0 = f_2(x, y) = 2x + 2y - 3$
- (c) $0 = f_1(x, y) = 2x - 4 \cos(y)$
 $0 = f_2(x, y) = 4x \operatorname{sen}(y)$
- (d) $0 = f_1(x, y, z) = x^2 + y^2 - z$
 $0 = f_2(x, y, z) = x^2 + y^2 + z^2 - 1$
 $0 = f_3(x, y, z) = x + y$
3. Determine una región en el plano XOY tal que la iteración de punto fijo aplicada al sistema

$$\begin{aligned}x &= g_1(x, y) = (x^2 - y^2 - x - 3)/3 \\y &= g_2(x, y) = (x + y + 1)/3.\end{aligned}$$

sea convergente para cualquier punto inicial (p_0, q_0) de dicha región (use un argumento similar al empleado en el Teorema 3.17).

4. Escriba el siguiente sistema de ecuaciones lineales como un problema de punto fijo y determine cotas para x , y y z de manera que la iteración de punto fijo sea convergente para cualquier punto inicial (p_0, q_0, r_0) que satisfaga dichas cotas.

$$\begin{aligned}6x + y + z &= 1 \\x + 4y + z &= 2 \\x + y + 5z &= 0\end{aligned}$$

5. Dado el sistema no lineal

$$\begin{aligned}x &= g_1(x, y) = \frac{8x - 4x^2 + y^2 + 1}{8} && \text{(hipérbola)} \\y &= g_2(x, y) = \frac{2x - x^2 + 4y - y^2 + 3}{4} && \text{(circunferencia)},\end{aligned}$$

use la aproximación inicial $(p_0, q_0) = (1.1, 2.0)$ y calcule las tres siguientes aproximaciones al punto fijo mediante (a) la iteración de punto fijo con las fórmulas (14) y (b) el método iterativo de Seidel con las fórmulas (18) (véase la Figura 3.7).

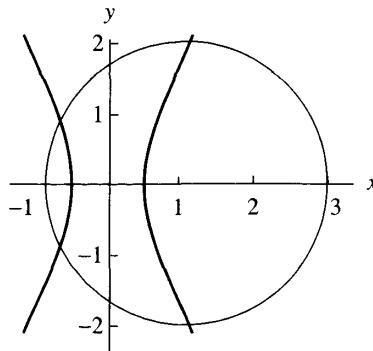


Figura 3.7 La hipérbola y la circunferencia del Ejercicio 5.

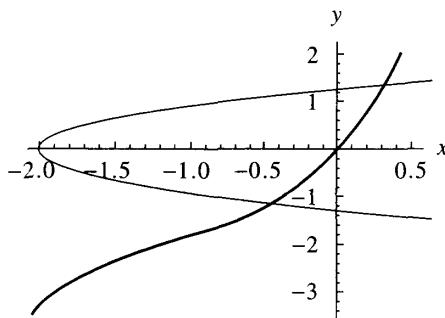


Figura 3.8 La cúbica y la parábola del Ejercicio 6.

6. Dado el sistema no lineal

$$\begin{aligned} x &= g_1(x, y) = \frac{y - x^3 + 3x^2 + 3x}{7} && \text{(cúbica)} \\ y &= g_2(x, y) = \frac{y^2 + 2y - x - 2}{2} && \text{(parábola),} \end{aligned}$$

use la aproximación inicial $(p_0, q_0) = (-0.3, -1.3)$ y calcule las tres siguientes aproximaciones al punto fijo mediante **(a)** la iteración de punto fijo con las fórmulas (14) y **(b)** el método iterativo de Seidel con las fórmulas (18) (véase la Figura 3.8).

7. Consideremos el sistema de ecuaciones no lineal

$$0 = f_1(x, y) = x^2 - y - 0.2$$

$$0 = f_2(x, y) = y^2 - x - 0.3.$$

Estas paráolas se cortan en dos puntos como se muestra en la Figura 3.9.

- (a)** Empiece con $(p_0, q_0) = (1.2, 1.2)$ las iteraciones del método de Newton-Raphson y calcule (p_1, q_1) y (p_2, q_2) .

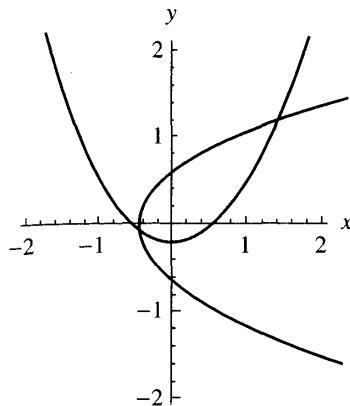


Figura 3.9 Las parábolas del Ejercicio 7.

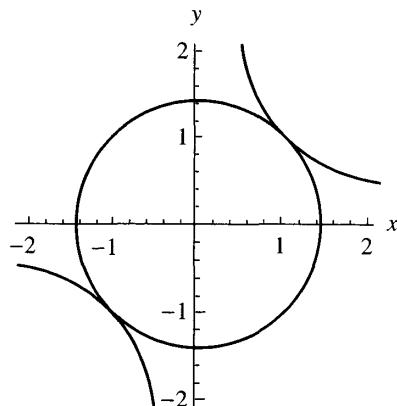


Figura 3.10 La circunferencia y la hipérbola del Ejercicio 8.

- (b) Empiece con $(p_0, q_0) = (-0.2, -0.2)$ las iteraciones del método de Newton-Raphson y calcule (p_1, q_1) y (p_2, q_2) .
8. Consideremos el sistema no lineal cuya gráfica se muestra en la Figura 3.10.
- $$0 = f_1(x, y) = x^2 + y^2 - 2$$
- $$0 = f_2(x, y) = xy - 1.$$
- (a) Compruebe que las soluciones son $(1, 1)$ y $(-1, -1)$.
(b) ¿Qué dificultades podrían aparecer si tratamos de aplicar el método de Newton-Raphson para hallar las soluciones?
9. Pruebe que el método iterativo de Jacobi para un sistema lineal de orden 3×3 es un caso especial del método de iteración de punto fijo (15). Más aún, pruebe que si la matriz de los coeficientes de un sistema lineal de orden 3×3 es de diagonal estrictamente dominante, entonces la condición (17) se cumple.
10. Pruebe que el método iterativo de Newton-Raphson para dos ecuaciones puede escribirse como un problema de punto fijo

$$x = g_1(x, y), \quad y = g_2(x, y),$$

siendo $g_1(x, y)$ y $g_2(x, y)$ las funciones

$$g_1(x, y) = x - \frac{f_1(x, y) \frac{\partial}{\partial y} f_2(x, y) - f_2(x, y) \frac{\partial}{\partial y} f_1(x, y)}{\det(\mathbf{J}(x, y))}$$

$$g_2(x, y) = y - \frac{f_2(x, y) \frac{\partial}{\partial x} f_1(x, y) - f_1(x, y) \frac{\partial}{\partial x} f_2(x, y)}{\det(\mathbf{J}(x, y))}.$$

11. Supongamos que se emplea el método de iteración de punto fijo para resolver el sistema (12). Pruebe, dando la secuencia de pasos que se relacionan más abajo, que las condiciones dadas en (16) son suficientes para garantizar que la sucesión $\{(p_k, q_k)\}$ converge a (p, q) . Supongamos que existe una constante K con $0 < K < 1$ tal que

$$\left| \frac{\partial}{\partial x} g_1(x, y) \right| + \left| \frac{\partial}{\partial y} g_1(x, y) \right| < K$$

y

$$\left| \frac{\partial}{\partial x} g_2(x, y) \right| + \left| \frac{\partial}{\partial y} g_2(x, y) \right| < K$$

para todo (x, y) en el rectángulo $R = \{(x, y) : a < x < b, c < y < d\}$. Supongamos también que $a < p_0 < b$ y que $c < q_0 < d$. Ahora se definen

$$e_k = p - p_k, \quad E_k = q - q_k \quad y \quad r_k = \max\{|e_k|, |E_k|\}.$$

Será necesario usar también la siguiente forma del teorema del valor medio para funciones de dos variables:

$$\begin{aligned} e_{k+1} &= \frac{\partial}{\partial x} g_1(a_k^*, q_k) e_k + \frac{\partial}{\partial y} g_1(p, c_k^*) E_k \\ E_{k+1} &= \frac{\partial}{\partial x} g_2(b_k^*, q_k) e_k + \frac{\partial}{\partial y} g_2(p, d_k^*) E_k, \end{aligned}$$

donde a_k^* y b_k^* están en $[a, b]$ y c_k^* y d_k^* están en $[c, d]$. Los pasos que hay que ir demostrando son:

- (a) $|e_1| \leq Kr_0$ y $|E_1| \leq Kr_0$
- (b) $|e_2| \leq Kr_1 \leq K^2r_0$ y $|E_2| \leq Kr_1 \leq K^2r_0$
- (c) $|e_k| \leq Kr_{k-1} \leq K^kr_0$ y $|E_k| \leq Kr_{k-1} \leq K^kr_0$
- (d) $\lim_{n \rightarrow \infty} p_k = p$ y $\lim_{n \rightarrow \infty} q_k = q$

12. Como se indicó más arriba, la matriz jacobiana del sistema (20) es el análogo bidimensional de la derivada. Escribamos el sistema (20) como una función vectorial $\mathbf{V} = \mathbf{F}(\mathbf{X})$ y denotemos por $\mathbf{J}(\mathbf{F})$ la matriz jacobiana de este sistema. Dados dos sistemas no lineales $\mathbf{V} = \mathbf{F}(\mathbf{X})$ y $\mathbf{V} = \mathbf{G}(\mathbf{X})$ y un número real c , pruebe que

- (a) $\mathbf{J}(c\mathbf{F}(\mathbf{X})) = c\mathbf{J}(\mathbf{F}(\mathbf{X}))$
- (b) $\mathbf{J}(\mathbf{F}(\mathbf{X}) + \mathbf{G}(\mathbf{X})) = \mathbf{J}(\mathbf{F}(\mathbf{X})) + \mathbf{J}(\mathbf{G}(\mathbf{X}))$

Algoritmos y programas

1. Use el Programa 3.6 para aproximar los puntos fijos de los Ejercicios 5 y 6 con una precisión de diez cifras decimales.

2. Use el Programa 3.7 para aproximar los puntos fijos de los Ejercicios 7 y 8 con una precisión de diez cifras decimales.
3. Construya un programa para hallar los puntos fijos de un sistema usando el método de iteración de punto fijo. Use su programa para aproximar los puntos fijos de los sistemas de los Ejercicios 5 y 6 con una precisión de ocho cifras decimales.
4. Use el Programa 3.7 para dar aproximaciones de los ceros de los siguientes sistemas con una precisión de diez cifras decimales.
 - (a) $0 = x^2 - x + y^2 + z^2 - 5$
 $0 = x^2 + y^2 - y + z^2 - 4$
 $0 = x^2 + y^2 + z^2 + z - 6$
 - (b) $0 = x^2 - x + 2y^2 + yz - 10$
 $0 = 5x - 6y + z$
 $0 = z - x^2 - y^2$
 - (c) $0 = (x + 1)^2 + (y + 1)^2 - z$
 $0 = (x - 1)^2 + y^2 - z$
 $0 = 4x^2 + 2y^2 + z^2 - 16$
 - (d) $0 = 9x^2 + 36y^2 + 4z^2 - 36$
 $0 = x^2 - 2y^2 - 20z$
 $0 = 16x - x^3 - 2y^2 - 16z^2$

5. Queremos resolver el sistema no lineal

$$\begin{aligned} 0 &= 7x^3 - 10x - y - 1 \\ 0 &= 8y^3 - 11y + x - 1. \end{aligned}$$

Use las instrucciones adecuadas del paquete de programas MATLAB para dibujar las gráficas de ambas curvas sobre un mismo sistema de coordenadas y compruebe que hay nueve puntos donde se cruzan ambas gráficas. Estime, a la vista del dibujo, cuáles son estos puntos. Utilice luego estas estimaciones como puntos iniciales en el Programa 3.7 para aproximar los puntos de intersección de las curvas con una precisión de nueve cifras decimales.

6. El sistema del Problema 5 puede escribirse como un problema de punto fijo

$$\begin{aligned} x &= \frac{7x^3 - y - 1}{10} \\ y &= \frac{8y^3 + x - 1}{11}. \end{aligned}$$

Realice experimentos con el computador para probar que, sea cual sea el punto de partida de la iteración, sólo uno de los nueve puntos puede ser逼近ado usando esta función particular de iteración de punto fijo. ¿Hay otras funciones de iteración que permitan逼近ar otras soluciones del sistema?

Interpolación y aproximación polinomial

Los procedimientos de cálculo que usan los computadores para evaluar una función ya incorporada, como $\sin(x)$, $\cos(x)$ o e^x , involucran aproximación mediante polinomios. Los métodos hoy por hoy efectivos utilizan funciones racionales (que son los cocientes de polinomios); sin embargo, la teoría de la aproximación polinomial es más adecuada para un primer curso de cálculo numérico y será la que consideraremos principalmente en este capítulo. Supongamos que queremos aproximar la función $f(x) = e^x$ mediante un polinomio de grado $n = 2$ en el intervalo $[-1, 1]$. El polinomio de Taylor de f en el origen

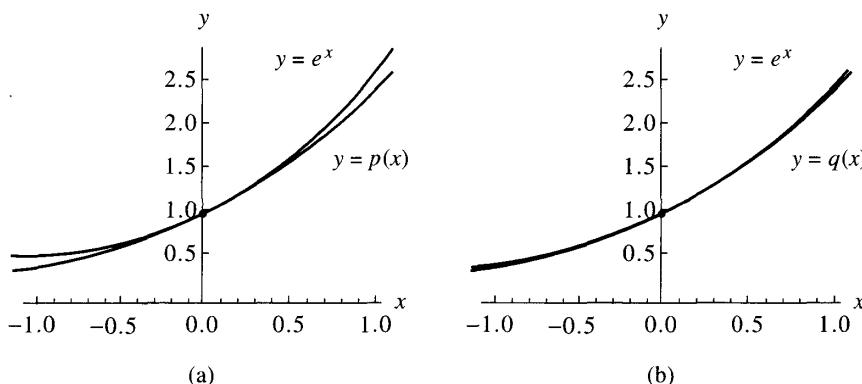


Figura 4.1 (a) El polinomio de Taylor $p(x) = 1 + x + 0.5x^2$ que aproxima a $f(x) = e^x$ en $[-1, 1]$. (b) La aproximación de Chebyshev $q(x) = 1 + 1.129772x + 0.532042x^2$ para $f(x) = e^x$ en $[-1, 1]$.

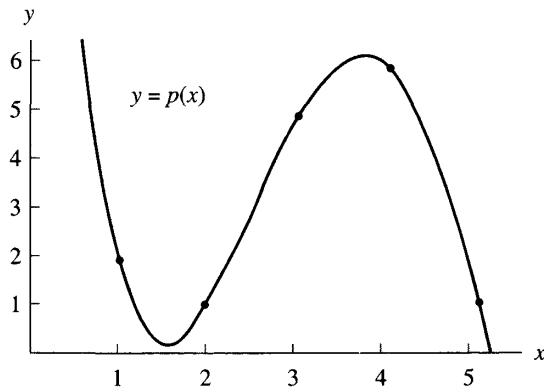


Figura 4.2 La gráfica del polinomio de interpolación que pasa por los puntos $(1, 2)$, $(2, 1)$, $(3, 5)$, $(4, 6)$ y $(5, 1)$.

se muestra en la Figura 4.1(a) y podemos compararlo con la aproximación de Chebyshev mostrada en la Figura 4.1(b). El error máximo en la aproximación por el polinomio de Taylor es 0.218282, mientras que el error máximo en la aproximación por el polinomio de Chebyshev es 0.056468. En este capítulo desarrollaremos la teoría básica necesaria para investigar estas cuestiones.

Un problema asociado es el de la interpolación¹ polinomial. Dados $n+1$ puntos en el plano (sin que haya dos en la misma recta vertical), el polinomio interpolador es el único polinomio de grado menor o igual que n que pasa por dichos puntos. Éste puede ser el caso de un conjunto de datos calculados con una cierta precisión. Son varios los métodos que podemos usar para construir el polinomio de interpolación: resolver un sistema de ecuaciones lineales para hallar sus coeficientes, usar los polinomios de Lagrange o construir una tabla de diferencias divididas para emplearla con los coeficientes del polinomio de Newton. Es importante que quienes deseen usar el cálculo numérico conozcan estas tres técnicas. Por ejemplo, el polinomio de interpolación de grado $n=4$ que pasa por los cinco puntos $(1, 2)$, $(2, 1)$, $(3, 5)$, $(4, 6)$ y $(5, 1)$ es

$$P(x) = \frac{5x^4 - 82x^3 + 427x^2 - 806x + 504}{24};$$

en la Figura 4.2 se muestran dichos puntos y la gráfica del polinomio.

¹ *N del T.* Los autores distinguen entre “collocation polynomial”, el que pasa por unos puntos dados cualesquiera, e “interpolation polynomial”, cuando los puntos dados están en la gráfica de una función fijada de antemano. En nuestra lengua no se suele hacer tal distinción.

Tabla 4.1 Desarrollos en serie de Taylor más comunes.

$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$	para todo x
$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots$	para todo x
$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots$	para todo x
$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots$	$-1 < x \leq 1$
$\arctan(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots$	$-1 < x \leq 1$
$(1+x)^p = 1 + px + \frac{p(p-1)}{2!}x^2 + \frac{p(p-1)(p-2)}{3!}x^3 + \dots$	para $ x < 1$

4.1 Series de Taylor y cálculo de los valores de una función

La noción de límite es la base del cálculo infinitesimal. Por ejemplo, la derivada

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

es el límite del cociente incremental cuando el numerador como el denominador tienden a cero. Una serie de Taylor ilustra otro tipo de límite; en este caso, se suman un número infinito de términos tomando el límite de sus sumas parciales. Una importante aplicación de las series de Taylor es su utilización para representar las funciones elementales: $\sin(x)$, $\cos(x)$, e^x , $\ln(x)$, etc.; en la Tabla 4.1 se muestran algunos de los desarrollos de Taylor más comunes. Las sumas parciales de la serie de Taylor se van calculando hasta que se consigue una aproximación a la función que tiene la precisión deseada. En otro orden de cosas, en los campos de las ingenierías y la física es habitual encontrarse con soluciones de los problemas expresadas como series.

¿Cómo podemos usar una suma finita para obtener una buena aproximación de la suma infinita? Vamos a usar, como ejemplo de ilustración, la serie de Taylor de la función exponencial dada en la Tabla 4.1 para calcular el número $e = e^1$, la base de los logaritmos neperianos y de la función exponencial. Tomando $x = 1$ y usando la serie obtenemos

$$e^1 = 1 + \frac{1}{1!} + \frac{1^2}{2!} + \frac{1^3}{3!} + \frac{1^4}{4!} + \dots + \frac{1^k}{k!} + \dots$$

Tabla 4.2 Sumas parciales S_n usadas para calcular e .

n	$S_n = 1 + \frac{1}{1!} + \frac{1}{2!} + \cdots + \frac{1}{n!}$
0	1.0
1	2.0
2	2.5
3	2.666666666666 ...
4	2.708333333333 ...
5	2.716666666666 ...
6	2.718055555555 ...
7	2.718253968254 ...
8	2.718278769841 ...
9	2.718281525573 ...
10	2.718281801146 ...
11	2.718281826199 ...
12	2.718281828286 ...
13	2.718281828447 ...
14	2.718281828458 ...
15	2.718281828459 ...

La definición de suma de una serie dada en la Sección 1.1 requiere que las sumas parciales S_N tiendan a un límite. Para el caso que nos atañe, las sumas parciales se muestran en la Tabla 4.2.

Una forma natural de pensar en la representación en serie de potencias de una función es considerarla como el límite de una sucesión de polinomios de grado creciente, de manera que si se añaden suficientes términos, entonces podemos conseguir una aproximación aceptable. Precisando más, ¿qué grado debe tener el polinomio? y ¿cómo calculamos los coeficientes de las potencias de x en el polinomio? El Teorema 4.1 responde estas cuestiones.

Teorema 4.1 (Aproximación por polinomios de Taylor). Consideremos una función $f \in C^{N+1}[a, b]$ y un punto $x_0 \in [a, b]$. Si $x \in [a, b]$, entonces

$$(1) \quad f(x) = P_N(x) + E_N(x),$$

donde $P_N(x)$ es un polinomio que podemos usar para aproximar $f(x)$:

$$(2) \quad f(x) \approx P_N(x) = \sum_{k=0}^N \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k,$$

llamado polinomio de Taylor de grado N de f alrededor de x_0 , y el término del

error $E_N(x)$ se puede escribir como

$$(3) \quad E_N(x) = \frac{f^{(N+1)}(c)}{(N+1)!} (x - x_0)^{N+1}$$

para algún valor $c = c(x)$ que está entre x y x_0 .

Demostración. Dejamos la demostración como ejercicio. •

La relación (2) indica cómo se calculan los coeficientes del polinomio de Taylor. Por otro lado, aunque el término del error (3) se expresa de forma similar, el hecho de que debamos evaluar $f^{(N+1)}(c)$ en un número c que depende de x y que no conocemos, nos lleva a que no tratemos de calcular $E_N(x)$; lo que hacemos es usar su expresión para determinar una cota de la precisión de la aproximación.

Ejemplo 4.1. Veamos por qué, en la Tabla 4.2, es suficiente con 15 términos para obtener una aproximación a $e = 2.718281828459$ que tiene una precisión de 13 cifras decimales.

Vamos a desarrollar $f(x) = e^x$ alrededor del punto $x_0 = 0$ en serie de potencias $(x - 0)^k = x^k$ hasta el polinomio de grado 15. Las derivadas que necesitamos son $f'(x) = f''(x) = \dots = f^{(16)}(x) = e^x$. Las quince primeras se usan para calcular los coeficientes $a_k = e^0/k!$ y escribir

$$(4) \quad P_{15}(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots + \frac{x^{15}}{15!}.$$

Poniendo $x = 1$ en (4) nos da la suma parcial $S_{15} = P_{15}(1)$. El término del error sirve para determinar la precisión de la aproximación:

$$(5) \quad E_{15}(x) = \frac{f^{(16)}(c)x^{16}}{16!}.$$

Puesto que hemos elegido $x_0 = 0$ y $x = 1$, el valor de c está entre ellos (o sea, $0 < c < 1$), por lo que $e^c < e^1$. Las sumas parciales de la Tabla 4.2 están acotadas superiormente por 3. Combinando estas desigualdades, obtenemos $e^c < 3$ que, a su vez, implica

$$|E_{15}(1)| = \frac{|f^{(16)}(c)|}{16!} \leq \frac{e^c}{16!} < \frac{3}{16!} < 1.433844 \times 10^{-13}.$$

En consecuencia, todas las cifras de la aproximación $e \approx 2.718281828459$ son exactas porque el error (valga lo que valga) debe ser menor que 2 en el decimotercer lugar decimal. ■

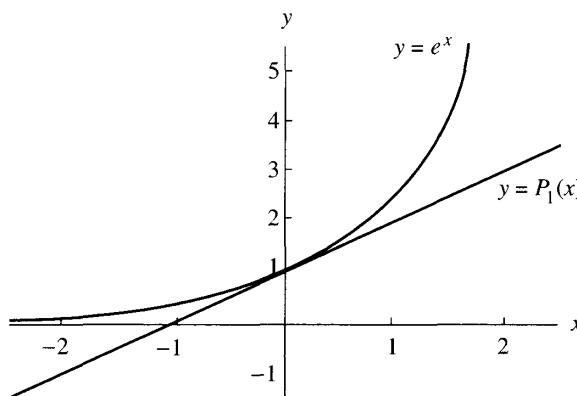


Figura 4.3 Gráficas de $y = e^x$ e $y = P_1(x) = 1 + x$.

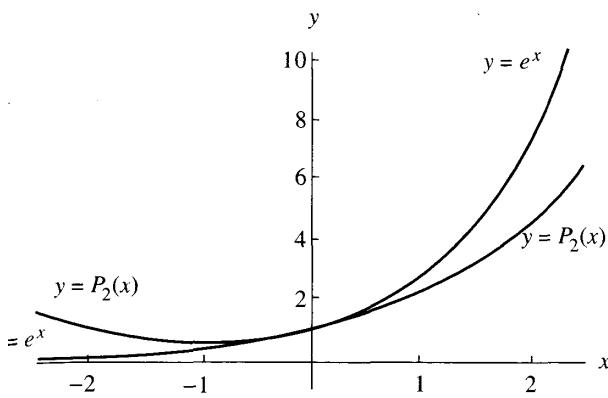
Mejor que dar una prueba rigurosa del Teorema 4.1, que puede encontrarse en casi cualquier texto de cálculo infinitesimal, vamos a discutir algunos aspectos interesantes de esta aproximación. Para ilustrarlos, seguimos usando como ejemplo la función $f(x) = e^x$ y el valor $x_0 = 0$. Recordando nuestros conocimientos de cálculo infinitesimal, sabemos que la pendiente de la curva $y = e^x$ en el punto (x, e^x) es $f'(x) = e^x$; por tanto, la pendiente de la curva en el punto $(0, 1)$ es $f'(0) = 1$. En consecuencia, la recta tangente a la curva en el punto $(0, 1)$ es $y = 1 + x$, que es la misma fórmula que se obtendría usando $N = 1$ en el Teorema 4.1; esto es, $P_1(x) = f(0) + f'(0)x/1! = 1 + x$. En otras palabras, $y = P_1(x)$ es la ecuación de la recta tangente a la curva en ese punto; las gráficas de estas curvas se muestran en la Figura 4.3.

Observemos que la aproximación $e^x \approx 1 + x$ es buena cerca del centro del desarrollo $x_0 = 0$ y que la distancia entre ambas curvas aumenta conforme x se aleja de 0. Hagamos notar que ambas tienen la misma pendiente en $(0, 1)$. En cálculo infinitesimal también aprendimos que la segunda derivada indica si una curva es convexa o cóncava. El estudio de la curvatura² muestra que si dos curvas $y = f(x)$ e $y = g(x)$ tienen la propiedad de que $f(x_0) = g(x_0)$, $f'(x_0) = g'(x_0)$ y $f''(x_0) = g''(x_0)$, entonces ambas tienen la misma curvatura en x_0 ; propiedad que sería deseable que tuviera una función polinomial que aproxime a $f(x)$. El Corolario 4.1 prueba que el polinomio de Taylor tiene esta propiedad para $N \geq 2$.

Corolario 4.1. Si $P_N(x)$ es el polinomio de Taylor de grado N dado en el Teorema 4.1, entonces

$$(6) \quad P_N^{(k)}(x_0) = f^{(k)}(x_0) \quad \text{para } k = 0, 1, \dots, N.$$

² La curvatura κ de una gráfica $y = f(x)$ en uno de sus puntos (x_0, y_0) se define como $\kappa = |f''(x_0)|/(1 + [f'(x_0)]^2)^{3/2}$.

Figura 4.4 Gráficas de $y = e^x$ e $y = P_2(x) = 1 + x + x^2/2$.

Demostración. Ponemos $x = x_0$ en las ecuaciones (2) y (3), y el resultado es $P_N(x_0) = f(x_0)$. Esto prueba que (6) es cierto para $k = 0$. Derivando el miembro derecho de (2) obtenemos

$$(7) \quad P'_N(x) = \sum_{k=1}^N \frac{f^{(k)}(x_0)}{(k-1)!} (x-x_0)^{k-1} = \sum_{k=0}^{N-1} \frac{f^{(k+1)}(x_0)}{k!} (x-x_0)^k.$$

Ponemos $x = x_0$ en (7) y obtenemos $P'_N(x_0) = f'(x_0)$. Esto prueba que (6) es cierto para $k = 1$. Derivando sucesivamente en (7) se obtienen las demás igualdades de (6); los detalles se dejan como ejercicio. •

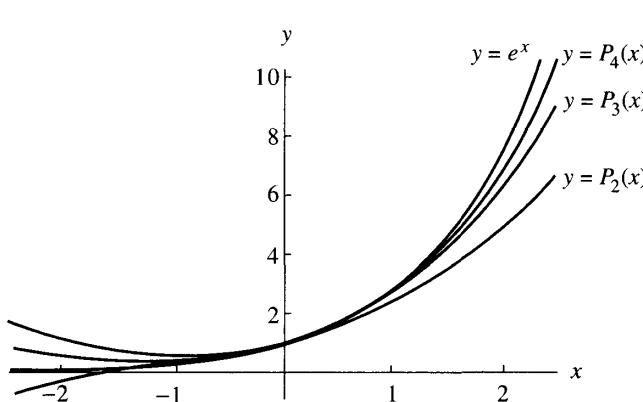
Aplicando el Corolario 4.1, vemos que para $y = P_2(x)$ se verifica que $f(x_0) = P_2(x_0)$, $f'(x_0) = P'_2(x_0)$ y $f''(x_0) = P''_2(x_0)$; por tanto, las gráficas de f y P_2 tienen la misma curvatura en x_0 . Volviendo a nuestro ejemplo, tenemos $f(x) = e^x$ y $P_2(x) = 1 + x + x^2/2$ cuyas gráficas se muestran en la Figura 4.4, donde vemos que ambas se doblan de la misma manera en $(0, 1)$.

En la teoría de la aproximación, el objetivo es encontrar una aproximación polinomial aceptablemente precisa a una función analítica³ $f(x)$ en un intervalo $[a, b]$. Ésta es una técnica que se emplea en el desarrollo del software de un computador. La precisión del polinomio de Taylor aumenta cuando elegimos N grande y, generalmente, decrece cuando el valor de x se aleja del centro del desarrollo x_0 ; por tanto, debemos elegir N suficientemente grande y restringir el valor máximo de $|x - x_0|$ para que el error no exceda una cota especificada. Así, si tomamos $2R$ como anchura de un intervalo centrado en x_0 (o sea, $|x - x_0| < R$),

³ Se dice que una función $f(x)$ es analítica en x_0 si admite derivadas continuas de todos los órdenes y puede representarse como una serie de potencias alrededor de x_0 en un intervalo centrado en dicho punto.

Tabla 4.3 Valores de la cota del error $|\text{error}| < e^R R^{N+1} / (N + 1)!$ mediante la aproximación $e^x \approx P_N(x)$ para $|x| \leq R$.

	$R = 2.0, x \leq 2.0$	$R = 1.5, x \leq 1.5$	$R = 1.0, x \leq 1.0$	$R = 0.5, x \leq 0.5$
$e^x \approx P_5(x)$	0.65680499	0.07090172	0.00377539	0.00003578
$e^x \approx P_6(x)$	0.18765857	0.01519323	0.00053934	0.00000256
$e^x \approx P_7(x)$	0.04691464	0.00284873	0.00006742	0.00000016
$e^x \approx P_8(x)$	0.01042548	0.00047479	0.00000749	0.00000001


Figura 4.5 Las gráficas de $y = e^x$, $y = P_2(x)$, $y = P_3(x)$ e $y = P_4(x)$.

entonces el valor absoluto del error satisface la relación

$$(8) \quad |\text{error}| = |E_N(x)| \leq \frac{MR^{N+1}}{(N + 1)!},$$

siendo $M = \max\{|f^{(N+1)}(z)| : x_0 - R \leq z \leq x_0 + R\}$. Si las derivadas están uniformemente acotadas, entonces la cota del error dada en (8) es proporcional a $R^{N+1}/(N + 1)!$ y decrece cuando N se hace grande, si R está fijo, o cuando R tiende a cero, si N está fijo. La Tabla 4.3 muestra el efecto de la elección de estos parámetros en la precisión de la aproximación $e^x \approx P_N(x)$ sobre el intervalo $|x| \leq R$. El error más pequeño se consigue para el mayor valor de N y el menor de R . Las gráficas de P_2 , P_3 y P_4 se muestran en la Figura 4.5.

Ejemplo 4.2. Vamos a establecer cotas del error para la aproximación $e^x \approx P_8(x)$ en los intervalos $|x| \leq 1.0$ y $|x| \leq 0.5$. Si $|x| \leq 1.0$, entonces tomando $R = 1.0$ y

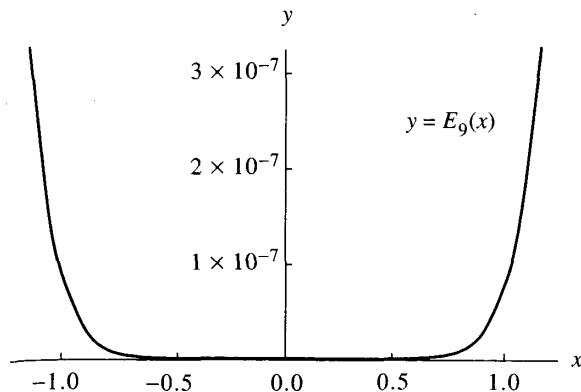


Figura 4.6 La gráfica del error $y = E_9(x) = e^x - P_9(x)$.

$|f^{(9)}(c)| = |e^c| \leq e^{1.0} = M$ en (8), obtenemos

$$|\text{error}| = |E_8(x)| \leq \frac{e^{1.0}(1.0)^9}{9!} \approx 0.00000749.$$

Si $|x| \leq 0.5$, entonces tomando $R = 0.5$ y $|f^{(9)}(c)| = |e^c| \leq e^{0.5} = M$ en (8), obtenemos

$$|\text{error}| = |E_8(x)| \leq \frac{e^{0.5}(0.5)^9}{9!} \approx 0.00000001.$$

Ejemplo 4.3. Para $f(x) = e^x$, vamos a probar que $N = 9$ es el menor número natural tal que $|\text{error}| = |E_N(x)| \leq 0.0000005$ para x en $[-1, 1]$, con lo cual podremos usar $P_9(x)$ para aproximar los valores de e^x con una precisión de seis cifras decimales.

Queremos hallar el menor número natural N tal que

$$|\text{error}| = |E_N(x)| \leq \frac{e^c(1)^{N+1}}{(N+1)!} < 0.0000005.$$

En el Ejemplo 4.2 vimos que $N = 8$ era demasiado pequeño, así que lo intentamos con $N = 9$ y descubrimos que $|E_9(x)| \leq e^1(1)^{9+1}/(9+1)! \leq 0.000000749$. Esta cota es ligeramente mayor que la deseada; así que cabría inclinarse por elegir $N = 10$. Sin embargo, hemos usado $e^c \leq e^1$, que es una estimación bastante cruda a la hora de hallar la cota del error, de forma que 0.000000749 es un poco mayor que el verdadero error. En la Figura 4.6 se muestra la gráfica de $E_9(x) = e^x - P_9(x)$; nótese que el valor máximo es, más o menos, 3×10^{-7} y que se alcanza en el extremo derecho ($1, E_9(1)$). De hecho, la cota del error en el intervalo es $E_9(1) = 2.718281828 - 2.718281526 \approx 3.024 \times 10^{-7}$, por lo que la elección de $N = 9$ está justificada.

Terminamos esta sección con el teorema que relaciona las series de Taylor de la Tabla 4.1 con los polinomios de Taylor del Teorema 4.1.

Teorema 4.2 (Serie de Taylor). Supongamos que $f(x)$ admite derivadas continuas de todos los órdenes en un intervalo (a, b) en el que está el punto x_0 . Supongamos que la sucesión de polinomios de Taylor (2) converge a $f(x)$, o sea,

$$(9) \quad f(x) = \lim_{N \rightarrow \infty} P_N(x) = \lim_{N \rightarrow \infty} \sum_{k=0}^N \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k$$

para todo $x \in (a, b)$, entonces f es analítica y puede desarrollarse en serie de Taylor alrededor de x_0

$$(10) \quad f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k.$$

Demuestração. Basta aplicar la definición de convergencia de series de la Sección 1.1. La condición de la existencia de límite se suele expresar diciendo que el término del error debe tender a cero cuando N tiende a infinito. Por tanto, una condición necesaria y suficiente para que se verifique (10) es que

$$(11) \quad \lim_{N \rightarrow \infty} E_N(x) = \lim_{N \rightarrow \infty} \frac{f^{(N+1)}(c)(x - x_0)^{N+1}}{(N + 1)!} = 0,$$

donde c depende de N y x .

El ejemplo típico de una función cuya serie de Taylor converge pero su suma no coincide con la propia función es

$$f(x) = \begin{cases} e^{-1/x^2} & \text{si } x \neq 0, \\ 0 & \text{si } x = 0. \end{cases}$$

Todas las derivadas de $f(x)$ son nulas en $x_0 = 0$, así que sus polinomios de Taylor son todos iguales al polinomio nulo.

Ejercicios

1. Aplique el Teorema 4.1 a la función $f(x) = \operatorname{sen}(x)$.

- (a) Use $x_0 = 0$ y calcule $P_5(x)$, $P_7(x)$ y $P_9(x)$.
 (b) Pruebe que si $|x| \leq 1$ entonces la aproximación

$$\operatorname{sen}(x) \approx x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!}$$

tiene como cota del error $|E_9(x)| < 1/10! \leq 2.75574 \times 10^{-7}$.

- (c) Use $x_0 = \pi/4$ y calcule $P_5(x)$, en el que aparecen potencias de $(x - \pi/4)$.

2. Aplique el Teorema 4.1 a la función $f(x) = \cos(x)$.

(a) Use $x_0 = 0$ y calcule $P_4(x)$, $P_6(x)$ y $P_8(x)$.

(b) Pruebe que si $|x| \leq 1$ entonces la aproximación

$$\cos(x) \approx 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!}$$

tiene como cota del error $|E_8(x)| < 1/9! \leq 2.75574 \times 10^{-6}$.

(c) Use $x_0 = \pi/4$ y calcule $P_4(x)$, en el que aparecen potencias de $(x - \pi/4)$.

3. ¿Tiene $f(x) = x^{1/2}$ desarrollo en serie de Taylor alrededor de $x_0 = 0$? Justifique su respuesta. ¿Y alrededor de $x_0 = 1$? Justifique su respuesta.

4. (a) Halle el polinomio de Taylor de grado $N = 5$ de la función definida por $f(x) = 1/(1+x)$ alrededor de $x_0 = 0$.

(b) Halle el término del error $E_5(x)$ para el polinomio del apartado (a).

5. Halle el polinomio de Taylor de grado $N = 3$ de la función $f(x) = e^{-x^2/2}$ alrededor de $x_0 = 0$.

6. Halle el polinomio de Taylor $P_3(x)$ de grado $N = 3$ de la función dada por $f(x) = x^3 - 2x^2 + 2x$ alrededor de $x_0 = 1$. Pruebe que $f(x) = P_3(x)$.

7. (a) Halle el polinomio de Taylor de grado $N = 5$ de la función $f(x) = x^{1/2}$ alrededor de $x_0 = 4$.

(b) Halle el polinomio de Taylor de grado $N = 5$ de la función $f(x) = x^{1/2}$ alrededor de $x_0 = 9$.

(c) ¿Cuál de los polinomios de los apartados (a) y (b) proporciona una aproximación mejor a $(6.5)^{1/2}$?

8. Aplique el Teorema 4.1 a la función $f(x) = (2+x)^{1/2}$.

(a) Halle el polinomio de Taylor $P_3(x)$ de $f(x)$ alrededor de $x_0 = 2$.

(b) Use $P_3(x)$ para hallar una aproximación a $3^{1/2}$.

(c) Halle el máximo valor de $|f^{(4)}(c)|$ sobre el intervalo $1 \leq c \leq 3$ y encuentre una cota de $|E_3(x)|$.

9. Determine el grado del polinomio de Taylor $P_N(x)$, desarrollado alrededor de $x_0 = 0$, que habría que usar para aproximar $e^{0.1}$ con un error menor que 10^{-6} .

10. Determine el grado del polinomio de Taylor $P_N(x)$, desarrollado alrededor de $x_0 = \pi$, que habría que usar para aproximar $\cos(33\pi/32)$ con un error menor que 10^{-6} .

11. (a) Halle el polinomio de Taylor de grado $N = 4$ de la función dada por $F(x) = \int_{-1}^x \cos(t^2) dt$ alrededor de $x_0 = 0$.

(b) Use el polinomio de Taylor para aproximar $F(0.1)$.

(c) Halle una cota del error cometido en la aproximación del apartado (b).

12. (a) Dada la serie geométrica

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + x^8 - \dots \quad \text{para } |x| < 1,$$

integre a ambos lados de la igualdad para obtener

$$\arctan(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots \quad \text{para } |x| < 1.$$

- (b) Use $\pi/6 = \arctan(3^{-1/2})$ y la serie del apartado (a) para probar que

$$\pi = 3^{1/2} \times 2 \left(1 - \frac{3^{-1}}{3} + \frac{3^{-2}}{5} - \frac{3^{-3}}{7} + \frac{3^{-4}}{9} - \dots \right).$$

- (c) Use la serie del apartado (b) para calcular π con una precisión de ocho cifras decimales.

Sepa que: $\pi \approx 3.141592653589793284\dots$

13. Aplique el Teorema 4.1 a la función $f(x) = \ln(1+x)$ con $x_0 = 0$.

- (a) Pruebe que $f^{(k)}(x) = (-1)^{k-1}((k-1)!)/(1+x)^k$.
 (b) Pruebe que el polinomio de Taylor de grado N es

$$P_N(x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots + \frac{(-1)^{N-1}x^N}{N}.$$

- (c) Pruebe que el término del error de $P_N(x)$ es

$$E_N(x) = \frac{(-1)^N x^{N+1}}{(N+1)(1+c)^{N+1}}.$$

- (d) Evalúe $P_3(0.5)$, $P_6(0.5)$ y $P_9(0.5)$ y compare estos valores con $\ln(1.5)$.
 (e) Pruebe que si $0.0 \leq x \leq 0.5$ entonces la aproximación

$$\ln(x) \approx x - \frac{x^2}{2} + \frac{x^3}{3} - \dots + \frac{x^7}{7} - \frac{x^8}{8} + \frac{x^9}{9}$$

tiene como cota del error $|E_9| \leq 0.00009765\dots$

14. Serie binomial. Sean $f(x) = (1+x)^p$ y $x_0 = 0$.

- (a) Pruebe que $f^{(k)}(x) = p(p-1)\cdots(p-k+1)(1+x)^{p-k}$.
 (b) Pruebe que el polinomio de Taylor de grado N es

$$P_N(x) = 1 + px + \frac{p(p-1)x^2}{2!} + \dots + \frac{p(p-1)\cdots(p-N+1)x^N}{N!}.$$

- (c) Pruebe que

$$E_N(x) = p(p-1)\cdots(p-N)x^{N+1}/((1+c)^{N+1-p}(N+1)!).$$

- (d) Para $p = 1/2$ calcule $P_2(0.5)$, $P_4(0.5)$ y $P_6(0.5)$ y compare estos valores con $(1.5)^{1/2}$.

- (e) Pruebe que si $0.0 \leq x \leq 0.5$, entonces la aproximación

$$(1+x)^{1/2} \approx 1 + \frac{x}{2} - \frac{x^2}{8} + \frac{x^3}{16} - \frac{5x^4}{128} + \frac{7x^5}{256}$$

tiene como cota del error $|E_5| \leq (0.5)^6(21/1024) = 0.0003204\dots$

- (f) Pruebe que si $p = N$ es un número natural, entonces

$$P_N(x) = 1 + Nx + \frac{N(N-1)x^2}{2!} + \cdots + Nx^{N-1} + x^N.$$

Nótese que ésta es la familiar fórmula del binomio.

15. En los siguientes casos, halle c tal que $|E_4| < 10^{-6}$ siempre que $|x - x_0| < c$.
- (a) $f(x) = \cos(x)$ y $x_0 = 0$.
 - (b) $f(x) = \operatorname{sen}(x)$ y $x_0 = \pi/2$.
 - (c) $f(x) = e^x$ y $x_0 = 0$.
16. (a) Supongamos que $y = f(x)$ es una función par (o sea, que $f(-x) = f(x)$ para todo x del dominio de f). ¿Qué puede decirse de $P_N(x)$ cuando $x_0 = 0$?
- (b) Supongamos que $y = f(x)$ es una función impar (o sea, que $f(-x) = -f(x)$ para todo x del dominio de f). ¿Qué puede decirse de $P_N(x)$ cuando $x_0 = 0$?
17. Sea $y = f(x)$ un polinomio de grado N tal que $f(x_0) > 0$ y, asimismo, que $f'(x_0), \dots, f^{(N)}(x_0) \geq 0$. Pruebe que todas las raíces reales de f son menores que x_0 . *Indicación.* Desarrolle f como un polinomio de Taylor de grado N alrededor de x_0 .
18. Dada $f(x) = e^x$, use el Teorema 4.1 para hallar $P_N(x)$, para $N = 1, 2, 3, \dots$, desarrollados alrededor de $x_0 = 0$. Pruebe que las raíces reales de $P_N(x)$ tienen multiplicidad menor o igual que uno. *Nota.* Si p es una raíz de multiplicidad M del polinomio $P(x)$, entonces p es una raíz de multiplicidad $M-1$ de $P'(x)$.
19. Termine la demostración del Corolario 4.1 escribiendo la expresión de $P_N^{(k)}(x)$ y probando que

$$P_N^{(k)}(x_0) = f^{(k)}(x_0) \quad \text{para } k = 2, 3, \dots, N.$$

Los Ejercicios 20 y 21 constituyen una demostración del Teorema de Taylor.

20. Supongamos que una función $g(t)$ y sus derivadas $g^{(k)}(t)$, para $k = 1, 2, \dots, N+1$, son continuas en un intervalo (a, b) que contiene un punto dado x_0 . Supongamos que existe otro punto x distinto de x_0 tal que para dichos puntos se tiene $g(x) = 0$ y $g(x_0) = g'(x_0) = \dots = g^{(N)}(x_0) = 0$. Pruebe que existe un valor c entre x_0 y x tal que $g^{(N+1)}(c) = 0$. *Observación.* Cuando escribimos $g(t)$ como función de t , los valores x y x_0 deben ser tratados como constantes con respecto a la variable t .

Indicación. Use el teorema de Rolle (Teorema 1.5, Sección 1.1) en el intervalo de extremos x_0 y x para hallar un punto c_1 tal que $g'(c_1) = 0$. Ahora, vuelva a usar el teorema de Rolle aplicado a la función $g'(t)$ en el intervalo de extremos x_0 y c_1 para hallar un punto c_2 tal que $g''(c_2) = 0$. Repita el proceso inductivamente hasta que obtenga un punto c_{N+1} tal que $g^{(N+1)}(c_{N+1}) = 0$.

- 21.** Use el resultado del Ejercicio 20 y la función particular

$$g(t) = f(t) - P_N(t) - E_N(x) \frac{(t - x_0)^{N+1}}{(x - x_0)^{N+1}},$$

donde $P_N(x)$ es el polinomio de Taylor de grado N , para probar que el término del error $E_N(x) = f(x) - P_N(x)$ puede escribirse como

$$E_N(x) = f^{(N+1)}(c) \frac{(x - x_0)^{N+1}}{(N + 1)!}.$$

Indicación. Halle $g^{(N+1)}(t)$ y evalúe esta función en $t = c$.

Algoritmos y programas

La naturaleza matricial del paquete de programas MATLAB nos permite evaluar funciones en un gran número de puntos. Por ejemplo, si $X=[-1 \ 0 \ 1]$, entonces $\cos(X)$ producirá $[\cos(-1) \ \cos(0) \ \cos(1)]$. De manera parecida, si $X=-1:0.1:1$, entonces $Y=\cos(X)$ producirá una matriz Y del mismo orden que X cuyas componentes son los valores correspondientes del coseno. Estas dos matrices pueden mostrarse en forma de tabla definiendo la matriz $D = [X' \ Y']$ (Nota. Las matrices X e Y deben tener el mismo orden.)

1. (a) Use la instrucción `plot` para dibujar, en una misma gráfica sobre el intervalo $[-1, 1]$, las funciones $\sin(x)$, $P_5(x)$, $P_7(x)$ y $P_9(x)$ del Ejercicio 1.
 (b) Obtenga una tabla cuyas columnas sean los valores de $\sin(x)$, $P_5(x)$, $P_7(x)$ y $P_9(x)$ evaluados en diez puntos x equiespaciados a lo largo del intervalo $[-1, 1]$.
2. (a) Use la instrucción `plot` para dibujar, en una misma gráfica sobre el intervalo $[-1, 1]$, las funciones $\cos(x)$, $P_4(x)$, $P_6(x)$ y $P_8(x)$ del Ejercicio 2.
 (b) Obtenga una tabla cuyas columnas sean los valores de $\cos(x)$, $P_4(x)$, $P_6(x)$ y $P_8(x)$ evaluados en 19 puntos x equiespaciados a lo largo de $[-1, 1]$.

4.2 Introducción a la interpolación

En la Sección 4.1 hemos visto cómo puede usarse un polinomio de Taylor para aproximar una función $f(x)$. La información necesaria para construir el polinomio de Taylor es el valor de f y los de sus derivadas en x_0 . Un inconveniente

de este procedimiento es que debemos conocer las derivadas de orden superior y, a menudo, suele ocurrir que o bien no están disponibles, o bien son difíciles de calcular.

Supongamos que conocemos $N + 1$ puntos $(x_0, y_0), (x_1, y_1), \dots, (x_N, y_N)$ de la curva $y = f(x)$, donde las abscisas x_k se distribuyen en un intervalo $[a, b]$ de manera que

$$a \leq x_0 < x_1 < \dots < x_N \leq b \quad \text{e} \quad y_k = f(x_k).$$

Construiremos un polinomio $P(x)$ de grado N que pase por estos $N + 1$ puntos. Para construirlo, únicamente necesitaremos conocer los valores x_k e y_k , así que las derivadas de orden superior no nos harán falta. El polinomio $P(x)$ puede luego usarse como una aproximación a $f(x)$ en todo el intervalo $[a, b]$; no obstante, si queremos conocer la función error $E(x) = f(x) - P(x)$, entonces sí necesitaremos conocer $f^{(N+1)}(x)$ o bien una cota de su tamaño como

$$M = \max\{|f^{(N+1)}(x)| : a \leq x \leq b\}.$$

Existen funciones especiales $y = f(x)$, que aparecen en análisis de tipo estadístico o científico, para las que sólo disponemos de una tabla de valores; es decir, sólo conocemos $N + 1$ puntos (x_k, y_k) y es necesario dar un método para aproximar $f(x)$ en abscisas que no están tabuladas. Si el error de los valores tabulados es significativo, entonces es mejor usar los métodos de ajuste de curvas que veremos en el Capítulo 5. Si, por el contrario, los puntos (x_k, y_k) tienen un grado alto de precisión, entonces podemos considerar el polinomio $y = P(x)$ que pasa por todos ellos. Cuando $x_0 < x < x_N$, la aproximación $P(x)$ se conoce como **valor interpolado**; si se tiene $x < x_0$ o bien $x_N < x$, entonces $P(x)$ se conoce como **valor extrapolado**. Los polinomios se utilizan para diseñar algoritmos de aproximación de funciones, para derivar e integrar numéricamente y para dibujar, mediante un computador, curvas que deben pasar por puntos especificados de antemano.

Recordemos brevemente que la forma eficiente de evaluar un polinomio $P(x)$ dado por:

$$(1) \quad P(x) = a_N x^N + a_{N-1} x^{N-1} + \dots + a_2 x^2 + a_1 x + a_0$$

es el método de Horner, o regla de Ruffini, de división sintética. La derivada de $P(x)$ es

$$(2) \quad P'(x) = N a_N x^{N-1} + (N-1) a_{N-1} x^{N-2} + \dots + 2 a_2 x + a_1$$

y la integral indefinida $I(x) = \int P(x) dx$, que verifica $I'(x) = P(x)$, es

$$(3) \quad I(x) = \frac{a_N x^{N+1}}{N+1} + \frac{a_{N-1} x^N}{N} + \dots + \frac{a_2 x^3}{3} + \frac{a_1 x^2}{2} + a_0 x + C,$$

donde C es la constante de integración. El Algoritmo 4.1 (que se encuentra al final de esta Sección 4.2) muestra cómo adaptar el método de Horner para calcular $P'(x)$ e $I(x)$.

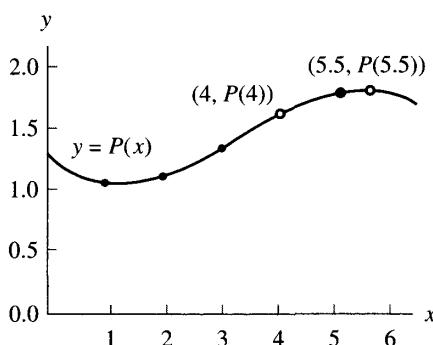


Figura 4.7 (a) El polinomio interpolador $P(x)$ puede usarse para interpolar en el punto $(4, P(4))$ y extrapolar en el punto $(5.5, P(5.5))$.

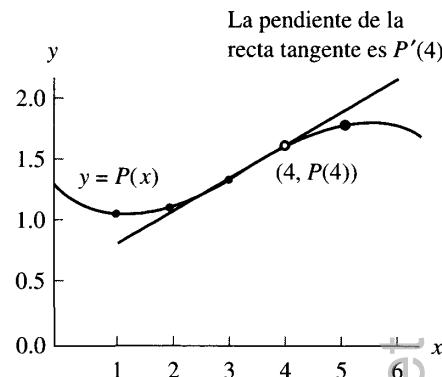


Figura 4.7 (b) El polinomio interpolador $P(x)$ se deriva y $P'(x)$ se usa para aproximar la pendiente en el punto $(4, P(4))$.

Ejemplo 4.4. El polinomio $P(x) = -0.02x^3 + 0.2x^2 - 0.4x + 1.28$ pasa por los cuatro puntos $(1, 1.06)$, $(2, 1.12)$, $(3, 1.34)$ y $(5, 1.78)$. Vamos a calcular (a) $P(4)$, (b) $P'(4)$, (c) $\int_1^4 P(x)dx$ y (d) $P(5.5)$. Finalmente mostraremos, en (e), cómo hallar los coeficientes de $P(x)$.

Usamos el Algoritmo 4.1(i)–(iii) (que es equivalente al proceso descrito en la Tabla 1.2) con $x = 4$.

$$\begin{aligned}
 \text{(a)} \quad & b_3 = a_3 = -0.02 \\
 & b_2 = a_2 + b_3 x = 0.2 + (-0.02)(4) = 0.12 \\
 & b_1 = a_1 + b_2 x = -0.4 + (0.12)(4) = 0.08 \\
 & b_0 = a_0 + b_1 x = 1.28 + (0.08)(4) = 1.60.
 \end{aligned}$$

El valor interpolado es $P(4) = 1.60$ (véase la Figura 4.7(a)):

$$\begin{aligned}
 \text{(b)} \quad & d_2 = 3a_3 = -0.06, \\
 & d_1 = 2a_2 + d_2 x = 0.4 + (-0.06)(4) = 0.16, \\
 & d_0 = a_1 + d_1 x = -0.4 + (0.16)(4) = 0.24.
 \end{aligned}$$

La derivada numérica es $P'(4) = 0.24$ (véase la Figura 4.7(b)). Para el cálculo de la integral evaluemos

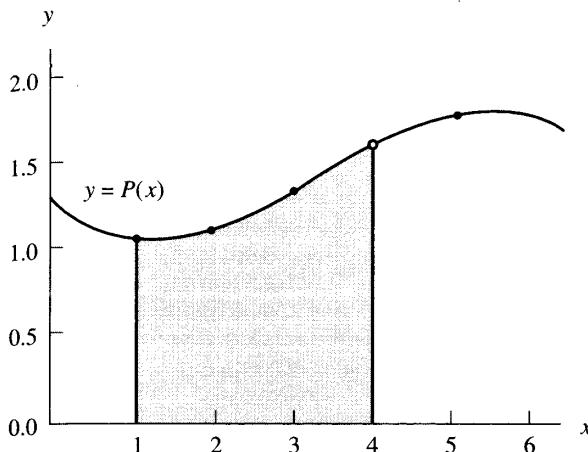


Figura 4.8 El polinomio interpolador $P(x)$ se integra y su primitiva se usa para hallar el área limitada por la curva para $1 \leq x \leq 4$.

$$\begin{aligned}
 (c) \quad i_4 &= \frac{a_3}{4} = -0.005, \\
 i_3 &= \frac{a_2}{3} + i_4 x = 0.06666667 + (-0.005)(4) = 0.04666667, \\
 i_2 &= \frac{a_1}{2} + i_3 x = -0.2 + (0.04666667)(4) = -0.01333333, \\
 i_1 &= a_0 + i_2 x = 1.28 + (-0.01333333)(4) = 1.22666667, \\
 i_0 &= 0 + i_1 x = 0 + (1.22666667)(4) = 4.90666667.
 \end{aligned}$$

Por tanto, $I(4) = 4.90666667$ y, de forma similar, $I(1) = 1.14166667$; luego, $\int_1^4 P(x) dx = I(4) - I(1) = 3.765$ (véase la Figura 4.8).

(d) Usamos el Algoritmo 4.1(i) con $x = 5.5$:

$$\begin{aligned}
 b_3 &= a_3 = -0.02, \\
 b_2 &= a_2 + b_3 x = 0.2 + (-0.02)(5.5) = 0.09, \\
 b_1 &= a_1 + b_2 x = -0.4 + (0.09)(5.5) = 0.095, \\
 b_0 &= a_0 + b_1 x = 1.28 + (0.095)(5.5) = 1.8025.
 \end{aligned}$$

El valor extrapolado es $P(5.5) = 1.8025$ (véase la Figura 4.7(a)).

(e) Podemos usar los métodos del Capítulo 3 para hallar los coeficientes. Supongamos que $P(x) = A + Bx + Cx^2 + Dx^3$, entonces en cada valor $x = 1, 2, 3$ y 5 obtenemos una ecuación para A, B, C y D .

$$\begin{aligned}
 (4) \quad \text{En } x = 1: A + 1B + 1C + 1D &= 1.06 \\
 \text{En } x = 2: A + 2B + 4C + 8D &= 1.12 \\
 \text{En } x = 3: A + 3B + 9C + 27D &= 1.34 \\
 \text{En } x = 5: A + 5B + 25C + 125D &= 1.78
 \end{aligned}$$

La solución de (4) es $A = 1.28, B = -0.4, C = 0.2$ y $D = -0.2$. ■

Tabla 4.4 Valores de la función $\ln(1 + x)$, de su polinomio de Taylor $T(x)$ de grado 5 y del error $\ln(1 + x) - T(x)$ en $[0, 1]$.

x	Función, $\ln(1 + x)$	Polinomio de Taylor, $T(x)$	Error, $\ln(1 + x) - T(x)$
0.0	0.00000000	0.00000000	0.00000000
0.2	0.18232156	0.18233067	-0.00000911
0.4	0.33647224	0.33698133	-0.00050909
0.6	0.47000363	0.47515200	-0.00514837
0.8	0.58778666	0.61380267	-0.02601601
1.0	0.69314718	0.78333333	-0.09018615

Este método para calcular los coeficientes es matemáticamente correcto, pero algunas veces es difícil resolver el sistema con la precisión adecuada. En este capítulo vamos a desarrollar algoritmos específicos para resolver este tipo de problemas.

Volvamos al problema de calcular los valores de una función dada mediante polinomios. En la Sección 4.1 vimos que el polinomio de Taylor de quinto grado de $f(x) = \ln(1 + x)$ es

$$(5) \quad T(x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5}.$$

Si usamos $T(x)$ para aproximar $\ln(1 + x)$ en el intervalo $[0, 1]$, entonces el error es 0 en $x = 0$ y alcanza su máximo en $x = 1$ (véase la Tabla 4.4); de hecho, el error entre $T(1)$ y el valor correcto $\ln(2)$ es del 13%. Nos gustaría encontrar un polinomio de grado 5 que se aproxime mejor a la función $\ln(1 + x)$ a lo largo del intervalo $[0, 1]$. Pues bien, el polinomio $P(x)$ del Ejemplo 4.5 es un polinomio interpolador que aproxima a $\ln(1 + x)$ con un error menor que 0.00002385 en el intervalo $[0, 1]$.

Ejemplo 4.5. Consideremos la función $f(x) = \ln(1 + x)$ y el polinomio

$$\begin{aligned} P(x) = & 0.02957206x^5 - 0.12895295x^4 + 0.28249626x^3 \\ & - 0.48907554x^2 + 0.99910735x \end{aligned}$$

construido sobre los seis nodos $x_k = k/5$ para $k = 0, 1, 2, 3, 4$ y 5. Podemos dar la siguiente descripción empírica de la aproximación $P(x) \approx \ln(1 + x)$.

1. $P(x_k) = f(x_k)$ en cada nodo (véase la Tabla 4.5).
2. El máximo error en el intervalo $[-0.1, 1.1]$ se da en $x = -0.1$ y se tiene $|\text{error}| \leq 0.00026334$ para $-0.1 \leq x \leq 1.1$ (véase la Figura 4.10). Como consecuencia, la gráfica de $y = P(x)$ parece idéntica a la de $y = \ln(1 + x)$ (véase la Figura 4.9).

Tabla 4.5 Valores de la función $f(x) = \ln(1 + x)$, de su polinomio aproximante $P(x)$ en el Ejemplo 4.5 y del error $E(x)$ en $[-0.1, 1.1]$.

x	Función, $f(x) = \ln(1 + x)$	Polinomio aproximante, $P(x)$	Error, $E(x) = f(x) - P(x)$
-0.1	-0.10536052	-0.10509718	-0.00026334
0.0	0.00000000	0.00000000	0.00000000
0.1	0.09531018	0.09528988	0.00002030
0.2	0.18232156	0.18232156	0.00000000
0.3	0.26236426	0.26237015	-0.00000589
0.4	0.33647224	0.33647224	0.00000000
0.5	0.40546511	0.40546139	0.00000372
0.6	0.47000363	0.47000363	0.00000000
0.7	0.53062825	0.53063292	-0.00000467
0.8	0.58778666	0.58778666	0.00000000
0.9	0.64185389	0.64184118	0.00001271
1.0	0.69314718	0.69314718	0.00000000
1.1	0.74193734	0.74206529	-0.00012795

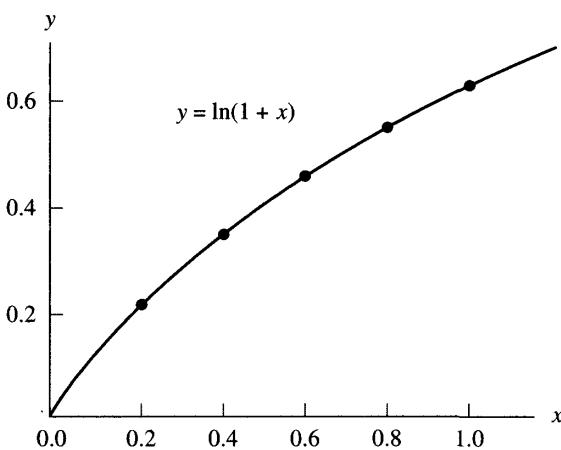


Figura 4.9 La gráfica de $y = P(x)$, que está prácticamente superpuesta a la de $y = \ln(1 + x)$.

3. El error máximo en el intervalo $[0, 1]$ se alcanza en $x = 0.06472456$ y se tiene $|error| \leq 0.00002385$ para $0 \leq x \leq 1$ (véase la Figura 4.10).

Observación. En cada nodo x_k tenemos $f(x_k) = P(x_k)$; por tanto, $E(x_k) = 0$ en dicho nodo. La gráfica de $E(x) = f(x) - P(x)$ se parece a la de una cuerda vibrante, siendo los nodos las abscisas en las que no hay desplazamiento. ■

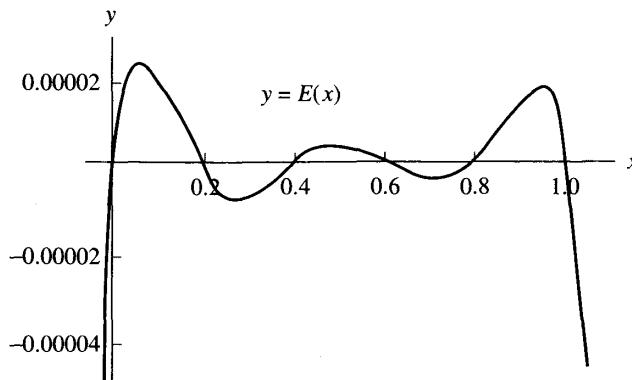


Figura 4.10 La gráfica del error

$$y = E(x) = \ln(1 + x) - P(x).$$

Algoritmo 4.1 (Cálculo con polinomios). Evaluación del polinomio $P(x)$, de su derivada $P'(x)$ y de su primitiva $\int P(x) dx + C$ mediante división sintética.

DATOS:	N	Grado de $P(x)$
	$A(0), A(1), \dots, A(N)$	Coeficientes de $P(x)$
	C	Constante
	X	Variable independiente

(i) Algoritmo para evaluar $P(x)$ $B(N) := A(N)$ DESDE $K = N - 1$ HASTA 0 $B(K) := A(K) + B(K + 1) * X$ IMPRIMIR “El valor de $P(x)$ es”, $B(0)$	Versión para ahorrar memoria: Poly := $A(N)$ DESDE $K = N - 1$ HASTA 0 Poly := $A(K) + Poly * X$ IMPRIMIR “El valor de $P(x)$ es”, Poly
(ii) Algoritmo para evaluar $P'(x)$ $D(N - 1) := N * A(N)$ DESDE $K = N - 1$ HASTA 1 $D(K - 1) := K * A(K) + D(K) * X$ IMPRIMIR “El valor de $P'(x)$ es”, $D(0)$	Versión para ahorrar memoria: Deriv := $N * A(N)$ DESDE $K = N - 1$ HASTA 1 Deriv := $K * A(K) + Deriv * X$ IMPRIMIR “El valor de $P'(x)$ es”, Deriv
(iii) Algoritmo para evaluar $I(x)$ $I(N + 1) := A(N)/(N + 1)$ DESDE $K = N$ HASTA 1 $I(K) := A(K - 1)/K + I(K + 1) * X$ $I(0) := C + I(1) * X$ IMPRIMIR “El valor de $I(x)$ es”, $I(0)$	Versión para ahorrar memoria: Integ := $A(N)/(N + 1)$ DESDE $K = N$ HASTA 1 Integ := $A(K - 1)/K + Integ * X$ Integ := $C + Integ * X$ IMPRIMIR “El valor de $I(x)$ es”, Integ

Ejercicios

1. Consideremos el polinomio $P(x) = -0.02x^3 + 0.1x^2 - 0.2x + 1.66$ que pasa por los puntos $(1, 1.54)$, $(2, 1.5)$, $(3, 1.42)$ y $(5, 0.66)$.
 - (a) Calcule $P(4)$.
 - (b) Calcule $P'(4)$.
 - (c) Halle la integral de $P(x)$ en el intervalo $[1, 4]$.
 - (d) Calcule el valor extrapolado $P(5.5)$.
 - (e) Muestre cómo se calculan los coeficientes de $P(x)$.
2. Consideremos el polinomio $P(x) = -0.04x^3 + 0.14x^2 - 0.16x + 2.08$ que pasa por los puntos $(0, 2.08)$, $(1, 2.02)$, $(2, 2.00)$ y $(4, 1.12)$.
 - (a) Calcule $P(3)$.
 - (b) Calcule $P'(3)$.
 - (c) Halle la integral de $P(x)$ en el intervalo $[0, 3]$.
 - (d) Calcule el valor extrapolado $P(4.5)$.
 - (e) Muestre cómo se calculan los coeficientes de $P(x)$.
3. Dado el polinomio $P(x) = -0.0292166667x^3 + 0.275x^2 - 0.570833333x - 1.375$ que pasa por los puntos $(1, 1.05)$, $(2, 1.10)$, $(3, 1.35)$ y $(5, 1.75)$.
 - (a) Pruebe que las ordenadas 1.05 , 1.10 , 1.35 y 1.75 difieren de las del Ejemplo 4.4 en menos del 1.8% y que, sin embargo, los coeficientes de x^3 y x difieren en más del 42% .
 - (b) Calcule $P(4)$ y compárela con lo que se obtuvo en el Ejemplo 4.4.
 - (c) Calcule $P'(4)$ y compárela con lo que se obtuvo en el Ejemplo 4.4.
 - (d) Halle la integral de $P(x)$ en el intervalo $[1, 4]$ y compárela con la que se obtuvo en el Ejemplo 4.4.
 - (e) Calcule el valor extrapolado $P(5.5)$ y compárela con el que se obtuvo en el Ejemplo 4.4.

Observación. El apartado (a) muestra que el cálculo de los coeficientes de un polinomio interpolador es un problema mal condicionado.

Algoritmos y programas

1. Escriba un programa en el lenguaje del paquete MATLAB para el Algoritmo 4.1. El programa debe aceptar como datos de entrada los coeficientes del polinomio $P(x) = a_Nx^N + a_{N-1}x^{N-1} + \cdots + a_2x^2 + a_1x + a_0$ como una matriz de orden $1 \times (N+1)$: $P = [a_N \ a_{N-1} \ \cdots \ a_2 \ a_1 \ a_0]$.
2. En este problema se trata de calcular, para cada una de las funciones relacionadas al final, el polinomio interpolador $P(x)$ de grado 5 que pasa por los puntos $(0, f(0))$, $(0.2, f(0.2))$, $(0.4, f(0.4))$, $(0.6, f(0.6))$, $(0.8, f(0.8))$ y

$(1, f(1))$. Los seis coeficientes de $P(x)$ son a_0, a_1, \dots, a_5 , de manera que

$$P(x) = a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0.$$

- (i) Determine los coeficientes de $P(x)$ resolviendo el sistema de ecuaciones lineales de orden 6×6

$$a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5 = f(x_j)$$

tomando $x_j = (j - 1)/5$ para $j = 1, 2, 3, 4, 5$ y 6.

- (ii) Utilice su programa del Problema 1 para calcular los valores interpolados $P(0.3)$, $P(0.4)$ y $P(0.5)$ y compare estos valores con $f(0.3)$, $f(0.4)$ y $f(0.5)$, respectivamente.
- (iii) Utilice su programa del Problema 1 para calcular los valores extrapolados $P(-0.1)$ y $P(1.1)$ y compare estos valores con $f(-0.1)$ y $f(1.1)$, respectivamente.
- (iv) Utilice su programa del Problema 1 para hallar la integral de $P(x)$ en el intervalo $[0, 1]$ y compare este valor con la integral de $f(x)$ en el intervalo $[0, 1]$.
- (v) Dibuje, en un mismo gráfico, $f(x)$ y $P(x)$ para $x \in [0, 1]$.
- (vi) Construya una tabla con los valores de $f(x_k)$, de $P(x_k)$ y los errores $E(x_k) = f(x_k) - P(x_k)$ siendo $x_k = k/100$ para $k = 0, 1, \dots, 100$.
- (a) $f(x) = e^x$
 (b) $f(x) = \sin(x)$
 (c) $f(x) = (x + 1)^{(x+1)}$
3. Se quiere diseñar una porción de la montaña rusa de un parque de atracciones usando tres polinomios. La primera sección debe ser un polinomio $P_1(x)$ de grado 1 que cubra una distancia horizontal de 30 metros, empezando a una altura de 32 metros y terminando a una altura de 20 metros. La tercera sección debe ser también un polinomio $Q_1(x)$ de grado 1 que cubra una distancia horizontal de 18 metros, empezando a una altura de 22 metros y terminando a una altura de 24 metros. La segunda sección debe ser un polinomio $P(x)$ (del menor grado posible) que cubra una distancia horizontal de 50 metros.
- (a) Halle las expresiones de $P(x)$, $P_1(x)$ y $Q_1(x)$ imponiendo las siguientes condiciones: por un lado, $P(30) = P_1(30)$, $P'(30) = P'_1(30)$, $P(80) = Q_1(80)$ y $P'(80) = Q'_1(80)$; por otro lado, las curvaturas de $P(x)$ deben coincidir en los extremos de su sección con, respectivamente, la de $P_1(x)$ en $x = 30$ y la de $Q_1(x)$ en $x = 80$.
- (b) Dibuje la gráfica de la porción completa.
 (c) Utilice el Algoritmo 4.1(iii) para hallar la altura media en esta porción de la montaña rusa a lo largo de su recorrido.

3 Interpolación de Lagrange

Interolar significa estimar el valor desconocido de una función en un punto, tomando una media ponderada de sus valores conocidos en puntos cercanos al dado. En la interpolación lineal —también conocida como la regla de tres— se utiliza un segmento rectilíneo que pasa por dos puntos que se conocen. La pendiente de la recta que pasa por dos puntos (x_0, y_0) y (x_1, y_1) viene dada por $m = (y_1 - y_0)/(x_1 - x_0)$; así que en la ecuación de la recta escrita como $y = m(x - x_0) + y_0$ podemos sustituir m y obtener

$$(1) \quad y = P(x) = y_0 + (y_1 - y_0) \frac{x - x_0}{x_1 - x_0}.$$

Si desarrollamos esta fórmula (1), el resultado es un polinomio de grado menor o igual que uno y la evaluación de $P(x)$ en x_0 y x_1 produce y_0 e y_1 , respectivamente:

$$(2) \quad \begin{aligned} P(x_0) &= y_0 + (y_1 - y_0)(0) = y_0, \\ P(x_1) &= y_0 + (y_1 - y_0)(1) = y_1. \end{aligned}$$

El matemático francés Joseph Louis Lagrange descubrió que se puede encontrar este polinomio usando un método ligeramente distinto. Si escribimos

$$(3) \quad y = P_1(x) = y_0 \frac{x - x_1}{x_0 - x_1} + y_1 \frac{x - x_0}{x_1 - x_0},$$

entonces cada uno de los sumandos del miembro derecho de esta relación es un término lineal, por lo que su suma será un polinomio de grado menor o igual que uno. Denotemos los cocientes de (3) por

$$(4) \quad L_{1,0}(x) = \frac{x - x_1}{x_0 - x_1} \quad \text{y} \quad L_{1,1}(x) = \frac{x - x_0}{x_1 - x_0}.$$

Un sencillo cálculo muestra que $L_{1,0}(x_0) = 1$, $L_{1,0}(x_1) = 0$, $L_{1,1}(x_0) = 0$ y $L_{1,1}(x_1) = 1$, así que el polinomio $P_1(x)$ definido en (3) también pasa por los dos puntos dados:

$$(5) \quad P_1(x_0) = y_0 + y_1(0) = y_0 \quad \text{y} \quad P_1(x_1) = y_0(0) + y_1 = y_1.$$

Los términos $L_{1,0}(x)$ y $L_{1,1}(x)$ definidos en (4) se llaman **polinomios coeficientes de Lagrange** para los nodos x_0 y x_1 . Usando esta notación, podemos escribir (3) como una suma

$$(6) \quad P_1(x) = \sum_{k=0}^1 y_k L_{1,k}(x).$$

Cuando las ordenadas y_k vienen dadas por $y_k = f(x_k)$, el proceso de utilizar $P_1(x)$ para aproximar $f(x)$ en el intervalo $[x_0, x_1]$ se conoce con el nombre de **interpolación lineal**. Si $x < x_0$ (o bien $x_1 < x$), entonces el uso de $P_1(x)$ para aproximar $f(x)$ se llama **extrapolación**. El siguiente ejemplo ilustra estos conceptos.

Ejemplo 4.6. Consideremos la gráfica de $y = f(x) = \cos(x)$ en $[0.0, 1.2]$.

- Vamos a usar los nodos $x_0 = 0.0$ y $x_1 = 1.2$ para construir un polinomio de interpolación lineal $P_1(x)$.
- Vamos a usar los nodos $x_0 = 0.2$ y $x_1 = 1.0$ para construir un polinomio de interpolación lineal $Q_1(x)$.
 - La fórmula (3) con las abscisas $x_0 = 0.0$ y $x_1 = 1.2$ y las ordenadas $y_0 = \cos(0.0) = 1.000000$ e $y_1 = \cos(1.2) = 0.362358$ proporciona

$$\begin{aligned} P_1(x) &= 1.000000 \frac{x - 1.2}{0.0 - 1.2} + 0.362358 \frac{x - 0.0}{1.2 - 0.0} \\ &= -0.833333(x - 1.2) + 0.301965(x - 0.0). \end{aligned}$$

- Cuando usamos los nodos $x_0 = 0.2$ y $x_1 = 1.0$ con los valores $y_0 = \cos(0.2) = 0.980067$ e $y_1 = \cos(1.0) = 0.540302$, el resultado es

$$\begin{aligned} Q_1(x) &= 0.980067 \frac{x - 1.0}{0.2 - 1.0} + 0.540302 \frac{x - 0.2}{1.0 - 0.2} \\ &= -1.225083(x - 1.0) + 0.675378(x - 0.2). \end{aligned}$$

Las Figuras 4.11 (a) y (b) muestran la gráfica de $y = \cos(x)$ junto con, respectivamente, las de $y = P_1(x)$ e $y = Q_1(x)$; estos dibujos y los resultados numéricos de la Tabla 4.6 nos sirven para comparar ambas aproximaciones y revelan que $Q_1(x)$ tiene un error menor en los puntos x_k que verifican $0.1 \leq x_k \leq 1.1$. El error más grande de los recogidos en la tabla correspondiente a P_1 , que es $f(0.6) - P_1(0.6) = 0.144157$, se reduce a $f(0.6) - Q_1(0.6) = 0.065151$ cuando se usa $Q_1(x)$. ■

La forma de generalizar la fórmula (6) para construir un polinomio $P_N(x)$ que tenga grado menor o igual que N y que pase por $N + 1$ puntos $(x_0, y_0), (x_1, y_1), \dots, (x_N, y_N)$ es la fórmula

$$(7) \quad P_N(x) = \sum_{k=0}^N y_k L_{N,k}(x),$$

donde $L_{N,k}$ es el polinomio coeficiente de Lagrange para los nodos x_0, x_1, \dots, x_N definido por

$$(8) \quad L_{N,k}(x) = \frac{(x - x_0) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_N)}{(x_k - x_0) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_N)},$$

donde se sobreentiende que los factores $(x - x_k)$ y $(x_k - x_k)$ no aparecen en el numerador del miembro derecho de la relación (8). Resulta cómodo introducir en (8) la notación compacta para el producto y escribir

$$(9) \quad L_{N,k}(x) = \frac{\prod_{\substack{j=0 \\ j \neq k}}^N (x - x_j)}{\prod_{\substack{j=0 \\ j \neq k}}^N (x_k - x_j)}.$$

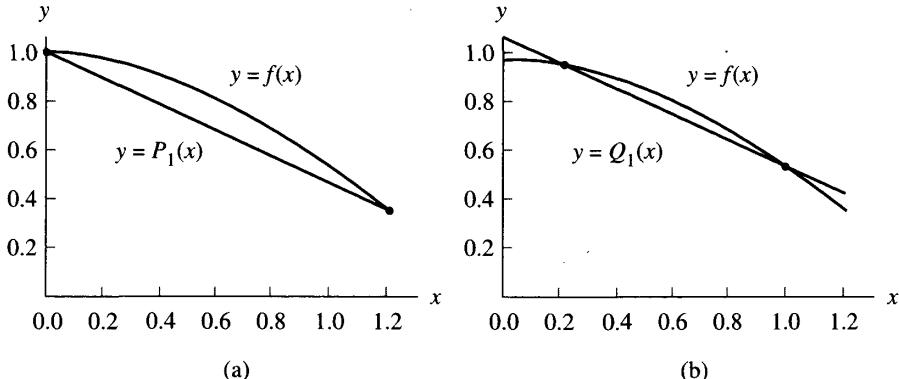


Figura 4.11 (a) La aproximación lineal $y = P_1(x)$ de $y = \cos(x)$ cuando los nodos $x_0 = 0.0$ y $x_1 = 1.2$ son los extremos del intervalo $[a, b]$. (b) La aproximación lineal $y = Q_1(x)$ de $y = \cos(x)$ cuando los nodos $x_0 = 0.2$ y $x_1 = 1.0$ están dentro del intervalo $[a, b]$.

La notación de (9) indica que en el numerador se forma el producto de todos los factores lineales $(x - x_j)$ pero sin incluir (saltándose) el factor $(x - x_k)$ (y lo análogo para el denominador).

Un cálculo directo prueba que, para cada k fijo, el polinomio coeficiente de Lagrange $L_{N,k}(x)$ tiene la siguiente propiedad:

$$(10) \quad L_{N,k}(x_j) = 1 \text{ si } j = k \quad \text{y} \quad L_{N,k}(x_j) = 0 \text{ si } j \neq k.$$

La sustitución directa de estos valores en la fórmula (7) permite probar que la curva polinomial $y = P_N(x)$ pasa por los puntos (x_i, y_i) :

$$(11) \quad P_N(x_j) = y_0 L_{N,0}(x_j) + \cdots + y_j L_{N,j}(x_j) + \cdots + y_N L_{N,N}(x_j) \\ = y_0(0) + \cdots + y_j(1) + \cdots + y_N(0) = y_j.$$

Para probar que $P_N(x)$ es único, aplicamos el teorema fundamental del álgebra, el cual establece que un polinomio no nulo $T(x)$ de grado menor o igual que N tiene, como mucho, N raíces; en otras palabras, si $T(x)$ es cero en $N + 1$ abscisas distintas, entonces es idénticamente cero. Supongamos, entonces que hay otro polinomio $Q_N(x)$ de grado menor o igual que N cuya gráfica pasa también por los $N + 1$ puntos dados. Formando el polinomio diferencia $T(x) = P_N(x) - Q_N(x)$, vemos que $T(x)$ es de grado menor o igual que N y que $T(x_j) = P_N(x_j) - Q_N(x_j) = y_j - y_j = 0$ para cada $j = 0, 1, \dots, N$; por tanto, $T(x) \equiv 0$ y, en consecuencia, $Q_N(x) = P_N(x)$.

Cuando se desarrolla la fórmula (7), lo que se obtiene es similar a (3). El polinomio interpolador de Lagrange cuadrático para los puntos (x_0, y_0) , (x_1, y_1)

Tabla 4.6 Comparación de $f(x) = \cos(x)$ con sus aproximaciones lineales $P_1(x)$ y $Q_1(x)$.

x_k	$f(x_k) = \cos(x_k)$	$P_1(x_k)$	$f(x_k) - P_1(x_k)$	$Q_1(x_k)$	$f(x_k) - Q_1(x_k)$
0.0	1.000000	1.000000	0.000000	1.090008	-0.090008
0.1	0.995004	0.946863	0.048141	1.035037	-0.040033
0.2	0.980067	0.893726	0.086340	0.980067	0.000000
0.3	0.955336	0.840589	0.114747	0.925096	0.030240
0.4	0.921061	0.787453	0.133608	0.870126	0.050935
0.5	0.877583	0.734316	0.143267	0.815155	0.062428
0.6	0.825336	0.681179	0.144157	0.760184	0.065151
0.7	0.764842	0.628042	0.136800	0.705214	0.059628
0.8	0.696707	0.574905	0.121802	0.650243	0.046463
0.9	0.621610	0.521768	0.099842	0.595273	0.026337
1.0	0.540302	0.468631	0.071671	0.540302	0.000000
1.1	0.453596	0.415495	0.038102	0.485332	-0.031736
1.2	0.362358	0.362358	0.000000	0.430361	-0.068003

y (x_2, y_2) es

(12)

$$P_2(x) = y_0 \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} + y_1 \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} + y_2 \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}.$$

El polinomio interpolador de Lagrange de grado $N = 3$ para los puntos $(x_0, y_0), (x_1, y_1), (x_2, y_2)$ y (x_3, y_3) es

$$(13) P_3(x) = y_0 \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} + y_1 \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} \\ + y_2 \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} + y_3 \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)}.$$

Ejemplo 4.7. Consideremos $y = f(x) = \cos(x)$ en $[0.0, 1.2]$.

- (a) Vamos a usar los tres nodos $x_0 = 0.0, x_1 = 0.6$ y $x_2 = 1.2$ para construir el polinomio interpolador cuadrático $P_2(x)$.
- (b) Vamos a usar los cuatro nodos $x_0 = 0.0, x_1 = 0.4, x_2 = 0.8$ y $x_3 = 1.2$ para construir el polinomio interpolador cúbico $P_3(x)$.

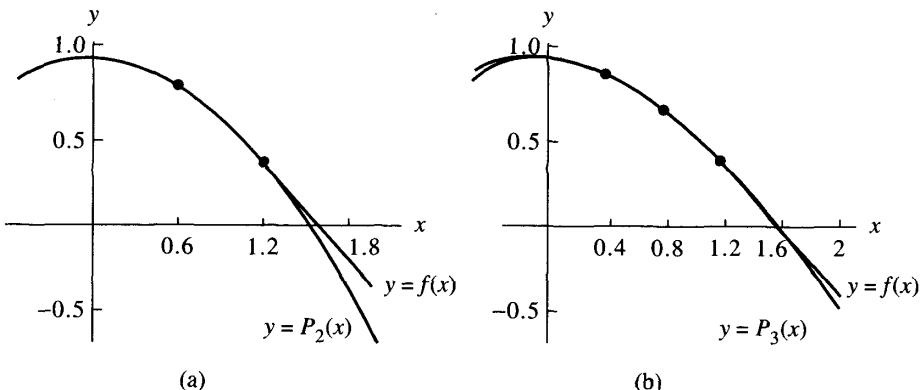


Figura 4.12 (a) El polinomio interpolador cuadrático $y = P_2(x)$ para los nodos $x_0 = 0.0$, $x_1 = 0.6$ y $x_2 = 1.2$. (b) El polinomio interpolador cúbico $y = P_3(x)$ para los nodos $x_0 = 0.0$, $x_1 = 0.4$, $x_2 = 0.8$ y $x_3 = 1.2$.

(a) Usando $x_0 = 0.0$, $x_1 = 0.6$, $x_2 = 1.2$ e $y_0 = \cos(0.0) = 1$, $y_1 = \cos(0.6) = 0.825336$, $y_2 = \cos(1.2) = 0.362358$ en la fórmula (12), obtenemos

$$\begin{aligned}
 P_2(x) &= 1.0 \frac{(x - 0.6)(x - 1.2)}{(0.0 - 0.6)(0.0 - 1.2)} + 0.825336 \frac{(x - 0.0)(x - 1.2)}{(0.6 - 0.0)(0.6 - 1.2)} \\
 &\quad + 0.362358 \frac{(x - 0.0)(x - 0.6)}{(1.2 - 0.0)(1.2 - 0.6)} \\
 &= 1.388889(x - 0.6)(x - 1.2) - 2.292599(x - 0.0)(x - 1.2) \\
 &\quad + 0.503275(x - 0.0)(x - 0.6).
 \end{aligned}$$

(b) Usando $x_0 = 0.0$, $x_1 = 0.4$, $x_2 = 0.8$, $x_3 = 1.2$ e $y_0 = \cos(0.0) = 1.0$, $y_1 = \cos(0.4) = 0.921061$, $y_2 = \cos(0.8) = 0.696707$, $y_3 = \cos(1.2) = 0.362358$ en la fórmula (13) obtenemos

$$\begin{aligned}
 P_3(x) &= 1.000000 \frac{(x - 0.4)(x - 0.8)(x - 1.2)}{(0.0 - 0.4)(0.0 - 0.8)(0.0 - 1.2)} \\
 &\quad + 0.921061 \frac{(x - 0.0)(x - 0.8)(x - 1.2)}{(0.4 - 0.0)(0.4 - 0.8)(0.4 - 1.2)} \\
 &\quad + 0.696707 \frac{(x - 0.0)(x - 0.4)(x - 1.2)}{(0.8 - 0.0)(0.8 - 0.4)(0.8 - 1.2)} \\
 &\quad + 0.362358 \frac{(x - 0.0)(x - 0.4)(x - 0.8)}{(1.2 - 0.0)(1.2 - 0.4)(1.2 - 0.8)} \\
 &= -2.604167(x - 0.4)(x - 0.8)(x - 1.2) \\
 &\quad + 7.195789(x - 0.0)(x - 0.8)(x - 1.2) \\
 &\quad - 5.443021(x - 0.0)(x - 0.4)(x - 1.2) \\
 &\quad + 0.943641(x - 0.0)(x - 0.4)(x - 0.8).
 \end{aligned}$$

230 CAP. 4 INTERPOLACIÓN Y APROXIMACIÓN POLINOMIAL

Las gráficas de $y = \cos(x)$ junto con las de los polinomios $y = P_2(x)$ e $y = P_3(x)$ se muestran en las Figuras 4.12 (a) y (b), respectivamente. ■

Términos y cotas del error

Es importante el entender la naturaleza del término del error que se comete cuando se utiliza un polinomio interpolador de Lagrange para aproximar una función $f(x)$. Este término, como veremos enseguida, es similar al término del error para el polinomio de Taylor; el único cambio es que el factor $(x - x_0)^{N+1}$ que allí aparecía se sustituye por el producto $(x - x_0)(x - x_1) \cdots (x - x_N)$, lo que resulta esperable ya que la interpolación es exacta en cada uno de los $N + 1$ nodos x_k , donde se verifica que $E_N(x_k) = f(x_k) - P_N(x_k) = y_k - y_k = 0$ para $k = 0, 1, 2, \dots, N$.

Teorema 4.3 (Polinomio interpolador de Lagrange). Supongamos que $f \in C^{N+1}[a, b]$ y que $x_0, x_1, \dots, x_N \in [a, b]$ son $N + 1$ nodos de interpolación. Si $x \in [a, b]$, entonces

$$(14) \quad f(x) = P_N(x) + E_N(x),$$

donde $P_N(x)$ es un polinomio que podemos usar para aproximar $f(x)$:

$$(15) \quad f(x) \approx P_N(x) = \sum_{k=0}^N f(x_k) L_{N,k}(x),$$

llamado **polinomio interpolador de Lagrange** de f para los nodos dados, y el término del error $E_N(x)$ se puede escribir como

$$(16) \quad E_N(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_N) f^{(N+1)}(c)}{(N + 1)!},$$

para algún valor $c = c(x)$ del intervalo $[a, b]$.

Demostración. Probaremos, a modo de ejemplo, la relación (16) cuando $N = 1$ y discutiremos el caso general en los ejercicios. Empezamos definiendo una función auxiliar $g(t)$ de la siguiente manera

$$(17) \quad g(t) = f(t) - P_1(t) - E_1(x) \frac{(t - x_0)(t - x_1)}{(x - x_0)(x - x_1)}.$$

Hagamos notar que x, x_0 y x_1 son constantes con respecto a la variable t y, también, que g vale cero en estos tres puntos; esto es,

$$g(x) = f(x) - P_1(x) - E_1(x) \frac{(x - x_0)(x - x_1)}{(x - x_0)(x - x_1)} = f(x) - P_1(x) - E_1(x) = 0,$$

$$g(x_0) = f(x_0) - P_1(x_0) - E_1(x) \frac{(x_0 - x_0)(x_0 - x_1)}{(x - x_0)(x - x_1)} = f(x_0) - P_1(x_0) = 0,$$

$$g(x_1) = f(x_1) - P_1(x_1) - E_1(x) \frac{(x_1 - x_0)(x_1 - x_1)}{(x - x_0)(x - x_1)} = f(x_1) - P_1(x_1) = 0.$$

Supongamos, sin pérdida de generalidad, que x está en el intervalo abierto (x_0, x_1) . Aplicando el teorema de Rolle a $g(t)$ en el intervalo $[x_0, x]$, obtenemos un punto d_0 , con $x_0 < d_0 < x$, tal que

$$(18) \quad g'(d_0) = 0.$$

Volviendo a aplicar el teorema de Rolle a $g(t)$ pero en el intervalo $[x, x_1]$, obtenemos un punto d_1 , con $x < d_1 < x_1$, tal que

$$(19) \quad g'(d_1) = 0.$$

Las relaciones (18) y (19) prueban que la función $g'(t)$ es cero en $t = d_0$ y $t = d_1$. Volviendo a usar el teorema de Rolle, pero esta vez aplicado a $g'(t)$ en $[d_0, d_1]$, obtenemos un valor $c \in [a, b]$ tal que

$$(20) \quad g''(c) = 0.$$

Volviendo a la relación (17) y calculando las derivadas $g'(t)$ y $g''(t)$, nos queda

$$(21) \quad g'(t) = f'(t) - P'_1(t) - E_1(x) \frac{(t - x_0) + (t - x_1)}{(x - x_0)(x - x_1)},$$

$$(22) \quad g''(t) = f''(t) - 0 - E_1(x) \frac{2}{(x - x_0)(x - x_1)},$$

donde hemos usado el hecho de que, por ser $P_1(t)$ es un polinomio de grado $N = 1$, su segunda derivada es $P''_1(t) \equiv 0$. Evaluando la expresión (22) en el punto $t = c$ y usando (20), se verifica

$$(23) \quad 0 = f''(c) - E_1(x) \frac{2}{(x - x_0)(x - x_1)},$$

de donde podemos despejar $E_1(x)$ para obtener la expresión dada en (16):

$$(24) \quad E_1(x) = \frac{(x - x_0)(x - x_1)f^{(2)}(c)}{2!},$$

lo que completa la prueba. •

El siguiente resultado aborda el caso especial en el que los nodos del polinomio interpolador de Lagrange están equiespaciados, $x_k = x_0 + hk$ para $k = 0, 1, \dots, N$, y el polinomio $P_N(x)$ se utiliza sólo para interpolar en el intervalo $[x_0, x_N]$.

Teorema 4.4 (Cotas del error para la interpolación de Lagrange con nodos equiespaciados). Supongamos que $f(x)$ está definida en un intervalo $[a, b]$, que contiene los nodos equiespaciados $x_k = x_0 + hk$. Supongamos además que $f(x)$ y sus derivadas, hasta la de orden 4, son continuas (por tanto, acotadas) en el subintervalo $[x_0, x_N]$; es decir,

$$(25) \quad |f^{(N+1)}(x)| \leq M_{N+1} \quad \text{para } x_0 \leq x \leq x_N,$$

para $N = 1, 2, 3$. Entonces los términos del error dados en (16) correspondientes a los casos $N = 1, 2$ y 3 admiten cotas de su tamaño expresables de manera cómoda por

$$(26) \quad |E_1(x)| \leq \frac{h^2 M_2}{8} \quad \text{válida para } x \in [x_0, x_1],$$

$$(27) \quad |E_2(x)| \leq \frac{h^3 M_3}{9\sqrt{3}} \quad \text{válida para } x \in [x_0, x_2],$$

$$(28) \quad |E_3(x)| \leq \frac{h^4 M_4}{24} \quad \text{válida para } x \in [x_0, x_3].$$

Demostración. Establecemos la cota dada en (26) y dejamos las otras como ejercicio. Haciendo el cambio de variable $x - x_0 = t$, con lo que $x - x_1 = t - h$, el término del error $E_1(x)$ puede escribirse como

$$(29) \quad E_1(x) = E_1(x_0 + t) = \frac{(t^2 - ht)f''(c)}{2!} \quad \text{para } 0 \leq t \leq h.$$

La cota de la derivada en este caso es

$$(30) \quad |f''(c)| \leq M_2 \quad \text{para } x_0 \leq c \leq x_1.$$

Ahora vamos a determinar una cota de la expresión $(t^2 - ht)$ del numerador de la fórmula (29); denotemos por $\Phi(t) = t^2 - ht$ esta expresión. Puesto que $\Phi'(t) = 2t - h$, sólo hay un punto crítico $t = h/2$ solución de $\Phi'(t) = 0$ y, por tanto, los valores extremos de $\Phi(t)$ en $[0, h]$ se alcanzan en los extremos del intervalo, donde $\Phi(0) = 0$ y $\Phi(h) = 0$, o en el punto crítico, donde $\Phi(h/2) = -h^2/4$. Como nos interesa el de mayor tamaño, obtenemos

$$(31) \quad |\Phi(t)| = |t^2 - ht| \leq \frac{|-h^2|}{4} = \frac{h^2}{4} \quad \text{para } 0 \leq t \leq h.$$

Al usar las cotas (30) y (31) para estimar la magnitud del producto que hay en el numerador de la fórmula (29), resulta

$$(32) \quad |E_1(x)| = \frac{|\Phi(t)||f^{(2)}(c)|}{2!} \leq \frac{h^2 M_2}{8}$$

lo que establece la fórmula (26). •

Comparación de la exactitud en términos de $O(h^{N+1})$

El Teorema 4.4 nos permite entender la simple relación que hay entre los términos del error para las interpolaciones lineal, cuadrática y cúbica. En los tres casos la cota del error $|E_N(x)|$ depende de h de dos maneras distintas; en primer lugar, el término h^{N+1} aparece explícitamente en la expresión de dicha cota, de manera que $|E_N(x)|$ es directamente proporcional a h^{N+1} . En segundo lugar, para N fijo, los valores de M_{N+1} dependen generalmente de h y tienden a $|f^{(N+1)}(x_0)|$ cuando h tiende a cero. Por consiguiente, cuando h tiende a cero, se tiene que $|E_N(x)|$ converge a cero a la misma velocidad que h^{N+1} . En la Definición 1.9 introdujimos la notación $O(h^N)$ que nos permite analizar cómodamente este tipo de comportamiento. Por ejemplo, la cota del error dada en (26) puede expresarse como

$$|E_1(x)| = O(h^2) \quad \text{válida para } x \in [x_0, x_1]$$

y, al emplear la notación $O(h^2)$ en vez de $h^2 M_2 / 8$ en la relación (26), lo que queremos expresar es la idea de que la cota del término del error es, aproximadamente, un múltiplo de h^2 ; es decir,

$$|E_1(x)| \leq Ch^2 \approx O(h^2).$$

Como consecuencia, si las derivadas de $f(x)$ están uniformemente acotadas en el intervalo $[a, b]$ y $|h| < 1$, entonces tomando N grande conseguiremos que h^{N+1} sea pequeño, de manera que el error del polinomio interpolador decrece con el grado.

Ejemplo 4.8. Consideremos $y = f(x) = \cos(x)$ en el intervalo $[0.0, 1.2]$. Vamos a usar las fórmulas (26), (27) y (28) para determinar las cotas del error de los polinomios interpoladores de Lagrange $P_1(x)$, $P_2(x)$ y $P_3(x)$ construidos en los Ejemplos 4.6 y 4.7.

En primer lugar, determinamos las cotas M_2 , M_3 y M_4 de las derivadas $|f^{(2)}(x)|$, $|f^{(3)}(x)|$ y $|f^{(4)}(x)|$, respectivamente, en el intervalo $[0.0, 1.2]$:

$$|f^{(2)}(x)| = |- \cos(x)| \leq |- \cos(0.0)| = 1.000000 = M_2,$$

$$|f^{(3)}(x)| = |\sin(x)| \leq |\sin(1.2)| = 0.932039 = M_3,$$

$$|f^{(4)}(x)| = |\cos(x)| \leq |\cos(0.0)| = 1.000000 = M_4.$$

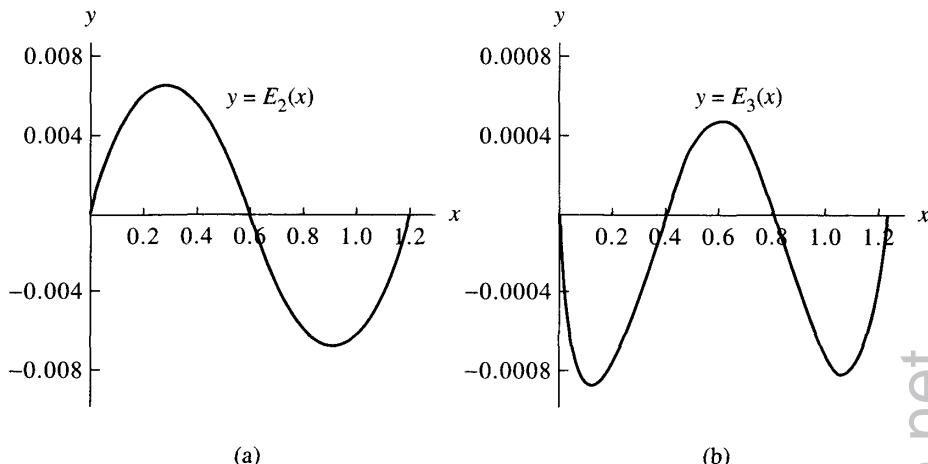


Figura 4.13 (a) La función del error $E_2(x) = \cos(x) - P_2(x)$. (b) La función del error $E_3(x) = \cos(x) - P_3(x)$.

Para $P_1(x)$ el espaciado entre los nodos es $h = 1.2$, luego su cota del error es

$$(33) \quad |E_1(x)| \leq \frac{h^2 M_2}{8} \leq \frac{(1.2)^2(1.000000)}{8} = 0.180000.$$

Para $P_2(x)$ el espaciado entre sus nodos es $h = 0.6$, luego su cota del error es

$$(34) \quad |E_2(x)| \leq \frac{h^3 M_3}{9\sqrt{3}} \leq \frac{(0.6)^3(0.932039)}{9\sqrt{3}} = 0.012915.$$

Para $P_3(x)$ el espaciado entre sus nodos es $h = 0.4$, luego su cota del error es

$$(35) \quad |E_3(x)| \leq \frac{h^4 M_4}{24} \leq \frac{(0.4)^4(1.000000)}{24} = 0.001067.$$

En el Ejemplo 4.6 vimos que $|E_1(0.6)| = |\cos(0.6) - P_1(0.6)| = 0.144157$, así que la cota 0.180000 de (33) es razonable. Las gráficas de las funciones $E_2(x) = \cos(x) - P_2(x)$ y $E_3(x) = \cos(x) - P_3(x)$ se muestran en las Figuras 4.13 (a) y (b), respectivamente, y los cálculos numéricos correspondientes se dan en la Tabla 4.7. Usando los valores de esta tabla, vemos que $|E_2(1.0)| = |\cos(1.0) - P_2(1.0)| = 0.008416$ y que $|E_3(0.2)| = |\cos(0.2) - P_3(0.2)| = 0.000855$, que concuerdan razonablemente bien con las cotas 0.012915 y 0.001067 obtenidas en (34) y (35), respectivamente.

MATLAB

El siguiente programa determina el polinomio interpolador cuya gráfica pasa por un conjunto dado de puntos, dando como resultado un vector cuyas componentes

Tabla 4.7 Comparación de $f(x) = \cos(x)$ y de las aproximaciones cuadrática y cúbica $P_2(x)$ y $P_3(x)$.

x_k	$f(x_k) = \cos(x_k)$	$P_2(x_k)$	$E_2(x_k)$	$P_3(x_k)$	$E_3(x_k)$
0.0	1.000000	1.000000	0.0	1.000000	0.0
0.1	0.995004	0.990911	0.004093	0.995835	-0.000831
0.2	0.980067	0.973813	0.006253	0.980921	-0.000855
0.3	0.955336	0.948707	0.006629	0.955812	-0.000476
0.4	0.921061	0.915592	0.005469	0.921061	0.0
0.5	0.877583	0.874468	0.003114	0.877221	0.000361
0.6	0.825336	0.825336	0.0	0.824847	0.00089
0.7	0.764842	0.768194	-0.003352	0.764491	0.000351
0.8	0.696707	0.703044	-0.006338	0.696707	0.0
0.9	0.621610	0.629886	-0.008276	0.622048	-0.000438
1.0	0.540302	0.548719	-0.008416	0.541068	-0.000765
1.1	0.453596	0.459542	-0.005946	0.454320	-0.000724
1.2	0.362358	0.362358	0.0	0.362358	0.0

son los coeficientes del polinomio interpolador de Lagrange. El programa utiliza las instrucciones `poly` y `conv`. La instrucción `poly` produce un vector cuyas componentes son los coeficientes de un polinomio del que se especifican las raíces. La instrucción `conv` proporciona un vector cuyas componentes son los coeficientes de un polinomio que es el producto de otros dos dados.

Ejemplo 4.9. Vamos a calcular el producto de los polinomios $P(x)$ y $Q(x)$ de primer grado cuyas raíces son, respectivamente, 2 y 3.

```
>>P=poly(2)
P=
    1 -2
>>Q=poly(3)
Q=
    1 -3
>>conv(P,Q)
ans=
    1 -5 6
```

Así que el producto de $P(x)$ por $Q(x)$ es $x^2 - 5x + 6$. ■

Programa 4.1 (Polinomio interpolador de Lagrange). Construcción del polinomio interpolador de Lagrange $P(x) = \sum_{k=0}^N y_k L_{N,k}(x)$ que pasa por los $N + 1$ puntos (x_k, y_k) para $k = 0, 1, \dots, N$.

```

function [C,L]=lagran(X,Y)
% Datos
%   - X es un vector que contiene la lista de las abscisas
%   - Y es un vector que contiene la lista de las ordenadas
% Resultados
%   - C es la matriz que contiene los coeficientes del
%     polinomio interpolador de Lagrange
%   - L es la matriz que contiene los coeficientes de los
%     polinomios coeficientes de Lagrange

w=length(X);
n=w-1;
L=zeros(w,w);

% Formación de los polinomios coeficientes de Lagrange
for k=1:n+1
    V=1;
    for j=1:n+1
        if k~=j
            V=conv(V,poly(X(j)))/(X(k)-X(j));
        end
    end
    L(k,:)=V;
end

% Cálculo de los coeficientes del polinomio
% interpolador de Lagrange
C=Y*L;

```

Ejercicios

- Determine, en los siguientes casos, el polinomio interpolador de Lagrange para aproximar la función $f(x) = x^3$.
 - El polinomio lineal $P_1(x)$ para los nodos $x_0 = -1$ y $x_1 = 0$.
 - El polinomio cuadrático $P_2(x)$ para los nodos $x_0 = -1$, $x_1 = 0$ y $x_2 = 1$.
 - El polinomio cúbico $P_3(x)$ para los nodos $x_0 = -1$, $x_1 = 0$, $x_2 = 1$ y $x_3 = 2$.
 - El polinomio lineal $Q_1(x)$ para los nodos $x_0 = 1$ y $x_1 = 2$.
 - El polinomio cuadrático $Q_2(x)$ para los nodos $x_0 = 0$, $x_1 = 1$ y $x_2 = 2$.
- Sea $f(x) = x + 2/x$.
 - Use el polinomio interpolador de Lagrange cuadrático con nodos $x_0 = 1$, $x_1 = 2$ y $x_2 = 2.5$ para aproximar $f(1.5)$ y $f(1.2)$.

- (b) Use el polinomio interpolador de Lagrange cuadrático con nodos $x_0 = 0.5$, $x_1 = 1$, $x_2 = 2$ y $x_3 = 2.5$ para aproximar $f(1.5)$ y $f(1.2)$.
3. Sea $f(x) = 2 \operatorname{sen}(\pi x/6)$ (con el ángulo medido en radianes).
- (a) Use el polinomio interpolador de Lagrange cuadrático con nodos $x_0 = 0$, $x_1 = 1$ y $x_2 = 3$ para aproximar $f(2)$ y $f(2.4)$.
- (b) Use el polinomio interpolador de Lagrange cuadrático con nodos $x_0 = 0$, $x_1 = 1$, $x_2 = 3$ y $x_3 = 5$ para aproximar $f(2)$ y $f(2.4)$.
4. Sea $f(x) = 2 \operatorname{sen}(\pi x/6)$ (con el ángulo medido en radianes).
- (a) Use el polinomio interpolador de Lagrange cuadrático con nodos $x_0 = 0$, $x_1 = 1$ y $x_2 = 3$ para aproximar $f(4)$ y $f(3.5)$.
- (b) Use el polinomio interpolador de Lagrange cúbico con nodos $x_0 = 0$, $x_1 = 1$, $x_2 = 3$ y $x_3 = 5$ para aproximar $f(4)$ y $f(3.5)$.
5. Escriba, para las siguientes funciones $f(x)$, el término del error $E_3(x)$ del polinomio interpolador de Lagrange cúbico con nodos $x_0 = -1$, $x_1 = 0$, $x_2 = 3$ y $x_3 = 4$.
- (a) $f(x) = 4x^3 - 3x + 2$
- (b) $f(x) = x^4 - 2x^3$
- (c) $f(x) = x^5 - 5x^4$
6. Sea $f(x) = x^x$.
- (a) Determine el polinomio interpolador de Lagrange cuadrático $P_2(x)$ para los nodos $x_0 = 1$, $x_1 = 1.25$ y $x_2 = 1.5$.
- (b) Use el polinomio calculado en el apartado (a) para estimar el valor medio de $f(x)$ en el intervalo $[1, 1.5]$.
- (c) Use la expresión (27) del Teorema 4.4 para hallar una cota del error que se produce al aproximar $f(x)$ mediante $P_2(x)$.
7. Consideremos los polinomios coeficientes de Lagrange $L_{2,k}(x)$ que se usan para calcular el polinomio interpolador de Lagrange cuadrático con nodos x_0 , x_1 y x_2 . Se define $g(x) = L_{2,0}(x) + L_{2,1}(x) + L_{2,2}(x) - 1$.
- (a) Pruebe que g es un polinomio de grado menor o igual que dos.
- (b) Pruebe que $g(x_k) = 0$ para $k = 0, 1, 2$.
- (c) Pruebe que $g(x) = 0$ para todo x . *Indicación.* Use el teorema fundamental del álgebra.
8. Sean $L_{N,0}(x)$, $L_{N,1}(x)$, ..., $L_{N,N}(x)$ los polinomios coeficientes de Lagrange para los $N + 1$ nodos x_0 , x_1 , ..., x_{N-1} y x_N . Pruebe que $\sum_{k=0}^N L_{N,k}(x) = 1$ para cualquier número real x .
9. Sea $f(x)$ un polinomio de grado menor o igual que N . Sea $P_N(x)$ el polinomio interpolador de Lagrange de $f(x)$ de grado menor o igual que N para los $N + 1$ nodos x_0 , x_1 , ..., x_N . Pruebe que $f(x) = P_N(x)$ para todo x . *Indicación.* Pruebe que el término del error $E_N(x)$ es idénticamente cero.

10. Consideremos la función $f(x) = \sin(x)$ en el intervalo $[0, 1]$. Use el Teorema 4.4 para determinar el tamaño de paso correspondiente h para el cual
- el polinomio de interpolación de Lagrange lineal tiene una precisión de 10^{-6} (o sea, halle h tal que $|E_1(x)| < 5 \times 10^{-7}$).
 - el polinomio de interpolación de Lagrange cuadrático tiene una precisión de 10^{-6} (o sea, halle h tal que $|E_2(x)| < 5 \times 10^{-7}$).
 - el polinomio de interpolación de Lagrange cúbico tiene una precisión de 10^{-6} (o sea, halle h tal que $|E_3(x)| < 5 \times 10^{-7}$).
11. Partiendo de la fórmula (16) con $N = 2$ pruebe la desigualdad (27). Sean $x_1 = x_0 + h$, $x_2 = x_0 + 2h$. Para ello, pruebe que si $x_0 \leq x \leq x_2$ entonces

$$|x - x_0||x - x_1||x - x_2| \leq \frac{2h^3}{3 \times 3^{1/2}}.$$

Indicación. Usando las sustituciones $t = x - x_1$, $t + h = x - x_0$ y $t - h = x - x_2$, considere la función $v(t) = t^3 - th^2$ en el intervalo $-h \leq t \leq h$; resuelva $v'(t) = 0$ para hallar el máximo y despeje, en función de h , el valor de t en el que se alcanza dicho máximo

12. *Interpolación lineal bidimensional.* Consideremos el plano cuya ecuación viene dada por $z = P(x, y) = A + Bx + Cy$ que pasa por los tres puntos (x_0, y_0, z_0) , (x_1, y_1, z_1) y (x_2, y_2, z_2) . Entonces los coeficientes A , B y C son la solución del sistema lineal

$$\begin{aligned} A + Bx_0 + Cy_0 &= z_0 \\ A + Bx_1 + Cy_1 &= z_1 \\ A + Bx_2 + Cy_2 &= z_2 \end{aligned}$$

- Determine A , B y C de manera que $z = P(x, y)$ pase por los puntos $(1, 1, 5)$, $(2, 1, 3)$ y $(1, 2, 9)$.
 - Determine A , B y C de manera que $z = P(x, y)$ pase por los puntos $(1, 1, 2.5)$, $(2, 1, 0)$ y $(1, 2, 4)$.
 - Determine A , B y C de manera que $z = P(x, y)$ pase por los puntos $(2, 1, 5)$, $(1, 3, 7)$ y $(3, 2, 4)$.
 - ¿Es posible hallar A , B y C de manera que $z = P(x, y)$ pase por los puntos $(1, 2, 5)$, $(3, 2, 7)$ y $(1, 2, 0)$? ¿Por qué?
13. Use el Teorema 1.7, el teorema de Rolle generalizado, y la función auxiliar

$$g(t) = f(t) - P_N(t) - E_N(x) \frac{(t - x_0)(t - x_1) \cdots (t - x_N)}{(x - x_0)(x - x_1) \cdots (x - x_N)},$$

donde $P_N(x)$ es el polinomio interpolador de Lagrange de grado N , para probar que el término del error $E_N(x) = f(x) - P_N(x)$ puede escribirse como

$$E_N(x) = (x - x_0)(x - x_1) \cdots (x - x_N) \frac{f^{(N+1)}(c)}{(N+1)!}.$$

Indicación. Calcule $g^{(N+1)}(t)$ y evalúe esta función en el punto $t = c$ cuya existencia garantiza el teorema citado.

Algoritmos y programas

1. Use el Programa 4.1 para hallar los coeficientes del polinomio interpolador del Problema 2(i), apartados (a), (b) y (c), de la subsección “Algoritmos y programas” de la Sección 4.2. Esboce en un mismo dibujo las gráficas de cada función y del correspondiente polinomio interpolador.
2. En la siguiente tabla se muestran temperaturas que fueron medidas cada hora, durante un lapso total de 5 horas, en Sevilla un día 27 de octubre.
 - (a) Use el Programa 4.1 para construir el polinomio interpolador de Lagrange correspondiente a los datos de la tabla.
 - (b) Use el Algoritmo 4.1(iii) para estimar la temperatura media durante el período de 5 horas dado.
 - (c) Dibuje los datos de la tabla y el polinomio del apartado (a) en el mismo gráfico y discuta el error que puede aparecer al usar dicho polinomio para estimar la temperatura media.

Hora	Grados (Celsius)
13	18
14	18
15	17
16	16
17	15
18	14

4 Polinomio interpolador de Newton

Hay ocasiones en las que resulta útil construir varios polinomios aproximantes $P_1(x)$, $P_2(x)$, ..., $P_N(x)$ y, después, elegir el más adecuado a nuestras necesidades. Si usamos los polinomios interpoladores de Lagrange, uno de los inconvenientes es que no hay relación entre la construcción de $P_{N-1}(x)$ y la de $P_N(x)$; cada polinomio debe construirse individualmente y el trabajo necesario para calcular polinomios de grado elevado requiere hacer muchas operaciones. Vamos a seguir ahora un camino de construcción distinto, en el cual los polinomios

interpoladores, que se llamarán de Newton, se calculan mediante un esquema recursivo

$$(1) \quad P_1(x) = a_0 + a_1(x - x_0),$$

$$(2) \quad P_2(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1),$$

$$(3) \quad P_3(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1)$$

$$+ a_3(x - x_0)(x - x_1)(x - x_2),$$

$$\vdots$$

$$(4) \quad P_N(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1)$$

$$+ a_3(x - x_0)(x - x_1)(x - x_2)$$

$$+ a_4(x - x_0)(x - x_1)(x - x_2)(x - x_3) + \cdots$$

$$+ a_N(x - x_0)(x - x_1)(x - x_2) \cdots (x - x_{N-1}).$$

El polinomio $P_N(x)$ se obtiene a partir de $P_{N-1}(x)$ usando la recurrencia

$$(5) \quad P_N(x) = P_{N-1}(x) + a_N(x - x_0)(x - x_1)(x - x_2) \cdots (x - x_{N-1}).$$

En este marco se dice que el polinomio $P_N(x)$ dado en la fórmula (4) es un **polinomio de Newton** con N **centros** x_0, x_1, \dots, x_{N-1} . Puesto que $P_N(x)$ involucra sumas de productos de factores lineales, siendo

$$a_N(x - x_0)(x - x_1)(x - x_2) \cdots (x - x_{N-1})$$

el de mayor grado, está claro que $P_N(x)$ es un polinomio de grado menor o igual que N .

Ejemplo 4.10. Dados los centros $x_0 = 1, x_1 = 3, x_2 = 4$ y $x_3 = 4.5$ y los coeficientes $a_0 = 5, a_1 = -2, a_2 = 0.5, a_3 = -0.1$ y $a_4 = 0.003$, vamos a calcular $P_1(x), P_2(x), P_3(x)$ y $P_4(x)$ y a evaluar $P_k(2.5)$ para $k = 1, 2, 3$ y 4 .

Usando las fórmulas (1) a (4), tenemos

$$P_1(x) = 5 - 2(x - 1),$$

$$P_2(x) = 5 - 2(x - 1) + 0.5(x - 1)(x - 3),$$

$$P_3(x) = P_2(x) - 0.1(x - 1)(x - 3)(x - 4),$$

$$P_4(x) = P_3(x) + 0.003(x - 1)(x - 3)(x - 4)(x - 4.5).$$

Ahora evaluamos estos polinomios en $x = 2.5$ y obtenemos

$$P_1(2.5) = 5 - 2(1.5) = 2,$$

$$P_2(2.5) = P_1(2.5) + 0.5(1.5)(-0.5) = 1.625,$$

$$P_3(2.5) = P_2(2.5) - 0.1(1.5)(-0.5)(-1.5) = 1.5125,$$

$$P_4(2.5) = P_3(2.5) + 0.003(1.5)(-0.5)(-1.5)(-2.0) = 1.50575. \blacksquare$$

Multiplicación encajada

Si N está fijo y tenemos que evaluar el polinomio $P_N(x)$ varias veces, entonces deberíamos usar multiplicaciones encajadas. El proceso es similar a la regla de Ruffini para polinomios escritos en su forma habitual; la diferencia reside en que a la variable independiente x hay que restarle los centros x_k . El esquema de multiplicaciones encajadas para $P_3(x)$ es

$$(6) \quad P_3(x) = ((a_3(x - x_2) + a_2)(x - x_1) + a_1)(x - x_0) + a_0;$$

de manera que, si deseamos evaluar $P_3(x)$ para un valor dado de x , entonces operamos desde dentro hacia afuera formando sucesivamente las cantidades:

$$(7) \quad \begin{aligned} S_3 &= a_3, \\ S_2 &= S_3(x - x_2) + a_2, \\ S_1 &= S_2(x - x_1) + a_1, \\ S_0 &= S_1(x - x_0) + a_0; \end{aligned}$$

esta última cantidad S_0 es $P_3(x)$.

Ejemplo 4.11. Vamos a calcular el valor de $P_3(2.5)$, que apareció en el Ejemplo 4.10, usando el esquema de multiplicaciones encajadas.

Usando la fórmula (6), escribimos

$$P_3(x) = ((-0.1(x - 4) + 0.5)(x - 3) - 2)(x - 1) + 5,$$

luego los valores de (7) son:

$$\begin{aligned} S_3 &= -0.1, \\ S_2 &= -0.1(2.5 - 4) + 0.5 = 0.65, \\ S_1 &= 0.65(2.5 - 3) - 2 = -2.325, \\ S_0 &= -2.325(2.5 - 1) + 5 = 1.5125 \end{aligned}$$

y, por tanto, $P_3(2.5) = 1.5125$. ■

Aproximación por polinomios: nodos y centros

Supongamos que queremos encontrar los coeficientes a_k de todos los polinomios $P_1(x), \dots, P_N(x)$ que nos sirven para aproximar una función dada $f(x)$. Entonces cada $P_k(x)$ es el polinomio de Newton que tiene como centros los puntos x_0, x_1, \dots, x_k y es también el polinomio de interpolación para los nodos x_0, x_1, \dots, x_{k+1} . Para el polinomio $P_1(x)$, los coeficientes a_0 y a_1 tienen un significado familiar; en este caso, se tiene que

$$(8) \quad P_1(x_0) = f(x_0) \quad \text{y} \quad P_1(x_1) = f(x_1).$$

De modo que, usando (1) y (8), podemos despejar a_0 y obtener

$$(9) \quad f(x_0) = P_1(x_0) = a_0 + a_1(x_0 - x_0) = a_0.$$

Por tanto, $a_0 = f(x_0)$. A continuación, usando (1), (8) y (9), tenemos

$$f(x_1) = P_1(x_1) = a_0 + a_1(x_1 - x_0) = f(x_0) + a_1(x_1 - x_0),$$

de donde podemos despejar a_1 :

$$(10) \quad a_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Por tanto, a_1 es la pendiente de la línea recta que pasa por los puntos $(x_0, f(x_0))$ y $(x_1, f(x_1))$.

Los coeficientes a_0 y a_1 son los mismos para $P_1(x)$ y $P_2(x)$ así que, para continuar, ahora evaluamos la expresión (2) en el nodo x_2 y obtenemos

$$(11) \quad f(x_2) = P_2(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1).$$

Usando en (11) los valores de a_0 y a_1 calculados en (9) y (10), nos queda

$$\begin{aligned} a_2 &= \frac{f(x_2) - a_0 - a_1(x_2 - x_0)}{(x_2 - x_0)(x_2 - x_1)} \\ &= \left(\frac{f(x_2) - f(x_0)}{x_2 - x_0} - \frac{f(x_1) - f(x_0)}{x_1 - x_0} \right) / (x_2 - x_1), \end{aligned}$$

que, por motivos computacionales, escribimos mejor como

$$(12) \quad a_2 = \left(\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0} \right) / (x_2 - x_0).$$

Es fácil probar que las dos fórmulas dadas para a_2 son equivalentes: basta escribir las fracciones poniendo su denominador común $(x_2 - x_1)(x_2 - x_0)(x_1 - x_0)$; detalles que dejamos como ejercicio. El numerador de (12) es la diferencia entre un cociente de diferencias; vamos a precisar esta idea de diferencias divididas, que será la herramienta con la cual podremos continuar el proceso recursivo.

Definición 4.1 (Diferencias divididas). Las *diferencias divididas* de una función $f(x)$ se definen como:

$$(13) \quad \begin{aligned} f[x_k] &= f(x_k), \\ f[x_{k-1}, x_k] &= \frac{f[x_k] - f[x_{k-1}]}{x_k - x_{k-1}}, \\ f[x_{k-2}, x_{k-1}, x_k] &= \frac{f[x_{k-1}, x_k] - f[x_{k-2}, x_{k-1}]}{x_k - x_{k-2}}, \\ f[x_{k-3}, x_{k-2}, x_{k-1}, x_k] &= \frac{f[x_{k-2}, x_{k-1}, x_k] - f[x_{k-3}, x_{k-2}, x_{k-1}]}{x_k - x_{k-3}}. \end{aligned}$$

Tabla 4.8 Tabla de diferencias divididas para $y = f(x)$.

x_k	$f[x_k]$	$f[,]$	$f[, ,]$	$f[, , ,]$	$f[, , , ,]$
x_0	$f[x_0]$				
x_1	$f[x_1]$	$f[x_0, x_1]$			
x_2	$f[x_2]$	$f[x_1, x_2]$	$f[x_0, x_1, x_2]$		
x_3	$f[x_3]$	$f[x_2, x_3]$	$f[x_1, x_2, x_3]$	$f[x_0, x_1, x_2, x_3]$	
x_4	$f[x_4]$	$f[x_3, x_4]$	$f[x_2, x_3, x_4]$	$f[x_1, x_2, x_3, x_4]$	$f[x_0, x_1, x_2, x_3, x_4]$

Las diferencias divididas de orden superior se forman de acuerdo con la siguiente regla recursiva:

$$(14) \quad f[x_{k-j}, x_{k-j+1}, \dots, x_k] = \frac{f[x_{k-j+1}, \dots, x_k] - f[x_{k-j}, \dots, x_{k-1}]}{x_k - x_{k-j}},$$

regla que se usa para construir la Tabla 4.8 de diferencias divididas. ▲

Los coeficientes a_k de los polinomios $P_N(x)$ dependen de los valores de interpolación $f(x_j)$ (con $j = 0, 1, \dots, k$); el siguiente teorema establece que a_k puede calcularse usando las diferencias divididas de $f(x)$:

$$(15) \quad a_k = f[x_0, x_1, \dots, x_k].$$

Teorema 4.5 (Polinomio interpolador de Newton). Supongamos que x_0, x_1, \dots, x_N son $N + 1$ números distintos en $[a, b]$. Entonces existe un único polinomio $P_N(x)$ de grado menor o igual que N tal que

$$f(x_j) = P_N(x_j) \quad \text{para } j = 0, 1, \dots, N.$$

La forma de Newton de este polinomio interpolador es

$$(16) \quad P_N(x) = a_0 + a_1(x - x_0) + \dots + a_N(x - x_0)(x - x_1)\dots(x - x_{N-1}),$$

siendo $a_k = f[x_0, x_1, \dots, x_k]$ para $k = 0, 1, \dots, N$.

Observación. Si $\{(x_j, y_j)\}_{j=0}^N$ es un conjunto de puntos cuyas abscisas son todas distintas, entonces los valores $f(x_j) = y_j$ pueden usarse para construir el único polinomio de grado menor o igual que N que pasa por los $N + 1$ puntos dados.

Corolario 4.2 (Aproximación de Newton). Supongamos que $P_N(x)$ es el polinomio interpolador de Newton dado en el Teorema 4.5 y que lo usamos para aproximar la función $f(x)$, esto es,

$$(17) \quad f(x) = P_N(x) + E_N(x).$$

Tabla 4.9 Tabla de diferencias divididas del polinomio del Ejemplo 4.12.

x_k	$f[x_k]$	Primera diferencia dividida	Segunda diferencia dividida	Tercera diferencia dividida	Cuarta diferencia dividida	Quinta diferencia dividida
$x_0 = 1$	-3					
$x_1 = 2$	0	3				
$x_2 = 3$	15	15	6			
$x_3 = 4$	48	33	9	1		
$x_4 = 5$	105	57	12	1	0	
$x_5 = 6$	192	87	15	1	0	0

Si $f \in C^{N+1}[a, b]$, entonces para cada $x \in [a, b]$ existe un número $c = c(x)$ en (a, b) , tal que el término del error puede escribirse como

$$(18) \quad E_N(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_N)f^{(N+1)}(c)}{(N + 1)!}.$$

Observación. El término del error $E_N(x)$ es, naturalmente, el mismo que el del polinomio interpolador de Lagrange que vimos en la fórmula (16) de la Sección 4.3.

Resulta interesante partir de una función polinomial $f(x)$ de grado N ya conocida y calcular su tabla de diferencias divididas; en este caso sabemos que $f^{(N+1)}(x) = 0$ para todo x , y el cálculo nos permitirá comprobar que la diferencia dividida $(N + 1)$ -ésima también es cero. La razón, aparte de la expresión del error, es que la diferencia dividida (14) es proporcional a una aproximación numérica de la derivada j -ésima.

Ejemplo 4.12. Sea $f(x) = x^3 - 4x$. Vamos a construir la tabla de diferencias divididas para los nodos $x_0 = 1, x_1 = 2, \dots, x_5 = 6$ y a calcular el polinomio interpolador de Newton $P_3(x)$ para los nodos x_0, x_1, x_2 y x_3 .

Los cálculos se muestran en la Tabla 4.9. Los coeficientes de $P_3(x)$ aparecen en la diagonal de la tabla de diferencias divididas y valen, respectivamente, $a_0 = -3, a_1 = 3, a_2 = 6$ y $a_3 = 1$. Los centros $x_0 = 1, x_1 = 2$ y $x_2 = 3$ son los valores dispuestos en la primera columna, así que, de acuerdo con la fórmula (3), podemos escribir

$$P_3(x) = -3 + 3(x - 1) + 6(x - 1)(x - 2) + (x - 1)(x - 2)(x - 3).$$

Ejemplo 4.13. Vamos a construir ahora la tabla de diferencias divididas de la función $f(x) = \cos(x)$ para los puntos $(k, \cos(k))$, con $k = 0, 1, 2, 3$ y 4 . Luego la

Tabla 4.10 Tabla de diferencias divididas usada para construir los polinomios interpoladores de Newton $P_k(x)$ en el Ejemplo 4.13.

x_k	$f[x_k]$	$f[,]$	$f[, ,]$	$f[, , ,]$	$f[, , , ,]$
$x_0 = 0.0$	1.0000000				
	0.5403023	-0.4596977			
	-0.4161468	-0.9564491	-0.2483757		
	-0.9899925	-0.5738457	0.1913017	0.1465592	
	-0.6536436	0.3363499	0.4550973	0.0879318	-0.0146568

usaremos para calcular los coeficientes a_k y los cuatro polinomios interpoladores de Newton $P_k(x)$, para $k = 1, 2, 3$ y 4 .

Hemos redondeado a siete cifras decimales, en aras de la simplicidad, los valores calculados, que se muestran en la Tabla 4.10. Podemos usar los nodos x_0, x_1, x_2 y x_3 y los elementos diagonales a_0, a_1, a_2, a_3 y a_4 de la Tabla 4.10 en la fórmula (16) para escribir los cuatro polinomios interpoladores de Newton

$$\begin{aligned}
 P_1(x) &= 1.0000000 - 0.4596977(x - 0.0), \\
 P_2(x) &= 1.0000000 - 0.4596977(x - 0.0) - 0.2483757(x - 0.0)(x - 1.0), \\
 P_3(x) &= 1.0000000 - 0.4596977(x - 0.0) - 0.2483757(x - 0.0)(x - 1.0) \\
 &\quad + 0.1465592(x - 0.0)(x - 1.0)(x - 2.0), \\
 P_4(x) &= 1.0000000 - 0.4596977(x - 0.0) - 0.2483757(x - 0.0)(x - 1.0) \\
 &\quad + 0.1465592(x - 0.0)(x - 1.0)(x - 2.0) \\
 &\quad - 0.0146568(x - 0.0)(x - 1.0)(x - 2.0)(x - 3.0).
 \end{aligned}$$

Veamos con un ejemplo cómo se calcula el coeficiente a_2 .

$$\begin{aligned}
 f[x_0, x_1] &= \frac{f[x_1] - f[x_0]}{x_1 - x_0} = \frac{0.5403023 - 1.0000000}{1.0 - 0.0} = -0.4596977, \\
 f[x_1, x_2] &= \frac{f[x_2] - f[x_1]}{x_2 - x_1} = \frac{-0.4161468 - 0.5403023}{2.0 - 1.0} = -0.9564491, \\
 a_2 &= f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \frac{-0.9564491 + 0.4596977}{2.0 - 0.0} = -0.2483757.
 \end{aligned}$$

Las gráficas de la curva $y = \cos(x)$ y las de los polinomios $y = P_1(x)$, $y = P_2(x)$ e $y = P_3(x)$ se muestran en las Figuras 4.14 (a), (b) y (c), respectivamente. ■

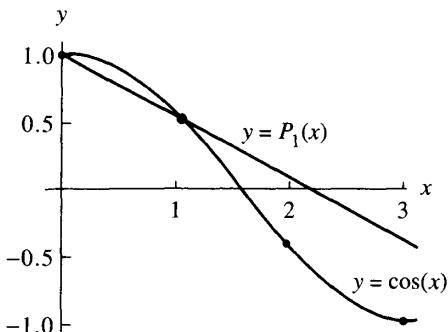


Figura 4.14 (a) Gráfica de $y = \cos(x)$ y del polinomio interpolador de Newton lineal $y = P_1(x)$ para los nodos $x_0 = 0.0$ y $x_1 = 1.0$.

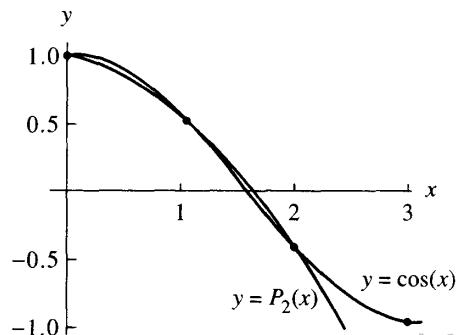


Figura 4.14 (b) Gráfica de $y = \cos(x)$ y del polinomio interpolador de Newton cuadrático $y = P_2(x)$ para los nodos $x_0 = 0.0$, $x_1 = 1.0$ y $x_2 = 2.0$.

MATLAB

Para realizar los cálculos con un computador, las diferencias divididas de la Tabla 4.8 se almacenan en una matriz cuyos elementos podemos llamar $D(k, j)$, de manera que

$$(19) \quad D(k, j) = f[x_{k-j}, x_{k-j+1}, \dots, x_k] \quad \text{para } j \leq k.$$

La relación (14) se emplea entonces para calcular recursivamente los elementos de la matriz:

$$(20) \quad D(k, j) = \frac{D(k, j-1) - D(k-1, j-1)}{x_k - x_{k-j}}.$$

Hagamos notar que el valor del coeficiente a_k dado en (15) es, por tanto, el elemento diagonal correspondiente $a_k = D(k, k)$. Con estos elementos damos a continuación el algoritmo para calcular las diferencias divididas y escribir $P_N(x)$ en la forma habitual. En el Problema 2 de la subsección “Algoritmos y programas” se sugiere la manera de modificar el algoritmo de forma que los coeficientes $\{a_k\}$ se calculen usando un vector, lo que es más conveniente, en términos del gasto de memoria del computador, que usar una matriz.

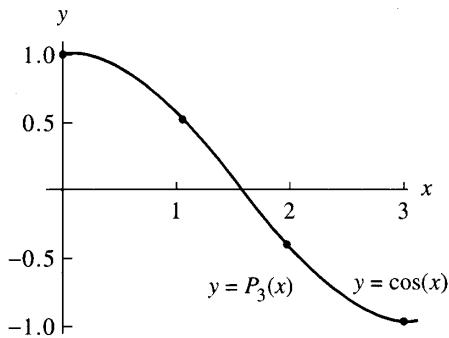


Figura 4.14 (c) Gráfica de $y = \cos(x)$ y del polinomio interpolador de Newton cúbico $y = P_3(x)$ para los nodos $x_0 = 0.0$, $x_1 = 1.0$, $x_2 = 2.0$ y $x_3 = 3.0$.

Programa 4.2 (Polinomio interpolador de Newton). Construcción del polinomio interpolador de Newton de grado menor o igual que N que pasa por los puntos $(x_k, y_k) = (x_k, f(x_k))$ para $k = 0, 1, \dots, N$:

$$(21) \quad P(x) = d_{0,0} + d_{1,1}(x - x_0) + d_{2,2}(x - x_0)(x - x_1) + \cdots + d_{N,N}(x - x_0)(x - x_1) \cdots (x - x_{N-1}),$$

siendo

$$d_{k,0} = y_k \quad \text{y} \quad d_{k,j} = \frac{d_{k,j-1} - d_{k-1,j-1}}{x_k - x_{k-j}}.$$

```
function [C,D]=newpoly(X,Y)
% Datos
%     - X es un vector con la lista de las abscisas
%     - Y es un vector con la lista de las ordenadas
% Resultados
%     - C es un vector que contiene los coeficientes
%         del polinomio interpolador de Newton, escrito
%         de forma habitual, en potencias decrecientes de x
%     - D es la tabla de diferencias divididas

n=length(X);
D=zeros(n,n);
D(:,1)=Y';
for j=2:n
    for k=j:n
        D(k,j)=(D(k,j-1)-D(k-1,j-1))/(X(k)-X(k-j+1));
    end
end
```

```

end

% Cálculo del vector que contiene los coeficientes
% del polinomio interpolador de Newton, escrito
% de forma habitual, en potencias decrecientes de x
C=D(n,n);
for k=(n-1):-1:1
    C=conv(C,poly(X(k)));
    m=length(C);
    C(m)=C(m)+D(k,k);
end

```

Ejercicios

En los Ejercicios 1 a 4, use los centros x_0, x_1, x_2 y x_3 y los coeficientes a_0, a_1, a_2, a_3 y a_4 que se dan para hallar los polinomios interpoladores de Newton $P_1(x)$, $P_2(x)$, $P_3(x)$ y $P_4(x)$ y calcule los valores de estos polinomios en $x = c$. *Indicación.* Use las relaciones (1)–(4) y las técnicas del Ejemplo 4.9.

1. $a_0 = 4 \quad a_1 = -1 \quad a_2 = 0.4 \quad a_3 = 0.01 \quad a_4 = -0.002$
 $x_0 = 1 \quad x_1 = 3 \quad x_2 = 4 \quad x_3 = 4.5 \quad c = 2.5$
2. $a_0 = 5 \quad a_1 = -2 \quad a_2 = 0.5 \quad a_3 = -0.1 \quad a_4 = 0.003$
 $x_0 = 0 \quad x_1 = 1 \quad x_2 = 2 \quad x_3 = 3 \quad c = 2.5$
3. $a_0 = 7 \quad a_1 = 3 \quad a_2 = 0.1 \quad a_3 = 0.05 \quad a_4 = -0.04$
 $x_0 = -1 \quad x_1 = 0 \quad x_2 = 1 \quad x_3 = 4 \quad c = 3$
4. $a_0 = -2 \quad a_1 = 4 \quad a_2 = -0.04 \quad a_3 = 0.06 \quad a_4 = 0.005$
 $x_0 = -3 \quad x_1 = -1 \quad x_2 = 1 \quad x_3 = 4 \quad c = 2$

En los Ejercicios 5 a 8:

- (a) Calcule la tabla de diferencias divididas para la función tabulada.
- (b) Escriba los polinomios interpoladores de Newton $P_1(x)$, $P_2(x)$, $P_3(x)$ y $P_4(x)$.
- (c) Calcule los valores de los polinomios hallados en el apartado (b) en los puntos x que se dan.
- (d) Compare los valores obtenidos en el apartado (c) con los valores de la función $f(x)$.

5. $f(x) = x^{1/2}$

$x = 4.5, 7.5$

k	x_k	$f(x_k)$
0	4.0	2.00000
1	5.0	2.23607
2	6.0	2.44949
3	7.0	2.64575
4	8.0	2.82843

6. $f(x) = 3.6/x$

$x = 2.5, 3.5$

k	x_k	$f(x_k)$
0	1.0	3.60
1	2.0	1.80
2	3.0	1.20
3	4.0	0.90
4	5.0	0.72

7. $f(x) = 3 \operatorname{sen}^2(\pi x/6)$
 $x = 1.5, 3.5$

k	x_k	$f(x_k)$
0	0.0	0.00
1	1.0	0.75
2	2.0	2.25
3	3.0	3.00
4	4.0	2.25

8. $f(x) = e^{-x}$
 $x = 0.5, 1.5$

k	x_k	$f(x_k)$
0	0.0	1.00000
1	1.0	0.36788
2	2.0	0.13534
3	3.0	0.04979
4	4.0	0.01832

9. Consideremos $M + 1$ puntos $(x_0, y_0), \dots, (x_M, y_M)$.

- (a) Pruebe que si las $(N + 1)$ -ésimas diferencias divididas son cero, entonces todas las siguientes hasta la M -ésima son también cero.
- (b) Pruebe que si las $(N + 1)$ -ésimas diferencias divididas son cero, entonces existe un polinomio $P_N(x)$ de grado N tal que

$$P_N(x_k) = y_k \quad \text{para } k = 0, 1, \dots, M.$$

En los Ejercicios 10 a 12, use el resultado del Ejercicio 9 para calcular el polinomio $P_N(x)$ que pasa por los $M + 1$ puntos ($N < M$).

x_k	y_k
0	-2
1	2
2	4
3	4
4	2
5	-2

x_k	y_k
1	8
2	17
3	24
4	29
5	32
6	33

x_k	y_k
0	5
1	5
2	3
3	5
4	17
5	45
6	95

13. Use el Corolario 4.2 para hallar una cota del máximo error
- $|E_2(x)|$
- que se comete en el intervalo
- $[0, \pi]$
- cuando usamos el polinomio interpolador de Newton

$P_2(x)$ para aproximar $f(x) = \cos(\pi x)$ en los centros $x_0 = 0$, $x_1 = \pi/2$ y $x_2 = \pi$.

Algoritmos y programas

- Use el Programa 4.2 y repita el Problema 2 de la subsección “Algoritmos y programas” de la Sección 4.3.
- En el Programa 4.2 la matriz D se emplea para almacenar la tabla de diferencias divididas.
 - Compruebe que la siguiente modificación del Programa 4.2 es una forma equivalente de calcular el polinomio interpolador de Newton.

```

for k=0:N
    A(k)=Y(k);
end
for j=1:N
    for k=N:-1:j
        A(k)=(A(k)-A(k-1))/(X(k)-X(k-j));
    end
end

```

- Repita el Problema 1 usando esta modificación del Programa 4.2.

4.5 Polinomios de Chebyshev (opcional)

Nos centramos ahora en el polinomio que interpola una función $f(x)$, definida en $[-1, 1]$, construido con los nodos $-1 \leq x_0 < x_1 < \dots < x_N \leq 1$. El polinomio interpolador, tanto en su forma de Lagrange como en su forma de Newton, satisface

$$f(x) = P_N(x) + E_N(x),$$

donde

$$(1) \quad E_N(x) = Q(x) \frac{f^{(N+1)}(c)}{(N+1)!}$$

y $Q(x)$ es el polinomio de grado $N+1$:

$$(2) \quad Q(x) = (x - x_0)(x - x_1) \cdots (x - x_N).$$

Ahora nos planteamos, usando la relación

$$|E_N(x)| \leq |Q(x)| \frac{\max\{|f^{(N+1)}(x)| : -1 \leq x \leq 1\}}{(N+1)!},$$

Tabla 4.11 Polinomios de Chebyshev desde $T_0(x)$ hasta $T_7(x)$.

$T_0(x) = 1$
$T_1(x) = x$
$T_2(x) = 2x^2 - 1$
$T_3(x) = 4x^3 - 3x$
$T_4(x) = 8x^4 - 8x^2 + 1$
$T_5(x) = 16x^5 - 20x^3 + 5x$
$T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1$
$T_7(x) = 64x^7 - 112x^5 + 56x^3 - 7x$

el problema, resuelto por P. L. Chebyshev, de cómo seleccionar el conjunto de nodos $\{x_k\}_{k=0}^N$ que haga mínimo el valor $\max\{|Q(x)| : -1 \leq x \leq 1\}$. Para ello, necesitamos saber qué son los polinomios de Chebyshev y cuáles son sus propiedades más importantes. Para empezar, los ocho primeros polinomios de Chebyshev se muestran en la Tabla 4.11.

Propiedades de los polinomios de Chebyshev

Propiedad 1. Relación de recurrencia

Los polinomios de Chebyshev pueden generarse de la siguiente manera: Tomamos $T_0(x) = 1$ y $T_1(x) = x$ y usamos la relación de recurrencia

$$(3) \quad T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x) \quad \text{para } k = 2, 3, \dots$$

Propiedad 2. Coeficiente líder

El coeficiente de x^N en $T_N(x)$ es 2^{N-1} para $N \geq 1$.

Propiedad 3. Simetría

Cuando $N = 2M$, el polinomio $T_{2M}(x)$ es una función par; esto es,

$$(4) \quad T_{2M}(-x) = T_{2M}(x).$$

Cuando $N = 2M + 1$, el polinomio $T_{2M+1}(x)$ es una función impar; esto es,

$$(5) \quad T_{2M+1}(-x) = -T_{2M+1}(x).$$

Propiedad 4. Representación trigonométrica en $[-1, 1]$

$$(6) \quad T_N(x) = \cos(N \arccos(x)) \quad \text{para } -1 \leq x \leq 1.$$

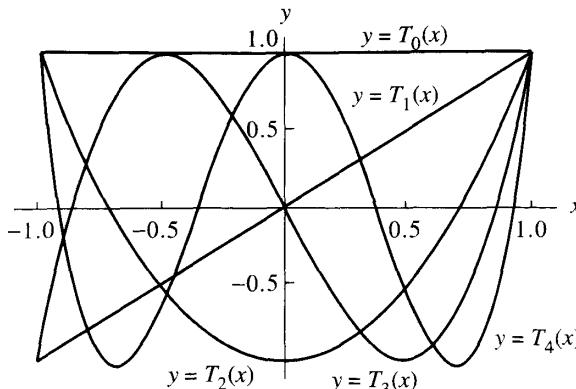


Figura 4.15 Gráficas de los polinomios de Chebyshev $T_0(x)$, $T_1(x)$, ..., $T_4(x)$ en $[-1, 1]$.

Propiedad 5. Ceros simples en $[-1, 1]$

$T_N(x)$ tiene N ceros distintos x_k que están todos en el intervalo $[-1, 1]$ (véase la Figura 4.15):

$$(7) \quad x_k = \cos\left(\frac{(2k+1)\pi}{2N}\right) \quad \text{para } k = 0, 1, \dots, N-1.$$

Estos valores se llaman *nodos (o abscisas) de Chebyshev*.

Propiedad 6. Valores extremos

$$(8) \quad |T_N(x)| \leq 1 \quad \text{para} \quad -1 \leq x \leq 1.$$

A menudo se utiliza la Propiedad 1 como definición de los polinomios de Chebyshev de grado alto. Veamos, por ejemplo, que $T_3(x) = 2xT_2(x) - T_1(x)$. Usando las expresiones de $T_1(x)$ y $T_2(x)$ dadas en la Tabla 4.11, obtenemos

$$2xT_2(x) - T_1(x) = 2x(2x^2 - 1) - x = 4x^3 - 3x = T_3(x).$$

La Propiedad 2 se prueba observando que la relación de recurrencia multiplica por 2 el coeficiente líder de $T_{N-1}(x)$ para obtener el de $T_N(x)$.

La Propiedad 3 puede establecerse probando (por inducción) que en $T_{2M}(x)$ sólo aparecen potencias pares de x y que en $T_{2M+1}(x)$ sólo aparecen potencias impares de x ; los detalles los dejamos como ejercicio.

La demostración de la Propiedad 4 puede hacerse usando la identidad trigonométrica

$$\cos(k\theta) = \cos(2\theta) \cos((k-2)\theta) - \sin(2\theta) \sin((k-2)\theta),$$

en la cual sustituimos $\cos(2\theta) = 2\cos^2(\theta) - 1$ y $\sin(2\theta) = 2\sin(\theta)\cos(\theta)$ para obtener

$\cos(k\theta) = 2\cos(\theta)(\cos(\theta)\cos((k-2)\theta) - \sin(\theta)\sin((k-2)\theta)) - \cos((k-2)\theta)$,
relación que se puede escribir de manera más simple como

$$\cos(k\theta) = 2\cos(\theta)\cos((k-1)\theta) - \cos((k-2)\theta).$$

Finalmente, sustituimos $\theta = \arccos(x)$ y queda

$$(9) \quad 2x\cos((k-1)\arccos(x)) - \cos((k-2)\arccos(x)) \\ = \cos(k\arccos(x)) \quad \text{para } -1 \leq x \leq 1.$$

Los dos primeros polinomios de Chebyshev son $T_0(x) = \cos(0\arccos(x)) = 1$ y $T_1(x) = \cos(1\arccos(x)) = x$. Supongamos que $T_k(x) = \cos(k\arccos(x))$ para $k = 2, 3, \dots, N-1$, entonces podemos usar la fórmula (3) junto con la igualdad (9) para establecer el caso general

$$\begin{aligned} T_N(x) &= 2xT_{N-1}(x) - T_{N-2}(x) \\ &= 2x\cos((N-1)\arccos(x)) - \cos((N-2)\arccos(x)) \\ &= \cos(N\arccos(x)) \quad \text{para } -1 \leq x \leq 1. \end{aligned}$$

Las Propiedades 5 y 6 son consecuencias de la Propiedad 4.

Minimax

El matemático ruso P. L. Chebyshev estudió el problema de minimizar la cota superior del término del error $|E_N(x)|$. La cota que venimos usando es el producto del valor máximo de $|Q(x)|$ (cuando x recorre el intervalo $[-1, 1]$) por el valor máximo de $|f^{(N+1)}(x)/(N+1)!|$ (cuando x recorre el intervalo $[-1, 1]$). Para minimizar el factor $\max\{|Q(x)|\}$, Chebyshev descubrió que x_0, x_1, \dots, x_N deben ser elegidos de manera que $Q(x) = (1/2^N)T_{N+1}(x)$.

Teorema 4.6. Supongamos que N está fijo. Entre todas las posibles elecciones del factor $Q(x)$ en la ecuación (2) (es decir, entre todas las posibles elecciones de nodos distintos $\{x_k\}_{k=0}^N$ en $[-1, 1]$), el polinomio $T(x) = T_{N+1}(x)/2^N$ es la única elección que verifica

$$\max\{|T(x)| : -1 \leq x \leq 1\} \leq \max\{|Q(x)| : -1 \leq x \leq 1\}.$$

Es más,

$$(10) \quad \max\{|T(x)| : -1 \leq x \leq 1\} = \frac{1}{2^N}.$$

Demostración. La demostración puede consultarse en la Referencia [29].

Tabla 4.12 Polinomios coeficientes de Lagrange usados para formar $P_3(x)$ con nodos equiespaciados $x_k = -1 + 2k/3$.

$$\begin{aligned}L_{3,0}(x) &= -0.06250000 + 0.06250000x + 0.56250000x^2 - 0.56250000x^3 \\L_{3,1}(x) &= 0.56250000 - 1.68750000x - 0.56250000x^2 + 1.68750000x^3 \\L_{3,2}(x) &= 0.56250000 + 1.68750000x - 0.56250000x^2 - 1.68750000x^3 \\L_{3,3}(x) &= -0.06250000 - 0.06250000x + 0.56250000x^2 + 0.56250000x^3\end{aligned}$$

Este resultado puede resumirse diciendo que, para la interpolación de Lagrange $f(x) = P_N(x) + E_N(x)$ en $[-1, 1]$, el menor valor de la cota del error

$$(\max\{|Q(x)|\})(\max\{|f^{(N+1)}(x)/(N+1)!\|})$$

se alcanza cuando los nodos $\{x_k\}$ son los nodos de Chebyshev de $T_{N+1}(x)$. Como ejemplo, vamos a ver qué pasa con los polinomios coeficientes de Lagrange que se usan para construir el polinomio interpolador cúbico $P_3(x)$; primero con nodos equiespaciados y luego con los nodos de Chebyshev. Recordemos que el polinomio interpolador de Lagrange de grado $N = 3$ se escribe como

$$(11) \quad P_3(x) = f(x_0)L_{3,0}(x) + f(x_1)L_{3,1}(x) + f(x_2)L_{3,2}(x) + f(x_3)L_{3,3}(x).$$

Nodos equiespaciados

Si aproximamos $f(x)$ por un polinomio de grado menor o igual que $N = 3$ en $[-1, 1]$, entonces los nodos equiespaciados son $x_0 = -1$, $x_1 = -1/3$, $x_2 = 1/3$ y $x_3 = 1$, que son valores cómodos de usar en los cálculos. Sustituyendo estos valores en la fórmula (8) de la Sección 4.3 y simplificando los resultados obtenemos los polinomios coeficientes $L_{3,k}(x)$ que se muestran en la Tabla 4.12.

Nodos de Chebyshev

Si queremos aproximar $f(x)$ por un polinomio de grado menor o igual que $N = 3$, usando los nodos de Chebyshev $x_0 = \cos(7\pi/8)$, $x_1 = \cos(5\pi/8)$, $x_2 = \cos(3\pi/8)$ y $x_3 = \cos(\pi/8)$, entonces el cálculo de los polinomios coeficientes puede ser tedioso si se hace a mano (pero podemos usar un computador). Después de hacer las simplificaciones oportunas, los resultados se muestran en la Tabla 4.13.

Ejemplo 4.14. Vamos a comparar los polinomios interpoladores de Lagrange de grado $N = 3$ de la función $f(x) = e^x$, que se obtienen usando los polinomios coeficientes de las Tablas 4.12 y 4.13, respectivamente.

Tabla 4.13 Polinomios coeficientes de Lagrange usados para formar $P_3(x)$ con los nodos de Chebyshev $x_k = \cos((7 - 2k)\pi/8)$.

$C_0(x) = -0.10355339 + 0.11208538x + 0.70710678x^2 - 0.76536686x^3$
$C_1(x) = 0.60355339 - 1.57716102x - 0.70710678x^2 + 1.84775906x^3$
$C_2(x) = 0.60355339 + 1.57716102x - 0.70710678x^2 - 1.84775906x^3$
$C_3(x) = -0.10355339 - 0.11208538x + 0.70710678x^2 + 0.76536686x^3$

Usando nodos equiespaciados, obtenemos el polinomio

$$P(x) = 0.99519577 + 0.99904923x + 0.54788486x^2 + 0.17615196x^3.$$

Para ello hemos calculado los valores de la función en los nodos

$$\begin{aligned} f(x_0) &= e^{(-1)} = 0.36787944, & f(x_1) &= e^{(-1/3)} = 0.71653131, \\ f(x_2) &= e^{(1/3)} = 1.39561243, & f(x_3) &= e^{(1)} = 2.71828183, \end{aligned}$$

y hemos usado los polinomios coeficientes $L_{3,k}(x)$ de la Tabla 4.12 formando la combinación lineal

$$\begin{aligned} P(x) &= 0.36787944L_{3,0}(x) + 0.71653131L_{3,1}(x) + 1.39561243L_{3,2}(x) \\ &\quad + 2.71828183L_{3,3}(x). \end{aligned}$$

De manera parecida, si usamos los nodos de Chebyshev, obtenemos

$$V(x) = 0.99461532 + 0.99893323x + 0.54290072x^2 + 0.17517569x^3.$$

Hagamos notar que los coeficientes son distintos de los de $P(x)$; la razón es que hemos usado nodos y valores de la función diferentes:

$$\begin{aligned} f(x_0) &= e^{-0.92387953} = 0.39697597, \\ f(x_1) &= e^{-0.38268343} = 0.68202877, \\ f(x_2) &= e^{0.38268343} = 1.46621380, \\ f(x_3) &= e^{0.92387953} = 2.51904417. \end{aligned}$$

Luego hemos usado el conjunto de polinomios coeficientes $C_k(x)$ que se muestran en la Tabla 4.13 para formar la combinación lineal:

$$V(x) = 0.39697597C_0(x) + 0.68202877C_1(x) + 1.46621380C_2(x) + 2.51904417C_3(x).$$

Para comparar las precisiones de $P(x)$ y $V(x)$, hemos dibujado las gráficas de los términos del error en las Figuras 4.16(a) y (b), respectivamente. El error máximo $|e^x - P(x)|$ se alcanza en $x = 0.75490129$, de manera que

$$|e^x - P(x)| \leq 0.00998481 \quad \text{para } -1 \leq x \leq 1.$$

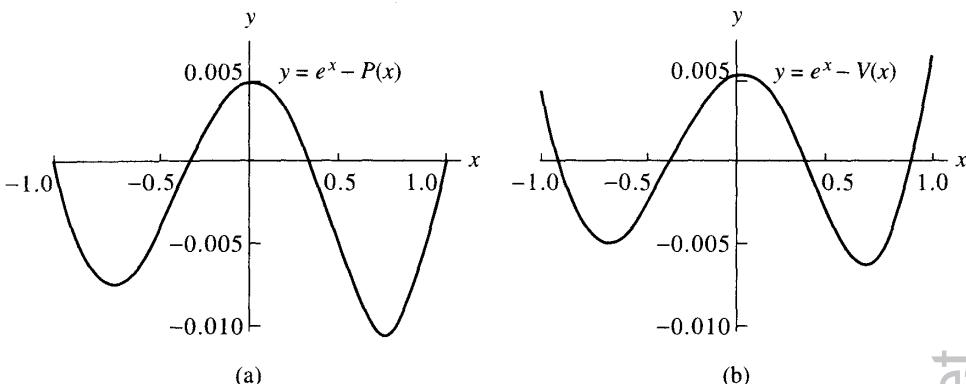


Figura 4.16 (a) El término del error $y = e^x - P(x)$ de la aproximación de Lagrange con nodos equiespaciados en $[-1, 1]$. (b) El término del error $y = e^x - V(x)$ de la aproximación de Lagrange con nodos de Chebyshev en $[-1, 1]$.

El error máximo $|e^x - V(x)|$ se alcanza en $x = 1$ y nos queda

$$|e^x - V(x)| \leq 0.00665687 \quad \text{para } -1 \leq x \leq 1.$$

Hagamos notar que el error máximo de $V(x)$ es como unos dos tercios del error máximo de $P(x)$ y, también, que el error de $V(x)$ está distribuido de forma más uniforme a lo largo del intervalo.

El fenómeno de Runge

Vamos a hacer un análisis algo más profundo de las ventajas que tiene el uso de los nodos de Chebyshev. Consideremos el polinomio interpolador de Lagrange de una función $f(x)$ en el intervalo $[-1, 1]$ para nodos equiespaciados, ¿tiende a cero el error $E_N(x) = f(x) - P_N(x)$ cuando N crece? Para funciones como $\sin(x)$ o e^x , cuyas derivadas están acotadas por la misma constante M , la respuesta es sí. En general, sin embargo, la respuesta a esa pregunta es no; es fácil hallar funciones para las que la sucesión $\{P_N(x)\}$ no converge. Por ejemplo, para $f(x) = 1/(1 + 12x^2)$, el máximo del término del error $E_N(x)$ crece cuando $N \rightarrow \infty$; esta falta de convergencia se conoce como **fenómeno de Runge** (véase la Referencia [90], págs. 275–278). El polinomio interpolador de Lagrange de grado 10 de esta función con 11 nodos equiespaciados se muestra en la Figura 4.17(a): cerca de los extremos del intervalo aparecen oscilaciones muy grandes y, si el número de nodos se incrementa, entonces las oscilaciones se hacen aún mayores. Este problema ocurre ¡porque los nodos están equiespaciados!

Si usamos los nodos de Chebyshev para construir un polinomio interpolador de grado 10 de $f(x) = 1/(1 + 12x^2)$, entonces el error es mucho menor, como se

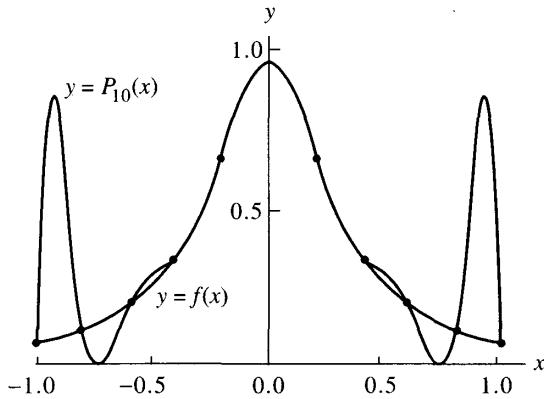


Figura 4.17 (a) La aproximación polinomial a $y = 1/(1 + 12x^2)$ para 11 nodos equiespaciados en $[-1, 1]$.

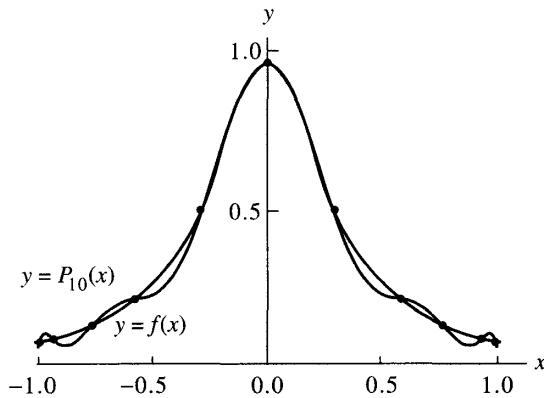


Figura 4.17 (b) La aproximación polinomial a $y = 1/(1 + 12x^2)$ para los 11 nodos de Chebyshev en $[-1, 1]$.

pone de manifiesto en la Figura 14.17(b). De hecho, cuando se usan los nodos de Chebyshev, el error $E_N(x)$ tiende a cero cuando $N \rightarrow \infty$. En general, puede probarse que si $f(x)$ y $f'(x)$ son continuas en $[-1, 1]$, entonces la interpolación con nodos de Chebyshev proporciona una sucesión de polinomios $\{P_N(x)\}$ que converge uniformemente a $f(x)$ en $[-1, 1]$.

Nodos de Chebyshev en otros intervalos

Algunas veces es necesario transformar un problema originalmente planteado en un intervalo $[a, b]$ y reformularlo como un problema en otro intervalo $[c, d]$ en el que se pueda resolver más cómodamente. En el caso que nos ocupa, si debemos obtener una aproximación $P_N(x)$ a $f(x)$ en un intervalo $[a, b]$, entonces hacemos

un cambio de variable para reformular el problema en $[-1, 1]$:

$$(12) \quad x = \left(\frac{b-a}{2} \right) t + \frac{a+b}{2} \quad \text{o bien} \quad t = 2 \frac{x-a}{b-a} - 1,$$

donde $a \leq x \leq b$ y $-1 \leq t \leq 1$.

Los nodos de Chebyshev correspondientes a $T_{N+1}(t)$ en $[-1, 1]$ son

$$(13) \quad t_k = \cos \left(\pi \frac{2k+1}{2N+2} \right) \quad \text{para } k = 0, 1, \dots, N,$$

de manera que los nodos de interpolación de Chebyshev en $[a, b]$ se obtienen usando el cambio de variable (12):

$$(14) \quad x_k = t_k \frac{b-a}{2} + \frac{a+b}{2} \quad \text{para } k = 0, 1, \dots, N.$$

Teorema 4.7 (Aproximación polinomial de Lagrange-Chebyshev). Sea $P_N(x)$ el polinomio interpolador de Lagrange para los nodos de Chebyshev dados en la fórmula (14). Si $f \in C^{N+1}[a, b]$, entonces

$$(15) \quad |f(x) - P_N(x)| \leq \frac{2(b-a)^{N+1}}{4^{N+1}(N+1)!} \max\{|f^{(N+1)}(x)| : a \leq x \leq b\}.$$

Ejemplo 4.15. Dada $f(x) = \sin(x)$ en el intervalo $[0, \pi/4]$, vamos a encontrar los nodos de Chebyshev para ese intervalo y la cota del error (15) del polinomio interpolador de Lagrange $P_5(x)$.

Usando las fórmulas (12)–(14) para hallar los nodos, obtenemos

$$x_k = \cos \left(\frac{(2k+1)\pi}{12} \right) \frac{\pi}{8} + \frac{\pi}{8} \quad \text{para } k = 0, 1, \dots, 5.$$

Usando la cota $|f^{(6)}(x)| \leq |- \sin(\pi/4)| = 2^{-1/2} = M$ en (15), nos queda

$$|f(x) - P_N(x)| \leq \left(\frac{\pi}{8} \right)^6 \left(\frac{2}{6!} \right) 2^{-1/2} \leq 0.00000720.$$

Propiedad de ortogonalidad

En el Ejemplo 4.14, hemos usado los nodos de Chebyshev para calcular un polinomio interpolador de Lagrange. En general, el polinomio interpolador de Chebyshev de grado N puede obtenerse usando el método de interpolación de Lagrange para los $N+1$ nodos que son las $N+1$ raíces de $T_{N+1}(x)$. Sin embargo, una forma más directa de abordar el problema de la aproximación polinomial sería expresar $P_N(x)$ como una combinación lineal de los polinomios $T_k(x)$ dados

en la Tabla 4.11 y, en general, en (3); es decir, escribir el polinomio interpolador de Chebyshev como

$$(16) \quad P_N(x) = \sum_{k=0}^N c_k T_k(x) = c_0 T_0(x) + c_1 T_1(x) + \cdots + c_N T_N(x).$$

Los coeficientes $\{c_k\}$ de la combinación (16) se pueden calcular fácilmente usando las siguientes propiedades de ortogonalidad: Dados

$$(17) \quad x_k = \cos \left(\pi \frac{2k+1}{2N+2} \right) \quad \text{para } k = 0, 1, \dots, N; \text{ entonces}$$

$$(18) \quad \sum_{k=0}^N T_i(x_k) T_j(x_k) = 0 \quad \text{cuando } i \neq j,$$

$$(19) \quad \sum_{k=0}^N T_i(x_k) T_j(x_k) = \frac{N+1}{2} \quad \text{cuando } i = j \neq 0,$$

$$(20) \quad \sum_{k=0}^N T_0(x_k) T_0(x_k) = N+1.$$

La Propiedad 4 y las identidades (18)–(20) permiten probar el siguiente teorema.

Teorema 4.8 (Aproximación de Chebyshev). El polinomio de aproximación de Chebyshev $P_N(x)$ de grado menor o igual que N para una función $f(x)$ dada en $[-1, 1]$ puede escribirse como una suma ponderada de los polinomios $\{T_j(x)\}$:

$$(21) \quad f(x) \approx P_N(x) = \sum_{j=1}^N c_j T_j(x)$$

en la que los coeficientes $\{c_j\}$ se calculan mediante las siguientes fórmulas

$$(22) \quad c_0 = \frac{1}{N+1} \sum_{k=0}^N f(x_k) T_0(x_k) = \frac{1}{N+1} \sum_{k=0}^N f(x_k)$$

y, para $j = 1, 2, \dots, N$,

$$(23) \quad c_j = \frac{2}{N+1} \sum_{k=0}^N f(x_k) T_j(x_k) = \frac{2}{N+1} \sum_{k=0}^N f(x_k) \cos \left(\frac{j\pi(2k+1)}{2N+2} \right).$$

Ejemplo 4.16. Vamos a calcular el polinomio de Chebyshev $P_3(x)$ que aproxima la función $f(x) = e^x$ en $[-1, 1]$.

Calculamos los coeficientes usando las fórmulas (22) y (23), sabiendo que los nodos son $x_k = \cos(\pi(2k+1)/8)$ para $k = 0, 1, 2$ y 3 :

$$\begin{aligned}c_0 &= \frac{1}{4} \sum_{k=0}^3 e^{x_k} T_0(x_k) = \frac{1}{4} \sum_{k=0}^3 e^{x_k} = 1.26606568, \\c_1 &= \frac{1}{2} \sum_{k=0}^3 e^{x_k} T_1(x_k) = \frac{1}{2} \sum_{k=0}^3 e^{x_k} x_k = 1.13031500, \\c_2 &= \frac{1}{2} \sum_{k=0}^3 e^{x_k} T_2(x_k) = \frac{1}{2} \sum_{k=0}^3 e^{x_k} \cos\left(2\pi \frac{2k+1}{8}\right) = 0.27145036, \\c_3 &= \frac{1}{2} \sum_{k=0}^3 e^{x_k} T_3(x_k) = \frac{1}{2} \sum_{k=0}^3 e^{x_k} \cos\left(3\pi \frac{2k+1}{8}\right) = 0.04379392.\end{aligned}$$

Por tanto, el polinomio de Chebyshev $P_3(x)$ para e^x es

$$(24) \quad P_3(x) = 1.26606568T_0(x) + 1.13031500T_1(x) + 0.27145036T_2(x) + 0.04379392T_3(x).$$

Al escribir el polinomio de Chebyshev de (24) en potencias de x , el resultado es

$$P_3(x) = 0.99461532 + 0.99893324x + 0.54290072x^2 + 0.17517568x^3,$$

que, naturalmente, coincide con el polinomio $V(x)$ del Ejemplo 4.14. Si nuestro objetivo es hallar el polinomio de aproximación de Chebyshev, entonces es preferible usar las fórmulas (22) y (23).

MATLAB

En el programa que damos a continuación se emplea la instrucción `eval` en vez de la instrucción `feval`, que ha sido empleada en programas anteriores. La instrucción `eval` interpreta una cadena de caracteres como una expresión funcional. Por ejemplo, las siguientes instrucciones permiten evaluar la función coseno en los puntos $x = k/10$ para $k = 0, 1, \dots, 5$:

```
>> x=0:.1:.5;
>> eval('cos(x)')
ans =
    1.0000 0.9950 0.9801 0.9553 0.9211 0.8776
```

Programa 4.3 (Aproximación de Chebyshev). Construcción y determinación del polinomio de interpolación de Chebyshev de grado N en el intervalo $[a, b]$ para los nodos

$$x_k = \frac{b-a}{2} \cos\left(\frac{(2k+1)\pi}{2N+2}\right) + \frac{a+b}{2}.$$

El polinomio viene dado por

$$P(x) = \sum_{j=0}^N c_j T_j \left(2 \frac{x-a}{b-a} - 1 \right).$$

```
function [C,X,Y]=cheby(fun,n,a,b)
```

```
% Datos
```

```
%     - fun es la función que deseamos aproximar,
%       dada como una cadena de caracteres
%     - n es el grado del polinomio de aproximación
%     - a es el extremo izquierdo
%     - b es el extremo derecho
```

```
% Resultados
```

```
%     - C es la lista de coeficientes del polinomio
%     - X contiene las abscisas de los nodos
%     - Y contiene los valores de fun en los nodos
```

```
if nargin==2, a=-1;b=1;end
```

```
d=pi/(2*n+2);
```

```
C=zeros(1,n+1);
```

```
for k=1:n+1
```

```
    X(k)=cos((2*k-1)*d);
```

```
end
```

```
X=(b-a)*X/2+(a+b)/2;
```

```
x=X;
```

```
Y=eval(fun);
```

```
for k =1:n+1
```

```
    z=(2*k-1)*d;
```

```
    for j=1:n+1
```

```
        C(j)=C(j)+Y(k)*cos((j-1)*z);
```

```
    end
```

```
end
```

```
C=2*C/(n+1);
```

```
C(1)=C(1)/2;
```

Ejercicios

1. Use la Propiedad 1 para
 - (a) construir $T_4(x)$ a partir de $T_3(x)$ y $T_2(x)$.
 - (b) construir $T_5(x)$ a partir de $T_4(x)$ y $T_3(x)$.
2. Use la Propiedad 1 para
 - (a) construir $T_6(x)$ a partir de $T_5(x)$ y $T_4(x)$.
 - (b) construir $T_7(x)$ a partir de $T_6(x)$ y $T_5(x)$.
3. Use el método de inducción matemática para probar la Propiedad 2.
4. Use el método de inducción matemática para probar la Propiedad 3.
5. Determine los valores máximo y mínimo de $T_2(x)$ en el intervalo $[-1, 1]$.
6. Determine los valores máximo y mínimo de $T_3(x)$ en el intervalo $[-1, 1]$.
Indicación. $T'_3(1/2) = 0$ y $T'_3(-1/2) = 0$.
7. Determine los valores máximo y mínimo de $T_4(x)$ en el intervalo $[-1, 1]$.
Indicación. $T'_4(0) = 0$, $T'_4(2^{-1/2}) = 0$ y $T'_4(-2^{-1/2}) = 0$.
8. Consideremos la función $f(x) = \sin(x)$ en el intervalo $[-1, 1]$.
 - (a) Use los polinomios coeficientes de la Tabla 4.13 para obtener el polinomio de aproximación de Lagrange-Chebyshev $P_3(x)$.
 - (b) Determine una cota del error $|\sin(x) - P_3(x)|$.
9. Consideremos la función $f(x) = \ln(x + 2)$ en el intervalo $[-1, 1]$.
 - (a) Use los polinomios coeficientes de la Tabla 4.13 para obtener el polinomio de aproximación de Lagrange-Chebyshev $P_3(x)$.
 - (b) Determine una cota del error $|\ln(x + 2) - P_3(x)|$.
10. El polinomio interpolador de Lagrange de grado $N = 2$ es

$$f(x) = f(x_0)L_{2,0}(x) + f(x_1)L_{2,1}(x) + f(x_2)L_{2,2}(x).$$

Pruebe que si se utilizan los nodos de Chebyshev $x_0 = \cos(5\pi/6)$, $x_1 = 0$ y $x_2 = \cos(\pi/6)$, entonces los polinomios coeficientes son:

$$\begin{aligned}L_{2,0}(x) &= -\frac{x}{\sqrt{3}} + \frac{2x^2}{3} \\L_{2,1}(x) &= 1 - \frac{4x^2}{3} \\L_{2,2}(x) &= \frac{x}{\sqrt{3}} + \frac{2x^2}{3}\end{aligned}$$

11. Consideremos la función $f(x) = \cos(x)$ en el intervalo $[-1, 1]$.
 - (a) Use los polinomios coeficientes del Ejercicio 10 para obtener el polinomio de aproximación de Lagrange-Chebyshev $P_2(x)$.
 - (b) Determine una cota del error $|\cos(x) - P_2(x)|$.

12. Consideremos la función $f(x) = e^x$ en el intervalo $[-1, 1]$.

(a) Use los polinomios coeficientes del Ejercicio 10 para obtener el polinomio de aproximación de Lagrange-Chebyshev $P_2(x)$.

(b) Determine una cota del error $|e^x - P_2(x)|$.

En los Ejercicios 13 a 15, compare, determinando las cotas del error, el polinomio de Taylor con el polinomio de aproximación de Lagrange-Chebyshev para la función $f(x)$ en $[-1, 1]$.

- 13.** $f(x) = \operatorname{sen}(x)$ y $N = 7$; el polinomio de Lagrange-Chebyshev es

$$\sin(x) \approx 0.99999998x - 0.16666599x^3 + 0.00832995x^5 - 0.00019297x^7.$$

14. $f(x) = \cos(x)$ y $N = 6$; el polinomio de Lagrange-Chebyshev es

$$\cos(x) \approx 1 - 0.49999734x^2 + 0.04164535x^4 - 0.00134608x^6$$

- 15.** $f(x) = e^x$ y $N = 7$; el polinomio de Lagrange-Chebyshev es

$$e^x \approx 0.99999980 + 0.99999998x + 0.50000634x^2 \\ + 0.16666737x^3 + 0.04163504x^4 + 0.00832984x^5 \\ + 0.00143925x^6 + 0.00020399x^7.$$

16. Pruebe la relación (18).

17. Pruebe la relación (19).

Algoritmos y programas

En los Problemas 1 a 6, use el Programa 4.3 para calcular los coeficientes $\{c_k\}$ del polinomio de aproximación de Chebyshev $P_N(x)$ para la función dada $f(x)$ en $[-1, 1]$ y los grados **(a)** $N = 4$, **(b)** $N = 5$, **(c)** $N = 6$ y **(d)** $N = 7$. En cada caso, dibuje conjuntamente $f(x)$ y $P_N(x)$.

- 1.** $f(x) = e^x$ **2.** $f(x) = \operatorname{sen}(x)$
3. $f(x) = \cos(x)$ **4.** $f(x) = \ln(x + 2)$
5. $f(x) = (x + 2)^{1/2}$ **6.** $f(x) = (x + 2)^{(x+2)}$

7. Use el Programa 4.3 (con $N = 5$) para aproximar el valor de $\int_0^1 \cos(x^2) dx$.

4.6 Aproximaciones de Padé

En esta sección introducimos la noción de aproximación racional a una función. Nos planteamos dar fórmulas que nos permitan aproximar una función $f(x)$.

sobre una porción pequeña de su dominio. Por ejemplo, es suficiente con aproximar $f(x) = \cos(x)$ en el intervalo $[0, \pi/2]$ ya que, en ese caso, podemos usar las identidades trigonométricas para calcular $\cos(x)$ para cualquier valor de x que no esté en el intervalo $[0, \pi/2]$.

Una aproximación racional a $f(x)$ en $[a, b]$ es un cociente de dos polinomios $P_N(x)$ y $Q_M(x)$ de grados N y M , respectivamente. Denotaremos dicho cociente por $R_{N,M}(x)$

$$(1) \quad R_{N,M}(x) = \frac{P_N(x)}{Q_M(x)} \quad \text{para } a \leq x \leq b.$$

Nuestro objetivo es hacer el error máximo tan pequeño como podamos. Para un número fijado de operaciones, suele ser posible hallar una aproximación racional cuyo error, medido a lo largo de $[a, b]$, es menor que el de cualquier aproximación polinomial. El desarrollo que haremos es introductorio y se limitará a las aproximaciones de Padé.

El **método de Padé** requiere que $f(x)$ y su derivada sean continuas en $x = 0$. Existen dos razones para trabajar en el punto $x = 0$: en primer lugar, las manipulaciones algebraicas suelen ser más simples y, en segundo lugar, siempre podemos usar un cambio de variable para reformular nuestro problema y trabajar sobre un intervalo en el que esté el punto cero. Los polinomios que se usan en (1) los escribiremos como

$$(2) \quad P_N(x) = p_0 + p_1x + p_2x^2 + \cdots + p_Nx^N$$

y

$$(3) \quad Q_M(x) = 1 + q_1x + q_2x^2 + \cdots + q_Mx^M.$$

Los polinomios que aparecen en (2) y (3) se construyen de manera que $f(x)$ y $R_{N,M}(x)$ coincidan en $x = 0$ y lo mismo ocurra con sus derivadas hasta las de orden $N + M$. En particular, cuando $Q_0(x) = 1$, la aproximación es simplemente el desarrollo de Maclaurin de $f(x)$. Una vez fijado el valor $N + M$, puede probarse que el error es menor cuando $P_N(x)$ y $Q_M(x)$ tienen el mismo grado o bien cuando el grado de $P_N(x)$ es una unidad más que el grado de $Q_M(x)$.

Hagamos notar que hemos impuesto que el término constante de Q_M sea $q_0 = 1$, lo que es permisible porque no puede ser cero y el cociente $R_{N,M}(x)$ no cambia si dividimos $P_N(x)$ y $Q_M(x)$ por la misma constante. En consecuencia, la función racional $R_{N,M}(x)$ tiene $N + M + 1$ coeficientes desconocidos. Supongamos que $f(x)$ es analítica y que su serie de Maclaurin es

$$(4) \quad f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_kx^k + \cdots,$$

y formemos la diferencia $f(x)Q_M(x) - P_N(x) = Z(x)$:

$$(5) \quad \left(\sum_{j=0}^{\infty} a_j x^j \right) \left(\sum_{j=0}^M q_j x^j \right) - \sum_{j=0}^N p_j x^j = \sum_{j=N+M+1}^{\infty} c_j x^j.$$

El subíndice que hemos puesto en el símbolo de la suma del miembro derecho de (5) es $j = M + N + 1$ porque las $N + M$ primeras derivadas de $f(x)$ y $R_{N,M}(x)$ deben coincidir en $x = 0$.

Si efectuamos la multiplicación en el miembro izquierdo de (5) e igualamos a cero los coeficientes de las potencias de x^j para $k = 0, 1, \dots, N + M$, obtenemos un sistema de $N + M + 1$ ecuaciones lineales:

$$(6) \quad \begin{aligned} a_0 - p_0 &= 0 \\ q_1 a_0 + a_1 - p_1 &= 0 \\ q_2 a_0 + q_1 a_1 + a_2 - p_2 &= 0 \\ q_3 a_0 + q_2 a_1 + q_1 a_2 + a_3 - p_3 &= 0 \\ q_M a_{N-M} + q_{M-1} a_{N-M+1} + \cdots + a_N - p_N &= 0 \end{aligned}$$

y

$$(7) \quad \begin{aligned} q_M a_{N-M+1} + q_{M-1} a_{N-M+2} + \cdots + q_1 a_N &+ a_{N+1} = 0 \\ q_M a_{N-M+2} + q_{M-1} a_{N-M+3} + \cdots + q_1 a_{N+1} &+ a_{N+2} = 0 \\ \vdots &\vdots \\ q_M a_N &+ q_{M-1} a_{N+1} + \cdots + q_1 a_{N+M-1} + a_{N+M} = 0. \end{aligned}$$

Hagamos notar que en cada ecuación se verifica que la suma de los subíndices de los factores de cada producto es el mismo y que su suma va creciendo desde 0 hasta $N + M$. Las M ecuaciones que aparecen en (7) sólo involucran las incógnitas q_1, q_2, \dots, q_M y son las que hay que resolver en primer lugar, luego se utiliza su solución en (6) para calcular p_0, p_1, \dots, p_N .

Ejemplo 4.17. Vamos a establecer la aproximación de Padé $R_{4,4}$ de la función coseno:

$$(8) \quad \cos(x) \approx R_{4,4}(x) = \frac{15120 - 6900x^2 + 313x^4}{15120 + 660x^2 + 13x^4}.$$

En la Figura 4.18 se muestran las gráficas de $\cos(x)$ y de $R_{4,4}(x)$ en $[-5, 5]$.

Si usamos el desarrollo de Maclaurin de $\cos(x)$ tendremos un sistema de nueve ecuaciones con nueve incógnitas. En vez de hacer eso, podemos darnos cuenta de que tanto $\cos(x)$ como $R_{4,4}(x)$ son funciones pares así que sus desarrollos sólo tienen potencias de x^2 y podemos simplificar los cálculos partiendo de $f(x) = \cos(x^{1/2})$:

$$(9) \quad f(x) = 1 - \frac{1}{2}x + \frac{1}{24}x^2 - \frac{1}{720}x^3 + \frac{1}{40320}x^4 - \cdots$$

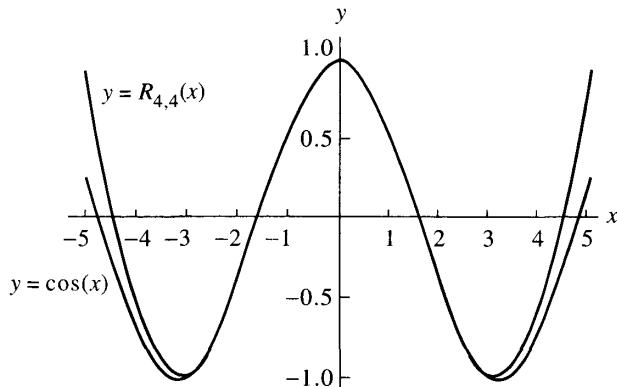


Figura 4.18 Las gráficas de $y = \cos(x)$ y de su aproximación de Padé $R_{4,4}(x)$.

Entonces, la ecuación (5) para este caso es

$$\begin{aligned} \left(1 - \frac{1}{2}x + \frac{1}{24}x^2 - \frac{1}{720}x^3 + \frac{1}{40320}x^4 - \dots\right) (1 + q_1x + q_2x^2) - p_0 - p_1x - p_2x^2 \\ = 0 + 0x + 0x^2 + 0x^3 + 0x^4 + c_5x^5 + c_6x^6 + \dots \end{aligned}$$

Comparando los coeficientes de las cinco primeras potencias de x , obtenemos el sistema de ecuaciones lineales:

$$(10) \quad \begin{aligned} 1 - p_0 &= 0 \\ -\frac{1}{2} + q_1 - p_1 &= 0 \\ \frac{1}{24} - \frac{1}{2}q_1 + q_2 - p_2 &= 0 \\ -\frac{1}{720} + \frac{1}{24}q_1 - \frac{1}{2}q_2 &= 0 \\ \frac{1}{40320} - \frac{1}{720}q_1 + \frac{1}{24}q_2 &= 0. \end{aligned}$$

Debemos empezar resolviendo las dos últimas ecuaciones de (10), para lo que las escribimos como

$$q_1 - 12q_2 = \frac{1}{30} \quad \text{y} \quad -q_1 + 30q_2 = \frac{-1}{56},$$

que resolvemos sumando ambas ecuaciones para despejar q_2 y hallando luego q_1 :

$$(11) \quad \begin{aligned} q_2 &= \frac{1}{18} \left(\frac{1}{30} - \frac{1}{56} \right) = \frac{13}{15120}, \\ q_1 &= \frac{1}{30} + \frac{156}{15120} = \frac{11}{252}. \end{aligned}$$

Ahora usamos las tres primeras ecuaciones de (10): Obviamente, $p_0 = 1$ y, con los valores q_1 y q_2 ya obtenidos en (11), podemos despejar p_1 y p_2 :

$$(12) \quad p_1 = -\frac{1}{2} + \frac{11}{252} = -\frac{115}{252},$$

$$p_2 = \frac{1}{24} - \frac{11}{504} + \frac{13}{15\,120} = \frac{313}{15\,120}.$$

Finalmente, construimos la aproximación racional a $f(x)$ con los coeficientes dados en (11) y (12):

$$(13) \quad f(x) \approx \frac{1 - 115x/252 + 313x^2/15\,120}{1 + 11x/252 + 13x^2/15\,120}.$$

Dado que $\cos(x) = f(x^2)$, sustituimos x por x^2 en la fórmula (13) y obtenemos la función deseada $R_{4,4}(x)$ que aparecía en (8). ■

Aproximación de Padé en forma de fracción continua

Para evaluar en un punto la aproximación de Padé $R_{4,4}(x)$ del Ejemplo 4.17 hace falta realizar un mínimo de doce operaciones aritméticas, pero es posible reducir este número a siete usando fracciones continuas. Para ello partimos de (8) y hallamos el cociente y el resto de la división:

$$\begin{aligned} R_{4,4}(x) &= \frac{15\,120/313 - (6900/313)x^2 + x^4}{15\,120/13 + (660/13)x^2 + x^4} \\ &= \frac{313}{13} - \left(\frac{296\,280}{169} \right) \left(\frac{12\,600/823 + x^2}{15\,120/13 + (600/13)x^2 + x^4} \right). \end{aligned}$$

Realizamos de nuevo el proceso usando el segundo factor del resto del paso anterior:

$$\begin{aligned} R_{4,4}(x) &= \frac{313}{13} - \frac{296\,280/169}{15\,120/13 + (660/13)x^2 + x^4} \\ &= \frac{313}{13} - \frac{296\,280/169}{\frac{379\,380}{10\,699} + x^2 + \frac{420\,078\,960/677\,329}{12\,600/823 + x^2}}. \end{aligned}$$

Escribiendo las fracciones como decimales para realizar los cálculos obtenemos:

$$(14) \quad R_{4,4}(x) = 24.07692308$$

$$-\frac{1753.13609467}{35.45938873 + x^2 + 620.19928277/(15.30984204 + x^2)}.$$

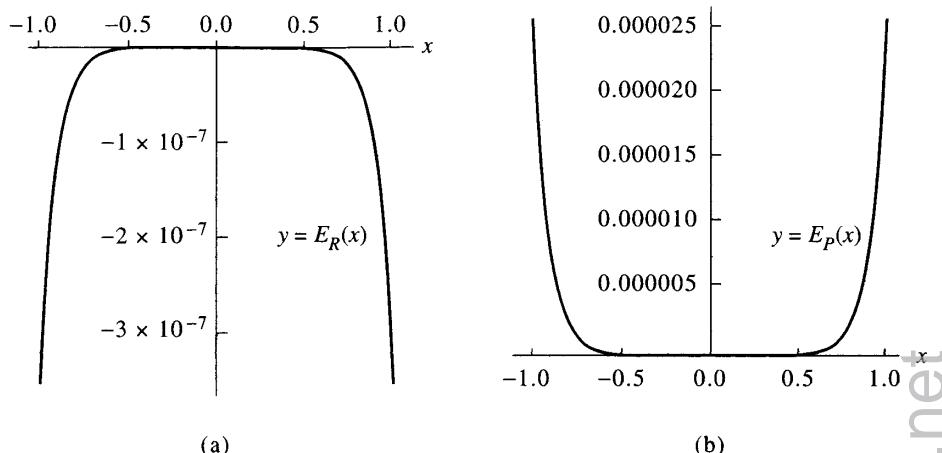


Figura 4.19 (a) Gráfica del error $E_R(x) = \cos(x) - R_{4,4}(x)$ de la aproximación de Padé $R_{4,4}(x)$. (b) Gráfica del error $E_P(x) = \cos(x) - P_6(x)$ de la aproximación de Taylor $P_6(x)$.

Para evaluar (14), empezamos calculando y y almacenando x^2 ; luego, procedemos desde la esquina inferior derecha del denominador con la secuencia: suma, división, suma, suma, división y resta. Por tanto, basta con siete operaciones aritméticas para evaluar $R_{4,4}(x)$ cuando se escribe como la fracción continua dada en (14).

Podemos comparar $R_{4,4}(x)$ con el polinomio de Taylor $P_6(x)$ de grado $N = 6$, que también se evalúa haciendo siete operaciones cuando se usa la regla de Ruffini:

$$(15) \quad P_6(x) = 1 + x^2 \left(-\frac{1}{2} + x^2 \left(\frac{1}{24} - \frac{1}{720} x^2 \right) \right) \\ = 1 + x^2 (-0.5 + x^2 (0.0416666667 - 0.0013888889 x^2)).$$

Las gráficas de $E_R(x) = \cos(x) - R_{4,4}(x)$ y $E_P(x) = \cos(x) - P_6(x)$ en $[-1, 1]$ se muestran en las Figuras 4.19(a) y (b), respectivamente; allí vemos que los errores más grandes se alcanzan en los extremos y valen $E_R(1) = -0.0000003599$ y $E_P(1) = 0.0000245281$, respectivamente. El tamaño del error más grande de $R_{4,4}(x)$ es, más o menos, el 1.467% del tamaño del error más grande de $P_6(x)$. La aproximación de Padé es mucho mejor aún en intervalos más pequeños; así, por ejemplo, en el intervalo $[-0.1, 0.1]$ encontramos que $E_R(0.1) = -0.0000000004$ y que $E_P(0.1) = 0.0000000966$, de manera que el tamaño del error $R_{4,4}(x)$ es, más o menos, el 0.384% del tamaño del error de $P_6(x)$.

Ejercicios

1. Establezca la aproximación de Padé:

$$e^x \approx R_{1,1}(x) = \frac{2+x}{2-x}.$$

2. (a) Determine la aproximación de Padé $R_{1,1}(x)$ para $f(x) = \ln(1+x)/x$.
Indicación. Trabaje a partir del desarrollo de Maclaurin:

$$f(x) = 1 - \frac{x}{2} + \frac{x^2}{3} - \dots$$

- (b) Use el resultado del apartado (a) para establecer la aproximación

$$\ln(1+x) \approx R_{2,1}(x) = \frac{6x+x^2}{6+4x}$$

3. (a) Determine $R_{1,1}(x)$ para $f(x) = \tan(x^{1/2})/x^{1/2}$. *Indicación.* Trabaje a partir del desarrollo de Maclaurin:

$$f(x) = 1 + \frac{x}{3} + \frac{2x^2}{15} + \dots$$

- (b) Use el resultado del apartado (a) para establecer la aproximación

$$\tan(x) \approx R_{3,2}(x) = \frac{15x-x^3}{15-6x^2}$$

4. (a) Determine $R_{1,1}(x)$ para $f(x) = \arctan(x^{1/2})/x^{1/2}$. *Indicación.* Trabaje a partir del desarrollo de Maclaurin:

$$f(x) = 1 - \frac{x}{3} + \frac{x^2}{5} - \dots$$

- (b) Use el resultado del apartado (a) para establecer la aproximación

$$\arctan(x) \approx R_{3,2}(x) = \frac{15x+4x^3}{15+9x^2}$$

- (c) Exprese la función racional $R_{3,2}(x)$ del apartado (b) como una fracción continua.

5. (a) Establezca la aproximación de Padé:

$$e^x \approx R_{2,2}(x) = \frac{12+6x+x^2}{12-6x+x^2}$$

- (b) Exprese la función racional $R_{2,2}(x)$ del apartado (a) como una fracción continua.

- 6. (a)** Determine la aproximación de Padé $R_{2,2}(x)$ para $f(x) = \ln(1 + x)/x$.
Indicación. Trabaje a partir del desarrollo de Maclaurin:

$$f(x) = 1 - \frac{x}{2} + \frac{x^2}{3} - \frac{x^3}{4} + \frac{x^4}{5} - \dots$$

- (b)** Use el resultado del apartado (a) para establecer la aproximación

$$\ln(1 + x) \approx R_{3,2}(x) = \frac{30x + 21x^2 + x^3}{30 + 36x + 9x^2}$$

- (c)** Exprese la función racional $R_{3,2}(x)$ del apartado (b) como una fracción continua.

- 7. (a)** Determine $R_{2,2}(x)$ para $f(x) = \tan(x^{1/2})/x^{1/2}$. *Indicación.* Trabaje a partir del desarrollo de Maclaurin:

$$f(x) = 1 + \frac{x}{3} + \frac{2x^2}{15} + \frac{17x^3}{315} + \frac{62x^4}{2835} + \dots$$

- (b)** Use el resultado del apartado (a) para establecer la aproximación

$$\tan(x) \approx R_{5,4}(x) = \frac{945x - 105x^3 + x^5}{945 - 420x^2 + 15x^4}$$

- (c)** Exprese la función racional $R_{5,4}(x)$ del apartado (b) como una fracción continua.

- 8. (a)** Determine $R_{2,2}(x)$ para $f(x) = \arctan(x^{1/2})/x^{1/2}$. *Indicación.* Trabaje a partir del desarrollo de Maclaurin:

$$f(x) = 1 - \frac{x}{3} + \frac{x^2}{5} - \frac{x^3}{7} + \frac{x^4}{9} - \dots$$

- (b)** Use el resultado del apartado (a) para establecer la aproximación

$$\arctan(x) \approx R_{5,4}(x) = \frac{945x + 735x^3 + 64x^5}{945 + 1050x^2 + 225x^4}$$

- (c)** Exprese la función racional $R_{5,4}(x)$ del apartado (b) como una fracción continua.

- 9.** Establezca la aproximación de Padé:

$$e^x \approx R_{3,3}(x) = \frac{120 + 60x + 12x^2 + x^3}{120 - 60x + 12x^2 + x^3}$$

- 10.** Establezca la aproximación de Padé:

$$e^x \approx R_{4,4}(x) = \frac{1680 + 840x + 180x^2 + 20x^3 + x^4}{1680 - 840x + 180x^2 - 20x^3 + x^4}$$

Algoritmos y programas

1. Compare las siguientes aproximaciones a $f(x) = e^x$.

$$\text{Taylor: } T_6(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24}$$

$$\text{Padé: } R_{2,2}(x) = \frac{12 + 6x + x^2}{12 - 6x + x^2}$$

- (a) Dibuje $f(x)$, $T_6(x)$ y $R_{2,2}(x)$ en un mismo gráfico.
 (b) Determine el valor del error más grande que se comete al aproximar $f(x)$ por $T_6(x)$ y $R_{2,2}(x)$, respectivamente, en el intervalo $[-1, 1]$.
 2. Compare las siguientes aproximaciones a $f(x) = \ln(1 + x)$.

$$\text{Taylor: } T_5(x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5}$$

$$\text{Padé: } R_{3,2}(x) = \frac{30x + 21x^2 + x^3}{30 + 36x + 9x^2}$$

- (a) Dibuje $f(x)$, $T_5(x)$ y $R_{3,2}(x)$ en un mismo gráfico.
 (b) Determine el valor del error más grande que se comete al aproximar $f(x)$ por $T_5(x)$ y $R_{3,2}(x)$, respectivamente, en el intervalo $[-1, 1]$.
 3. Compare las siguientes aproximaciones a $f(x) = \tan(x)$.

$$\text{Taylor: } T_9(x) = x + \frac{x^3}{3} + \frac{2x^5}{15} + \frac{17x^7}{315} + \frac{62x^9}{2835}$$

$$\text{Padé: } R_{5,4}(x) = \frac{945x - 105x^3 + x^5}{945 - 420x^2 + 15x^4}$$

- (a) Dibuje $f(x)$, $T_9(x)$ y $R_{5,4}(x)$ en un mismo gráfico.
 (b) Determine el valor del error más grande que se comete al aproximar $f(x)$ por $T_9(x)$ y $R_{5,4}(x)$, respectivamente, en el intervalo $[-1, 1]$.
 4. Compare las siguientes aproximaciones de Padé a $f(x) = \sin(x)$ en el intervalo $[-1.2, 1.2]$.

$$R_{5,4}(x) = \frac{166\,320x - 22\,260x^3 + 551x^5}{15(11\,088 + 364x^2 + 5x^4)}$$

$$R_{7,6}(x) = \frac{11\,511\,339\,840x - 1\,640\,635\,920x^2 + 52\,785\,432x^5 - 479\,249x^7}{7(1\,644\,477\,120 + 39\,702\,960x^2 + 453\,960x^4 + 2623x^6)}$$

- (a) Dibuje $f(x)$, $R_{5,4}(x)$ y $R_{7,6}(x)$ en un mismo gráfico.
 (b) Determine el valor del error más grande que se comete al aproximar $f(x)$ por $R_{5,4}(x)$ y $R_{7,6}(x)$, respectivamente, en el intervalo $[-1.2, 1.2]$.

272 CAP. 4 INTERPOLACIÓN Y APROXIMACIÓN POLINOMIAL

5. (a) Use las ecuaciones (6) y (7) para construir las aproximaciones de Padé $R_{6,6}(x)$ y $R_{8,8}(x)$ para la función $f(x) = \cos(x)$ en el intervalo $[-1.2, 1.2]$.
(b) Dibuje $f(x)$, $R_{6,6}(x)$ y $R_{8,8}(x)$ en un mismo gráfico.
(c) Determine el valor del error más grande que se comete al aproximar $f(x)$ por $R_{6,6}(x)$ y $R_{8,8}(x)$, respectivamente, sobre el intervalo $[-1.2, 1.2]$.

Ajuste de curvas

Las aplicaciones de las técnicas numéricas a la ciencia y la ingeniería consisten, a menudo, en ajustar una curva a datos experimentales. Por ejemplo, en 1601 el astrónomo alemán Johannes Kepler formuló su tercera ley del movimiento planetario, $T = Cx^{3/2}$, donde x es la distancia al Sol medida en millones de kilómetros, T es el período orbital medido en días y C es una constante. Las parejas de datos (x, T) observados para los primeros cuatro planetas, Mercurio, Venus, La Tierra y Marte, son $(58, 88)$, $(108, 225)$, $(150, 365)$ y $(228, 687)$ y el coeficiente C obtenido por el método de los mínimos cuadrados es $C = 0.199769$. La curva $T = 0.199769x^{3/2}$ y las parejas de datos se muestran en la Figura 5.1.

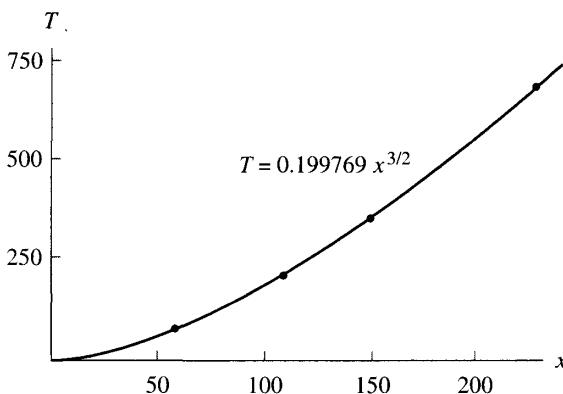


Figura 5.1 El ajuste en mínimos cuadrados para los cuatro primeros planetas, $T = 0.199769x^{3/2}$, usando la tercera ley de Kepler del movimiento planetario.

5.1 Rectas de regresión en mínimos cuadrados

En la ciencia y la ingeniería se da, a menudo, el caso de que un experimento produce un conjunto de datos $(x_1, y_1), \dots, (x_N, y_N)$, siendo las abscisas $\{x_k\}$ distintas entre sí. Uno de los objetivos del cálculo numérico es la determinación de una fórmula $y = f(x)$ que relacione las variables. Normalmente se dispone de una clase de fórmulas previamente establecidas, y lo que hay que hallar son los valores más adecuados de unos coeficientes o de unos parámetros para estas fórmulas. Aunque hay muchos tipos distintos de funciones que podemos usar, suele ocurrir que existe un modelo matemático subyacente, basado en la situación física que se esté estudiando, que determina la forma de la función salvo algunos coeficientes. En esta sección haremos hincapié en la clase de las funciones lineales de la forma

$$(1) \quad y = f(x) = Ax + B.$$

En el Capítulo 4 hemos visto cómo se construye un polinomio cuya gráfica pase por todos los puntos de un conjunto dado. Si todos los valores $\{x_k\}, \{y_k\}$ se conocen con una precisión de varias cifras significativas, entonces la interpolación polinomial produce buenos resultados; lo que no ocurre en otras circunstancias. Algunos experimentos se llevan a cabo con una maquinaria especializada que permite obtener los datos con varias cifras significativas de precisión; sin embargo, muchos experimentos se realizan con un equipamiento que proporciona los datos con una precisión de, como mucho, dos o tres cifras significativas. A esto se añade, a menudo, un cierto error experimental en las mediciones de forma que, aunque se calculen tres o cuatro cifras de los valores $\{x_k\}$ e $\{y_k\}$, sucede que el valor exacto $f(x_k)$ verifica

$$(2) \quad f(x_k) = y_k + e_k,$$

donde e_k es el error de medición.

¿Cómo encontramos la mejor aproximación lineal de la forma dada en (1) que pase cerca (no por encima de cada uno) de los puntos? Para responder esta pregunta, hay que considerar los *errores* (también llamados *desviaciones* o *residuos*):

$$(3) \quad e_k = f(x_k) - y_k \quad \text{para} \quad 1 \leq k \leq N.$$

Hay varias normas que podemos usar con los residuos dados en (3) para medir la distancia entre la curva $y = f(x)$ y los datos.

Tabla 5.1 Cálculos para hallar $E_1(f)$ y $E_2(f)$ en el Ejemplo 5.1.

x_k	y_k	$f(x_k) = 8.6 - 1.6x_k$	$ e_k $	e_k^2
-1	10.0	10.2	0.2	0.04
0	9.0	8.6	0.4	0.16
1	7.0	7.0	0.0	0.00
2	5.0	5.4	0.4	0.16
3	4.0	3.8	0.2	0.04
4	3.0	2.2	0.8	0.64
5	0.0	0.6	0.6	0.36
6	-1.0	-1.0	0.0	0.00
			2.6	1.40

(4) Error máximo: $E_\infty(f) = \max\{|f(x_k) - y_k| : 1 \leq k \leq N\},$

(5) Error medio: $E_1(f) = \frac{1}{N} \sum_{k=1}^N |f(x_k) - y_k|,$

(6) Error cuadrático medio: $E_2(f) = \left(\frac{1}{N} \sum_{k=1}^N |f(x_k) - y_k|^2 \right)^{1/2}.$

El siguiente ejemplo muestra cómo se aplican estas normas cuando tenemos una función y un conjunto de puntos dados.

Ejemplo 5.1. Vamos a comparar el error máximo, el error medio y el error cuadrático medio de la aproximación lineal $y = f(x) = 8.6 - 1.6x$ con respecto al conjunto de datos $(-1, 10), (0, 9), (1, 7), (2, 5), (3, 4), (4, 3), (5, 0)$ y $(6, -1)$.

Los errores se calculan a partir de los valores $f(x_k)$ y e_k dados en la Tabla 5.1.

(7) $E_\infty(f) = \max\{0.2, 0.4, 0.0, 0.4, 0.2, 0.8, 0.6, 0.0\} = 0.8,$

(8) $E_1(f) = \frac{1}{8}(2.6) = 0.325,$

(9) $E_2(f) = \left(\frac{1.4}{8} \right)^{1/2} \approx 0.41833.$

Podemos ver que el error máximo es el más grande; de manera que si el error en un punto es grande, entonces el valor de este error es el que determina $E_\infty(f)$. El error medio $E_1(f)$ es simplemente la media aritmética de los valores absolutos de los errores en los puntos; se usa a menudo porque es fácil de calcular. El error

cuadrático medio $E_2(f)$ se suele usar cuando se considera la naturaleza aleatoria de los errores.

Una línea de ajuste óptimo es aquella que minimiza una de las cantidades de las expresiones (4) a (6); por tanto, hay tres líneas de ajuste óptimo que podríamos encontrar. La que corresponde a la tercera norma, el error cuadrático medio, es la elección tradicional porque es mucho más fácil de minimizar computacionalmente que las otras. ■

Determinación de la recta de regresión

Sea $\{(x_k, y_k)\}_{k=1}^N$ un conjunto de N puntos cuyas abscisas $\{x_k\}$ son todas distintas. La **recta de regresión** o **recta óptima en (el sentido de los) mínimos cuadrados** es la recta de ecuación $y = f(x) = Ax + B$ que minimiza el error cuadrático medio $E_2(f)$.

Observemos que la cantidad $E_2(f)$ será mínima si, y sólo si, lo es el valor $N(E_2(f))^2 = \sum_{k=1}^N (Ax_k + B - y_k)^2$, valor que puede visualizarse geométricamente como la suma de los cuadrados de las distancias verticales desde los puntos hasta la recta. El siguiente resultado explica el proceso completo.

Teorema 5.1 (Recta de regresión en mínimos cuadrados). Supongamos que $\{(x_k, y_k)\}_{k=1}^N$ son N puntos cuyas abscisas $\{x_k\}_{k=1}^N$ son distintas. Entonces, los coeficientes de la recta de regresión

$$y = Ax + B$$

son la solución del siguiente sistema lineal, conocido como las **ecuaciones normales de Gauss**:

$$(10) \quad \begin{aligned} \left(\sum_{k=1}^N x_k^2 \right) A + \left(\sum_{k=1}^N x_k \right) B &= \sum_{k=1}^N x_k y_k, \\ \left(\sum_{k=1}^N x_k \right) A + N B &= \sum_{k=1}^N y_k. \end{aligned}$$

Demostración. Si empezamos con la recta $y = Ax + B$, entonces la distancia vertical d_k desde el punto (x_k, y_k) hasta el punto $(x_k, Ax_k + B)$ de la recta es $d_k = |Ax_k + B - y_k|$ (véase la Figura 5.2). Debemos minimizar la suma de los cuadrados de las distancias verticales d_k :

$$(11) \quad E(A, B) = \sum_{k=1}^N (Ax_k + B - y_k)^2 = \sum_{k=1}^N d_k^2.$$

El valor mínimo de la función $E(A, B)$ se determina igualando a cero las derivadas parciales $\partial E / \partial A$ y $\partial E / \partial B$ y resolviendo las ecuaciones que resultan

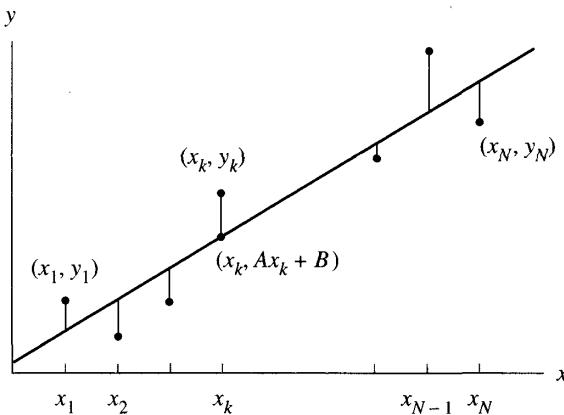


Figura 5.2 Distancias verticales entre los puntos $\{(x_k, y_k)\}$ y la recta $y = Ax + B$.

en A y B . Nótese que $\{x_k\}$ e $\{y_k\}$ son constantes en la expresión (11) ¡y que A y B son las variables! Fijando B y derivando $E(A, B)$ respecto de A , obtenemos

$$(12) \quad \frac{\partial E(A, B)}{\partial A} = \sum_{k=1}^N 2(Ax_k + B - y_k)(x_k) = 2 \sum_{k=1}^N (Ax_k^2 + Bx_k - x_k y_k).$$

Ahora fijamos A y, derivando $E(A, B)$ respecto de B , obtenemos

$$(13) \quad \frac{\partial E(A, B)}{\partial B} = \sum_{k=1}^N 2(Ax_k + B - y_k) = 2 \sum_{k=1}^N (Ax_k + B - y_k).$$

Igualando a cero las derivadas parciales obtenidas en (12) y (13) y usando la propiedad distributiva de la suma, obtenemos

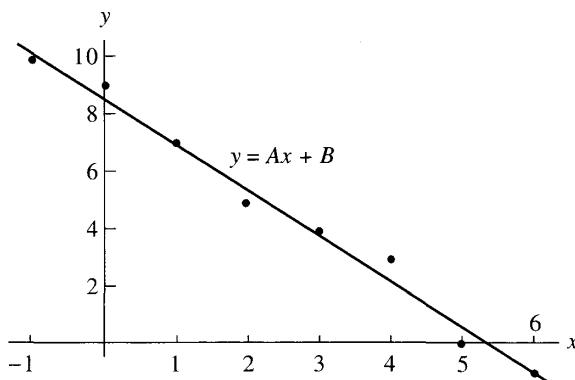
$$(14) \quad 0 = \sum_{k=1}^N (Ax_k^2 + Bx_k - x_k y_k) = A \sum_{k=1}^N x_k^2 + B \sum_{k=1}^N x_k - \sum_{k=1}^N x_k y_k,$$

$$(15) \quad 0 = \sum_{k=1}^N (Ax_k + B - y_k) = A \sum_{k=1}^N x_k + NB - \sum_{k=1}^N y_k. \quad \bullet$$

Escribiendo las ecuaciones (14) y (15) como un sistema estándar, formamos las ecuaciones normales de Gauss (10). La solución de este sistema puede obtenerse mediante cualquiera de las técnicas para resolver sistemas de ecuaciones lineales que vimos en el Capítulo 3; sin embargo, en el método empleado en el Programa 5.1 se hace una traslación de los datos para que la matriz resultante esté bien condicionada (véanse los ejercicios).

Tabla 5.2 Cálculo de los coeficientes de la ecuaciones normales de Gauss.

x_k	y_k	x_k^2	$x_k y_k$
-1	10	1	-10
0	9	0	0
1	7	1	7
2	5	4	10
3	4	9	12
4	3	16	12
5	0	25	0
<u>6</u>	<u>-1</u>	<u>36</u>	<u>-6</u>
<u>20</u>	<u>37</u>	<u>92</u>	<u>25</u>

**Figura 5.3** La recta de regresión $y = -1.6071429x + 8.6428571$.

Ejemplo 5.2. Vamos a calcular la recta de regresión para los datos del Ejemplo 5.1.

Las sumas necesarias para establecer las ecuaciones normales (10) se obtienen fácilmente usando los valores de la Tabla 5.2. El sistema para A y B es, entonces,

$$92A + 20B = 25$$

$$20A + 8B = 37$$

cuya solución es $A \approx -1.6071429$ y $B \approx 8.6428571$. Por tanto, la recta de regresión en mínimos cuadrados es (véase la Figura 5.3)

$$y = -1.6071429x + 8.6428571.$$

El ajuste potencial $y = Ax^M$

Algunas situaciones se modelan mediante una función del tipo $f(x) = Ax^M$, donde M es una constante conocida, el ejemplo del movimiento planetario dado

en la Figura 5.1 ilustra este hecho. En estos casos sólo hay que determinar un parámetro.

Teorema 5.2 (Ajuste potencial). Supongamos que tenemos N puntos $\{(x_k, y_k)\}_{k=1}^N$ cuyas abscisas son distintas. Entonces, el coeficiente A de la curva potencial óptima en mínimos cuadrados $y = Ax^M$ viene dado por

$$(16) \quad A = \left(\sum_{k=1}^N x_k^M y_k \right) \Big/ \left(\sum_{k=1}^N x_k^{2M} \right).$$

Demuestração. Usando la técnica de los mínimos cuadrados, lo que hacemos es buscar el mínimo de la función:

$$(17) \quad E(A) = \sum_{k=1}^N (Ax_k^M - y_k)^2,$$

para lo que, en este caso, bastará con resolver $E'(A) = 0$. La derivada es

$$(18) \quad E'(A) = 2 \sum_{k=1}^N (Ax_k^M - y_k)(x_k^M) = 2 \sum_{k=1}^N (Ax_k^{2M} - x_k^M y_k),$$

luego el coeficiente A es la solución de la ecuación

$$(19) \quad 0 = A \sum_{k=1}^N x_k^{2M} - \sum_{k=1}^N x_k^M y_k,$$

que se reduce a la fórmula dada en la expresión (16).

Ejemplo 5.3. Con el objetivo de medir la aceleración de la gravedad, se han recogido en la Tabla 5.3 unos datos experimentales sobre el tiempo que tarda en llegar al suelo un cuerpo, según la altura desde la que se lo deja caer. La relación funcional es $d = \frac{1}{2}gt^2$, donde d es la distancia de caída medida en metros y t es el tiempo medido en segundos. Vamos a aproximar con estos datos el valor de la aceleración de la gravedad g .

Usamos los valores dados en la Tabla 5.3 para calcular las sumas que necesitamos en la fórmula (16), siendo $M = 2$ el exponente que hemos tomado.

El coeficiente es $A = 7.68680/1.5664 = 4.9073$ y obtenemos la curva de ajuste $d = 4.9073t^2$ con lo cual $g \approx 2A = 9.7146 \text{ m/s}^2$.

Damos a continuación un programa para construir la recta de regresión que es computacionalmente estable: proporciona resultados fiables incluso cuando las ecuaciones normales de Gauss (10) están mal condicionadas. El desarrollo del algoritmo en el que se basa este programa se sugiere en los Ejercicios 4 a 7.

Tabla 5.3 Cálculo de los coeficientes para un ajuste potencial.

Tiempo, t_k	Distancia, d_k	$d_k t_k^2$	t_k^4
0.200	0.1960	0.00784	0.0016
0.400	0.7850	0.12560	0.0256
0.600	1.7665	0.63594	0.1296
0.800	3.1405	2.00992	0.4096
1.000	4.9075	4.90750	1.0000
		7.68680	1.5664

Programa 5.1 (Recta de regresión). Construcción de la recta de regresión $y = Ax + B$ que mejor se ajusta en el sentido de los mínimos cuadrados a los N datos $(x_1, y_1), \dots, (x_N, y_N)$.

```
function [A,B]=lsline(X,Y)
% Datos
% - X es el vector de abscisas 1 x n
% - Y es el vector de ordenadas 1 x n
% Resultados
% - A es el coeficiente de x en Ax + B
% - B es el término independiente en Ax + B
xmean=mean(X);
ymean=mean(Y);
sumx2=(X-xmean)*(X-xmean)';
sumxy=(Y-ymean)*(X-xmean)';
A=sumxy/sumx2;
B=ymean-A*xmean;
```

Ejercicios

En los Ejercicios 1 y 2, determine la recta de regresión $y = f(x) = Ax + B$ correspondiente a los datos y calcule $E_2(f)$

1. (a)

x_k	y_k	$f(x_k)$
-2	1	1.2
-1	2	1.9
0	3	2.6
1	3	3.3
2	4	4.0

(b)

x_k	y_k	$f(x_k)$
-6	7	7.0
-2	5	4.6
0	3	3.4
2	2	2.2
6	0	-0.2

2. (a)

x_k	y_k	$f(x_k)$
-4	1.2	0.44
-2	2.8	3.34
0	6.2	6.24
2	7.8	9.14
4	13.2	12.04

(b)

x_k	y_k	$f(x_k)$
-6	-5.3	-6.00
-2	-3.5	-2.84
0	-1.7	-1.26
2	0.2	0.32
6	4.0	3.48

(c)

x_k	y_k	$f(x_k)$
-8	6.8	7.32
-2	5.0	3.81
0	2.2	2.64
4	0.5	0.30
6	-1.3	-0.87

(d)

x_k	y_k	$f(x_k)$
-4	-3	-3.0
-1	-1	-0.9
0	0	-0.2
2	1	1.2
3	2	1.9

3. Determine el ajuste potencial óptimo en mínimos cuadrados $y = Ax$ (o sea, $M = 1$ con lo que la línea es una recta que pasa por el origen) para los siguientes datos y calcule $E_2(f)$.

(a)

x_k	y_k	$f(x_k)$
-4	-3	-2.8
-1	-1	-0.7
0	0	0.0
2	1	1.4
3	2	2.1

(b)

x_k	y_k	$f(x_k)$
3	1.6	1.722
4	2.4	2.296
5	2.9	2.870
6	3.4	3.444
8	4.6	4.592

(c)

x_k	y_k	$f(x_k)$
1	1.6	1.58
2	2.8	3.16
3	4.7	4.74
4	6.4	6.32
5	8.0	7.90

4. Se definen las medias aritméticas \bar{x} e \bar{y} para los puntos $\{(x_k, y_k)\}_{k=1}^N$ como

$$\bar{x} = \frac{1}{N} \sum_{k=1}^N x_k \quad \text{e} \quad \bar{y} = \frac{1}{N} \sum_{k=1}^N y_k.$$

Pruebe que el punto (\bar{x}, \bar{y}) está en la recta de regresión determinada por el conjunto de puntos dado.

5. Pruebe que la solución del sistema dado en (10) es

$$A = \frac{1}{D} \left(N \sum_{k=1}^N x_k y_k - \sum_{k=1}^N x_k \sum_{k=1}^N y_k \right),$$

$$B = \frac{1}{D} \left(\sum_{k=1}^N x_k^2 \sum_{k=1}^N y_k - \sum_{k=1}^N x_k \sum_{k=1}^N x_k y_k \right),$$

siendo

$$D = N \sum_{k=1}^N x_k^2 - \left(\sum_{k=1}^N x_k \right)^2.$$

Indicación. Use eliminación gaussiana para resolver el sistema.

6. Pruebe que el valor del denominador D del Ejercicio 5 no es cero.
Indicación. Pruebe que $D = N \sum_{k=1}^N (x_k - \bar{x})^2$.

7. Pruebe que los coeficientes A y B de la recta de regresión pueden calcularse mediante el proceso que describimos a continuación. En primer lugar se calculan las medias aritméticas \bar{x} e \bar{y} definidas en el Ejercicio 4 y, en segundo lugar, se realizan los cálculos siguientes:

$$C = \sum_{k=1}^N (x_k - \bar{x})^2, \quad A = \frac{1}{C} \sum_{k=1}^N (x_k - \bar{x})(y_k - \bar{y}), \quad B = \bar{y} - A\bar{x}.$$

Indicación. Use las variables $X_k = x_k - \bar{x}$ e $Y_k = y_k - \bar{y}$ y empiece determinando la recta $Y = AX$ correspondiente a estas parejas de datos.

8. Determine los ajustes potenciales $y = Ax^2$ e $y = Bx^3$ para las siguientes parejas de datos y use $E_2(f)$ para decidir cuál se ajusta mejor.

(a)	x_k	y_k
	2.0	5.1
	2.3	7.5
	2.6	10.6
	2.9	14.4
	3.2	19.0

(b)	x_k	y_k
	2.0	5.9
	2.3	8.3
	2.6	10.7
	2.9	13.7
	3.2	17.0

9. Determine los ajustes potenciales $y = A/x$ e $y = B/x^2$ para las siguientes parejas de datos y use $E_2(f)$ para decidir qué curva se ajusta mejor.

(a)	x_k	y_k
	0.5	7.1
	0.8	4.4
	1.1	3.2
	1.8	1.9
	4.0	0.9

(b)	x_k	y_k
	0.7	8.1
	0.9	4.9
	1.1	3.3
	1.6	1.6
	3.0	0.5

10. (a) Deducza las ecuaciones normales de Gauss para calcular la recta que pasa por el origen $y = Ax$ que mejor se ajusta en el sentido de los mínimos cuadrados.
- (b) Deducza las ecuaciones normales de Gauss para calcular la parábola que pasa por el origen $y = Ax^2$ que mejor se ajusta en el sentido de los mínimos cuadrados.
- (c) Deducza las ecuaciones normales de Gauss para calcular la parábola con vértice en el eje de ordenadas $y = Ax^2 + B$ que mejor se ajusta en el sentido de los mínimos cuadrados.
11. Consideremos las rectas de regresión que se obtienen para cada uno de los conjuntos de datos $S_N = \{(\frac{k}{N}, (\frac{k}{N})^2)\}_{k=1}^N$, donde $N = 2, 3, 4, \dots$. Nótese que, para cada valor de N , todos los puntos de S_N están en la gráfica de $f(x) = x^2$ para x en $[0, 1]$. Sean \bar{x}_N e \bar{y}_N las medias aritméticas de las parejas de datos de S_N (véase el Ejercicio 4). Sea \hat{x} la media de los valores de x en el intervalo $[0, 1]$ y sea \hat{y} el valor medio de $f(x) = x^2$ en el intervalo $[0, 1]$.
- (a) Pruebe que $\lim_{N \rightarrow \infty} \bar{x}_N = \hat{x}$.
- (b) Pruebe que $\lim_{N \rightarrow \infty} \bar{y}_N = \hat{y}$.
12. Consideremos las rectas de regresión que se obtienen para cada uno de los conjuntos de datos

$$S_N = \left\{ \left((b-a) \frac{k}{N} + a, f \left((b-a) \frac{k}{N} + a \right) \right) \right\}_{k=1}^N$$

con $N = 2, 3, 4, \dots$. Sea $y = f(x)$ una función integrable en el intervalo cerrado $[a, b]$. Repita, en este caso, los apartados (a) y (b) del Ejercicio 11.

Algoritmos y programas

- La ley de Hooke establece que $F = kx$, siendo F la fuerza (en dinas) que es necesario ejercer para conseguir que un muelle se estire una longitud x (en centímetros) desde su posición de equilibrio. Use el Programa 5.1 para hallar una aproximación a la constante del muelle k para los siguientes conjuntos de datos.

(a)	x_k	F_k
	0.2	3.6
	0.4	7.3
	0.6	10.9
	0.8	14.5
	1.0	18.2

(b)	x_k	F_k
	0.2	5.3
	0.4	10.6
	0.6	15.9
	0.8	21.2
	1.0	26.4

2. Escriba un programa para aproximar la constante gravitacional g para los siguientes conjuntos de datos; utilice el ajuste potencial mostrado en el Ejemplo 5.3.

(a)	Tiempo, t_k	Distancia, d_k
	0.200	0.1960
	0.400	0.7835
	0.600	1.7630
	0.800	3.1345
	1.000	4.8975

(b)	Tiempo, t_k	Distancia, d_k
	0.200	0.1965
	0.400	0.7855
	0.600	1.7675
	0.800	3.1420
	1.000	4.9095

3. Los siguientes datos proporcionan las distancias desde los nueve planetas al Sol y su período orbital —el tiempo que tardan en completar una órbita— en días.

Planeta	Distancia al Sol ($\text{km} \times 10^6$)	Período orbital (días)
Mercurio	57.59	87.99
Venus	108.11	224.70
La Tierra	149.57	365.26
Marte	227.84	686.98
Júpiter	778.14	4 332.4
Saturno	1427.0	10 759
Urano	2870.3	30 684
Neptuno	4499.9	60 188
Plutón	5909.0	90 710

Modifique su programa del Problema 2, de manera que también calcule $E_2(f)$, y úselo para determinar el ajuste potencial óptimo en mínimos cuadrados de la forma $y = Cx^{3/2}$ para (a) los cuatro primeros planetas y (b) los nueve planetas.

4. (a) Determine la recta de regresión para los datos $\{(x_k, y_k)\}_{k=1}^{50}$, siendo $x_k = (0.1)k$ e $y_k = x_k + \cos(k^{1/2})$.
- (b) Calcule $E_2(f)$.
- (c) Dibuje los datos y la recta de regresión sobre un mismo gráfico.

Ajuste de curvas

El método de linealización de los datos para $y = Ce^{Ax}$

Supongamos que queremos ajustar una curva exponencial de la forma

$$(1) \quad y = Ce^{Ax}$$

a un conjunto de puntos $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$ dado de antemano. Para afrontar este problema, empezamos tomando logaritmos en (1):

$$(2) \quad \ln(y) = Ax + \ln(C)$$

y, a continuación, hacemos un cambio de variable (y uno de constante)

$$(3) \quad Y = \ln(y), \quad X = x \quad \text{y} \quad B = \ln(C).$$

Lo que se obtiene es una relación lineal entre las nuevas variables X e Y :

$$(4) \quad Y = AX + B.$$

Los datos originales (x_k, y_k) se han transformado, con el cambio de variables, en $(X_k, Y_k) = (x_k, \ln(y_k))$; a este proceso lo llamamos **método de linealización de los datos**. El problema ahora es calcular la recta de regresión (4) para los puntos $\{(X_k, Y_k)\}$, para lo que planteamos las correspondientes ecuaciones normales de Gauss

$$(5) \quad \begin{aligned} \left(\sum_{k=1}^N X_k^2 \right) A + \left(\sum_{k=1}^N X_k \right) B &= \sum_{k=1}^N X_k Y_k, \\ \left(\sum_{k=1}^N X_k \right) A + NB &= \sum_{k=1}^N Y_k. \end{aligned}$$

Una vez calculados A y B , hallamos el parámetro C de la relación (1):

$$(6) \quad C = e^B.$$

Ejemplo 5.4. Vamos a usar el método de linealización de los datos para hallar un ajuste exponencial $y = Ce^{Ax}$ a los cinco datos $(0, 1.5)$, $(1, 2.5)$, $(2, 3.5)$, $(3, 5.0)$ y $(4, 7.5)$.

Aplicando el cambio de variables (4) a los datos originales obtenemos

$$(7)$$

$$\begin{aligned} \{(X_k, Y_k)\} &= \{(0, \ln(1.5)), (1, \ln(2.5)), (2, \ln(3.5)), (3, \ln(5.0)), (4, \ln(7.5))\} \\ &= \{(0, 0.40547), (1, 0.91629), (2, 1.25276), (3, 1.60944), (4, 2.01490)\}. \end{aligned}$$

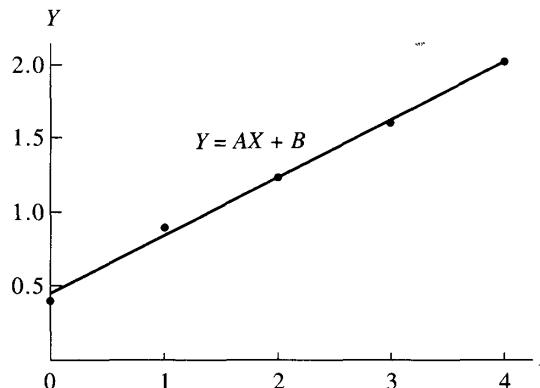


Figura 5.4 Los datos linealizados $\{(X_k, Y_k)\}$.

Tabla 5.4 Cálculo de los coeficientes de las ecuaciones normales de Gauss para los datos linealizados $\{(X_k, Y_k)\}$.

x_k	y_k	X_k	$Y_k = \ln(y_k)$	X_k^2	$X_k Y_k$
0.0	1.5	0.0	0.405465	0.0	0.000000
1.0	2.5	1.0	0.916291	1.0	0.916291
2.0	3.5	2.0	1.252763	4.0	2.505526
3.0	5.0	3.0	1.609438	9.0	4.828314
4.0	7.5	4.0	2.014903	16.0	8.059612
		10.0 $= \sum X_k$	6.198860 $= \sum Y_k$	30.0 $= \sum X_k^2$	16.309743 $= \sum X_k Y_k$

Los puntos transformados aparecen alineados, como se muestra en la Figura 5.4. Vamos ahora a comprobar que la recta de regresión $Y = AX + B$ para los puntos calculados en (7) es:

$$(8) \quad Y = 0.391202X + 0.457367.$$

Las operaciones necesarias para el cálculo de los coeficientes de las ecuaciones normales de Gauss (5) se muestran en la Tabla 5.4.

Las ecuaciones normales son, entonces,

$$(9) \quad \begin{aligned} 30A + 10B &= 16.309742 \\ 10A + 5B &= 6.198860, \end{aligned}$$

cuya solución es, como ya se ha dicho, $A = 0.3912023$ y $B = 0.457367$. Finalmente, obtenemos el valor de C que es $C = e^{0.457367} = 1.579910$. Sustituyendo los valores de A y C en la relación (1) obtenemos el ajuste exponencial (véase la Figura 5.5):

$$(10) \quad y = 1.579910e^{0.3912023x} \quad (\text{por el método de linealización de los datos}). \blacksquare$$

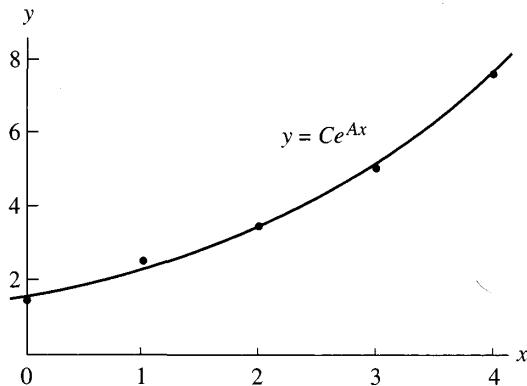


Figura 5.5 El ajuste exponencial $y = 1.579910e^{0.3912023x}$ obtenido con el método de linearización de los datos.

El método no lineal de los mínimos cuadrados para $y = Ce^{Ax}$

Supongamos, como antes, que queremos ajustar una curva exponencial de la forma

$$(11) \quad y = Ce^{Ax}$$

a un conjunto de puntos $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$ dado de antemano. El método no lineal de los mínimos cuadrados consiste en hallar el mínimo de la función

$$(12) \quad E(A, C) = \sum_{k=1}^N (Ce^{Ax_k} - y_k)^2.$$

Para ello, hallamos las derivadas parciales de $E(A, C)$ respecto de A y C ,

$$(13) \quad \frac{\partial E}{\partial A} = 2 \sum_{k=1}^N (Ce^{Ax_k} - y_k)(Cx_k e^{Ax_k})$$

y

$$(14) \quad \frac{\partial E}{\partial C} = 2 \sum_{k=1}^N (Ce^{Ax_k} - y_k)(e^{Ax_k}).$$

Al igualar a cero estas derivadas parciales (13) y (14) obtenemos, después de simplificar, las correspondientes ecuaciones normales

$$(15) \quad \begin{aligned} C \sum_{k=1}^N x_k e^{2Ax_k} - \sum_{k=1}^N x_k y_k e^{Ax_k} &= 0, \\ C \sum_{k=1}^N e^{Ax_k} - \sum_{k=1}^N y_k e^{Ax_k} &= 0. \end{aligned}$$

La diferencia con el método de linealización de los datos reside en el hecho de que las ecuaciones (15) no son lineales para las incógnitas A y C . Podríamos resolver las ecuaciones normales (15) usando el método iterativo de Newton-Raphson, pero ello conlleva un número alto de operaciones, por no mencionar el hecho de que necesitaríamos buenas aproximaciones iniciales de A y C . Muchos paquetes de programas incluyen, como subprogramas ya construidos, instrucciones que permiten minimizar funciones de varias variables y que podríamos utilizar para hallar el mínimo de la función $E(A, C)$ directamente. Uno de esos métodos es el de Nelder-Mead, que veremos en el Capítulo 8 y que tiene la ventaja de que no hace necesario el cálculo de las derivadas parciales.

Ejemplo 5.5. Vamos a usar el método no lineal de los mínimos cuadrados para hallar el ajuste exponencial óptimo $y = Ce^{Ax}$ para los cinco datos $(0, 1.5)$, $(1, 2.5)$, $(2, 3.5)$, $(3, 5.0)$ y $(4, 7.5)$.

Tenemos que minimizar la cantidad $E(A, C)$ dada por

$$(16) \quad E(A, C) = (C - 1.5)^2 + (Ce^A - 2.5)^2 + (Ce^{2A} - 3.5)^2 \\ + (Ce^{3A} - 5.0)^2 + (Ce^{4A} - 7.5)^2.$$

Para aproximar los valores de A y C que minimizan $E(A, C)$, usaremos la instrucción `fmins` del paquete de programas MATLAB. Primero hay que definir la función $E(A, C)$ como un fichero de texto, digamos `E.m`, para que pueda ser usado con el paquete MATLAB.

```
function z=E(u)
A=u(1);
C=u(2);
z=(C-1.5).^2+(C.*exp(A)-2.5).^2+(C.*exp(2*A)-3.5).^2+...
(C.*exp(3*A)-5.0).^2+(C.*exp(4*A)-7.5).^2;
```

Usando la instrucción `fmins` con los valores iniciales $A = 1.0$ y $C = 1.0$, obtenemos

```
>>fmins('E',[1 1])
ans =
0.38357046980073 1.61089952247928
```

De manera que el ajuste exponencial obtenido es

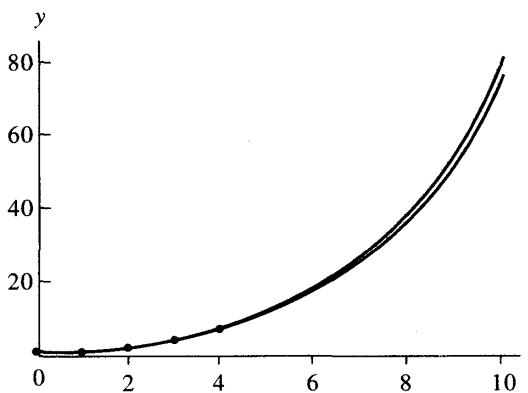
$$(17) \quad y = 1.6108995e^{0.3835705}$$

(ajuste por el método no lineal de los mínimos cuadrados).

Las soluciones obtenidas usando el método de linealización de datos y el método no lineal de los mínimos cuadrados se comparan en la Tabla 5.5. Hay una ligera discrepancia en los coeficientes; si deseamos interpolar, es posible comprobar que ambas aproximaciones difieren en no más del 2% a lo largo del intervalo $[0, 4]$ (véanse la Tabla 5.5 y la Figura 5.6). Si los errores de los datos siguen una distribución normal, entonces suele ser preferible usar (17). Cuando se realiza una extrapolación

Tabla 5.5 Comparación de los dos ajustes exponenciales.

x_k	y_k	$1.5799e^{0.39120x}$	$1.6109e^{0.38357x}$
0.0	1.5	1.5799	1.6109
1.0	2.5	2.3363	2.3640
2.0	3.5	3.4548	3.4692
3.0	5.0	5.1088	5.0911
4.0	7.5	7.5548	7.4713
5.0		11.1716	10.9644
6.0		16.5202	16.0904
7.0		24.4293	23.6130
8.0		36.1250	34.6527
9.0		53.4202	50.8535
10.0		78.9955	74.6287

**Figura 5.6** Comparación gráfica de las dos curvas exponenciales.

en un punto exterior al rango de los datos, las dos soluciones se separan y su discrepancia aumenta hasta casi un 6% para $x = 10$. ■

Cambios de variable que linealizan los datos

La técnica de linealizar los datos ha sido empleada frecuentemente para ajustar curvas tales como $y = Ce^{Ax}$, $y = A \ln(x) + B$ e $y = A/x + B$ a un conjunto de datos. Una vez elegido el tipo de curva, hay que realizar un cambio de variable adecuado de manera que las nuevas variables se relacionen linealmente. Por ejemplo, puede verificarse fácilmente que una relación del tipo $y = D/(x + C)$ se transforma en una relación lineal $Y = AX + B$ usando el cambio de variables

(y constantes) $X = xy$, $Y = y$, $C = -1/A$ y $D = -B/A$. En la Figura 5.7 se muestran las gráficas de varios tipos de curvas que pueden usarse y en la Tabla 5.6 se muestran algunos cambios de variable usados para linealizar datos.

Combinaciones lineales en mínimos cuadrados

El problema de las combinaciones lineales en mínimos cuadrados puede formularse de la siguiente manera: Dados N puntos $\{(x_k, y_k)\}$ y un conjunto de M funciones linealmente independientes $\{f_j(x)\}$, se trata de encontrar M coeficientes $\{c_j\}$ tales que la función $f(x)$ definida como la combinación lineal

$$(18) \quad f(x) = \sum_{j=1}^M c_j f_j(x)$$

minimice la suma de los cuadrados de los errores

$$(19) \quad E(c_1, c_2, \dots, c_M) = \sum_{k=1}^N (f(x_k) - y_k)^2 = \sum_{k=1}^N \left(\left(\sum_{j=1}^M c_j f_j(x_k) \right) - y_k \right)^2.$$

Para que E alcance un mínimo en un punto, es necesario que cada derivada parcial en dicho punto sea cero (o sea, $\partial E / \partial c_i = 0$ para $i = 1, 2, \dots, M$), lo que equivale a que $\{c_j\}$ sea la solución del sistema de ecuaciones lineales

$$(20) \quad \sum_{k=1}^N \left(\left(\sum_{j=1}^M c_j f_j(x_k) \right) - y_k \right) (f_i(x_k)) = 0 \quad \text{para } i = 1, 2, \dots, M.$$

Intercambiando el orden de las sumas que aparecen en (20) obtenemos un sistema de ecuaciones lineales de orden $M \times M$ en el que las incógnitas son los coeficientes $\{c_j\}$; estas ecuaciones reciben el nombre de ecuaciones normales (o, también, ecuaciones normales de Gauss):

$$(21) \quad \sum_{j=1}^M \left(\sum_{k=1}^N f_i(x_k) f_j(x_k) \right) c_j = \sum_{k=1}^N f_i(x_k) y_k \quad \text{para } i = 1, 2, \dots, M.$$

Formulación matricial

Aunque (21) es fácilmente reconocible como un sistema de M ecuaciones lineales con M incógnitas, hay que prestar un poco de atención para evitar realizar operaciones aritméticas innecesarias a la hora de escribir este sistema en notación

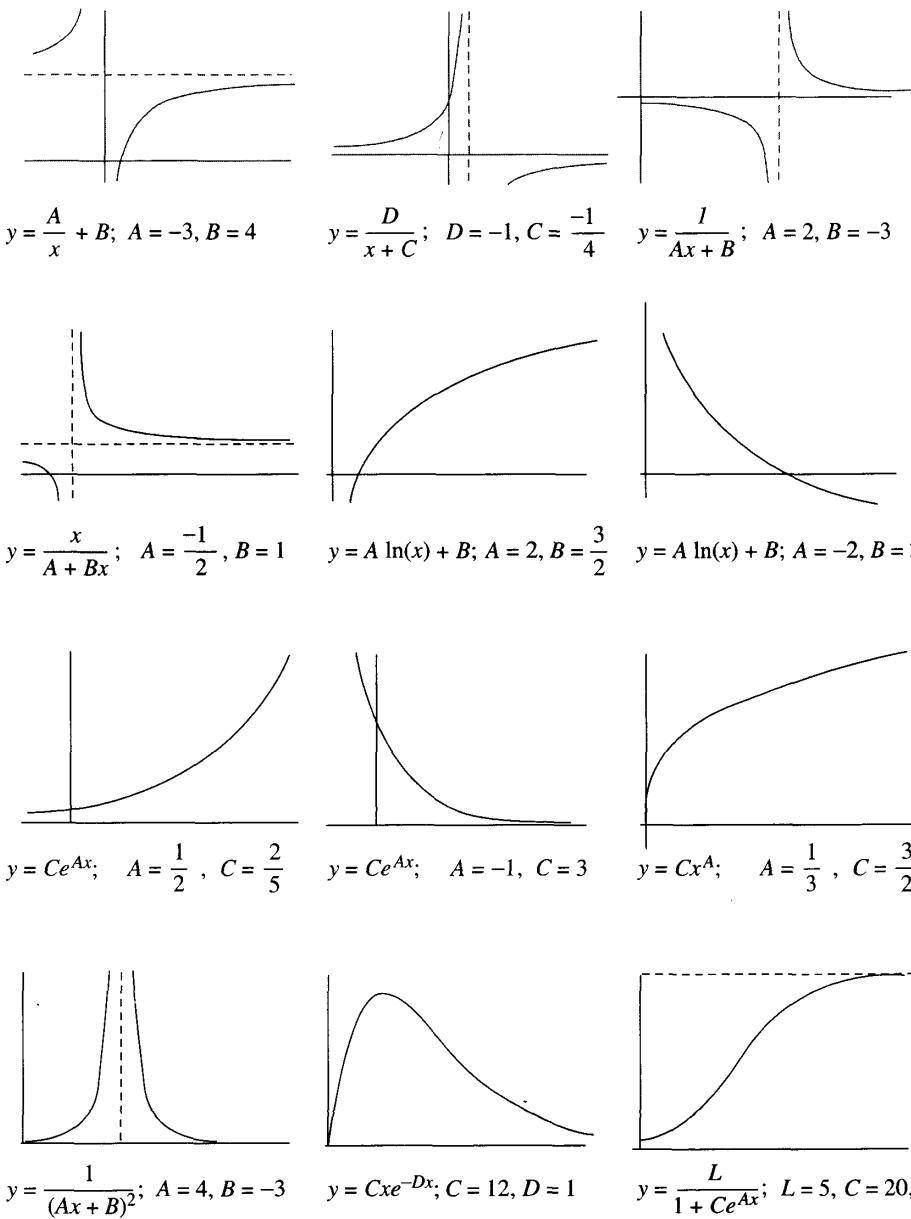


Figura 5.7 Tipos de curvas con las que puede usarse el método de linealización de los datos.

Tabla 5.6 Cambios de variable para linealizar los datos.

Función, $y = f(x)$	Linealización, $Y = Ax + B$	Cambios
$y = \frac{A}{x} + B$	$y = A\frac{1}{x} + B$	$X = \frac{1}{x}, Y = y$
$y = \frac{D}{x + C}$	$y = \frac{-1}{C}(xy) + \frac{D}{C}$	$X = xy, Y = y$
$y = \frac{1}{Ax + B}$	$\frac{1}{y} = Ax + B$	$C = \frac{-1}{A}, D = \frac{-B}{A}$
$y = \frac{x}{Ax + B}$	$\frac{1}{y} = A\frac{1}{x} + B$	$X = x, Y = \frac{1}{y}$
$y = A \ln(x) + B$	$y = A \ln(x) + B$	$X = \ln(x), Y = y$
$y = Ce^{Ax}$	$\ln(y) = Ax + \ln(C)$	$X = x, Y = \ln(y),$ $C = e^B$
$y = Cx^A$	$\ln(y) = A \ln(x) + \ln(C)$	$X = \ln(x), Y = \ln(y),$ $C = e^B$
$y = (Ax + B)^{-2}$	$y^{-1/2} = Ax + B$	$X = x, Y = y^{-1/2}$
$y = Cxe^{-Dx}$	$\ln\left(\frac{y}{x}\right) = -Dx + \ln(C)$	$X = x, Y = \ln\left(\frac{y}{x}\right)$ $C = e^B, D = -A$
$y = \frac{L}{1 + Ce^{Ax}}$	$\ln\left(\frac{L}{y} - 1\right) = Ax + \ln(C)$	$X = x, Y = \ln\left(\frac{L}{y} - 1\right),$ $C = e^B$

matricial. La clave está en darse cuenta de que la matriz \mathbf{F} y su traspuesta \mathbf{F}' , que damos a continuación, juegan un papel fundamental:

$$\mathbf{F} = \begin{bmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_M(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_M(x_2) \\ f_1(x_3) & f_2(x_3) & \cdots & f_M(x_3) \\ \vdots & \vdots & & \vdots \\ f_1(x_N) & f_2(x_N) & \cdots & f_M(x_N) \end{bmatrix},$$

$$\mathbf{F}' = \begin{bmatrix} f_1(x_1) & f_1(x_2) & f_1(x_3) & \cdots & f_1(x_N) \\ f_2(x_1) & f_2(x_2) & f_2(x_3) & \cdots & f_2(x_N) \\ \vdots & \vdots & \vdots & & \vdots \\ f_M(x_1) & f_M(x_2) & f_M(x_3) & \cdots & f_M(x_N) \end{bmatrix}.$$

Consideremos el producto de \mathbf{F}' por la matriz columna \mathbf{Y} :

$$(22) \quad \mathbf{F}'\mathbf{Y} = \begin{bmatrix} f_1(x_1) & f_1(x_2) & f_1(x_3) & \cdots & f_1(x_N) \\ f_2(x_1) & f_2(x_2) & f_2(x_3) & \cdots & f_2(x_N) \\ \vdots & \vdots & \vdots & & \vdots \\ f_M(x_1) & f_M(x_2) & f_M(x_3) & \cdots & f_M(x_N) \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}.$$

El elemento de la i -ésima fila del producto $\mathbf{F}'\mathbf{Y}$ en (22) coincide con el elemento i -ésimo de la matriz columna que contiene los términos independientes en el sistema (21); esto es,

$$(23) \quad \sum_{k=1}^N f_i(x_k) y_k = \text{fila}_i \mathbf{F}' \cdot [y_1 \ y_2 \ \dots \ y_N]'$$

Ahora consideremos el producto $\mathbf{F}'\mathbf{F}$, que es la matriz de orden $M \times M$:

$$\mathbf{F}'\mathbf{F} = \begin{bmatrix} f_1(x_1) & f_1(x_2) & f_1(x_3) & \cdots & f_1(x_N) \\ f_2(x_1) & f_2(x_2) & f_2(x_3) & \cdots & f_2(x_N) \\ \vdots & \vdots & \vdots & & \vdots \\ f_M(x_1) & f_M(x_2) & f_M(x_3) & \cdots & f_M(x_N) \end{bmatrix} \begin{bmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_M(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_M(x_2) \\ f_1(x_3) & f_2(x_3) & \cdots & f_M(x_3) \\ \vdots & \vdots & & \vdots \\ f_1(x_N) & f_2(x_N) & \cdots & f_M(x_N) \end{bmatrix}.$$

El elemento que ocupa la posición (i, j) en la matriz $\mathbf{F}'\mathbf{F}$ coincide con el coeficiente de c_j en la i -ésima ecuación del sistema (21); esto es,

$$(24) \quad \sum_{k=1}^N f_i(x_k) f_j(x_k) = f_i(x_1) f_j(x_1) + f_i(x_2) f_j(x_2) + \cdots + f_i(x_N) f_j(x_N).$$

Cuando M es pequeño, una forma computacionalmente eficiente de calcular los coeficientes óptimos en mínimos cuadrados de la combinación (18) es almacenar la matriz \mathbf{F} , calcular $\mathbf{F}'\mathbf{F}$ y $\mathbf{F}'\mathbf{Y}$ y, finalmente, resolver el sistema lineal

$$(25) \quad \mathbf{F}'\mathbf{F}\mathbf{C} = \mathbf{F}'\mathbf{Y} \quad \text{cuya incógnita es } \mathbf{C}.$$

Ajuste polinomial

Cuando el método que hemos descrito inmediatamente antes se aplica al caso en el que tenemos $M + 1$ funciones dadas por $\{f_j(x) = x^{j-1}\}$, la función $f(x)$ será un polinomio de grado menor o igual que M :

$$(26) \quad f(x) = c_1 + c_2x + c_3x^2 + \cdots + c_{M+1}x^M.$$

Vamos a mostrar ahora cómo se calcula la **parábola óptima en el sentido de los mínimos cuadrados**; la extensión a polinomios de grado más alto es sencilla y queda como ejercicio.

Teorema 5.3 (Parábola óptima en mínimos cuadrados). Supongamos que tenemos N puntos $\{(x_k, y_k)\}_{k=1}^N$ cuyas abscisas son todas distintas. Los coeficientes de la parábola de ecuación

$$(27) \quad y = f(x) = Ax^2 + Bx + C$$

que mejor se ajusta a dichos puntos en el sentido de los mínimos cuadrados son las soluciones A , B y C del sistema de ecuaciones lineales

$$(28) \quad \begin{aligned} \left(\sum_{k=1}^N x_k^4 \right) A + \left(\sum_{k=1}^N x_k^3 \right) B + \left(\sum_{k=1}^N x_k^2 \right) C &= \sum_{k=1}^N y_k x_k^2, \\ \left(\sum_{k=1}^N x_k^3 \right) A + \left(\sum_{k=1}^N x_k^2 \right) B + \left(\sum_{k=1}^N x_k \right) C &= \sum_{k=1}^N y_k x_k, \\ \left(\sum_{k=1}^N x_k^2 \right) A + \left(\sum_{k=1}^N x_k \right) B + NC &= \sum_{k=1}^N y_k. \end{aligned}$$

Demostración. Los coeficientes A , B y C de la parábola óptima deben minimizar la función:

$$(29) \quad E(A, B, C) = \sum_{k=1}^N (Ax_k^2 + Bx_k + C - y_k)^2.$$

Las derivadas parciales $\partial E / \partial A$, $\partial E / \partial B$ y $\partial E / \partial C$ deben ser todas cero, así que:

$$(30) \quad \begin{aligned} 0 &= \frac{\partial E(A, B, C)}{\partial A} = 2 \sum_{k=1}^N (Ax_k^2 + Bx_k + C - y_k)^1(x_k^2), \\ 0 &= \frac{\partial E(A, B, C)}{\partial B} = 2 \sum_{k=1}^N (Ax_k^2 + Bx_k + C - y_k)^1(x_k), \\ 0 &= \frac{\partial E(A, B, C)}{\partial C} = 2 \sum_{k=1}^N (Ax_k^2 + Bx_k + C - y_k)^1(1). \end{aligned}$$

Tabla 5.7 Cálculo de los coeficientes de las ecuaciones normales para la parábola óptima del Ejemplo 5.6.

x_k	y_k	x_k^2	x_k^3	x_k^4	$x_k y_k$	$x_k^2 y_k$
-3	3	9	-27	81	-9	27
0	1	0	0	0	0	0
2	1	4	8	16	2	4
4	3	16	64	256	12	48
3	8	29	45	353	5	79

Usando la propiedad distributiva de la suma, podemos sacar cada coeficiente A , B y C como factor común en las sumas de (30) en las que aparecen, lo que resulta en las ecuaciones normales escritas en (28).

Ejemplo 5.6. Vamos a determinar la parábola óptima en mínimos cuadrados para los cuatro puntos $(-3, 3)$, $(0, 1)$, $(2, 1)$ y $(4, 3)$.

En la Tabla 5.7 se muestran los cálculos necesarios para construir el sistema lineal (28) que viene dado por

$$353A + 45B + 29C = 79$$

$$45A + 29B + 3C = 5$$

$$29A + 3B + 4C = 8$$

y cuya solución es $A = 585/3278$, $B = -631/3278$ y $C = 1394/1639$, y la parábola que buscamos es (véase la Figura 5.8)

$$y = \frac{585}{3278}x^2 - \frac{631}{3278}x + \frac{1394}{1639} = 0.178462x^2 - 0.192495x + 0.850519.$$

El fenómeno de la oscilación polinomial

No deja de ser tentadora la posibilidad de utilizar un polinomio óptimo en el sentido de los mínimos cuadrados para ajustar datos que no son lineales. Pero si los datos no muestran una naturaleza polinomial, puede ocurrir que la curva resultante presente oscilaciones grandes. Este fenómeno, llamado **oscilación polinomial**, se hace más pronunciado conforme aumenta el grado del polinomio y, por esta razón, no se suelen usar polinomios de grado seis o mayor; a no ser que se sepa que la función de la que provienen los datos es un polinomio.

Por ejemplo, vamos a usar la función $f(x) = 1.44/x^2 + 0.24x$ para generar seis parejas de datos $(0.25, 23.1)$, $(1.0, 1.68)$, $(1.5, 1.0)$, $(2.0, 0.84)$, $(2.4, 0.826)$ y

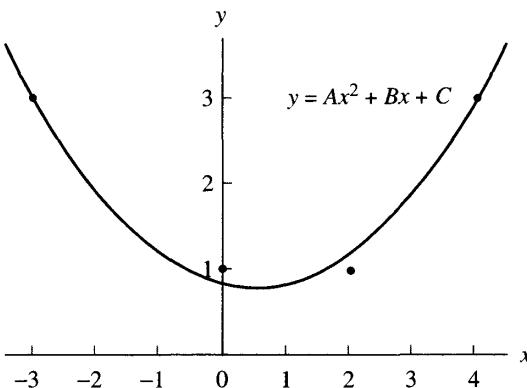


Figura 5.8 La parábola óptima en mínimos cuadrados del Ejemplo 5.6.

(5.0, 1.2576). Los ajustes mediante polinomios óptimos en mínimos cuadrados que se obtienen, para diferentes grados, son

$$P_2(x) = 22.93 - 16.96x + 2.553x^2,$$

$$P_3(x) = 33.04 - 46.51x + 19.51x^2 - 2.296x^3,$$

$$P_4(x) = 39.92 - 80.93x + 58.39x^2 - 17.15x^3 + 1.680x^4,$$

y

$$P_5(x) = 46.02 - 118.1x + 119.4x^2 - 57.51x^3 + 13.03x^4 - 1.085x^5$$

cuyas gráficas se muestran en las Figuras 5.9(a)–(d). Hagamos notar que $P_3(x)$, $P_4(x)$ y $P_5(x)$ presentan oscilaciones grandes en el intervalo [2, 5]; es más, aunque $P_5(x)$ pasa incluso por los seis puntos, es la que peor se aproxima a la función. Si insistimos en elegir un polinomio que se ajuste a los datos y se aproxime a esta función, deberíamos elegir $P_2(x)$.

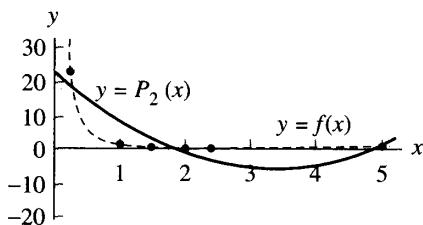
El programa que damos a continuación emplea la matriz \mathbf{F} cuyos términos son los valores $f_j(x_k) = x_k^{j-1}$.

Programa 5.2 (Polinomio óptimo en mínimos cuadrados). Construcción del polinomio de grado M dado por

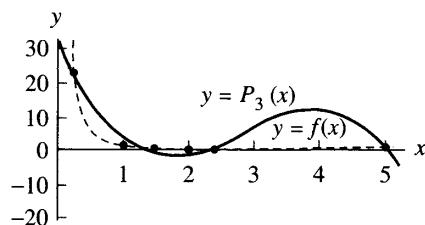
$$P_M(x) = c_1 + c_2x + c_3x^2 + \cdots + c_Mx^{M-1} + c_{M+1}x^M$$

que mejor se ajusta en el sentido de los mínimos cuadrados a las N parejas de datos $\{(x_k, y_k)\}_{k=1}^N$.

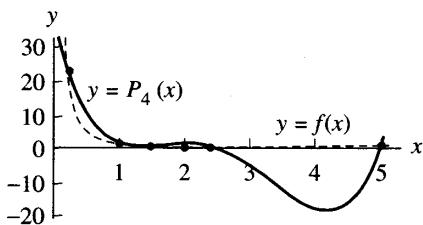
```
function C = lspoly(X,Y,M)
% Datos
%      - X es el vector de orden 1 x n de las abscisas
```



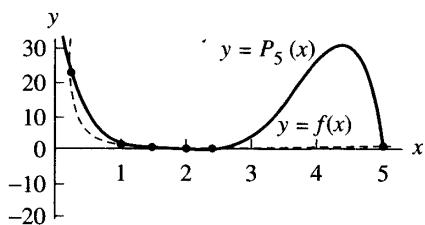
(a)



(b)



(c)



(d)

Figura 5.9 (a) Ajuste de $P_2(x)$ a los datos. (b) Ajuste de $P_3(x)$ a los datos.
 (c) Ajuste de $P_4(x)$ a los datos. (d) Ajuste de $P_5(x)$ a los datos.

```
% - Y es el vector de orden 1 x n de las ordenadas
% - M es el grado del polinomio óptimo
% Resultado
% - C es el vector de coeficientes del polinomio
%     en potencias decrecientes de x

n=length(X);
B=zeros(1:M+1);
F=zeros(n,M+1);

% Se rellenan las columnas de F con las potencias de X
for k=1:M+1
    F(:,k)=X'.^(k-1);
end

% Resolución de las ecuaciones normales
A=F'*F;
B=F'*Y';
C=A\B;
C=flipud(C);
```

Ejercicios

1. Determine la parábola óptima en el sentido de los mínimos cuadrados de la forma $f(x) = Ax^2 + Bx + C$ para cada conjunto de datos.

(a)	x_k	y_k
	-3	15
	-1	5
	1	1
	3	5

(b)	x_k	y_k
	-3	-1
	-1	25
	1	25
	3	1

2. Determine la parábola óptima en el sentido de los mínimos cuadrados de la forma $f(x) = Ax^2 + Bx + C$ para cada conjunto de datos.

(a)	x_k	y_k
	-2	-5.8
	-1	1.1
	0	3.8
	1	3.3
	2	-1.5

(b)	x_k	y_k
	-2	2.8
	-1	2.1
	0	3.25
	1	6.0
	2	11.5

(c)	x_k	y_k
	-2	10.0
	-1	1.0
	0	0.0
	1	2.0
	2	9.0

3. Para el conjunto de datos que se muestra al final del enunciado, determine la curva de cada familia que mejor se les ajusta en el sentido de los mínimos cuadrados.

- (a) $f(x) = Ce^{Ax}$, usando el cambio de variables $X = x$, $Y = \ln(y)$ y $C = e^B$, dado en la Tabla 5.6, para linealizar los datos.
- (b) $f(x) = Cx^A$, usando el cambio de variables $X = \ln(x)$, $Y = \ln(y)$ y $C = e^B$, dado en la Tabla 5.6, para linealizar los datos.
- (c) Use $E_2(f)$ para determinar cuál de las dos curvas se les ajusta mejor.

x_k	y_k
1	0.6
2	1.9
3	4.3
4	7.6
5	12.6

4. Para el conjunto de datos que se muestra al final del enunciado, determine la curva de cada familia que mejor se les ajusta en el sentido de los mínimos cuadrados.

- (a) $f(x) = Ce^{Ax}$, usando el cambio de variables $X = x$, $Y = \ln(y)$ y $C = e^B$, dado en la Tabla 5.6, para linealizar los datos.

- (b) $f(x) = 1/(Ax + B)$, usando el cambio de variables $X = x$ e $Y = 1/y$, dado en la Tabla 5.6, para linealizar los datos.
- (c) Use $E_2(f)$ para determinar cuál de las dos curvas se les ajusta mejor.

x_k	y_k
-1	6.62
0	3.94
1	2.17
2	1.35
3	0.89

5. Para los dos conjuntos de datos que se muestran al final del enunciado, determine la curva de cada familia que mejor se les ajusta en el sentido de los mínimos cuadrados.

- (a) $f(x) = Ce^{Ax}$, usando el cambio de variables $X = x$, $Y = \ln(y)$ y $C = e^B$, dado en la Tabla 5.6, para linealizar los datos.
- (b) $f(x) = (Ax + B)^{-2}$, usando el cambio $X = x$ e $Y = y^{-1/2}$, dado en la Tabla 5.6, para linealizar los datos.
- (c) Use $E_2(f)$ para determinar cuál de las curvas se les ajusta mejor.

(i)

x_k	y_k
-1	13.45
0	3.01
1	0.67
2	0.15

(ii)

x_k	y_k
-1	13.65
0	1.38
1	0.49
3	0.15

6. Crecimiento logístico de poblaciones. Cuando una población $P(t)$ no puede crecer más halla de un cierto valor límite L , la gráfica de la función $P(t)$ es una curva, llamada curva logística, de ecuación $y = L/(1 + Ce^{At})$. Calcule A y C para los siguientes datos, siendo L un valor conocido.

- (a) $(0, 200)$, $(1, 400)$, $(2, 650)$, $(3, 850)$, $(4, 950)$ y $L = 1000$.
- (b) $(0, 500)$, $(1, 1000)$, $(2, 1800)$, $(3, 2800)$, $(4, 3700)$ y $L = 5000$.
7. Use los siguientes conjuntos de datos sobre la población de los Estados Unidos de América para hallar la curva logística $P(t)$ correspondiente y estime la población en el año 2000.

(a) Suponga que $L = 8 \times 10^8$

Año	t_k	P_k
1800	-10	5.3
1850	-5	23.2
1900	0	76.1
1950	5	152.3

(b) Suponga que $L = 8 \times 10^8$

Año	t_k	P_k
1900	0	76.1
1920	2	106.5
1940	4	132.6
1960	6	180.7
1980	8	226.5

En los Ejercicios 8–15, realice el cambio de variables indicado en la Tabla 5.6 y deduzca la correspondiente relación lineal entre las nuevas variables.

8. $y = \frac{A}{x} + B$

9. $y = \frac{D}{x+C}$

10. $y = \frac{1}{Ax+B}$

11. $y = \frac{x}{A+Bx}$

12. $y = A \ln(x) + B$

13. $y = Cx^A$

14. $y = (Ax+B)^{-2}$

15. $y = Cxe^{-Dx}$

16. (a) Siga el procedimiento descrito en la demostración del Teorema 5.3 para deducir las ecuaciones normales que permiten hallar la curva de la forma $f(x) = A \cos(x) + B \sin(x)$ que mejor se ajusta a un conjunto de datos en el sentido de los mínimos cuadrados.

- (b) Utilice los resultados del apartado (a) para hallar la curva de ecuación $y = f(x) = A \cos(x) + B \sin(x)$ que mejor se ajusta al conjunto de datos

x_k	y_k
-3.0	-0.1385
-1.5	-2.1587
0.0	0.8330
1.5	2.2774
3.0	-0.5110

17. El plano $z = Ax + By + C$ que mejor se ajusta en el sentido de los mínimos cuadrados a un conjunto de N puntos $(x_1, y_1, z_1), (x_2, y_2, z_2), \dots, (x_N, y_N, z_N)$ se obtiene minimizando

$$E(A, B, C) = \sum_{k=1}^N (Ax_k + By_k + C - z_k)^2.$$

Deduzca las correspondientes ecuaciones normales:

$$\begin{aligned} \left(\sum_{k=1}^N x_k^2 \right) A + \left(\sum_{k=1}^N x_k y_k \right) B + \left(\sum_{k=1}^N x_k \right) C &= \sum_{k=1}^N z_k x_k, \\ \left(\sum_{k=1}^N x_k y_k \right) A + \left(\sum_{k=1}^N y_k^2 \right) B + \left(\sum_{k=1}^N y_k \right) C &= \sum_{k=1}^N z_k y_k, \\ \left(\sum_{k=1}^N x_k \right) A + \left(\sum_{k=1}^N y_k \right) B + NC &= \sum_{k=1}^N z_k. \end{aligned}$$

18. Determine los planos que mejor se ajustan a los siguientes conjuntos de datos en el sentido de los mínimos cuadrados.
- (a) $(1, 1, 7), (1, 2, 9), (2, 1, 10), (2, 2, 11)$ y $(2, 3, 12)$
 - (b) $(1, 2, 6), (2, 3, 7), (1, 1, 8), (2, 2, 8)$ y $(2, 1, 9)$
 - (c) $(3, 1, -3), (2, 1, -1), (2, 2, 0), (1, 1, 1)$ y $(1, 2, 3)$
19. Consideremos la siguiente tabla de datos

x_k	y_k
1.0	2.0
2.0	5.0
3.0	10.0
4.0	17.0
5.0	26.0

Si, para linealizar los datos, hacemos el cambio de variables $X = xy$ e $Y = 1/y$ en la función $y = D/(x + C)$ entonces el ajuste que se obtiene es

$$y = \frac{-17.719403}{x - 5.476617}.$$

Si lo que se hace es el cambio de variables $X = x$ e $Y = 1/y$ en la función $y = 1/(Ax + B)$, también para linealizar los datos, entonces el ajuste que se obtiene es

$$y = \frac{1}{-0.1064253x + 0.4987330}.$$

Determine cuál de las curvas de ajuste es mejor y por qué una de las soluciones es completamente absurda.

Algoritmos y programas

1. La relación, hora a hora, de temperaturas en Puerto Real (Cádiz) durante un día de noviembre se da en la tabla que figura más abajo.

- (a) Siguiendo el procedimiento descrito en el Ejemplo 5.5 (utilice la instrucción `fmins` del paquete MATLAB), calcule la función de la forma $f(x) = A \cos(Bx) + C \sin(Dx)$ que mejor se ajusta a los datos de la tabla en el sentido de los mínimos cuadrados.
- (b) Determine $E_2(f)$.
- (c) Dibuje los datos y la curva obtenida en una misma gráfica.

Hora	Grados	Hora	Grados
1	8	13	16
2	8	14	16
3	8	15	15
4	8	16	14
5	7	17	13
6	7	18	13
7	7	19	12
8	8	20	11
9	10	21	10
10	14	22	10
11	14	23	9
12	15	24	8

5.3 Interpolación polinomial a trozos

Frecuentemente, la interpolación polinomial para un conjunto numeroso de $N+1$ datos $\{(x_k, y_k)\}_{k=0}^N$ resulta ser muy poco satisfactoria. Como hemos visto en la Sección 5.2, dado que un polinomio de grado N puede tener $N - 1$ extremos relativos, es posible que su gráfica presente oscilaciones grandes al hacerla pasar por los puntos dados. Otra opción es ir enlazando, una detrás de otra, las gráficas de unos polinomios de grado bajo $S_k(x)$ que sólo interpolan entre dos nodos consecutivos (x_k, y_k) y (x_{k+1}, y_{k+1}) (véase la Figura 5.10). Las porciones adyacentes de la curva $y = S_k(x)$ e $y = S_{k+1}(x)$, que se construyen sobre los intervalos $[x_k, x_{k+1}]$ y $[x_{k+1}, x_{k+2}]$, respectivamente, se enlazan una con la otra en el punto (x_{k+1}, y_{k+1}) y el conjunto de funciones $\{S_k(x)\}$ forma una **curva polinomial a trozos** o **cercha**, que denotaremos por $S(x)$.

Interpolación lineal a trozos

El polinomio más simple que podemos usar, un polinomio de grado 1, produce una línea quebrada que consta de los segmentos rectilíneos que unen los puntos consecutivamente. Si usamos el polinomio interpolador de Lagrange de la

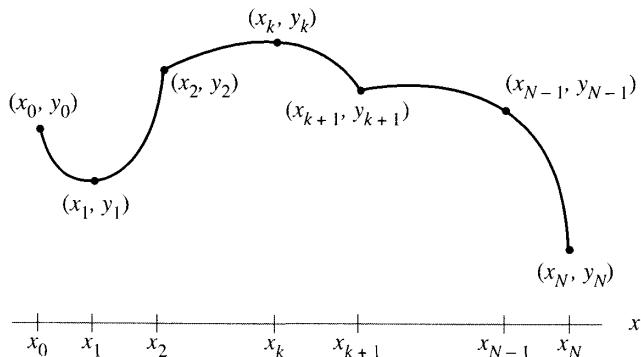


Figura 5.10 Interpolación polinomial a trozos.

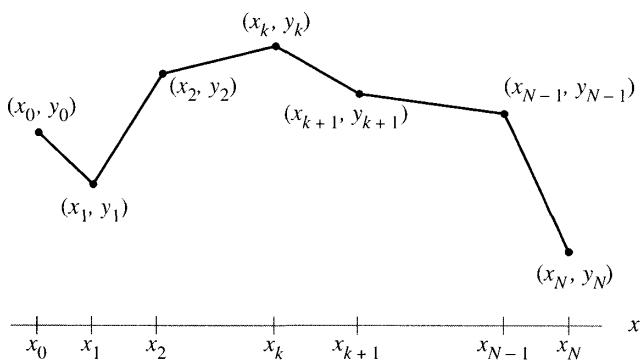


Figura 5.11 Interpolación lineal a trozos (cercha lineal).

Sección 4.3 para determinar cada trozo de esta línea quebrada obtenemos

$$(1) \quad S_k(x) = y_k \frac{x - x_{k+1}}{x_k - x_{k+1}} + y_{k+1} \frac{x - x_k}{x_{k+1} - x_k} \quad \text{para} \quad x_k \leq x \leq x_{k+1}.$$

El resultado se muestra en la Figura 5.11.

Obtenemos una expresión equivalente usando la ecuación de la recta cuando se conocen un punto y su pendiente:

$$S_k(x) = y_k + d_k(x - x_k),$$

siendo $d_k = (y_{k+1} - y_k)/(x_{k+1} - x_k)$. La *cercha* (en inglés, *spline*) *lineal* que

se obtiene se puede escribir como

$$(2) \quad S(x) = \begin{cases} y_0 + d_0(x - x_0) & \text{para } x \text{ en } [x_0, x_1], \\ y_1 + d_1(x - x_1) & \text{para } x \text{ en } [x_1, x_2], \\ \vdots & \vdots \\ y_k + d_k(x - x_k) & \text{para } x \text{ en } [x_k, x_{k+1}], \\ \vdots & \vdots \\ y_{N-1} + d_{N-1}(x - x_{N-1}) & \text{para } x \text{ en } [x_{N-1}, x_N]. \end{cases}$$

La expresión (2) es mejor que la expresión (1) para calcular $S(x)$ explícitamente. Suponiendo que las abscisas están ordenadas de manera creciente $x_0 < x_1 < \dots < x_{N-1} < x_N$ y dado un valor de x , el intervalo $[x_k, x_{k+1}]$ al que x pertenece puede hallarse calculando de manera sucesiva las diferencias $x - x_1, \dots, x - x_k, x - x_{k+1}$ hasta llegar a $k + 1$, que es el menor entero tal que $x - x_{k+1} < 0$. Una vez encontrado el valor de k tal que $x_k \leq x \leq x_{k+1}$, el valor de la función $S(x)$ es

$$(3) \quad S(x) = S_k(x) = y_k + d_k(x - x_k).$$

La técnica descrita puede extenderse a polinomios de mayor grado; por ejemplo, si tenemos un número impar de nodos x_0, x_1, \dots, x_{2M} , entonces podemos construir un polinomio cuadrático en cada subintervalo $[x_{2k}, x_{2k+2}]$, para $k = 0, 1, \dots, M - 1$ y obtener una cercha parabólica (esto es, un polinomio cuadrático a trozos). Un inconveniente de este método es que la curvatura de la cercha resultante puede cambiar abruptamente en los nodos pares x_{2k} , con lo cual su gráfica puede presentar esquinas y distorsiones no deseadas en algunas aplicaciones. La derivada segunda de una cercha cuadrática es, en general, discontinua en los nodos pares; sin embargo, si usamos polinomios cúbicos a trozos, entonces podemos conseguir que tanto la primera derivada como la segunda sean continuas en el intervalo.

Cerchas cúbicas

El ajuste de una curva polinomial a trozos a un conjunto de puntos dados tiene aplicaciones en los campos del diseño asistido por computador, de la fabricación asistida por computador y de los sistemas de generación de gráficas por computador. Lo que habitualmente se desea es dibujar una curva que pase por una serie de puntos, cuyas coordenadas conocemos con bastante precisión, y que sea suave. Tradicionalmente esto se hacía usando una cercha, una regla flexible o un conjunto de plantillas rígidas, con las que se podía dibujar a mano alzada una curva que parece suave a la mirada. Matemáticamente, es posible construir una función cúbica $S_k(x)$ en cada intervalo $[x_k, x_{k+1}]$ de manera que la

curva definida a trozos $y = S(x)$ que resulta es dos veces derivable y la segunda derivada es continua en el intervalo completo $[x_0, x_N]$. La continuidad de $S'(x)$ significa que la curva de ecuación $y = S(x)$ no tiene esquinas; la continuidad de $S''(x)$ significa que el radio de curvatura está definido en cada punto.

Definición 5.1 (Cercha cúbica interpoladora). Supongamos que tenemos $N + 1$ puntos $\{(x_k, y_k)\}_{k=0}^N$ cuyas abscisas están ordenadas de manera creciente $a = x_0 < x_1 < \dots < x_N = b$. Se dice que una función $S(x)$ es una **cercha cúbica interpoladora** para dichos datos si existen N polinomios cúbicos $S_k(x)$, que podemos escribir en términos de unos coeficientes $s_{k,0}$, $s_{k,1}$, $s_{k,2}$ y $s_{k,3}$ como

$$\text{I. } S(x) = S_k(x) = s_{k,0} + s_{k,1}(x - x_k) + s_{k,2}(x - x_k)^2 + s_{k,3}(x - x_k)^3$$

para $x \in [x_k, x_{k+1}]$ y $k = 0, 1, \dots, N-1$, que verifican las siguientes propiedades:

$$\text{II. } S(x_k) = y_k \quad \text{para } k = 0, 1, \dots, N,$$

$$\text{III. } S_k(x_{k+1}) = S_{k+1}(x_{k+1}) \quad \text{para } k = 0, 1, \dots, N-2,$$

$$\text{IV. } S'_k(x_{k+1}) = S'_{k+1}(x_{k+1}) \quad \text{para } k = 0, 1, \dots, N-2,$$

$$\text{V. } S''_k(x_{k+1}) = S''_{k+1}(x_{k+1}) \quad \text{para } k = 0, 1, \dots, N-2.$$

Las relaciones I significan que $S(x)$ es un polinomio cúbico a trozos. Las relaciones II significan que $S(x)$ interpola los datos. Las relaciones III y IV significan que $S(x)$ es una función derivable y con derivada continua. Finalmente, las relaciones V significan que la derivada segunda de $S(x)$ también existe y es continua.

Existencia de cerchas cúbicas interpoladoras

Vamos determinar si es posible construir una cercha cúbica que verifique las propiedades I a V. Cada polinomio cúbico $S_k(x)$ tiene cuatro coeficientes ($s_{k,0}$, $s_{k,1}$, $s_{k,2}$ y $s_{k,3}$) desconocidos a priori, lo que hace un total de $4N$ coeficientes por determinar; esto significa que deben especificarse $4N$ condiciones. Las relaciones II sobre los datos proporcionan $N + 1$ condiciones y cada una de las relaciones III, IV y V proporcionan $N - 1$; por tanto, la definición de cercha cúbica interpoladora especifica $N + 1 + 3(N - 1) = 4N - 2$ condiciones que deben verificarse, lo que nos deja dos grados de libertad para calcular los coeficientes. Estos dos grados de libertad se llaman **restricciones en los extremos** porque normalmente involucran los valores de $S'(x)$ o los de $S''(x)$ en los extremos del intervalo x_0 y x_N ; esto lo discutiremos luego, ahora vamos a pasar a la construcción de la cercha.

Puesto que $S(x)$ es un polinomio cúbico a trozos, su derivada segunda $S''(x)$ es lineal a trozos en $[x_0, x_N]$. En consecuencia, la fórmula de interpolación lineal

de Lagrange nos proporciona para $x_k \leq x \leq x_{k+1}$ la siguiente representación de $S''(x) = S''_k(x)$:

$$(4) \quad S''_k(x) = S''(x_k) \frac{x - x_{k+1}}{x_k - x_{k+1}} + S''(x_{k+1}) \frac{x - x_k}{x_{k+1} - x_k}.$$

Usando ahora la notación $m_k = S''(x_k)$, $m_{k+1} = S''(x_{k+1})$ y $h_k = x_{k+1} - x_k$, la expresión (4) queda

$$(5) \quad S''_k(x) = \frac{m_k}{h_k}(x_{k+1} - x) + \frac{m_{k+1}}{h_k}(x - x_k)$$

para $x_k \leq x \leq x_{k+1}$ y $k = 0, 1, \dots, N - 1$. Al integrar la relación (5) dos veces aparecen dos constantes de integración. Organizando de manera conveniente el polinomio lineal que involucra estas constantes, el resultado de estas integraciones lo podemos escribir de la siguiente manera

$$(6) \quad S_k(x) = \frac{m_k}{6h_k}(x_{k+1} - x)^3 + \frac{m_{k+1}}{6h_k}(x - x_k)^3 + p_k(x_{k+1} - x) + q_k(x - x_k).$$

Ahora evaluamos $S_k(x_k)$ y $S_k(x_{k+1})$ usando la ecuación (6) y los valores de interpolación dados $y_k = S_k(x_k)$ e $y_{k+1} = S_k(x_{k+1})$, obteniendo:

$$(7) \quad y_k = \frac{m_k}{6}h_k^2 + p_k h_k \quad \text{e} \quad y_{k+1} = \frac{m_{k+1}}{6}h_k^2 + q_k h_k.$$

Podemos despejar fácilmente p_k y q_k en estas ecuaciones y, al sustituir los valores obtenidos en la ecuación (6), lo que obtenemos es la siguiente expresión de la función cúbica $S_k(x)$:

$$(8) \quad S_k(x) = \frac{m_k}{6h_k}(x_{k+1} - x)^3 + \frac{m_{k+1}}{6h_k}(x - x_k)^3 + \left(\frac{y_k}{h_k} - \frac{m_k h_k}{6} \right) (x_{k+1} - x) + \left(\frac{y_{k+1}}{h_k} - \frac{m_{k+1} h_k}{6} \right) (x - x_k).$$

Lo que hemos conseguido en la representación (8) es reducir el problema de forma que ahora sólo nos falta determinar los coeficientes desconocidos $\{m_k\}$. Para hallar estos valores, hagamos notar que no hemos impuesto aún las condiciones sobre las derivadas primeras en los nodos. Derivamos, pues, en (8) y obtenemos:

$$(9) \quad S'_k(x) = -\frac{m_k}{2h_k}(x_{k+1} - x)^2 + \frac{m_{k+1}}{2h_k}(x - x_k)^2 - \left(\frac{y_k}{h_k} - \frac{m_k h_k}{6} \right) + \frac{y_{k+1}}{h_k} - \frac{m_{k+1} h_k}{6}.$$

Ahora evaluamos (9) en el punto x_k para hallar la derivada por la derecha de S en x_k y, simplificando el resultado, nos queda:

$$(10) \quad S'_k(x_k) = -\frac{m_k}{3}h_k - \frac{m_{k+1}}{6}h_k + d_k, \quad \text{siendo} \quad d_k = \frac{y_{k+1} - y_k}{h_k}.$$

De manera parecida, reemplazando k por $k - 1$ en (9) para obtener la expresión de $S'_{k-1}(x)$ y evaluando esta expresión en x_k , para hallar la derivada por la izquierda de S en x_k , obtenemos:

$$(11) \quad S'_{k-1}(x_k) = \frac{m_k}{3}h_{k-1} + \frac{m_{k-1}}{6}h_{k-1} + d_{k-1}.$$

Finalmente, usando las relaciones IV y las expresiones (10) y (11), obtenemos una relación entre m_{k-1} , m_k y m_{k+1} que es crucial en la construcción de la cercha:

$$(12) \quad h_{k-1}m_{k-1} + 2(h_{k-1} + h_k)m_k + h_km_{k+1} = u_k$$

siendo $u_k = 6(d_k - d_{k-1})$ para $k = 1, 2, \dots, N - 1$.

Construcción de las cerchas cúbicas interpoladoras

Observemos que las incógnitas de las ecuaciones de (12) son los valores deseados $\{m_k\}$ y que los demás términos se obtienen a partir de los datos $\{(x_k, y_k)\}$ mediante operaciones aritméticas muy simples. Por tanto, lo que aparece en (12) es, en realidad, un sistema de $N - 1$ ecuaciones lineales con $N + 1$ incógnitas y, en consecuencia, para que tenga solución única habría que añadirle dos ecuaciones adicionales, que llamamos restricciones en los extremos. Lo habitual es que estas dos ecuaciones sirvan para eliminar m_0 de la primera ecuación del sistema (12) y m_N de la última, la $(N - 1)$ -ésima. En la Tabla 5.8 se relacionan las estrategias más usuales para elegir las restricciones en los extremos.

Consideremos la estrategia (v) de la Tabla 5.8. Si fijamos el valor de m_0 , entonces podemos calcular h_0m_0 y escribir la primera ecuación (cuando $k = 1$) de (12) como

$$(13) \quad 2(h_0 + h_1)m_1 + h_1m_2 = u_1 - h_0m_0.$$

De manera parecida, si fijamos m_N , entonces podemos calcular $h_{N-1}m_N$ y la última ecuación (cuando $k = N - 1$) de (12) resulta

$$(14) \quad h_{N-2}m_{N-2} + 2(h_{N-2} + h_{N-1})m_{N-1} = u_{N-1} - h_{N-1}m_N.$$

Las ecuaciones (13) y (14) junto con las que aparecen en (12) para $k = 2, 3, \dots, N - 2$ forman un sistema lineal de $N - 1$ ecuaciones con $N - 1$ incógnitas: los coeficientes m_1, m_2, \dots, m_{N-1} .

Independientemente de la estrategia concreta que elijamos de la Tabla 5.8, podemos escribir las ecuaciones primera ($k = 1$) y última ($k = N - 1$) de (12) de manera que el sistema para las incógnitas m_1, m_2, \dots, m_{N-1} sea un sistema

Tabla 5.8 Restricciones en los extremos para cerchas cúbicas.

Descripción de la estrategia	Ecuaciones para m_0 y m_N
(i) <i>Cercha cónica sujeta:</i> se especifican $S'(x_0)$, $S'(x_n)$ (es la “mejor elección” si se conocen las derivadas)	$m_0 = \frac{3}{h_0}(d_0 - S'(x_0)) - \frac{m_1}{2}$, $m_N = \frac{3}{h_{N-1}}(S'(x_N) - d_{N-1}) - \frac{m_{N-1}}{2}$
(ii) <i>Cercha cónica natural</i> (una “curva relajada”)	$m_0 = 0, m_N = 0$
(iii) Se extrae $S''(x)$ a los extremos	$m_0 = m_1 - \frac{h_0(m_2 - m_1)}{h_1}$, $m_N = m_{N-1} + \frac{h_{N-1}(m_{N-1} - m_{N-2})}{h_{N-2}}$
(iv) $S''(x)$ es constante cerca de los extremos	$m_0 = m_1, m_N = m_{N-1}$
(v) Se especifica $S''(X)$ en cada extremo	$m_0 = S''(x_0), m_N = S''(x_N)$

lineal tridiagonal $\mathbf{HM} = \mathbf{V}$ dado por:

$$(15) \quad \begin{bmatrix} b_1 & c_1 & & & \\ a_1 & b_2 & c_2 & & \\ & \ddots & & & \\ & & a_{N-3} & b_{N-2} & c_{N-2} \\ & & & a_{N-2} & b_{N-1} \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_{N-2} \\ m_{N-1} \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{N-2} \\ v_{N-1} \end{bmatrix}.$$

El sistema dado en (15) es de diagonal estrictamente dominante, por lo que tiene solución única (véanse los detalles en el Capítulo 3). Una vez calculados los coeficientes $\{m_k\}$, los coeficientes de la cercha $\{s_{k,j}\}$ para el trozo $S_k(x)$ vienen dados por las fórmulas

$$(16) \quad \begin{aligned} s_{k,0} &= y_k, & s_{k,1} &= d_k - \frac{h_k(2m_k + m_{k+1})}{6}, \\ s_{k,2} &= \frac{m_k}{2}, & s_{k,3} &= \frac{m_{k+1} - m_k}{6h_k}. \end{aligned}$$

Podemos escribir cada polinomio cúbico $S_k(x)$ en forma de multiplicaciones encajadas para que la computación sea eficiente en términos del número de operaciones:

$$(17) \quad S_k(x) = ((s_{k,3}w + s_{k,2})w + s_{k,1})w + y_k, \quad \text{siendo } w = x - x_k,$$

recordando que $S_k(x)$ se usa cuando $x_k \leq x \leq x_{k+1}$.

En resumen: las ecuaciones de (12) junto con una estrategia de la Tabla 5.8 nos permiten construir una cercha cúbica que tiene propiedades específicas distintivas en los extremos del intervalo. Concretamente, los valores de m_0 y m_N que se dan en la Tabla 5.8 se utilizan para transformar la primera y la última ecuación de (12) y formar el sistema tridiagonal de $N - 1$ ecuaciones lineales que aparece desarrollado en (15). Luego se calculan los coeficientes m_1, m_2, \dots, m_{N-1} resolviendo este sistema tridiagonal y, finalmente, las fórmulas dadas en la relación (16) se usan para determinar los coeficientes de cada trozo cúbico de la cercha. Para que sirvan de referencia vamos a mostrar, a continuación, cómo deben prepararse las ecuaciones para cada tipo diferente de cercha.

Restricciones en los extremos

Los cinco lemas siguientes indican, para cada una de las restricciones en los extremos de la Tabla 5.8, cómo debe formarse el sistema lineal tridiagonal que hay que resolver para construir cada uno de los correspondientes tipos de cercha.

Lema 5.1 (Cercha sujeta). Existe una única cercha cúbica sujeta que verifica las condiciones sobre la derivada primera en la frontera dadas por $S'(a) = d_0$ y $S'(b) = d_N$.

Demostración. Basta resolver el sistema lineal tridiagonal

$$\left(\frac{3}{2}h_0 + 2h_1 \right) m_1 + h_1 m_2 = u_1 - 3(d_0 - S'(x_0))$$

$$h_{k-1}m_{k-1} + 2(h_{k-1} + h_k)m_k + h_k m_{k+1} = u_k \quad (k = 2, 3, \dots, N-2)$$

$$h_{N-2}m_{N-2} + \left(2h_{N-2} + \frac{3}{2}h_{N-1} \right) m_{N-1} = u_{N-1} - 3(S'(x_N) - d_{N-1}). \quad \bullet$$

Observación. La cercha sujeta involucra las derivadas en los extremos; esta cercha puede visualizarse como la curva que se obtiene cuando hacemos pasar un segmento flexible, que hemos sujetado por sus extremos con una inclinación determinada, por los puntos del conjunto de datos. Esta cercha sería útil para dibujar una curva suave que deba pasar por varios puntos dados de antemano.

Lema 5.2 (Cercha natural). Existe una única cercha cúbica natural verificando las condiciones de frontera libre dadas por $S''(a) = 0$ y $S''(b) = 0$.

Demostración. Basta resolver el sistema lineal tridiagonal

$$2(h_0 + h_1)m_1 + h_1 m_2 = u_1$$

$$h_{k-1}m_{k-1} + 2(h_{k-1} + h_k)m_k + h_k m_{k+1} = u_k \quad (k = 2, 3, \dots, N-2)$$

$$h_{N-2}m_{N-2} + 2(h_{N-2} + h_{N-1})m_{N-1} = u_{N-1}. \quad \bullet$$

Observación. La cercha natural es la curva que se obtendría al hacer pasar un segmento flexible a través de los puntos de un conjunto de datos, pero dejando libre la inclinación en los extremos para que adquiera la posición que minimiza la conducta osculatoria de la curva. Este tipo de cercha es útil para ajustar una curva a datos experimentales que tienen varias cifras significativas de precisión.

Lema 5.3 (Cercha extrapolada). Existe una única cercha cúbica que utiliza extrapolación desde los nodos x_1 y x_2 para determinar $S''(a)$ y extrapolación desde los nodos x_{N-1} y x_{N-2} para determinar $S''(b)$.

Demostración. Basta resolver el sistema lineal tridiagonal

$$\begin{aligned} \left(3h_0 + 2h_1 + \frac{h_0^2}{h_1}\right)m_1 + \left(h_1 - \frac{h_0^2}{h_1}\right)m_2 &= u_1 \\ h_{k-1}m_{k-1} + 2(h_{k-1} + h_k)m_k + h_km_{k+1} &= u_k \quad (k = 2, 3, \dots, N-2) \\ \left(h_{N-2} - \frac{h_{N-1}^2}{h_{N-2}}\right)m_{N-2} + \left(2h_{N-2} + 3h_{N-1} + \frac{h_{N-1}^2}{h_{N-2}}\right)m_{N-1} &= u_{N-1}. \end{aligned}$$

Observación. La cercha extrapolada se obtiene al exigir que las cúbicas del primer y del último intervalo sean extensiones de las adyacentes; es decir, la cercha forma una sola cónica en el intervalo $[x_0, x_2]$ y una sola cónica en el intervalo $[x_{N-2}, x_N]$.

Lema 5.4 (Cercha con terminación parabólica). Existe una única cercha cónica tal que $S'''(x) \equiv 0$ en el intervalo $[x_0, x_1]$ y $S'''(x) \equiv 0$ en el intervalo $[x_{N-1}, x_N]$.

Demostración. Basta resolver el sistema lineal tridiagonal

$$\begin{aligned} (3h_0 + 2h_1)m_1 + h_1m_2 &= u_1 \\ h_{k-1}m_{k-1} + 2(h_{k-1} + h_k)m_k + h_km_{k+1} &= u_k \quad (k = 2, 3, \dots, N-2) \\ h_{N-2}m_{N-2} + (2h_{N-2} + 3h_{N-1})m_{N-1} &= u_{N-1}. \end{aligned}$$

Observación. La hipótesis de que $S'''(x) \equiv 0$ en el intervalo $[x_0, x_1]$ impone que la cónica degenera en una parábola en dicho intervalo; y lo mismo ocurre en $[x_{N-1}, x_N]$.

Lema 5.5 (Cercha con curvatura dada en los extremos). Existe una única cercha cónica verificando las condiciones sobre la derivada segunda en la frontera dadas al especificar los valores de $S''(a)$ y $S''(b)$.

Demostración. Basta resolver el sistema lineal tridiagonal

$$\begin{aligned} 2(h_0 + h_1)m_1 + h_1m_2 &= u_1 - h_0S''(x_0) \\ h_{k-1}m_{k-1} + 2(h_{k-1} + h_k)m_k + h_km_{k+1} &= u_k \quad (k = 2, 3, \dots, N-2) \\ h_{N-2}m_{N-2} + 2(h_{N-2} + h_{N-1})m_{N-1} &= u_{N-1} - h_{N-1}S''(x_N). \end{aligned}$$

•

Observación. Al imponer los valores $S''(a)$ y $S''(b)$, podemos ajustar la curvatura en cada extremo.

Los cinco ejemplos que siguen ilustran la conducta de estas cerchas. Es posible mezclar las condiciones en los extremos para obtener una variedad aún más amplia de cerchas que se pueden construir, dejamos estas variaciones como motivo de investigación.

Ejemplo 5.7. Vamos a determinar la cercha sujeta que pasa por los puntos $(0, 0)$, $(1, 0.5)$, $(2, 2.0)$ y $(3, 1.5)$ y que verifica las condiciones sobre la primera derivada en la frontera dadas por $S'(0) = 0.2$ y $S'(3) = -1$.

Primero hay que calcular las siguientes cantidades

$$\begin{aligned} h_0 &= h_1 = h_2 = 1 \\ d_0 &= (y_1 - y_0)/h_0 = (0.5 - 0.0)/1 = 0.5 \\ d_1 &= (y_2 - y_1)/h_1 = (2.0 - 0.5)/1 = 1.5 \\ d_2 &= (y_3 - y_2)/h_2 = (1.5 - 2.0)/1 = -0.5 \\ u_1 &= 6(d_1 - d_0) = 6(1.5 - 0.5) = 6.0 \\ u_2 &= 6(d_2 - d_1) = 6(-0.5 - 1.5) = -12.0. \end{aligned}$$

Ahora usamos el Lema 5.1 y obtenemos el sistema

$$\begin{aligned} \left(\frac{3}{2} + 2\right)m_1 + m_2 &= 6.0 - 3(0.5 - 0.2) = 5.1, \\ m_1 + \left(2 + \frac{3}{2}\right)m_2 &= -12.0 - 3(-1.0 - (-0.5)) = -10.5. \end{aligned}$$

Simplificando estas ecuaciones y escribiéndolas en forma matricial, tenemos

$$\begin{bmatrix} 3.5 & 1.0 \\ 1.0 & 3.5 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} 5.1 \\ -10.5 \end{bmatrix}.$$

La solución, que se calcula fácilmente, es $m_1 = 2.25$ y $m_2 = -3.72$. Ahora aplicamos las ecuaciones de la fila (i) de la Tabla 5.8 para determinar los coeficientes m_0 y m_3 :

$$\begin{aligned} m_0 &= 3(0.5 - 0.2) - \frac{2.52}{2} = -0.36, \\ m_3 &= 3(-1.0 + 0.5) - \frac{-3.72}{2} = 0.36. \end{aligned}$$

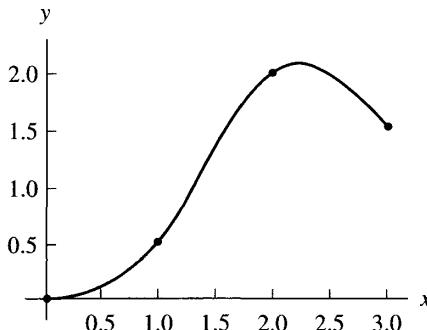


Figura 5.12 La cercha cúbica sujeta con las condiciones sobre la derivada: \$S'(0) = 0.2\$ y \$S'(3) = -1\$.

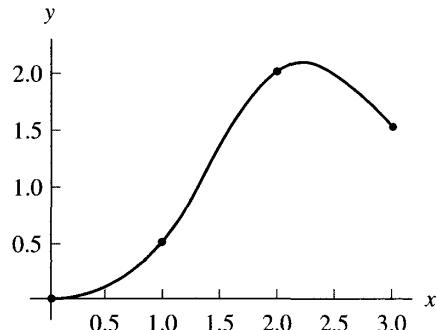


Figura 5.13 La cercha cúbica natural con \$S''(0) = 0\$ y \$S''(3) = 0\$.

Finalmente, los valores \$m_0 = -0.36\$, \$m_1 = 2.25\$, \$m_2 = -3.72\$ y \$m_3 = 0.36\$ se sustituyen en la relación (16) para determinar los coeficientes de la cercha, que es:

$$(18) \quad \begin{aligned} S_0(x) &= 0.48x^3 - 0.18x^2 + 0.2x && \text{para } 0 \leq x \leq 1, \\ S_1(x) &= -1.04(x-1)^3 + 1.26(x-1)^2 \\ &\quad + 1.28(x-1) + 0.5 && \text{para } 1 \leq x \leq 2, \\ S_2(x) &= 0.68(x-2)^3 - 1.86(x-2)^2 \\ &\quad + 0.68(x-2) + 2.0 && \text{para } 2 \leq x \leq 3. \end{aligned}$$

Esta cercha cúbica sujeta se muestra en la Figura 5.12.

Ejemplo 5.8. Vamos a determinar la cercha cúbica natural que pasa por los puntos \$(0,0.0)\$, \$(1,0.5)\$, \$(2,2.0)\$ y \$(3,1.5)\$ y verifica las condiciones de frontera libre \$S''(x) = 0\$ y \$S''(3) = 0\$.

Los valores \$\{h_k\}\$, \$\{d_k\}\$ y \$\{u_k\}\$ son los mismos que los calculados en el Ejemplo 5.7. Ahora usamos el Lema 5.2 para obtener las ecuaciones

$$\begin{aligned} 2(1+1)m_1 + m_2 &= 6.0, \\ m_1 + 2(1+1)m_2 &= -12.0, \end{aligned}$$

que escribimos de forma matricial como

$$\begin{bmatrix} 4.0 & 1.0 \\ 1.0 & 4.0 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} 6.0 \\ -12.0 \end{bmatrix}.$$

La solución es fácil de encontrar: \$m_1 = 2.4\$ y \$m_2 = -3.6\$. Puesto que \$m_0 = S''(0) = 0\$ y \$m_3 = S''(3) = 0\$, al usar la relación (16) para hallar los coeficientes de la cercha,

el resultado es

$$(19) \quad \begin{aligned} S_0(x) &= 0.4x^3 + 0.1x && \text{para } 0 \leq x \leq 1, \\ S_1(x) &= -(x-1)^3 + 1.2(x-1)^2 \\ &\quad + 1.3(x-1) + 0.5 && \text{para } 1 \leq x \leq 2, \\ S_2(x) &= 0.6(x-2)^3 - 1.8(x-2)^2 \\ &\quad + 0.7(x-2) + 2.0 && \text{para } 2 \leq x \leq 3. \end{aligned}$$

Esta cercha cúbica natural se muestra en la Figura 5.13.

Ejemplo 5.9. Vamos a determinar la cercha cúbica extrapolada que pasa por los puntos $(0, 0.0)$, $(1, 0.5)$, $(2, 2.0)$ y $(3, 1.5)$.

Usando los valores $\{h_k\}$, $\{d_k\}$ y $\{u_k\}$ ya calculados en el Ejemplo 5.7 junto con el Lema 5.3, obtenemos el sistema lineal

$$\begin{aligned} (3+2+1)m_1 + (1-1)m_2 &= 6.0, \\ (1-1)m_1 + (2+3+1)m_2 &= -12.0, \end{aligned}$$

o, en forma matricial,

$$\begin{bmatrix} 6.0 & 0.0 \\ 0.0 & 6.0 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} 6.0 \\ -12.0 \end{bmatrix},$$

cuya solución $m_1 = 1.0$ y $m_2 = -2.0$ se obtiene trivialmente. Ahora aplicamos las ecuaciones de la fila (iii) de la Tabla 5.8 para calcular m_0 y m_3 :

$$\begin{aligned} m_0 &= 1.0 - (-2.0 - 1.0) = 4.0, \\ m_3 &= -2.0 + (-2.0 - 1.0) = -5.0. \end{aligned}$$

Finalmente, los valores de $\{m_k\}$ se emplean en la relación (16) para hallar los coeficientes de la cercha, que resulta ser

$$(20) \quad \begin{aligned} S_0(x) &= -0.5x^3 + 2.0x^2 - x && \text{para } 0 \leq x \leq 1, \\ S_1(x) &= -0.5(x-1)^3 + 0.5(x-1)^2 \\ &\quad + 1.5(x-1) + 0.5 && \text{para } 1 \leq x \leq 2, \\ S_2(x) &= -0.5(x-2)^3 - (x-2)^2 \\ &\quad + (x-2) + 2.0 && \text{para } 2 \leq x \leq 3. \end{aligned}$$

La cercha cúbica extrapolada se muestra en la Figura 5.14.

Ejemplo 5.10. Vamos a determinar la cercha con terminación parabólica que pasa por los puntos $(0, 0.0)$, $(1, 0.5)$, $(2, 2.0)$ y $(3, 1.5)$.

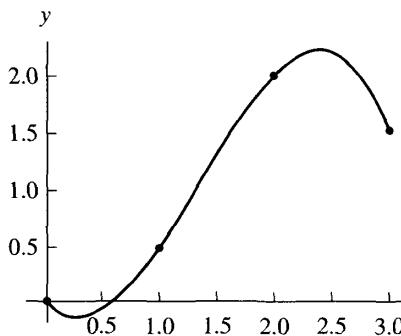


Figura 5.14 La cercha cúbica extrapolada.

Usando los valores $\{h_k\}$, $\{d_k\}$ y $\{u_k\}$ calculados en el Ejemplo 5.7 y aplicando el Lema 5.4, obtenemos

$$\begin{aligned} (3+2)m_1 + m_2 &= 6.0, \\ m_1 + (2+3)m_2 &= -12.0, \end{aligned}$$

cuya forma matricial es

$$\begin{bmatrix} 5.0 & 1.0 \\ 1.0 & 5.0 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} 6.0 \\ -12.0 \end{bmatrix}.$$

La solución ahora es $m_1 = 1.75$ y $m_2 = -2.75$. Puesto que $S'''(x) \equiv 0$ en los subintervalos primero y último, las fórmulas de la fila (iv) de la Tabla 5.8 nos dicen que $m_0 = m_1 = 1.75$ y $m_3 = m_2 = -2.75$. Sustituimos los valores de $\{m_k\}$ en la relación (16) y obtenemos la cercha

$$(21) \quad \begin{aligned} S_0(x) &= 0.875x^2 - 0.375x && \text{para } 0 \leq x \leq 1, \\ S_1(x) &= -0.75(x-1)^3 + 0.875(x-1)^2 \\ &\quad + 1.375(x-1) + 0.5 && \text{para } 1 \leq x \leq 2, \\ S_2(x) &= -1.375(x-2)^2 + 0.875(x-2) + 2.0 && \text{para } 2 \leq x \leq 3. \end{aligned}$$

Esta cercha con terminación parabólica se muestra en la Figura 5.15. ■

Ejemplo 5.11. Vamos a determinar la cercha cúbica con curvatura dada en los extremos que pasa por los puntos $(0, 0.0)$, $(1, 0.5)$, $(2, 2.0)$ y $(3, 1.5)$ y verifica las restricciones en los extremos $S''(0) = -0.3$ y $S''(3) = 3.3$.

Usando de nuevo los valores $\{h_k\}$, $\{d_k\}$ y $\{u_k\}$ ya calculados en el Ejemplo 5.7 y aplicando luego el Lema 5.5, obtenemos

$$\begin{aligned} 2(1+1)m_1 + m_2 &= 6.0 - (-0.3) = 6.3, \\ m_1 + 2(1+1)m_2 &= -12.0 - (3.3) = -15.3. \end{aligned}$$

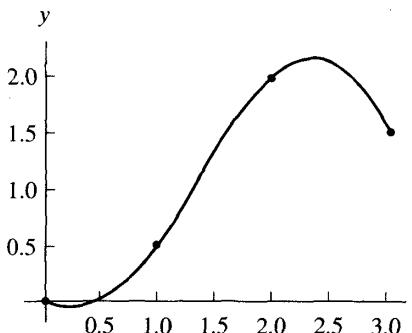


Figura 5.15 La cercha cúbica con terminación parabólica.

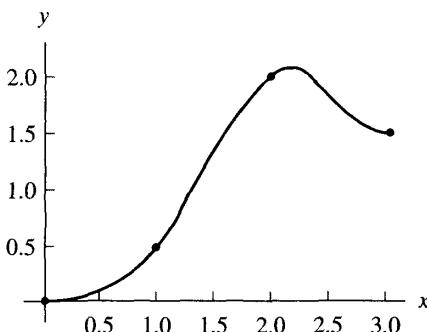


Figura 5.16 La cercha cúbica con curvatura dada en los extremos por $S''(0) = -0.3$ y $S''(3) = 3.3$.

La forma matricial del sistema es ahora

$$\begin{bmatrix} 4.0 & 1.0 \\ 1.0 & 4.0 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} 6.3 \\ -15.3 \end{bmatrix},$$

cuya solución es $m_1 = 2.7$ y $m_2 = -4.5$. Las restricciones en los extremos nos proporcionan $m_0 = S''(0) = -0.3$ y $m_3 = S''(3) = 3.3$. Sustituyendo los valores de $\{m_k\}$ en la relación (16) obtenemos la solución:

$$(22) \quad \begin{aligned} S_0(x) &= 0.5x^3 - 0.15x^2 + 0.15x && \text{para } 0 \leq x \leq 1, \\ S_1(x) &= -1.2(x-1)^3 + 1.35(x-1)^2 \\ &\quad + 1.35(x-1) + 0.5 && \text{para } 1 \leq x \leq 2, \\ S_2(x) &= 1.3(x-2)^3 - 2.25(x-2)^2 \\ &\quad + 0.45(x-2) + 2.0 && \text{para } 2 \leq x \leq 3. \end{aligned}$$

Esta cercha cúbica con curvatura dada en los extremos se muestra en la Figura 5.16.

La propiedad de oscilación mínima

Una propiedad de las cerchas que es interesante desde el punto de vista de su utilización en la práctica es que presentan una conducta oscilatoria mínima; es decir, entre todas las funciones que son dos veces derivables y con derivada segunda continua en el intervalo $[a, b]$ y que, además, interpolan un conjunto de datos $\{(x_k, y_k)\}_{k=0}^N$, las cerchas son las que oscilan menos. El siguiente resultado explica este fenómeno.

Teorema 5.4 (Propiedad de mínima oscilación de las cerchas). Supongamos que $f \in C^2[a, b]$ y que $S(x)$ es la única cercha cúbica sujeta que interpola $f(x)$ pasando por los puntos $\{(x_k, f(x_k))\}_{k=0}^N$ y verifica las restricciones en los extremos $S'(a) = f'(a)$ y $S'(b) = f'(b)$. Entonces

$$(23) \quad \int_a^b (S''(x))^2 dx \leq \int_a^b (f''(x))^2 dx.$$

Demostración. Usando la fórmula de integración por partes y teniendo en cuenta las condiciones en los extremos, se deduce

$$\begin{aligned} & \int_a^b S''(x)(f''(x) - S''(x)) dx \\ &= S''(x)(f'(x) - S'(x)) \Big|_{x=a}^{x=b} - \int_a^b S'''(x)(f'(x) - S'(x)) dx \\ &= 0 - 0 - \int_a^b S'''(x)(f'(x) - S'(x)) dx. \end{aligned}$$

Como $S'''(x) = 6s_{k,3}$ en el subintervalo $[x_k, x_{k+1}]$, entonces

$$\int_{x_k}^{x_{k+1}} S'''(x)(f'(x) - S'(x)) dx = 6s_{k,3}(f(x) - S(x)) \Big|_{x=x_k}^{x=x_{k+1}} = 0$$

para $k = 0, 1, \dots, N - 1$. Por tanto, $\int_a^b S''(x)(f''(x) - S''(x)) dx = 0$, y de aquí deducimos que

$$(24) \quad \int_a^b S''(x)f''(x) dx = \int_a^b (S''(x))^2 dx.$$

Puesto que $0 \leq (f''(x) - S''(x))^2$, obtenemos la siguiente relación entre las integrales:

$$\begin{aligned} (25) \quad 0 &\leq \int_a^b (f''(x) - S''(x))^2 dx \\ &= \int_a^b (f''(x))^2 dx - 2 \int_a^b f''(x)S''(x) dx + \int_a^b (S''(x))^2 dx. \end{aligned}$$

Sustituyendo, finalmente, el resultado de (24) en (25), obtenemos

$$0 \leq \int_a^b (f''(x))^2 dx - \int_a^b (S''(x))^2 dx,$$

que es equivalente a la relación (23) y prueba el teorema. •

El programa que damos a continuación sirve para construir la cercha cúbica sujeta que interpola los datos $\{(x_k, y_k)\}_{k=0}^N$. Los coeficientes, en orden descendiente, de la cúbica $S_k(x)$ (para $k = 0, 1, \dots, N - 1$) se almacenan en la fila $(k - 1)$ -ésima de la matriz S que se obtiene. En los ejercicios se plantea que se lleven a cabo las modificaciones oportunas para obtener las cerchas cúbicas correspondientes a las otras restricciones en los extremos dadas en la Tabla 5.8 y descritas en los Lemas 5.2 a 5.5.

Programa 5.3 (Cercha cúbica sujeta). Construcción y determinación de la cercha cúbica sujeta $S(x)$ que interpola los $N + 1$ datos $\{(x_k, y_k)\}_{k=0}^N$.

```

function S=csfit(X,Y,dx0,dxn)
% Datos
%     - X es un vector 1 x n que contiene las abscisas
%     - Y es un vector 1 x n que contiene las ordenadas
%     - dx0 = S'(x0) es la derivada en el primer extremo
%     - dxn = S'(xn) es la derivada en el segundo extremo
% Resultado
%     - S: las filas de S son los coeficientes, en orden
%       descendiente de cada cúbica de la cercha
N=length(X)-1;
H=diff(X);
D=diff(Y)./H;
A=H(2:N-1);
B=2*(H(1:N-1)+H(2:N));
C=H(2:N);
U=6*diff(D);
% Restricciones en los extremos para la cercha sujeta
B(1)=B(1)-H(1)/2;
U(1)=U(1)-3*(D(1)-dx0);
B(N-1)=B(N-1)-H(N)/2;
U(N-1)=U(N-1)-3*(dxn-D(N));
for k=2:N-1
    temp=A(k-1)/B(k-1);
    B(k)=B(k)-temp*C(k-1);
    U(k)=U(k)-temp*U(k-1);
end

```

```

end
M(N)=U(N-1)/B(N-1);
for k=N-2:-1:1
    M(k+1)=(U(k)-C(k)*M(k+2))/B(k);
end
M(1)=3*(D(1)-dx0)/H(1)-M(2)/2;
M(N+1)=3*(dxn-D(N))/H(N)-M(N)/2;
for k=0:N-1
    S(k+1,1)=(M(k+2)-M(k+1))/(6*H(k+1));
    S(k+1,2)=M(k+1)/2;
    S(k+1,3)=D(k+1)-H(k+1)*(2*M(k+1)+M(k+2))/6;
    S(k+1,4)=Y(k+1);
end

```

Ejemplo 5.12. Vamos a determinar la cercha cúbica sujeta que pasa por los puntos $(0, 0.0)$, $(1, 0.5)$, $(2, 2.0)$ y $(3, 1.5)$ y verifica las condiciones sobre la primera derivada en los extremos dadas por $S'(0) = 0.2$ y $S'(3) = -1$.

Usando el paquete de programas MATLAB:

```

>>X=[0 1 2 3]; Y=[0 0.5 2.0 1.5]; dx0=0.2; dxn=-1;
>>S=csfit(X,Y,dx0,dxn)
S =
    0.4800 -0.1800 0.2000 0
   -1.0400  1.2600 1.2800 0.5000
    0.6800 -1.8600 0.6800 2.0000

```

Hagamos notar que las filas de la matriz S son, precisamente, los coeficientes de la cercha cúbica sujeta que obtuvimos en la relación (18) del Ejemplo 5.7. Las siguientes instrucciones muestran cómo podemos dibujar la cercha cúbica interpolante usando la instrucción `polyval` de MATLAB. La gráfica que se obtiene es la misma que se muestra en la Figura 5.12.

```

>>x1=0:.01:1; y1=polyval(S(1,:),x1-X(1));
>>x2=1:.01:2; y2=polyval(S(2,:),x2-X(2));
>>x3=2:.01:3; y3=polyval(S(3,:),x3-X(3));
>>plot(x1,y1,x2,y2,x3,y3,X,Y,'.')

```

Ejercicios

- Consideremos el polinomio $S(x) = a_0 + a_1x + a_2x^2 + a_3x^3$.
 - Pruebe que las condiciones $S(1) = 1$, $S'(1) = 0$, $S(2) = 2$ y $S'(2) = 0$

producen el sistema de ecuaciones:

$$\begin{aligned} a_0 + a_1 + a_2 + a_3 &= 1 \\ a_1 + 2a_2 + 3a_3 &= 0 \\ a_0 + 2a_1 + 4a_2 + 8a_3 &= 2 \\ a_1 + 4a_2 + 12a_3 &= 0 \end{aligned}$$

- (b) Resuelva el sistema del apartado (a) y dibuje la cúbica obtenida.
2. Consideremos el polinomio $S(x) = a_0 + a_1x + a_2x^2 + a_3x^3$.
- (a) Pruebe que las condiciones $S(1) = 3$, $S'(1) = -4$, $S(2) = 1$ y $S'(2) = 2$ producen el sistema de ecuaciones
- $$\begin{aligned} a_0 + a_1 + a_2 + a_3 &= 3 \\ a_1 + 2a_2 + 3a_3 &= -4 \\ a_0 + 2a_1 + 4a_2 + 8a_3 &= 1 \\ a_1 + 4a_2 + 12a_3 &= 2 \end{aligned}$$
- (b) Resuelva el sistema del apartado (a) y dibuje la cúbica obtenida.
3. Determine cuál de las siguientes funciones es una cercha cúbica. *Indicación.* ¿Alguna de las cinco partes de la Definición 5.1 no se verifica?
- (a) $f(x) = \begin{cases} \frac{19}{2} - \frac{81}{4}x + 15x^2 - \frac{13}{4}x^3 & \text{para } 1 \leq x \leq 2 \\ \frac{-77}{2} + \frac{207}{4}x - 21x^2 + \frac{11}{4}x^3 & \text{para } 2 \leq x \leq 3 \end{cases}$
- (b) $f(x) = \begin{cases} 11 - 24x + 18x^2 - 4x^3 & \text{para } 1 \leq x \leq 2 \\ -54 + 72x - 30x^2 + 4x^3 & \text{para } 2 \leq x \leq 3 \end{cases}$
- (c) $f(x) = \begin{cases} 18 - \frac{75}{2}x + 26x^2 - \frac{11}{2}x^3 & \text{para } 1 \leq x \leq 2 \\ -70 + \frac{189}{2}x - 40x^2 + \frac{11}{2}x^3 & \text{para } 2 \leq x \leq 3 \end{cases}$
- (d) $f(x) = \begin{cases} 13 - 31x + 23x^2 - 5x^3 & \text{para } 1 \leq x \leq 2 \\ -35 + 51x - 22x^2 + 3x^3 & \text{para } 2 \leq x \leq 3 \end{cases}$
4. Determine la cercha cúbica sujeta que pasa por los puntos $(-3, 2)$, $(-2, 0)$, $(1, 3)$ y $(4, 1)$ y verifica las condiciones sobre la derivada primera en los extremos dadas por $S'(-3) = -1$ y $S'(4) = 1$.
5. Halle la cercha cúbica natural que pasa por los puntos $(-3, 2)$, $(-2, 0)$, $(1, 3)$ y $(4, 1)$ y verifica las condiciones de frontera libre $S''(-3) = 0$ y $S''(4) = 0$.
6. Determine la cercha cúbica extrapolada que pasa por los puntos $(-3, 2)$, $(-2, 0)$, $(1, 3)$ y $(4, 1)$.
7. Determine la cercha cúbica con terminación parabólica que pasa por los puntos $(-3, 2)$, $(-2, 0)$, $(1, 3)$ y $(4, 1)$.

8. Determine la cercha cúbica con curvatura dada en los extremos que pasa por los puntos $(-3, 2)$, $(-2, 0)$, $(1, 3)$ y $(4, 1)$ y verifica las condiciones sobre la derivada segunda en los extremos dadas por $S''(-3) = -1$ y $S''(4) = 2$.

9. (a) Halle la cercha cúbica sujeta que pasa por los puntos $\{(x_k, f(x_k))\}_{k=0}^3$, que están en la gráfica de $f(x) = x + \frac{2}{x}$, usando los nodos $x_0 = 1/2$, $x_1 = 1$, $x_2 = 3/2$ y $x_3 = 2$. Utilice las condiciones sobre la primera derivada en los extremos dadas por $S'(x_0) = f'(x_0)$ y $S'(x_3) = f'(x_3)$. Dibuje f y la cercha cúbica sujeta interpoladora en un mismo gráfico.
(b) Halle la cercha cúbica natural que pasa por los puntos $\{(x_k, f(x_k))\}_{k=0}^3$, que están en la gráfica de $f(x) = x + \frac{2}{x}$, usando los nodos $x_0 = 1/2$, $x_1 = 1$, $x_2 = 3/2$ y $x_3 = 2$. Utilice las condiciones de frontera libre $S''(x_0) = 0$ y $S''(x_3) = 0$. Dibuje f y la cercha cúbica natural interpoladora en un mismo gráfico.

10. (a) Halle la cercha cúbica sujeta que pasa por los puntos $\{(x_k, f(x_k))\}_{k=0}^3$, que están en la gráfica de $f(x) = \cos(x^2)$, usando los nodos $x_0 = 0$, $x_1 = \sqrt{\pi/2}$, $x_2 = \sqrt{3\pi/2}$ y $x_3 = \sqrt{5\pi/2}$. Utilice las condiciones sobre la derivada en los extremos dadas por $S'(x_0) = f'(x_0)$ y $S'(x_3) = f'(x_3)$. Dibuje f y la cercha cúbica sujeta interpoladora en un mismo gráfico.
(b) Halle la cercha cúbica natural que pasa por los puntos $\{(x_k, f(x_k))\}_{k=0}^3$, que están en la gráfica de $f(x) = \cos(x^2)$, usando los nodos $x_0 = 0$, $x_1 = \sqrt{\pi/2}$, $x_2 = \sqrt{3\pi/2}$ y $x_3 = \sqrt{5\pi/2}$. Utilice las condiciones de frontera libre $S''(x_0) = 0$ y $S''(x_3) = 0$. Dibuje f y la cercha cúbica natural interpoladora en un mismo gráfico.

11. Utilice las sustituciones

$$x_{k+1} - x = h_k + (x_k - x)$$

y

$$(x_{k+1} - x)^3 = h_k^3 + 3h_k^2(x_k - x) + 3h_k(x_k - x)^2 + (x_k - x)^3$$

para probar que cuando la relación (8) se desarrolla en potencias de $(x_k - x)$, entonces sus coeficientes son los dados en la relación (16).

12. Consideremos cada cúbica $S_k(x)$ en el intervalo $[x_k, x_{k+1}]$.

(a) Proporcione una fórmula para $\int_{x_k}^{x_{k+1}} S_k(x) dx$.
Luego evalúe $\int_{x_0}^{x_3} S(x) dx$ para la cercha obtenida en el apartado (a) del Ejercicio 10

(c) Ejercicio 11

13. Muestre cómo la estrategia (i) de la Tabla 5.8 y el sistema (12) se combinan para obtener las ecuaciones dadas en el Lema 5.1.

14. Muestre cómo la estrategia (iii) de la Tabla 5.8 y el sistema (12) se combinan para obtener las ecuaciones dadas en el Lema 5.3.

- 15.** (a) Usando los nodos $x_0 = -2$ y $x_1 = 0$, pruebe que $f(x) = x^3 - x$ es su propia cercha cúbica sujeta en el intervalo $[-2, 0]$.
- (b) Usando los nodos $x_0 = -2$, $x_1 = 0$ y $x_2 = 2$, pruebe que $f(x) = x^3 - x$ es su propia cercha cúbica sujeta en el intervalo $[-2, 2]$. Nota. f tiene un punto de inflexión en x_1 .
- (c) Use los resultados de los apartados (a) y (b) para probar que cualquier polinomio de grado tres $f(x) = a_0 + a_1x + a_2x^2 + a_3x^3$, es su propia cercha cúbica sujeta en cualquier intervalo cerrado $[a, b]$.
- (d) ¿Qué puede decirse, si es que puede decirse algo, de las otras cuatro clases de cerchas cúbicas descritas en los Lemas 5.2 a 5.5?

Algoritmos y programas

1. Las distancias recorridas d_k por un coche en los instantes t_k se dan en la siguiente tabla. Use el Programa 5.3 con las condiciones sobre la primera derivada dadas por $S'(0) = 0$ y $S'(8) = 98$ para determinar la cercha cúbica sujeta que interpola estos puntos.

Tiempo, t_k	0	2	4	6	8
Distancia, d_k	0	40	160	300	480

2. Modifique el Programa 5.3 para hallar las cerchas cúbicas **(a)** natural, **(b)** extrapolada, **(c)** con terminación parabólica y **(d)** de curvatura dada en los extremos, correspondientes a un conjunto dado de puntos.
3. Use sus programas del Problema 2 para hallar los cinco tipos distintos de cerchas cúbicas que interpolan los puntos $(0, 1)$, $(1, 0)$, $(2, 0)$, $(3, 1)$, $(4, 2)$, $(5, 2)$ y $(6, 1)$, siendo $S'(0) = -0.6$ y $S'(6) = -1.8$ en el caso de la cercha sujeta y $S''(0) = 1$ y $S''(6) = -1$ en el caso de la cercha con curvatura dada en los extremos. Dibuje las cinco curvas y los datos sobre un mismo gráfico.
4. Use sus programas del Problema 2 para hallar los cinco tipos distintos de cerchas cúbicas que interpolan los puntos $(0, 0)$, $(1, 4)$, $(2, 8)$, $(3, 9)$, $(4, 9)$, $(5, 8)$ y $(6, 6)$, siendo $S'(0) = 1$ y $S'(6) = -2$ en el caso de la cercha sujeta y $S''(0) = 1$ y $S''(6) = -1$ en el caso de la cercha con curvatura dada en los extremos. Dibuje las cinco curvas y los datos sobre un mismo gráfico.
5. Considere la tabla de temperaturas dada en el Problema 1 de la subsección “Algoritmos y programas” de la Sección 3.4. Determine la cercha cúbica natural que interpola los datos de dicha tabla, dibuje esta cercha y los datos sobre un mismo gráfico y utilice dicha cercha y los resultados del apartado (a) del Ejercicio 12 para aproximar la temperatura media durante el día que se tomaron dichos datos.

6. Aproxime la gráfica de la función $f(x) = x - \cos(x^3)$ sobre el intervalo $[-3, 3]$ usando una cerca cúbica sujeta.

5.4 Series de Fourier y polinomios trigonométricos

En ciencia y en ingeniería se estudian a menudo fenómenos físicos, como la luz y el sonido, que tienen un carácter periódico. Estos fenómenos se describen mediante funciones $g(x)$ que son periódicas:

$$(1) \quad g(x + P) = g(x) \quad \text{para todo } x,$$

en cuyo caso se dice que P es un **período** de la función.

Será suficiente con que consideremos funciones de período 2π , ya que si $g(x)$ tiene período P , entonces $f(x) = g(Px/2\pi)$ es periódica y tiene período 2π . Para verificar esto, basta observar que

$$(2) \quad f(x + 2\pi) = g\left(\frac{Px}{2\pi} + P\right) = g\left(\frac{Px}{2\pi}\right) = f(x).$$

En consecuencia, a lo largo de toda esta sección supondremos que $f(x)$ es una función periódica con período 2π , esto es,

$$(3) \quad f(x + 2\pi) = f(x) \quad \text{para todo } x.$$

Observemos que la gráfica de $y = f(x)$ se obtiene repitiendo la porción de dicha gráfica que corresponde a cualquier intervalo de longitud 2π , como se muestra en la Figura 5.17.

Ejemplos de funciones con período 2π son $\sin(jx)$ y $\cos(jx)$, siendo j un número entero. Esto plantea la siguiente pregunta: ¿podemos representar una función periódica como una suma de términos del tipo $a_j \cos(jx)$ y $b_j \sin(jx)$? Veremos enseguida que la respuesta es, en general, afirmativa.

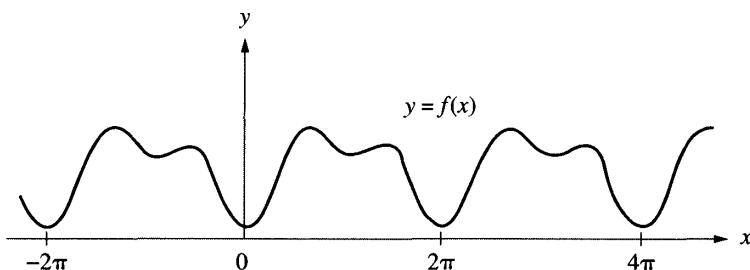


Figura 5.17 Una función continua y periódica $f(x)$ con período 2π .

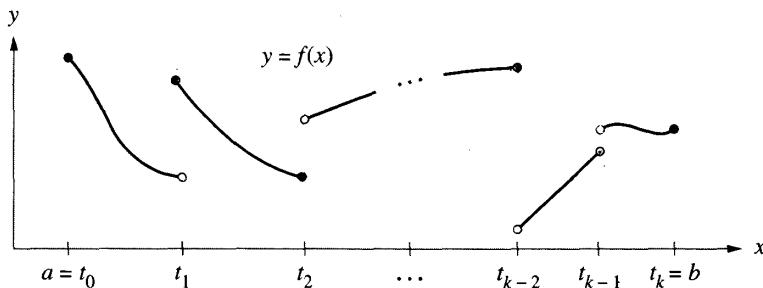


Figura 5.18 Una función continua a trozos en $[a, b]$.

Definición 5.2 (Continuidad a trozos). Diremos que una función $f(x)$ es **continua a trozos** en $[a, b]$ si existe una partición t_0, t_1, \dots, t_K de $[a, b]$, siendo $a = t_0 < t_1 < \dots < t_K = b$, tal que $f(x)$ es continua en cada intervalo abierto $t_{i-1} < x < t_i$ para $i = 1, 2, \dots, K$ y, además, existen los límites laterales de $f(x)$ por la izquierda y por la derecha en cada uno de los puntos t_i ; esta situación se ilustra en la Figura 5.18.

▲

Definición 5.3 (Series de Fourier). Supongamos que $f(x)$ es periódica con período 2π y que es continua a trozos en $[-\pi, \pi]$. La **serie de Fourier** $S(x)$ de $f(x)$ es

$$(4) \quad S(x) = \frac{a_0}{2} + \sum_{j=1}^{\infty} (a_j \cos(jx) + b_j \sin(jx)),$$

donde los coeficientes a_j y b_j vienen dados por las fórmulas de Euler

$$(5) \quad a_j = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(jx) dx \quad \text{para } j = 0, 1, \dots$$

y

$$(6) \quad b_j = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(jx) dx \quad \text{para } j = 1, 2, \dots$$

▲

El factor $\frac{1}{2}$ que aparece en el término constante $a_0/2$ de la serie de Fourier (4) se ha introducido por comodidad: así podemos obtener a_0 usando la fórmula general (5) para $j = 0$. La convergencia, bajo ciertas condiciones, de la serie de Fourier se establece en el siguiente resultado.

Teorema 5.5 (Desarrollo en serie de Fourier). Supongamos que $S(x)$ es la serie de Fourier de $f(x)$ en $[-\pi, \pi]$. Si su derivada $f'(x)$ es continua a trozos en $[-\pi, \pi]$ y tiene derivadas laterales por la izquierda y por la derecha en cada punto de dicho intervalo, entonces $S(x)$ converge para todo $x \in [-\pi, \pi]$ y la relación

$$S(x) = f(x)$$

se verifica en todos los puntos $x \in [-\pi, \pi]$ en los que f es continua. Además, si $x = a$ es un punto de discontinuidad de f , entonces

$$S(a) = \frac{f(a^-) + f(a^+)}{2},$$

donde $f(a^-)$ y $f(a^+)$ denotan los límites laterales por la izquierda y por la derecha, respectivamente. En estas condiciones y con el significado que acabamos de dar, obtenemos el desarrollo en serie de Fourier

$$(7) \quad f(x) = \frac{a_0}{2} + \sum_{j=1}^{\infty} (a_j \cos(jx) + b_j \operatorname{sen}(jx)).$$

Daremos un breve esquema de cómo se obtienen las fórmulas (5) y (6) al final de esta subsección.

Ejemplo 5.13. Vamos a probar que la función definida por $f(x) = x/2$ para $-\pi < x < \pi$ y extendida periódicamente por la relación $f(x + 2\pi) = f(x)$ admite el desarrollo en serie de Fourier

$$f(x) = \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j} \operatorname{sen}(jx) = \operatorname{sen}(x) - \frac{\operatorname{sen}(2x)}{2} + \frac{\operatorname{sen}(3x)}{3} - \dots$$

Usando las fórmulas de Euler y el método de integración por partes, obtenemos

$$a_j = \frac{1}{\pi} \int_{-\pi}^{\pi} \frac{x}{2} \cos(jx) dx = \frac{x \operatorname{sen}(jx)}{2\pi j} + \frac{\cos(jx)}{2\pi j^2} \Big|_{-\pi}^{\pi} = 0$$

para $j = 1, 2, 3, \dots$ y

$$b_j = \frac{1}{\pi} \int_{-\pi}^{\pi} \frac{x}{2} \operatorname{sen}(jx) dx = \frac{-x \cos(jx)}{2\pi j} + \frac{\operatorname{sen}(jx)}{2\pi j^2} \Big|_{-\pi}^{\pi} = \frac{(-1)^{j+1}}{j}$$

para $j = 1, 2, 3, \dots$ El coeficiente a_0 lo obtenemos haciendo un cálculo aparte:

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} \frac{x}{2} dx = \frac{x^2}{4\pi} \Big|_{-\pi}^{\pi} = 0.$$

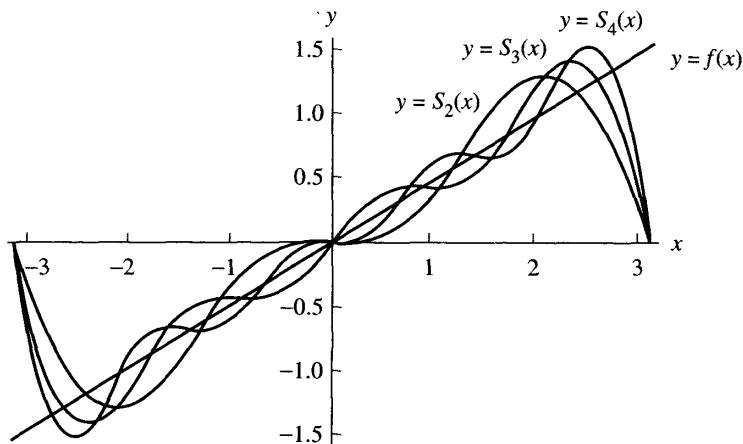


Figura 5.19 La función $f(x) = x/2$ en $[-\pi, \pi]$ y sus aproximaciones trigonométricas $S_2(x)$, $S_3(x)$ y $S_4(x)$.

Estos cálculos prueban que todos los coeficientes de las funciones coseno son cero. La gráfica de $f(x)$ y de las sumas parciales

$$S_2(x) = \sin(x) - \frac{\sin(2x)}{2},$$

$$S_3(x) = \sin(x) - \frac{\sin(2x)}{2} + \frac{\sin(3x)}{3},$$

y

$$S_4(x) = \sin(x) - \frac{\sin(2x)}{2} + \frac{\sin(3x)}{3} - \frac{\sin(4x)}{4}$$

se muestran en la Figura 5.19.

Vamos a enunciar ahora algunas de las propiedades generales de las series de Fourier; sus demostraciones quedan como ejercicios.

Teorema 5.6 (Series de cosenos). Supongamos que $f(x)$ es una función par; o sea, supongamos que se verifica $f(-x) = f(x)$ para todo x . Si $f(x)$ tiene período 2π y si $f(x)$ y $f'(x)$ son continuas a trozos, entonces la serie de Fourier de f tiene solamente los términos de los cosenos:

$$(8) \quad f(x) = \frac{a_0}{2} + \sum_{j=1}^{\infty} a_j \cos(jx),$$

donde

$$(9) \quad a_j = \frac{2}{\pi} \int_0^{\pi} f(x) \cos(jx) dx \quad \text{para } j = 0, 1, \dots$$

Teorema 5.7 (Series de senos). Supongamos que $f(x)$ es una función impar; o sea, supongamos que se verifica $f(-x) = -f(x)$ para todo x . Si $f(x)$ tiene período 2π y si $f(x)$ y $f'(x)$ son continuas a trozos, entonces la serie de Fourier de f tiene solamente los términos de los senos:

$$(10) \quad f(x) = \sum_{j=1}^{\infty} b_j \operatorname{sen}(jx),$$

donde

$$(11) \quad b_j = \frac{2}{\pi} \int_0^\pi f(x) \operatorname{sen}(jx) dx \quad \text{para } j = 1, 2, \dots$$

Ejemplo 5.14. Vamos a probar que la función dada por $f(x) = |x|$ si $-\pi < x < \pi$ y extendida periódicamente por la relación $f(x + 2\pi) = f(x)$ tiene la representación en serie de Fourier de senos dada por

$$(12) \quad \begin{aligned} f(x) &= \frac{\pi}{2} - \frac{4}{\pi} \sum_{j=1}^{\infty} \frac{\cos((2j-1)x)}{(2j-1)^2} \\ &= \frac{\pi}{2} - \frac{4}{\pi} \left(\cos(x) + \frac{\cos(3x)}{3^2} + \frac{\cos(5x)}{5^2} + \dots \right). \end{aligned}$$

La función $f(x)$ es par, así que podemos usar el Teorema 5.6 de manera que sólo necesitamos calcular los coeficientes $\{a_j\}$:

$$\begin{aligned} a_j &= \frac{2}{\pi} \int_0^\pi x \cos(jx) dx = \frac{2x \operatorname{sen}(jx)}{\pi j} + \frac{2 \cos(jx)}{\pi j^2} \Big|_0^\pi \\ &= \frac{2 \cos(j\pi) - 2}{\pi j^2} = \frac{2((-1)^j - 1)}{\pi j^2} \quad \text{para } j = 1, 2, 3, \dots \end{aligned}$$

Puesto que $((-1)^j - 1) = 0$ cuando j es par, la series de senos tiene sólo sus términos impares y sus coeficientes correspondientes siguen el patrón:

$$a_1 = \frac{-4}{\pi}, \quad a_3 = \frac{-4}{\pi 3^2}, \quad a_5 = \frac{-4}{\pi 5^2}, \quad \dots$$

El coeficiente a_0 lo obtenemos mediante un cálculo aparte:

$$a_0 = \frac{2}{\pi} \int_0^\pi x dx = \frac{x^2}{\pi} \Big|_0^\pi = \pi.$$

Por tanto, hemos hallado los coeficientes que anunciábamos en (12). ■

Demostración de las fórmulas de Euler dadas en el Teorema 5.5. Damos un argumento heurístico en el que suponemos la existencia y la convergencia de la

representación en serie de Fourier. Para determinar a_0 , integramos ambos lados de (7) y obtenemos

$$\begin{aligned}
 \int_{-\pi}^{\pi} f(x) dx &= \int_{-\pi}^{\pi} \left(\frac{a_0}{2} + \sum_{j=1}^{\infty} (a_j \cos(jx) + b_j \sin(jx)) \right) dx \\
 (13) \quad &= \int_{-\pi}^{\pi} \frac{a_0}{2} dx + \sum_{j=1}^{\infty} a_j \int_{-\pi}^{\pi} \cos(jx) dx + \sum_{j=1}^{\infty} b_j \int_{-\pi}^{\pi} \sin(jx) dx \\
 &= \pi a_0 + 0 + 0.
 \end{aligned}$$

La justificación de que se puede intercambiar el orden de las operaciones de sumar e integrar requiere un análisis detallado de las propiedades de la convergencia uniforme que puede encontrarse en textos de carácter más avanzado. Lo que hemos probado es, entonces, que

$$(14) \quad a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) dx.$$

Para determinar a_m , supongamos que fijamos un número entero $m > 0$. Multiplicando por $\cos(mx)$ ambos miembros de (7) e integrando, tenemos

$$\begin{aligned}
 (15) \quad \int_{-\pi}^{\pi} f(x) \cos(mx) dx &= \frac{a_0}{2} \int_{-\pi}^{\pi} \cos(mx) dx + \sum_{j=1}^{\infty} a_j \int_{-\pi}^{\pi} \cos(jx) \cos(mx) dx \\
 &\quad + \sum_{j=1}^{\infty} b_j \int_{-\pi}^{\pi} \sin(jx) \cos(mx) dx.
 \end{aligned}$$

La relación (15) puede simplificarse usando las propiedades de ortogonalidad de las funciones trigonométricas que establecemos a continuación. El valor del primer sumando del miembro derecho de (15) es

$$(16) \quad \frac{a_0}{2} \int_{-\pi}^{\pi} \cos(mx) dx = \frac{a_0 \sin(mx)}{2m} \Big|_{-\pi}^{\pi} = 0.$$

El valor del término que contiene el producto $\cos(jx) \cos(mx)$ se calcula usando la identidad trigonométrica

$$(17) \quad \cos(jx) \cos(mx) = \frac{1}{2} \cos((j+m)x) + \frac{1}{2} \cos((j-m)x).$$

Cuando $j \neq m$, usamos (17) y obtenemos

$$\begin{aligned}
 (18) \quad a_j \int_{-\pi}^{\pi} \cos(jx) \cos(mx) dx &= \frac{1}{2} a_j \int_{-\pi}^{\pi} \cos((j+m)x) dx \\
 &\quad + \frac{1}{2} a_j \int_{-\pi}^{\pi} \cos((j-m)x) dx = 0 + 0 = 0.
 \end{aligned}$$

Mientras que, cuando $j = m$, el valor de la integral es

$$(19) \quad a_m \int_{-\pi}^{\pi} \cos(jx) \cos(mx) dx = a_m \pi.$$

El valor del término del último miembro derecho de (15), el que contiene el producto $\sin(jx) \cos(mx)$, se calcula usando la identidad trigonométrica:

$$(20) \quad \sin(jx) \cos(mx) = \frac{1}{2} \sin((j+m)x) + \frac{1}{2} \sin((j-m)x).$$

Cualesquiera que sean los valores j y m en (20), obtenemos

$$(21) \quad b_j \int_{-\pi}^{\pi} \sin(jx) \cos(mx) dx = \frac{1}{2} b_j \int_{-\pi}^{\pi} \sin((j+m)x) dx \\ + \frac{1}{2} b_j \int_{-\pi}^{\pi} \sin((j-m)x) dx = 0 + 0 = 0,$$

Por consiguiente, usando los resultados de (16), (18), (19) y (21) en la relación (15), concluimos que

$$(22) \quad \pi a_m = \int_{-\pi}^{\pi} f(x) \cos(mx) dx, \quad \text{para } m = 1, 2, \dots,$$

lo que establece la fórmula de Euler (5). La fórmula de Euler (6) se prueba de manera análoga.

Aproximación mediante polinomios trigonométricos

Definición 5.4 (Polinomio trigonométrico). Una serie (finita) de la forma

$$(23) \quad T_M(x) = \frac{a_0}{2} + \sum_{j=1}^M (a_j \cos(jx) + b_j \sin(jx))$$

se llama **polinomio trigonométrico** de grado M . ▲

Teorema 5.8 (Serie de Fourier discreta). Si $\{(x_j, y_j)\}_{j=0}^N$ son $N+1$ puntos tales que sus abscisas

$$(24) \quad x_j = -\pi + \frac{2j\pi}{N} \quad \text{para } j = 0, 1, \dots, N$$

están equiespaciadas y sus ordenadas son de la forma $y_j = f(x_j)$ para una función $f(x)$ que es periódica de período 2π y si $2M < N$, entonces existe

un polinomio trigonométrico $T_M(x)$ de la forma dada en (23) que minimiza la cantidad

$$(25) \quad \sum_{k=1}^N (f(x_k) - T_M(x_k))^2.$$

Los coeficientes a_j y b_j de este polinomio vienen dados por las fórmulas

$$(26) \quad a_j = \frac{2}{N} \sum_{k=1}^N f(x_k) \cos(jx_k) \quad \text{para } j = 0, 1, \dots, M,$$

y

$$(27) \quad b_j = \frac{2}{N} \sum_{k=1}^N f(x_k) \sin(jx_k) \quad \text{para } j = 1, 2, \dots, M.$$

Hagamos notar que aunque las fórmulas (26) y (27) se obtienen a partir de un procedimiento de mínimos cuadrados, también pueden verse como aproximaciones numéricas a las integrales de las fórmulas de Euler (5) y (6); estas fórmulas de Euler proporcionan los coeficientes de Fourier de una función continua, mientras que las fórmulas (26) y (27) proporcionan los coeficientes del polinomio trigonométrico que interpola los datos. En el ejemplo siguiente usaremos datos generados con la función $f(x) = x/2$ en un número finito de puntos. Conforme aumenta el número de puntos, los coeficientes del polinomio trigonométrico se acercan a los coeficientes de la serie de Fourier.

Ejemplo 5.15. Vamos a usar los doce puntos $x_k = -\pi + k\pi/6$, para $k = 1, 2, \dots, 12$ para encontrar el polinomio trigonométrico de aproximación de grado $M = 5$ correspondiente a los doce datos $\{(x_k, f(x_k))\}_{k=1}^{12}$, siendo $f(x) = x/2$. También compararemos los resultados que se obtienen cuando se usan 60 y 360 puntos y cuando se usan los primeros cinco términos de la serie de Fourier de $f(x)$ dada en el Ejemplo 5.13.

Puesto que se supone que la función se extiende periódicamente, el valor $f(\pi)$ debemos calcularlo usando la fórmula

$$(28) \quad f(\pi) = \frac{f(\pi^-) + f(\pi^+)}{2} = \frac{\pi/2 - \pi/2}{2} = 0.$$

La función es impar, así que los coeficientes de los términos de los cosenos deben ser cero (es decir, $a_j = 0$ para todo j). El polinomio trigonométrico de grado $M = 5$ sólo contiene los términos de los senos, de manera que cuando usamos la fórmula (27) con el valor dado en (28), obtenemos

$$(29) \quad T_5(x) = 0.9770486 \sin(x) - 0.4534498 \sin(2x) + 0.26179938 \sin(3x) \\ - 0.1511499 \sin(4x) + 0.0701489 \sin(5x).$$

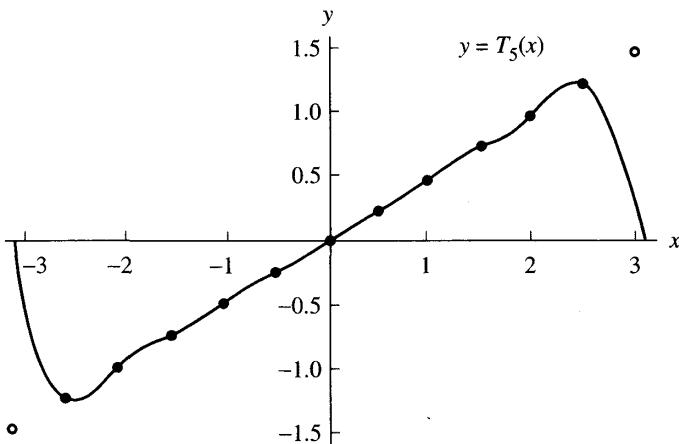


Figura 5.20 El polinomio trigonométrico $T_5(x)$ de grado $M = 5$ que pasa por 11 puntos que están sobre la recta $y = x/2$.

Tabla 5.9 Comparación de los coeficientes de los polinomios trigonométricos que aproximan $f(x) = x/2$ en $[-\pi, \pi]$.

	Coeficientes de los polinomios trigonométricos			Coeficientes de la serie de Fourier
	12 puntos	60 puntos	360 puntos	
b_1	0.97704862	0.99908598	0.99997462	1.0
b_2	-0.45344984	-0.49817096	-0.49994923	-0.5
b_3	0.26179939	0.33058726	0.33325718	0.33333333
b_4	-0.15114995	-0.24633386	-0.24989845	-0.25
b_5	0.07014893	0.19540972	0.19987306	0.2

La gráfica de $T_5(x)$ se muestra en la Figura 5.20.

Los coeficientes del polinomio trigonométrico de grado $M = 5$ cambian sólo ligeramente cuando el número de nodos de interpolación se incrementa hasta 60 y hasta 360; conforme el número de puntos crece, los coeficientes se acercan a los coeficientes de la serie de Fourier de $f(x)$. Estos resultados se recogen en la Tabla 5.9. ■

El siguiente programa permite construir sendas matrices \mathbf{A} y \mathbf{B} que contienen los coeficientes a_j y b_j , respectivamente, del polinomio trigonométrico (23) de grado M .

Programa 5.4 (Polinomios Trigonométricos). Construcción del polinomio trigonométrico de grado M de la forma

$$P(x) = \frac{a_0}{2} + \sum_{j=1}^M (a_j \cos(jx) + b_j \sin(jx))$$

correspondiente a N nodos equiespaciados $x_k = -\pi + 2\pi k/N$, para $k = 1, 2, \dots, N$. Esta construcción puede efectuarse si $2M + 1 \leq N$.

```
function [A,B]=tpcoeff(X,Y,M)
% Datos
%     - X es un vector de abscisas equiespaciadas en [-pi,pi]
%     - Y es un vector de ordenadas
%     - M es el grado del polinomio trigonométrico
% Resultados
%     - A es el vector de los coeficientes de los cos(jx)
%     - B es el vector de los coeficientes de los sen(jx)
N=length(X)-1;
max1=fix((N-1)/2);
if M>max1
    M=max1;
end
A=zeros(1,M+1);
B=zeros(1,M+1);
Yends=(Y(1)+Y(N+1))/2;
Y(1)=Yends;
Y(N+1)=Yends;
A(1)=sum(Y);
for j=1:M
    A(j+1)=cos(j*X)*Y';
    B(j+1)=sin(j*X)*Y';
end
A=2*A/N;
B=2*B/N;
A(1)=A(1)/2;
```

El programita siguiente permite evaluar el polinomio trigonométrico $P(x)$ de grado M obtenido con el Programa 5.4 en un valor concreto de x .

```
function z=tp(A,B,x,M)
z=A(1);
for j= 1:M
    z=z+A(j+1)*cos(j*x)+B(j+1)*sin(j*x);
```

end

Por ejemplo, la sucesión de instrucciones del paquete MATLAB que se relaciona a continuación permite dibujar una gráfica análoga a la de la Figura 5.20.

```
>>x=-pi:.01:pi;
>>y=tp(A,B,x,M);
>>plot(x,y,X,Y,'o')
```

Ejercicios

En los Ejercicios 1 a 4, halle la representación en serie de Fourier de la función dada. Indicación. Proceda como se hizo en los Ejemplos 5.13 y 5.14 y dibuje sobre un mismo sistema de coordenadas las gráficas de la función y de las sumas parciales $S_2(x)$, $S_3(x)$ y $S_4(x)$ de su serie de Fourier (como en la Figura 5.19).

$$1. f(x) = \begin{cases} -1 & \text{si } -\pi < x < 0 \\ 1 & \text{si } 0 < x < \pi \end{cases}$$

$$2. f(x) = \begin{cases} \frac{\pi}{2} + x & \text{si } -\pi \leq x < 0 \\ \frac{\pi}{2} - x & \text{si } 0 \leq x < \pi \end{cases}$$

$$3. f(x) = \begin{cases} 0 & \text{si } -\pi \leq x < 0 \\ x & \text{si } 0 \leq x < \pi \end{cases}$$

$$4. f(x) = \begin{cases} -1 & \text{si } \frac{\pi}{2} < x < \pi \\ 1 & \text{si } -\frac{\pi}{2} < x < \frac{\pi}{2} \\ -1 & \text{si } -\pi < x < -\frac{\pi}{2} \end{cases}$$

$$5. f(x) = \begin{cases} -\pi - x & \text{si } -\pi \leq x < -\frac{\pi}{2} \\ x & \text{si } -\frac{\pi}{2} \leq x < \frac{\pi}{2} \\ \pi - x & \text{si } \frac{\pi}{2} \leq x < \pi \end{cases}$$

6. En el Ejercicio 1, ponga $x = \pi/2$ y pruebe que

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots$$

7. En el Ejercicio 2, ponga $x = 0$ y pruebe que

$$\frac{\pi^2}{8} = 1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \dots$$

8. Halle la representación en serie de Fourier de cosenos de la función periódica que en un período se define como $f(x) = x^2/4$ para $-\pi \leq x < \pi$.
9. Sea $f(x)$ una función periódica de período $2P$; es decir, $f(x+2P) = f(x)$ para todo x . Haciendo los cambios adecuados, demuestre que las fórmulas de Euler

(5) y (6) para f son

$$a_0 = \frac{1}{P} \int_{-P}^P f(x) dx$$

$$a_j = \frac{1}{P} \int_{-P}^P f(x) \cos\left(\frac{j\pi x}{P}\right) dx \quad \text{para } j = 1, 2, \dots$$

$$b_j = \frac{1}{P} \int_{-P}^P f(x) \sin\left(\frac{j\pi x}{P}\right) dx \quad \text{para } j = 1, 2, \dots$$

En los Ejercicios 10 a 12, use los resultados del Ejercicio 9 para hallar las representaciones en serie de Fourier de la función dada y dibuje $f(x)$, $S_4(x)$ y $S_6(x)$ en un mismo gráfico.

10. $f(x) = \begin{cases} 0 & \text{si } -2 \leq x < 0 \\ 1 & \text{si } 0 \leq x < 2 \end{cases}$

11. $f(x) = \begin{cases} -1 & \text{si } -3 \leq x < -1 \\ |x| & \text{si } -1 \leq x < 1 \\ 1 & \text{si } 1 \leq x < 3 \end{cases}$

12. $f(x) = -x^2 + 9 \quad \text{para } -3 \leq x < 3.$

13. Demuestre los Teoremas 5.6 y 5.7.

Algoritmos y programas

- Utilice el Programa 5.4 con $N = 12$ puntos y proceda como en el Ejemplo 5.15 para hallar el polinomio trigonométrico de grado $M = 5$ para los puntos equiespaciados $\{(x_k, f(x_k))\}_{k=1}^{12}$, siendo $f(x)$ la función que aparece en el (a) Ejercicio 1, (b) Ejercicio 2, (c) Ejercicio 3 y (d) Ejercicio 4. En cada caso, dibuje las funciones $f(x)$, $T_5(x)$ y los puntos $\{(x_k, f(x_k))\}_{k=1}^{12}$ en un mismo gráfico.
- Use el Programa 5.4 para hallar los coeficientes del polinomio $T_5(x)$ del Ejemplo 5.15 cuando se usan 60 y, respectivamente, 360 nodos equiespaciados.
- Modifique el Programa 5.4 de manera que sirva para calcular polinomios trigonométricos de período $2P = b - a$ para nodos equiespaciados en un intervalo $[a, b]$.
- Use su modificación del Programa 5.4 para hallar el polinomio trigonométrico $T_5(x)$ correspondiente a las funciones (a) $f(x)$ del Ejercicio 10, con 12 nodos equiespaciados y (b) $f(x)$ del Ejercicio 12, con 60 nodos equiespaciados. En cada caso, dibuje la función, el polinomio $T_5(x)$ y los datos sobre un mismo gráfico.
- Considere la tabla de temperaturas dada en el Problema 1 de la subsección “Algoritmos y programas” de la Sección 3.4; hay un total de 24 datos.
 - Determine el polinomio trigonométrico correspondiente $T_7(x)$.

334 CAP. 5 AJUSTE DE CURVAS

Tabla 5.10 Datos del Problema 6.

Fecha	Temperatura media
Ene. 1	-14
Ene. 29	-9
Feb. 26	2
Mar. 26	15
Abr. 23	35
May. 21	52
Jun. 18	62
Jul. 16	63
Ago. 13	58
Sep. 10	50
Oct. 8	34
Nov. 5	12
Dic. 3	-5

- (b) Dibuje $T_7(x)$ y los 24 puntos en un mismo gráfico.
 (c) Repita los apartados (a) y (b) usando las temperaturas de su localidad.
6. En la Tabla 5.10 se recoge una muestra de la temperatura media diaria en Fairbanks, Alaska, medida en grados Fahrenheit. Dicha muestra se extiende a lo largo de un año y consta de 13 datos tomados en intervalos iguales de 28 días.
- (a) Determine el polinomio trigonométrico correspondiente $T_6(x)$.
 (b) Dibuje $T_6(x)$ y los 13 puntos en un mismo gráfico.

Derivación numérica

Las fórmulas de derivación numérica son importantes en el desarrollo de algoritmos para resolver problemas de contorno de ecuaciones diferenciales ordinarias y ecuaciones en derivadas parciales (véanse los Capítulos 9 y 10). El uso de funciones conocidas como ejemplos de aplicación de las técnicas de derivación numérica permite comparar la aproximación numérica con la respuesta exacta. Como ilustración usemos la función de Bessel $J_1(x)$, cuyos valores pueden hallarse tabulados en libros de referencia habituales. Tomamos ocho puntos equiespaciados

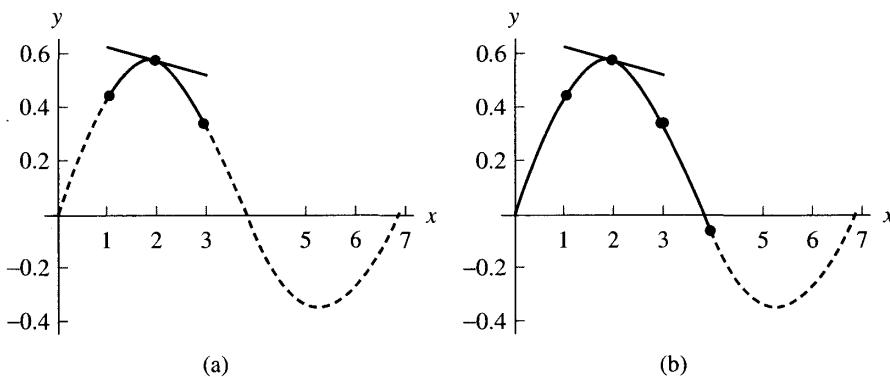


Figura 6.1 (a) La recta tangente a $p_2(x)$ en $(2, 0.5767)$ tiene pendiente $p'_2(2) = -0.0505$. (b) La recta tangente a $p_4(x)$ en $(2, 0.5767)$ tiene pendiente $p'_4(2) = -0.0618$.

ciados en el intervalo $[0, 7]$, que nos darían los puntos de la gráfica $(0, 0.0000)$, $(1, 0.4400)$, $(2, 0.5767)$, $(3, 0.3391)$, $(4, -0.0660)$, $(5, -0.3276)$, $(6, -0.2767)$ y $(7, -0.004)$. El principio subyacente es la derivación de un polinomio interpolador; para verlo centremos nuestra atención en el cálculo de $J'_1(2)$. El polinomio interpolador $p_2(x) = -0.0710 + 0.6982x - 0.1872x^2$ pasa por los puntos $(1, 0.4400)$, $(2, 0.5767)$ y $(3, 0.3391)$ y lo usamos para obtener $J'_1(2) \approx p'_2(2) = -0.0505$. Este polinomio cuadrático $p_2(x)$ y su recta tangente en el punto $(2, J_1(2))$ se muestran en la Figura 6.1(a). Si usamos cinco nodos de interpolación, podemos obtener una aproximación mejor: El polinomio $p_4(x) = 0.4986x + 0.011x^2 - 0.0813x^3 + 0.0116x^4$ pasa por los puntos $(0, 0.0000)$, $(1, 0.4400)$, $(2, 0.5767)$, $(3, 0.3391)$ y $(4, -0.0660)$ y, derivándolo, obtenemos $J'_1(2) \approx p'_4(2) = -0.0618$. El polinomio $p_4(x)$ y su recta tangente en el punto $(2, J_1(2))$ se muestran en la Figura 6.1(b). El valor exacto de la derivada es $J'_1(2) = -0.0645$ y los errores cometidos al usar $p_2(x)$ y $p_4(x)$ son -0.0140 y -0.0026 , respectivamente. En este capítulo desarrollamos la teoría introductoria necesaria para investigar la exactitud de los métodos de derivación numérica.

6.1 Aproximaciones a la derivada

El límite del cociente incremental

Vamos a afrontar ahora el problema de aproximar numéricamente la derivada de $f(x)$:

$$(1) \quad f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}.$$

El método parece claro: elegimos una sucesión $\{h_k\}$ tal que $h_k \rightarrow 0$ y calculamos el límite de la sucesión

$$(2) \quad D_k = \frac{f(x+h_k) - f(x)}{h_k} \quad \text{para } k = 1, 2, \dots$$

Puesto que sólo calcularemos un número finito de términos D_1, D_2, \dots, D_N de la sucesión (2) y usaremos el último D_N como respuesta, la pregunta es obvia ¿por qué calculamos D_1, D_2, \dots, D_{N-1} ? Equivalentemente, podríamos preguntar ¿qué valor de h_N hay que elegir para asegurar que D_N es una buena aproximación a la derivada $f'(x)$? Para responder esta pregunta, empezaremos por analizar un ejemplo para ver por qué no hay una solución simple del problema.

Por ejemplo, consideremos la función $f(x) = e^x$ y usemos los incrementos $h = 1, 1/2$ y $1/4$ para construir las rectas secantes que pasan por los correspondientes puntos $(0, 1)$ y $(h, f(h))$. Conforme h disminuye, la recta secante se aproxima a la recta tangente como se muestra en la Figura 6.2. Aunque

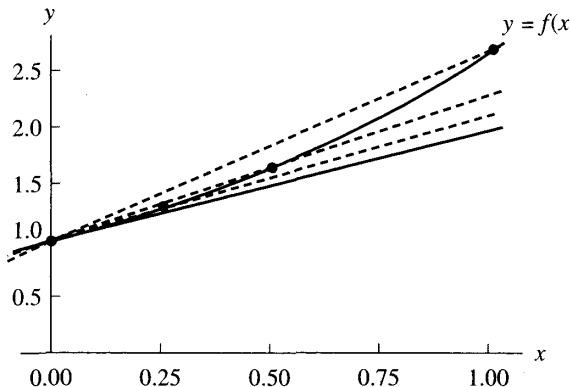


Figura 6.2 Varias rectas secantes a $y = e^x$.

Tabla 6.1 Cálculo de los cocientes incrementales $D_k = (e^{1+h_k} - e)/h_k$.

h_k	$f_k = f(1 + h_k)$	$f_k - e$	$D_k = (f_k - e)/h_k$
$h_1 = 0.1$	3.004166024	0.285884196	2.858841960
$h_2 = 0.01$	2.745601015	0.027319187	2.731918700
$h_3 = 0.001$	2.721001470	0.002719642	2.719642000
$h_4 = 0.0001$	2.718553670	0.000271842	2.718420000
$h_5 = 0.00001$	2.718309011	0.000027183	2.718300000
$h_6 = 10^{-6}$	2.718284547	0.000002719	2.719000000
$h_7 = 10^{-7}$	2.718282100	0.000000272	2.720000000
$h_8 = 10^{-8}$	2.718281856	0.000000028	2.800000000
$h_9 = 10^{-9}$	2.718281831	0.000000003	3.000000000
$h_{10} = 10^{-10}$	2.718281828	0.000000000	0.000000000

la Figura 6.2 proporciona una buena visualización del proceso descrito en (1), desde el punto de vista numérico hay que usar $h = 0.00001$ para obtener una buena respuesta y para este valor de h las gráficas de las rectas tangente y secante son indistinguibles.

Ejemplo 6.1. Para $f(x) = e^x$ y $x = 1$, vamos a calcular los cocientes incrementales D_k usando los incrementos $h_k = 10^{-k}$ para $k = 1, 2, \dots, 10$; arrastraremos nueve cifras decimales en todas las operaciones.

En la Tabla 6.1 se muestran los valores $f(1 + h_k)$ y $(f(1 + h_k) - f(1))/h_k$ que se utilizan para calcular D_k . ■

El incremento mayor $h_1 = 0.1$ no proporciona una buena aproximación $D_1 \approx f'(1)$ porque h_1 es demasiado grande; el cociente incremental es la pendiente de una recta secante que pasa por dos puntos que no están suficientemente

cerca. Por otro lado, cuando usamos la fórmula (2) trabajando con una precisión fija de nueve cifras decimales, h_9 proporciona la aproximación $D_9 = 3$ y h_{10} proporciona $D_{10} = 0$. O sea, si h_k es demasiado pequeño, entonces los valores de la función $f(x + h_k)$ y $f(x)$ que hay que calcular están demasiado cerca y al hacer diferencia $f(x + h_k) - f(x)$ puede aparecer el problema de la pérdida de cifras significativas debido a la substracción de cantidades que son casi iguales. El valor $h_{10} = 10^{-10}$ es tan pequeño que los valores $f(x + h_{10})$ y $f(x)$ almacenados por el computador son iguales y, en consecuencia, el cociente incremental calculado es cero. En el Ejemplo 6.1, el valor exacto del límite es $f'(1) \approx 2.718281828$ y puede observarse que el valor $h_5 = 10^{-5}$ es el que da la mejor aproximación $D_5 = 2.7183$.

El Ejemplo 6.1 muestra que no es fácil hallar numéricamente el límite que aparece en (2): La sucesión empieza a converger a e , llega a D_5 que es el valor que más se acerca y luego sus términos se alejan de e . En el Programa 6.1 se sugiere que se vayan calculando los términos de la sucesión $\{D_k\}$ hasta que $|D_{N+1} - D_N| \geq |D_N - D_{N-1}|$; la intención es tratar de determinar la mejor aproximación antes de que los términos empiecen a alejarse del límite. Cuando aplicamos este criterio al Ejemplo 6.1, tenemos $0.0007 = |D_6 - D_5| > |D_5 - D_4| = 0.00012$; por tanto, D_5 es la respuesta elegida. Vamos ahora a desarrollar fórmulas de aproximación que proporcionan un grado de precisión razonable para valores de h no demasiado pequeños.

Las fórmulas de diferencias centradas

Si la función $f(x)$ puede evaluarse en puntos que están a ambos lados de x , entonces la mejor fórmula que involucra dos puntos es la que utiliza abscisas situadas simétricamente a izquierda y derecha de x .

Teorema 6.1 (Fórmula centrada de orden $O(h^2)$). Supongamos que $f \in C^3[a, b]$ y que $x - h, x, x + h \in [a, b]$. Entonces

$$(3) \quad f'(x) \approx \frac{f(x + h) - f(x - h)}{2h}.$$

Es más, existe un número $c = c(x) \in [a, b]$ tal que

$$(4) \quad f'(x) = \frac{f(x + h) - f(x - h)}{2h} + E_{\text{trunc}}(f, h),$$

siendo

$$E_{\text{trunc}}(f, h) = -\frac{h^2 f^{(3)}(c)}{6} = O(h^2).$$

El término $E(f, h)$ se llama **error de truncamiento**.

Demostración. Usamos la fórmula de Taylor de orden dos de f , alrededor de x , para $f(x+h)$ y $f(x-h)$:

$$(5) \quad f(x+h) = f(x) + f'(x)h + \frac{f^{(2)}(x)h^2}{2!} + \frac{f^{(3)}(c_1)h^3}{3!}$$

y

$$(6) \quad f(x-h) = f(x) - f'(x)h + \frac{f^{(2)}(x)h^2}{2!} - \frac{f^{(3)}(c_2)h^3}{3!}.$$

Restando (6) de (5) obtenemos

$$(7) \quad f(x+h) - f(x-h) = 2f'(x)h + \frac{(f^{(3)}(c_1) + f^{(3)}(c_2))h^3}{3!}.$$

Como $f^{(3)}(x)$ es continua, podemos usar el teorema del valor intermedio para deducir que existe un valor c tal que

$$(8) \quad \frac{f^{(3)}(c_1) + f^{(3)}(c_2)}{2} = f^{(3)}(c),$$

igualdad que sustituimos en (7) para, tras ordenar un poco los términos, obtener

$$(9) \quad f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{f^{(3)}(c)h^2}{3!}.$$

El primer término del miembro derecho de (9) es la fórmula de diferencia centrada (3) y el segundo es el error de truncamiento, así que el teorema ya está demostrado. •

Si los valores de la tercera derivada $f^{(3)}(c)$ no cambian muy rápidamente, entonces el error de truncamiento que aparece en (4) tiende a cero a la misma velocidad que h^2 , lo que expresamos mediante la notación $O(h^2)$. Cuando hacemos los cálculos con un computador, no es aconsejable elegir h demasiado pequeño; por eso sería útil disponer de una fórmula que aproxime $f'(x)$ y que tenga un error de truncamiento de orden $O(h^4)$.

Teorema 6.2 (Fórmula centrada de orden $O(h^4)$). Supongamos que $f \in C^5[a, b]$ y que $x - 2h, x - h, x, x + h, x + 2h \in [a, b]$. Entonces

$$(10) \quad f'(x) \approx \frac{-f(x+2h) + 8f(x+h) - 8f(x-h) + f(x-2h)}{12h}.$$

Es más, existe un número $c = c(x) \in [a, b]$ tal que

$$(11) \quad f'(x) = \frac{-f(x+2h) + 8f(x+h) - 8f(x-h) + f(x-2h)}{12h} + E_{\text{trunc}}(f, h),$$

siendo

$$E_{\text{trunc}}(f, h) = \frac{h^4 f^{(5)}(c)}{30} = \mathcal{O}(h^4).$$

Demostración. Esta vez usamos la fórmula de Taylor de cuarto orden de f , alrededor de x , para $f(x+h)$ y $f(x-h)$:

$$(12) \quad f(x+h) - f(x-h) = 2f'(x)h + \frac{2f^{(3)}(x)h^3}{3!} + \frac{2f^{(5)}(c_1)h^5}{5!}.$$

Ahora usamos como incremento $2h$, en vez de h , y escribimos la correspondiente aproximación:

$$(13) \quad f(x+2h) - f(x-2h) = 4f'(x)h + \frac{16f^{(3)}(x)h^3}{3!} + \frac{64f^{(5)}(c_2)h^5}{5!}.$$

A continuación multiplicamos por 8 los términos de la relación (12) y le restamos la relación (13), con ello se simplifican los términos que contienen $f^{(3)}(x)$ y obtenemos

$$(14) \quad \begin{aligned} & -f(x+2h) + 8f(x+h) - 8f(x-h) + f(x-2h) \\ &= 12f'(x)h + \frac{(16f^{(5)}(c_1) - 64f^{(5)}(c_2))h^5}{120}. \end{aligned}$$

Si $f^{(5)}(x)$ tiene signo constante y no cambia muy rápidamente cerca de x , podemos encontrar un punto c en $[x-2h, x+2h]$ tal que

$$(15) \quad 16f^{(5)}(c_1) - 64f^{(5)}(c_2) = -48f^{(5)}(c).$$

Sustituyendo (15) en (14) y despejando luego $f'(x)$, obtenemos

$$(16) \quad f'(x) = \frac{-f(x+2h) + 8f(x+h) - 8f(x-h) + f(x-2h)}{12h} + \frac{f^{(5)}(c)h^4}{30}$$

El primer término del miembro derecho de (16) es la fórmula de diferencia centrada (10) y el segundo término es el error de truncamiento, lo que finaliza la demostración del teorema. •

Supongamos que $|f^{(5)}(c)|$ está acotada cuando c recorre $[a, b]$, entonces el error de truncamiento de la expresión (11) converge a cero a la misma velocidad que h^4 , lo que se expresa mediante la notación $\mathcal{O}(h^4)$ introducida en la Definición 1.3. Ahora podemos comparar las fórmulas (3) y (10). Supongamos que $f(x)$ admite cinco derivadas continuas y que $|f^{(3)}(c)|$ y $|f^{(5)}(c)|$ valen más o menos lo mismo, entonces el error de truncamiento de la fórmula (10) es de orden $\mathcal{O}(h^4)$ y convergerá a cero más rápidamente que el error de truncamiento, que es de orden $\mathcal{O}(h^2)$, de la fórmula (3); esto permite usar un incremento mayor para lograr la misma precisión.

Tabla 6.2 Derivación numérica mediante las fórmulas (3) y (10).

Incremento	Aproximación con la fórmula (3)	Error con la fórmula (3)	Aproximación con la fórmula (10)	Error con la fórmula (10)
0.1	-0.716161095	-0.001194996	-0.717353703	-0.000002389
0.01	-0.717344150	-0.000011941	-0.717356108	0.000000017
0.001	-0.717356000	-0.000000091	-0.717356167	0.000000076
0.0001	-0.717360000	-0.000003909	-0.717360833	0.000004742

Ejemplo 6.2. Sea $f(x) = \cos(x)$.

- (a) Vamos a usar las fórmulas (3) y (10) con incrementos $h = 0.1, 0.01, 0.001$ y 0.0001 para calcular aproximaciones a $f'(0.8)$. Trabajaremos con nueve cifras decimales significativas.
- (b) Compararemos los valores obtenidos con el exacto $f'(0.8) = -\sin(0.8)$.

(a) Usando la fórmula (3) con $h = 0.01$, obtenemos

$$f'(0.8) \approx \frac{f(0.81) - f(0.79)}{0.02} \approx \frac{0.689498433 - 0.703845316}{0.02} \approx -0.717344150.$$

Usando la fórmula (10) con $h = 0.01$, obtenemos

$$\begin{aligned} f'(0.8) &\approx \frac{-f(0.82) + 8f(0.81) - 8f(0.79) + f(0.78)}{0.12} \\ &\approx \frac{-0.682221207 + 8(0.689498433) - 8(0.703845316) + 0.710913538}{0.12} \\ &\approx -0.717356108. \end{aligned}$$

(b) El error en las aproximaciones proporcionadas por las fórmulas (3) y (10) resulta ser -0.000011941 y 0.000000017 , respectivamente. Vemos que, en este ejemplo, la fórmula (10) proporciona una aproximación a $f'(0.8)$ mejor que la que proporciona la fórmula (3) cuando $h = 0.01$ pero no cuando $h = 0.0001$ (véase la Tabla 6.2). El análisis del error que haremos a continuación nos permitirá entender por qué ocurre esto. El resto de los cálculos se recogen en la Tabla 6.2. ■

Análisis del error e incremento óptimo

Un aspecto importante en el estudio de la derivación numérica es el efecto de los errores de redondeo cuando los cálculos se hacen con un computador. Vamos a examinar las fórmulas de los términos del error con mayor detalle. Supongamos

que usamos un computador para hacer los cálculos de manera que podemos escribir

$$f(x_0 - h) = y_{-1} + e_{-1} \quad \text{y} \quad f(x_0 + h) = y_1 + e_1,$$

donde hemos aproximado $f(x_0 - h)$ y $f(x_0 + h)$, respectivamente, por los números del computador y_{-1} e y_1 , siendo los errores de redondeo e_{-1} y e_1 . El siguiente resultado explica la compleja naturaleza del análisis del error de los métodos de derivación numérica.

Corolario 6.1(a). Supongamos que f verifica las hipótesis del Teorema 6.1 y que usamos la **fórmula computacional**

$$(17) \quad f'(x_0) \approx \frac{y_1 - y_{-1}}{2h}.$$

Entonces el término del error de esta fórmula viene dado por las siguientes relaciones:

$$(18) \quad f'(x_0) = \frac{y_1 - y_{-1}}{2h} + E(f, h)$$

donde

$$(19) \quad \begin{aligned} E(f, h) &= E_{\text{red}}(f, h) + E_{\text{trunc}}(f, h) \\ &= \frac{e_1 - e_{-1}}{2h} - \frac{h^2 f^{(3)}(c)}{6}; \end{aligned}$$

o sea, el **término del error total** $E(f, h)$ consta de una parte debida a los errores de redondeo más otra debida al error de truncamiento.

Corolario 6.1(b). Supongamos que f verifica las hipótesis del Teorema 6.1 y que hacemos los cálculos con un computador de manera que $|e_{-1}| \leq \varepsilon$, $|e_1| \leq \varepsilon$ y $M = \max\{|f^{(3)}(x)| : a \leq x \leq b\}$. Entonces

$$(20) \quad |E(f, h)| \leq \frac{\varepsilon}{h} + \frac{Mh^2}{6}.$$

Además, el valor de h que minimiza la expresión del miembro derecho de la desigualdad (20) es

$$(21) \quad h = \left(\frac{3\varepsilon}{M}\right)^{1/3}.$$

Cuando h es pequeño, la porción de (19) dada por $(e_1 - e_{-1})/2h$ puede ser relativamente grande; eso es lo que pasa en el Ejemplo 6.2 cuando $h = 0.0001$: Los errores de redondeo correspondientes son

$$\begin{aligned} f(0.8001) &= 0.696634970 + e_1 & \text{siendo } e_1 \approx -0.0000000003 \\ f(0.7999) &= 0.696778442 + e_{-1} & \text{siendo } e_{-1} \approx 0.0000000005 \end{aligned}$$

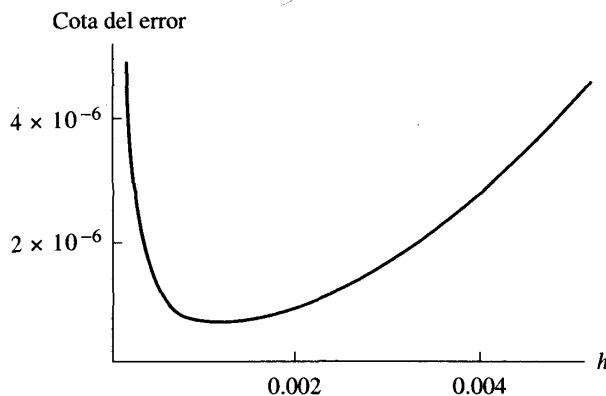


Figura 6.3 Determinación del incremento óptimo $h = 0.001144714$ cuando aplicamos la fórmula (21) a la función $f(x) = \cos(x)$ en el Ejemplo 6.2.

y el término del error de truncamiento es

$$\frac{-h^2 f^{(3)}(c)}{6} \approx -(0.0001)^2 \left(\frac{\sin(0.8)}{6} \right) \approx 0.000000001,$$

por tanto, ya podemos estimar el término del error total $E(f, h)$ en la fórmula (19):

$$\begin{aligned} E(f, h) &\approx \frac{-0.0000000003 - 0.0000000005}{0.0002} - 0.000000001 \\ &= -0.000004001. \end{aligned}$$

De hecho, la aproximación numérica a la derivada que se obtiene con el incremento $h = 0.0001$ es, haciendo los cálculos,

$$\begin{aligned} f'(0.8) &\approx \frac{f(0.8001) - f(0.7999)}{0.0002} = \frac{0.696634970 - 0.696778442}{0.0002} \\ &= -0.717360000, \end{aligned}$$

donde es evidente la pérdida de unas cuatro cifras significativas. El error real es -0.000003909 que está cerca del error predicho -0.000004001 .

Si aplicamos la fórmula (21) al Ejemplo 6.2, con la cota $|f^{(3)}(x)| \leq |\sin(x)| \leq 1 = M$ y el valor $\varepsilon = 0.5 \times 10^{-9}$ como magnitud del error de redondeo, el valor óptimo del incremento es $h = (1.5 \times 10^{-9}/1)^{1/3} = 0.001144714$. De los cuatro incrementos propuestos para usar la fórmula (3), es $h = 0.001$ el que está más cerca del valor óptimo 0.001144714 y es el que proporciona la mejor aproximación a $f'(0.8)$ (véanse la Tabla 6.2 y la Figura 6.3).

El análisis del error de la fórmula (10) es parecido; supongamos que usamos un computador para hacer los cálculos de manera que $f(x_0 + kh) = y_k + e_k$.

Corolario 6.2(a). Supongamos que f verifica las hipótesis del Teorema 6.2 y que usamos la **fórmula computacional**

$$(22) \quad f'(x_0) \approx \frac{-y_2 + 8y_1 - 8y_{-1} + y_{-2}}{12h}.$$

Entonces, el término del error de esta fórmula viene dado por las siguientes expresiones

$$(23) \quad f'(x_0) = \frac{-y_2 + 8y_1 - 8y_{-1} + y_{-2}}{12h} + E(f, h)$$

donde

$$(24) \quad \begin{aligned} E(f, h) &= E_{\text{red}}(f, h) + E_{\text{trunc}}(f, h) \\ &= \frac{-e_2 + 8e_1 - 8e_{-1} + e_{-2}}{12h} + \frac{h^4 f^{(5)}(c)}{30}, \end{aligned}$$

o sea, el **término del error total** $E(f, h)$ consta de una parte debida a los errores de redondeo más otra debida al error de truncamiento.

Corolario 6.2(b). Supongamos que f verifica las hipótesis del Teorema 6.2 y que hacemos los cálculos con un computador de manera que $|e_k| \leq \varepsilon$ y $M = \max\{|f^{(5)}(x)| : a \leq x \leq b\}$. Entonces

$$(25) \quad |E(f, h)| \leq \frac{3\varepsilon}{2h} + \frac{Mh^4}{30}.$$

Además, el valor de h que minimiza la expresión del miembro derecho de la desigualdad (25) es

$$(26) \quad h = \left(\frac{45\varepsilon}{4M}\right)^{1/5}.$$

Si aplicamos la fórmula (25) al Ejemplo 6.2, usando la cota $|f^{(5)}(x)| \leq |\operatorname{sen}(x)| \leq 1 = M$ y el valor $\varepsilon = 0.5 \times 10^{-9}$ como magnitud del error de redondeo, el incremento óptimo es $h = (22.5 \times 10^{-9}/4)^{1/5} = 0.022388475$. De los cuatro incrementos propuestos para usar la fórmula (10), es $h = 0.01$ el que está más cerca del valor óptimo 0.022388475 y es el que proporciona la mejor aproximación a $f'(0.8)$ (véanse la Tabla 6.2 y la Figura 6.4).

No deberíamos terminar la discusión del Ejemplo 6.2 sin mencionar que las fórmulas de derivación numérica pueden obtenerse deduciéndolas de otra manera: derivando un polinomio de interpolación. Así, por ejemplo, el polinomio interpolador de Lagrange que pasa por los puntos $(0.7, \cos(0.7))$, $(0.8, \cos(0.8))$ y $(0.9, \cos(0.9))$ es

$$\begin{aligned} p_2(x) &= 38.2421094(x - 0.8)(x - 0.9) - 69.6706709(x - 0.7)(x - 0.9) \\ &\quad + 31.0804984(x - 0.7)(x - 0.8), \end{aligned}$$

Cota del error

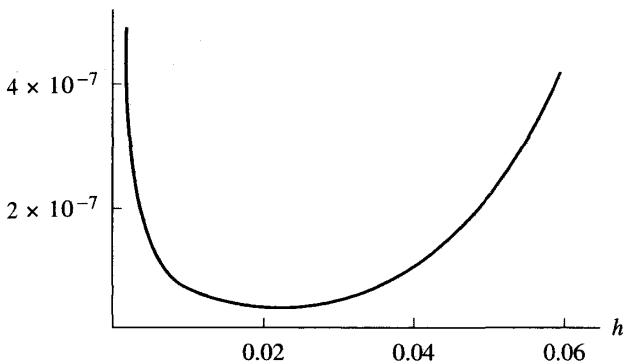


Figura 6.4 Determinación del incremento óptimo $h = 0.022388475$ cuando aplicamos la fórmula (26) a la función $f(x) = \cos(x)$ en el Ejemplo 6.2.

que podemos desarrollar para escribirlo en la forma habitual:

$$p_2(x) = 1.046875165 - 0.159260044x - 0.348063157x^2.$$

Realizando un cálculo similar podemos obtener el polinomio de grado cuatro $p_4(x)$ que pasa por los puntos $(0.6, \cos(0.6))$, $(0.7, \cos(0.7))$, $(0.8, \cos(0.8))$, $(0.9, \cos(0.9))$ y $(1.0, \cos(1.0))$:

$$\begin{aligned} p_4(x) = & 0.998452927 + 0.009638391x - 0.523291341x^2 \\ & + 0.026521229x^3 + 0.028981100x^4. \end{aligned}$$

Derivando estos polinomios, obtenemos $p'_2(0.8) = -0.716161095$ y $p'_4(0.8) = -0.717353703$, que se corresponden con los valores obtenidos para $h = 0.1$ que se muestran en la Tabla 6.2. En las Figuras 6.5(a) y (b) se muestran las gráficas de $p_2(x)$ y de $p_4(x)$ y de sus rectas tangentes en el punto $(0.8, \cos(0.8))$, respectivamente.

El método de extrapolación de Richardson

En esta subsección vamos a profundizar en la relación que hay entre las fórmulas (3) y (10). Definimos $f_k = f(x_k) = f(x_0 + kh)$ y usamos la notación $D_0(h)$ y $D_0(2h)$ para denotar las aproximaciones a $f'(x_0)$ que se obtienen al aplicar la fórmula (3) con incrementos h y $2h$, respectivamente:

$$(27) \quad f'(x_0) \approx D_0(h) + Ch^2$$

y

$$(28) \quad f'(x_0) \approx D_0(2h) + 4Ch^2.$$

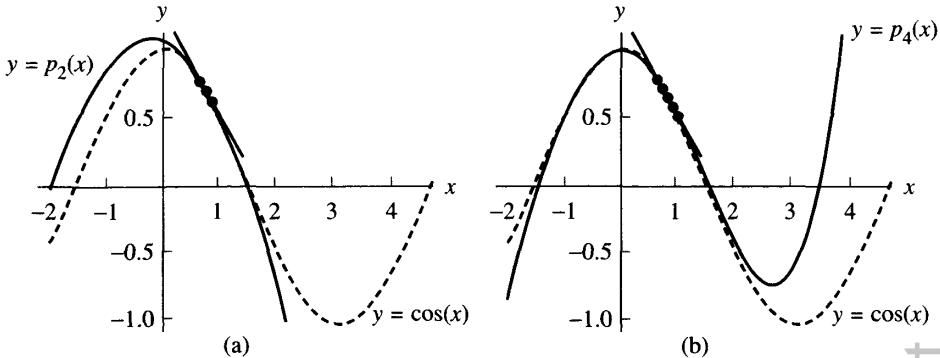


Figura 6.5 (a) Gráficas de $y = \cos(x)$ y del polinomio interpolador $p_2(x)$ que se usa para estimar $f'(0.8) \approx p'_2(0.8) = -0.716161095$. (b) Gráficas de $y = \cos(x)$ y del polinomio interpolador $p_4(x)$ usado para estimar $f'(0.8) \approx p'_4(0.8) = -0.717353703$.

Multiplicando la relación (27) por 4 y restando la relación (28) del producto resultante, los términos que contienen C se simplifican y nos queda

$$(29) \quad 3f'(x_0) \approx 4D_0(h) - D_0(2h) = \frac{4(f_1 - f_{-1})}{2h} - \frac{f_2 - f_{-2}}{4h}.$$

Ahora despejamos $f'(x_0)$ en (29) y obtenemos

$$(30) \quad f'(x_0) \approx \frac{4D_0(h) - D_0(2h)}{3} = \frac{-f_2 + 8f_1 - 8f_{-1} + f_{-2}}{12h}.$$

La expresión del miembro derecho de (30) es, precisamente, la fórmula de diferencia centrada (10).

Ejemplo 6.3. Sea $f(x) = \cos(x)$. Vamos a usar las relaciones (27) y (28) con $h = 0.01$ para mostrar cómo se usa la combinación lineal $(4D_0(h) - D_0(2h))/3$ dada en (30) para obtener la aproximación a $f'(0.8)$ dada en la fórmula (10). Trabajamos con nueve cifras decimales significativas en todas las operaciones.

Usando (27) y (28) con $h = 0.01$ obtenemos

$$\begin{aligned} D_0(h) &\approx \frac{f(0.81) - f(0.79)}{0.02} \approx \frac{0.689498433 - 0.703845316}{0.02} \\ &\approx -0.717344150 \end{aligned}$$

y

$$\begin{aligned} D_0(2h) &\approx \frac{f(0.82) - f(0.78)}{0.04} \approx \frac{0.682221207 - 0.710913538}{0.04} \\ &\approx -0.717308275. \end{aligned}$$

Ahora calculamos la combinación lineal dada en (30):

$$\begin{aligned} f'(0.8) &\approx \frac{4D_0^\bullet(h) - D_0(2h)}{3} \approx \frac{4(-0.717344150) - (-0.717308275)}{3} \\ &\approx -0.717356108, \end{aligned}$$

que es exactamente la solución obtenida para aproximar $f'(0.8)$ en el Ejemplo 6.2 al usar directamente la fórmula (10). ■

El método de obtener una fórmula de mayor orden para aproximar $f'(x_0)$ a partir de una fórmula de menor orden se llama **extrapolación**. Para llevarlo a cabo, hace falta que el término del error dado en (3) pueda desarrollarse en serie de potencias de h que contenga sólo exponentes pares. Ya hemos visto cómo podemos usar los incrementos h y $2h$ para eliminar el término que contiene h^2 . Para ver ahora cómo podemos eliminar h^4 , denotemos por $D_1(h)$ y $D_1(2h)$ las aproximaciones a $f'(x_0)$ de orden $\mathcal{O}(h^4)$ que se obtienen con la fórmula (16) usando los incrementos h y $2h$, respectivamente. Entonces

$$(31) \quad f'(x_0) = \frac{-f_2 + 8f_1 - 8f_{-1} + f_{-2}}{24h} + \frac{h^4 f^{(5)}(c_1)}{30} \approx D_1(h) + Ch^4$$

y

$$(32) \quad f'(x_0) = \frac{-f_4 + 8f_2 - 8f_{-2} + f_{-4}}{12h} + \frac{16h^4 f^{(5)}(c_2)}{30} \approx D_1(2h) + 16Ch^4.$$

Suponiendo que $f^{(5)}(x)$ tiene signo constante y no cambia demasiado rápidamente, podemos utilizar la hipótesis de que $f^{(5)}(c_1) \approx f^{(5)}(c_2)$ para eliminar los términos que contienen h^4 en las fórmulas (31) y (32); el resultado es

$$(33) \quad f'(x_0) \approx \frac{16D_1(h) - D_1(2h)}{15}.$$

El modelo general para ir mejorando los cálculos se recoge en el siguiente resultado.

Teorema 6.3 (Método de extrapolación de Richardson). Supongamos que $D_{k-1}(h)$ es una aproximación de orden $\mathcal{O}(h^{2k})$ a $f'(x_0)$ que verifica:

$$(34) \quad f'(x_0) = D_{k-1}(h) + c_1 h^{2k} + c_2 h^{2k+2} + \dots,$$

con lo cual

$$(35) \quad f'(x_0) = D_{k-1}(2h) + 4^k c_1 h^{2k} + 4^{k+1} c_2 h^{2k+2} + \dots$$

Entonces podemos construir la siguiente aproximación mejorada:

$$(36) \quad f'(x_0) = D_k(h) + \mathcal{O}(h^{2k+2}) = \frac{4^k D_{k-1}(h) - D_{k-1}(2h)}{4^k - 1} + \mathcal{O}(h^{2k+2}).$$

MATLAB

El programa que damos a continuación usa la fórmula (3) de diferencias centradas de orden $O(h^2)$ para aproximar la derivada de una función en un punto dado. Lo que se hace es generar una sucesión de aproximaciones $\{D_k\}$ en las que el intervalo centrado para D_{k+1} mide un décimo de la longitud del intervalo centrado para D_k . El programa produce una matriz $L=[H' D' E']$, en la que H es un vector que contiene los incrementos, D es un vector que contiene las aproximaciones a la derivada y E es un vector que contiene las cotas del error. Nota. Hay que introducir la función f como una cadena de caracteres; esto es, 'f'.

Programa 6.1 (Derivación numérica mediante límites). Construcción de las aproximaciones numéricas a $f'(x)$ mediante la generación de una sucesión

$$f'(x) \approx D_k = \frac{f(x + 10^{-k}h) - f(x - 10^{-k}h)}{2(10^{-k}h)} \quad \text{para } k = 0, \dots, n$$

que continúa hasta que $|D_{n+1} - D_n| \geq |D_n - D_{n-1}|$ o bien $|D_n - D_{n-1}|$ se hace menor que la tolerancia, que es el criterio con el que se trata de encontrar la mejor aproximación $D_n \approx f'(x)$.

```
function [L,n]=difflim(f,x,toler)
% Datos
%     - f es la función, introducida como una
%       cadena de caracteres 'f'
%     - x es el punto en el que se deriva
%     - toler es la tolerancia para el error
% Resultados
%     - L=[H' D' E']:
%       H es el vector de los incrementos
%       D es el vector de las aproximaciones a la derivada
%       E es el vector de las cotas del error
%     - n es la coordenada de la "mejor aproximación"
max1=15;
h=1;
H(1)=h;
D(1)=(feval(f,x+h)-feval(f,x-h))/(2*h);
E(1)=0;
R(1)=0;
for n=1:2
    h=h/10;
    H(n+1)=h;
```

```

D(n+1)=(feval(f,x+h)-feval(f,x-h))/(2*h);
E(n+1)=abs(D(n+1)-D(n));
R(n+1)=2*E(n+1)*(abs(D(n+1))+abs(D(n))+eps);
end
n=2;
while((E(n)>E(n+1))&(R(n)>toler))&n<max1
    h=h/10;
    H(n+2)=h;
    D(n+2)=(feval(f,x+h)-feval(f,x-h))/(2*h);
    E(n+2)=abs(D(n+2)-D(n+1));
    R(n+2)=2*E(n+2)*(abs(D(n+2))+abs(D(n+1))+eps);
    n=n+1;
end
n=length(D)-1;
L=[H' D' E'];

```

En el Programa 6.2 que damos a continuación se usa el método de extrapolación de Richardson dado en el Teorema 6.3, hagamos notar que la expresión que se utiliza para calcular los elementos de la fila j -ésima es algebraicamente equivalente a la fórmula (36).

Programa 6.2 (Derivación numérica usando extrapolación). Construcción de una tabla $D(j, k)$ (con $k \leq j$) de aproximaciones numéricas a $f'(x)$ en la que se utiliza $f'(x) \approx D(n, n)$ como respuesta final. Las aproximaciones $D(j, k)$ se almacenan en una matriz triangular inferior cuya primera columna viene dada por

$$D(j, 1) = \frac{f(x + 2^{-j}h) - f(x - 2^{-j}h)}{2^{-j+1}h}$$

y cuyos elementos en la fila j -ésima, para $j \geq 2$, son

$$D(j, k) = D(j, k - 1) + \frac{D(j, k - 1) - D(j - 1, k - 1)}{4^k - 1} \quad (2 \leq k \leq j).$$

```

function [D,err,relerr,n]=diffext(f,x,delta,toler)
% Datos
%     - f es la función, introducida como
%       una cadena de caracteres 'f'
%     - x es el punto en el que se deriva
%     - delta es la tolerancia para el error
%     - toler es la tolerancia para el error relativo
% Resultados
%     - D es la matriz de las aproximaciones a la derivada

```

```

%      - err es la cota del error
%      - relerr es la cota del error relativo
%      - n es la coordenada de la ‘‘mejor aproximación’’
err=1;
relerr=1;
h=1;
j=1;
D(1,1)=(feval(f,x+h)-feval(f,x-h))/(2*h);
while relerr>toler & err>delta & j<12
    h=h/2;
    D(j+1,1)=(feval(f,x+h)-feval(f,x-h))/(2*h);
    for k=1:j
        D(j+1,k+1)=D(j+1,k)+(D(j+1,k)-D(j,k))/((4^k)-1);
    end
    err=abs(D(j+1,j+1)-D(j,j));
    relerr=2*err/(abs(D(j+1,j+1))+abs(D(j,j))+eps);
    j=j+1;
end
[n,n]=size(D);

```

Ejercicios

1. Sea $f(x) = \sin(x)$, con x medido en radianes.

- (a) Calcule aproximaciones a $f'(0.8)$ usando la fórmula (3) y tomando $h = 0.1$, $h = 0.01$ y $h = 0.001$. Realice las operaciones con ocho o nueve cifras decimales significativas.
- (b) Compare los valores obtenidos con $f'(0.8) = \cos(0.8)$.
- (c) Calcule las cotas del error de truncamiento (4) tomando

$$|f^{(3)}(c)| \leq \cos(0.7) \approx 0.764842187.$$

2. Sea $f(x) = e^x$.

- (a) Calcule aproximaciones a $f'(2.3)$ usando fórmula (3) y tomando $h = 0.1$, $h = 0.01$ y $h = 0.001$. Realice las operaciones con ocho o nueve cifras decimales significativas.
- (b) Compare los valores obtenidos con $f'(2.3) = e^{2.3}$.
- (c) Calcule las cotas del error de truncamiento (4) tomando

$$|f^{(3)}(c)| \leq e^{2.4} \approx 11.02317638.$$

3. Sea $f(x) = \sin(x)$, con x medido en radianes.

- (a) Calcule aproximaciones a $f'(0.8)$ usando la fórmula (10) con $h = 0.1$ y $h = 0.01$ y compare los valores obtenidos con $f'(0.8) = \cos(0.8)$.

- (b) Calcule las aproximaciones a $f'(0.8)$ del apartado (a) usando la fórmula de extrapolación dada en (29).
- (c) Calcule las cotas del error de truncamiento (11) tomando

$$|f^{(5)}(c)| \leq \cos(0.6) \approx 0.825335615.$$

4. Sea $f(x) = e^x$.

- (a) Calcule aproximaciones a $f'(2.3)$ usando la fórmula (10) con $h = 0.1$ y $h = 0.01$ y compare los valores obtenidos con $f'(2.3) = e^{2.3}$.
- (b) Calcule las aproximaciones a $f'(2.3)$ del apartado (a) usando la fórmula de extrapolación dada en (29).
- (c) Calcule las cotas del error de truncamiento (11) tomando

$$|f^{(5)}(c)| \leq e^{2.5} \approx 12.18249396.$$

5. Compare las fórmulas de derivación numérica (3) y (10). Para ello, considere $f(x) = x^3$ y determine aproximaciones a $f'(2)$:

- (a) Usando la fórmula (3) con $h = 0.05$,
- (b) Usando la fórmula (10) con $h = 0.05$.
- (c) Calcule las cotas del error de truncamiento (4) y (11).

6. (a) Use el teorema de Taylor para probar que

$$f(x+h) = f(x) + hf'(x) + \frac{h^2 f''(c)}{2}, \quad \text{siendo } |c-x| < h.$$

- (b) Use el apartado (a) para probar que el cociente incremental de la expresión (2) tiene un error de orden $O(h) = -hf^{(2)}(c)/2$.
- (c) ¿Por qué es mejor usar la fórmula (3) que la fórmula (2)?
7. Fórmulas de derivación parcial. La derivada parcial $f_x(x, y)$ de $f(x, y)$ con respecto a x se obtiene manteniendo y fijo y derivando con respecto a x . Análogamente, $f_y(x, y)$ se obtiene manteniendo x fijo y derivando con respecto a y . La fórmula (3) puede adaptarse para calcular derivadas parciales:

$$(i) \quad f_x(x, y) = \frac{f(x+h, y) - f(x-h, y)}{2h} + O(h^2),$$

$$f_y(x, y) = \frac{f(x, y+h) - f(x, y-h)}{2h} + O(h^2).$$

- (a) Sea $f(x, y) = xy/(x+y)$. Calcule aproximaciones a $f_x(2, 3)$ y $f_y(2, 3)$ usando las fórmulas de la relación (i) con $h = 0.1, 0.01$ y 0.001 . Compare los valores obtenidos con los exactos.
- (b) Sea $z = f(x, y) = \arctan(y/x)$ con z medido en radianes. Calcule aproximaciones a $f_x(3, 4)$ y $f_y(3, 4)$ usando las fórmulas dadas en (i) con $h = 0.1, 0.01$ y 0.001 . Compare los resultados obtenidos con los valores exactos.

8. Complete los detalles que faltan para deducir la relación (33) a partir de las relaciones (31) y (32).
9. (a) Pruebe que el valor dado en (21) es el valor de h que minimiza la expresión del miembro derecho de (20).
- (b) Pruebe que el valor dado en (26) es el valor de h que minimiza la expresión del miembro derecho de (25).
10. El voltaje $E = E(t)$ en un circuito eléctrico obedece la ecuación $E(t) = L(dI/dt) + RI(t)$, donde R es la resistencia, L es la inductancia e I es la intensidad de corriente. Consideremos $L = 0.05$ henrios, $R = 2$ ohmios y los valores de la intensidad $I(t)$, en amperios, que se relacionan en la tabla siguiente.

t	$I(t)$
1.0	8.2277
1.1	7.2428
1.2	5.9908
1.3	4.5260
1.4	2.9122

- (a) Determine $I'(1.2)$ mediante derivación numérica y use este valor para calcular $E(1.2)$.
- (b) Compare su respuesta con la que se obtiene sabiendo que la expresión de $I(t)$ es $I(t) = 10e^{-t/10} \operatorname{sen}(2t)$.
11. La distancia $D = D(t)$ recorrida por un móvil se muestra en la siguiente tabla:

t	$D(t)$
8.0	17.453
9.0	21.460
10.0	25.752
11.0	30.301
12.0	35.084

- (a) Determine la velocidad $V(10)$ mediante derivación numérica.
- (b) Compare su respuesta con la que se obtiene sabiendo que la expresión de $D(t)$ es $D(t) = -70 + 7t + 70e^{-t/10}$.

12. Consideremos la siguiente tabulación de la función coseno en la que el error de redondeo está acotado por $|e_k| \leq 5 \times 10^{-6}$:

x	$f(x) = \cos(x)$
1.100	0.45360
1.190	0.37166
1.199	0.36329
1.200	0.36236
1.201	0.36143
1.210	0.35302
1.300	0.26750

Haciendo sus cálculos con los valores de la tabla:

- (a) Determine aproximaciones a $f'(1.2)$ usando la fórmula (17) con $h = 0.1$, $h = 0.01$ y $h = 0.001$.
- (b) Compare sus resultados con $f'(1.2) = -\operatorname{sen}(1.2) \approx -0.93204$.
- (c) Determine el término del error total dado en la expresión (19) para cada uno de los tres casos del apartado (a).
13. Consideremos la siguiente tabulación de la función logaritmo neperiano en la que el error de redondeo está acotado por $|e_k| \leq 5 \times 10^{-6}$:

x	$f(x) = \ln(x)$
2.900	1.06471
2.990	1.09527
2.999	1.09828
3.000	1.09861
3.001	1.09895
3.010	1.10194
3.100	1.13140

Haciendo sus cálculos con los valores de la tabla:

- (a) Determine aproximaciones a $f'(3.0)$ usando la fórmula (17) con los incrementos $h = 0.1$, $h = 0.01$ y $h = 0.001$.
- (b) Compare sus resultados con $f'(3.0) = \frac{1}{3} \approx 0.33333$.
- (c) Determine el término del error total dado en la expresión (19) para los tres casos del apartado (a).
14. Supongamos que se construye una tabla de valores de una función $f(x_k)$ que se redondean con tres cifras decimales de manera que su error de redondeo está acotado por 5×10^{-4} . Supongamos también que $|f^{(3)}(c)| \leq 1.5$ y $|f^{(5)}(c)| \leq 1.5$.
- (a) Determine el incremento óptimo h para la fórmula (17).
- (b) Determine el incremento óptimo h para la fórmula (22).

15. Consideremos la siguiente tabulación de la función coseno en la que el error de redondeo está acotado por $|e_k| \leq 5 \times 10^{-6}$.

x	$f(x) = \cos(x)$
1.000	0.54030
1.100	0.45360
1.198	0.36422
1.199	0.36329
1.200	0.36236
1.201	0.36143
1.202	0.36049
1.300	0.26750
1.400	0.16997

- (a) Haciendo sus cálculos con los valores de la tabla, aproxime $f'(1.2)$ usando la fórmula (22) con $h = 0.1$ y $h = 0.001$.
 (b) Determine el término del error total dado en la expresión (24) para los dos casos del apartado (a).
16. Consideremos la siguiente tabulación de la función logaritmo neperiano en la que el error de redondeo está acotado por $|e_k| \leq 5 \times 10^{-6}$:

x	$f(x) = \ln(x)$
2.800	1.02962
2.900	1.06471
2.998	1.09795
2.999	1.09828
3.000	1.09861
3.001	1.09895
3.002	1.09928
3.100	1.13140
3.200	1.16315

- (a) Haciendo sus cálculos con los valores de la tabla, aproxime $f'(3.0)$ la fórmula (22) con $h = 0.1$ y $h = 0.001$.
 (b) Determine el término del error total dado en la expresión (24) para los dos casos del apartado (a).

Algoritmos y programas

1. Use el Programa 6.1 para aproximar la derivada de cada una de las siguientes funciones en el punto dado x ; las aproximaciones deberían tener una precisión

de 13 cifras decimales. *Nota.* Puede que sea necesario cambiar los valores de **max1** y del punto inicial **h** del programa.

(a) $f(x) = 60x^{45} - 32x^{33} + 233x^5 - 47x^2 - 77$; $x = 1/\sqrt{3}$

(b) $f(x) = \tan\left(\cos\left(\frac{\sqrt{5} + \operatorname{sen}(x)}{1+x^2}\right)\right)$; $x = \frac{1+\sqrt{5}}{3}$

(c) $f(x) = \operatorname{sen}(\cos(1/x))$; $x = 1/\sqrt{2}$

(d) $f(x) = \operatorname{sen}(x^3 - 7x^2 + 6x + 8)$; $x = \frac{1-\sqrt{5}}{2}$

(e) $f(x) = x^{x^x}$; $x = 0.0001$

2. Modifique el Programa 6.1 de manera que se construyan las aproximaciones dadas por la fórmula (10) de orden $O(h^4)$. Use este programa para aproximar las derivadas de las funciones dadas en el Problema 1. De nuevo, la precisión debería ser de 13 cifras decimales.
3. Use el Programa 6.2 para aproximar las derivadas de las funciones dadas en el Problema 1. De nuevo, la precisión debería ser de 13 cifras decimales. *Nota.* Puede que necesite cambiar los valores iniciales de **err**, **reterr** y **h**.

6.2 Fórmulas de derivación numérica

Otras fórmulas de diferencias centradas

Las fórmulas de aproximación a $f'(x)$ de la sección anterior requieren que la función se pueda evaluar en abscisas situadas simétricamente a ambos lados del punto x , por eso se llaman fórmulas de diferencias centradas. Para obtener fórmulas de diferencias centradas que aproximen las derivadas de orden superior se puede emplear el teorema de Taylor; las elecciones más habituales son las

Tabla 6.3 Fórmulas de diferencias centradas de orden $O(h^2)$.

$$f'(x_0) \approx \frac{f_1 - f_{-1}}{2h}$$

$$f''(x_0) \approx \frac{f_1 - 2f_0 + f_{-1}}{h^2}$$

$$f^{(3)}(x_0) \approx \frac{f_2 - f_1 + 2f_{-1} - f_{-2}}{2h^3}$$

$$f^{(4)}(x_0) \approx \frac{f_2 - 4f_1 + 6f_0 - 4f_{-1} + f_{-2}}{h^4}$$

fórmulas de orden $O(h^2)$ y las de orden $O(h^4)$ que se dan en las Tablas 6.3 y 6.4. En esta sección usaremos indistintamente x y x_0 para denotar el punto donde se quiere derivar ya que la notación x_0 es más cómoda para describir las fórmulas y, así, usaremos la notación $f_k = f(x_0 + kh)$ para $k = -3, -2, -1, 0, 1, 2, 3$; notación que empleamos por primera vez en las Tablas 6.3 y 6.4.

A modo de ejemplo, vamos a deducir la fórmula de orden $O(h^2)$ para $f''(x)$ que se muestra en la Tabla 6.3. Escribimos los desarrollos en serie de Taylor:

$$(1) \quad f(x+h) = f(x) + hf'(x) + \frac{h^2 f''(x)}{2} + \frac{h^3 f^{(3)}(x)}{6} + \frac{h^4 f^{(4)}(x)}{24} + \dots$$

y

$$(2) \quad f(x-h) = f(x) - hf'(x) + \frac{h^2 f''(x)}{2} - \frac{h^3 f^{(3)}(x)}{6} + \frac{h^4 f^{(4)}(x)}{24} - \dots$$

Sumando los desarrollos (1) y (2) eliminamos los términos que contienen las derivadas impares $f'(x), f^{(3)}(x), f^{(5)}(x), \dots$:

$$(3) \quad f(x+h) + f(x-h) = 2f(x) + \frac{2h^2 f''(x)}{2} + \frac{2h^4 f^{(4)}(x)}{24} + \dots$$

Ahora despejamos $f''(x)$ de la expresión (3) y obtenemos

$$(4) \quad f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} - \frac{2h^2 f^{(4)}(x)}{4!} - \frac{2h^4 f^{(6)}(x)}{6!} - \dots - \frac{2h^{2k-2} f^{(2k)}(x)}{(2k)!} - \dots$$

Si truncamos el desarrollo en serie (4) en la cuarta derivada, entonces existe un valor c en $[x-h, x+h]$ tal que

$$(5) \quad f''(x_0) = \frac{f_1 - 2f_0 + f_{-1}}{h^2} - \frac{h^2 f^{(4)}(c)}{12},$$

Tabla 6.4 Fórmulas de diferencias centradas de orden $O(h^4)$.

$$f'(x_0) \approx \frac{-f_2 + 8f_1 - 8f_{-1} + f_{-2}}{12h}$$

$$f''(x_0) \approx \frac{-f_2 + 16f_1 - 30f_0 + 16f_{-1} - f_{-2}}{12h^2}$$

$$f^{(3)}(x_0) \approx \frac{-f_3 + 8f_2 - 13f_1 + 13f_{-1} - 8f_{-2} + f_{-3}}{8h^3}$$

$$f^{(4)}(x_0) \approx \frac{-f_3 + 12f_2 - 39f_1 + 56f_0 - 39f_{-1} + 12f_{-2} - f_{-3}}{6h^4}$$

Tabla 6.5 Aproximaciones numéricas a $f''(x)$ en el Ejemplo 6.4

Incre- mento	Aproximación con la fórmula (6)	Error con la fórmula (6)
$h = 0.1$	-0.696126300	-0.000580409
$h = 0.01$	-0.696690000	-0.000016709
$h = 0.001$	-0.696000000	-0.000706709

que es la fórmula de aproximación a $f''(x)$ deseada:

$$(6) \quad f''(x_0) \approx \frac{f_1 - 2f_0 + f_{-1}}{h^2}.$$

Ejemplo 6.4. Sea $f(x) = \cos(x)$.

- (a) Vamos a usar la fórmula (6) con $h = 0.1, 0.01$ y 0.001 para calcular aproximaciones a $f''(0.8)$ trabajando con nueve cifras decimales.
- (b) Después, compararemos estas aproximaciones con el valor exacto de la derivada segunda, que es $f''(0.8) = -\cos(0.8)$.
- (a) Los cálculos cuando $h = 0.01$ son

$$\begin{aligned} f''(0.8) &\approx \frac{f(0.81) - 2f(0.80) + f(0.79)}{0.001} \\ &\approx \frac{0.689498433 - 2(0.696706709) + 0.703845316}{0.0001} \\ &\approx -0.696690000. \end{aligned}$$

- (b) El error de la aproximación obtenida es -0.000016709 . El resto de los cálculos se resumen en la Tabla 6.5. Como antes, el análisis del error nos enseñará por qué $h = 0.01$ es la mejor elección. ■

Análisis del error

Sea $f_k = y_k + e_k$, donde e_k es el error que se tiene al calcular $f(x_k)$, incluyendo los errores de medida y el error de redondeo. Entonces podemos escribir la fórmula (6) como

$$(7) \quad f''(x_0) = \frac{y_1 - 2y_0 + y_{-1}}{h^2} + E(f, h).$$

El término del error $E(h, f)$ de la fórmula de derivación numérica dada en (7) tendrá una parte debida al error de redondeo y otra debida al error de truncamiento:

$$(8) \quad E(f, h) = \frac{e_1 - 2e_0 + e_{-1}}{h^2} - \frac{h^2 f^{(4)}(c)}{12}.$$

Si suponemos que cada error e_k es de tamaño ε , que los errores se acumulan independientemente de los signos y que $|f^{(4)}(x)| \leq M$, entonces obtenemos la siguiente cota del error:

$$(9) \quad |E(f, h)| \leq \frac{4\varepsilon}{h^2} + \frac{Mh^2}{12}.$$

Si h es pequeño, entonces el término $4\varepsilon/h^2$ debido a los errores de redondeo puede ser grande. Cuando h es grande, entonces el término $Mh^2/12$ es grande; el incremento óptimo es el valor de h que minimiza

$$(10) \quad g(h) = \frac{4\varepsilon}{h^2} + \frac{Mh^2}{12}.$$

Al igualar $g'(h) = 0$ obtenemos $-8\varepsilon/h^3 + Mh/6 = 0$, con lo cual $h^4 = 48\varepsilon/M$ de donde obtenemos el valor óptimo del incremento:

$$(11) \quad h = \left(\frac{48\varepsilon}{M}\right)^{1/4}.$$

Si aplicamos la fórmula (11) en el caso del Ejemplo 6.4, usando la cota $|f^{(4)}(x)| \leq |\cos(x)| \leq 1 = M$ y el valor $\varepsilon = 0.5 \times 10^{-9}$, entonces el incremento óptimo es $h = (24 \times 10^{-9}/1)^{1/4} = 0.01244666$, con lo que $h = 0.01$ es el incremento más próximo al óptimo de los tres propuestos en dicho ejemplo.

Puesto que la porción del error debida al redondeo es inversamente proporcional al cuadrado de h , este término aumenta conforme h disminuye; esto es lo que a veces se llama el **dilema del incremento**. Una solución parcial de este problema es el usar una fórmula de orden superior, de manera que con un valor más grande de h obtengamos la misma precisión. La fórmula de aproximación a $f''(x_0)$ de orden $O(h^4)$ en la Tabla 6.4 es

$$(12) \quad f''(x_0) = \frac{-f_2 + 16f_1 - 30f_0 + 16f_{-1} - f_{-2}}{12h^2} + E(f, h).$$

El término del error para la fórmula (12) viene dado por

$$(13) \quad E(f, h) = \frac{16\varepsilon}{3h^2} + \frac{h^4 f^{(6)}(c)}{90},$$

Tabla 6.6 Aproximaciones numéricas a $f''(x)$ en el Ejemplo 6.5.

Incre- mento	Aproximación con la fórmula (12)	Error con la fórmula (12)
$h = 1.0$	-0.689625413	-0.007081296
$h = 0.1$	-0.696705958	-0.000000751
$h = 0.01$	-0.696690000	-0.000016709

siendo c un punto del intervalo $[x - 2h, x + 2h]$. Una cota del error $|E(f, h)|$ es, entonces,

$$(14) \quad |E(f, h)| \leq \frac{16\varepsilon}{3h^2} + \frac{h^4 M}{90},$$

donde $|f^{(6)}(x)| \leq M$; de aquí se deduce fácilmente que el valor óptimo de h es

$$(15) \quad h = \left(\frac{240\varepsilon}{M} \right)^{1/6}.$$

Ejemplo 6.5. Sea $f(x) = \cos(x)$.

- (a) Vamos a usar la fórmula (12) con $h = 1.0, 0.1$ y 0.01 para calcular aproximaciones a $f''(0.8)$ trabajando con nueve cifras decimales significativas.
- (b) Despues, compararemos los valores obtenidos con el valor exacto de la derivada segunda, que es $f''(0.8) = -\cos(0.8)$.
- (c) Finalmente, determinaremos el valor óptimo del incremento.

Hacemos los cálculos con $h = 0.1$; el resto se recoge en la Tabla 6.6.

- (a) Los cálculos para $h = 0.1$ son

$$\begin{aligned} f''(0.8) & \\ &\approx \frac{-f(1.0) + 16f(0.9) - 30f(0.8) + 16f(0.7) - f(0.6)}{0.12} \\ &\approx \frac{-0.540302306 + 9.945759488 - 20.90120127 + 12.23747499 - 0.825335615}{0.12} \\ &\approx -0.696705958. \end{aligned}$$

- (b) El error de esta aproximación es -0.000000751 .

(c) Para aplicar la fórmula (15), podemos usar la cota $|f^{(6)}(x)| \leq |\cos(x)| \leq 1 = M$ y el valor $\varepsilon = 0.5 \times 10^{-9}$, con los que podemos determinar el tamaño óptimo del incremento $h = (120 \times 10^{-9}/1)^{1/6} = 0.070231219$. ■

Tabla 6.7 Fórmulas de diferencias progresivas y regresivas de orden $O(h^2)$.

$f'(x_0) \approx \frac{-3f_0 + 4f_1 - f_2}{2h}$	(diferencia progresiva)
$f'(x_0) \approx \frac{3f_0 - 4f_{-1} + f_{-2}}{2h}$	(diferencia regresiva)
$f''(x_0) \approx \frac{2f_0 - 5f_1 + 4f_2 - f_3}{h^2}$	(diferencia progresiva)
$f''(x_0) \approx \frac{2f_0 - 5f_{-1} + 4f_{-2} - f_{-3}}{h^2}$	(diferencia regresiva)
$f^{(3)}(x_0) \approx \frac{-5f_0 + 18f_1 - 24f_2 + 14f_3 - 3f_4}{2h^3}$	
$f^{(3)}(x_0) \approx \frac{5f_0 - 18f_{-1} + 24f_{-2} - 14f_{-3} + 3f_{-4}}{2h^3}$	
$f^{(4)}(x_0) \approx \frac{3f_0 - 14f_1 + 26f_2 - 24f_3 + 11f_4 - 2f_5}{h^4}$	
$f^{(4)}(x_0) \approx \frac{3f_0 - 14f_{-1} + 26f_{-2} - 24f_{-3} + 11f_{-4} - 2f_{-5}}{h^4}$	

Generalmente, a la hora de calcular numéricamente una derivada, lo que se consigue es una precisión que suele ser, aproximadamente, la mitad de la precisión del propio computador. Esta pérdida severa de cifras significativas tendrá lugar casi siempre, a no ser que tengamos la suerte de hallar el valor óptimo del incremento; por tanto, hay que proceder con cuidado cuando realizamos derivaciones numéricas. Las dificultades se agudizan cuando trabajamos con datos experimentales, ya que los valores de las funciones se han redondeado y sólo tienen unas pocas cifras significativas; en este caso, si tenemos que obtener una derivada numérica a partir de los datos, es preferible considerar una curva de ajuste en mínimos cuadrados y derivar la fórmula obtenida para dicha curva.

Derivada del polinomio interpolador de Lagrange

Si sólo podemos evaluar la función en abscisas que están a un lado de x_0 , entonces las fórmulas de diferencias centradas no pueden usarse. Las fórmulas que utilizan abscisas equiespaciadas que están todas a la derecha (o izquierda) de x_0 se llaman fórmulas de diferencias progresivas (o regresivas). Estas fórmulas pueden deducirse derivando el polinomio interpolador de Lagrange y algunas de las más comunes se relacionan en la Tabla 6.7.

Ejemplo 6.6. Vamos a deducir la fórmula de diferencia progresiva

$$f''(x_0) \approx \frac{2f_0 - 5f_1 + 4f_2 - f_3}{h^2}.$$

Empezamos con el polinomio interpolador de Lagrange de $f(t)$ para los cuatro nodos x_0, x_1, x_2 y x_3 :

$$\begin{aligned} f(t) \approx & f_0 \frac{(t-x_1)(t-x_2)(t-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + f_1 \frac{(t-x_0)(t-x_2)(t-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \\ & + f_2 \frac{(t-x_0)(t-x_1)(t-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + f_3 \frac{(t-x_0)(t-x_1)(t-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}. \end{aligned}$$

Derivando dos veces los productos de los numeradores obtenemos:

$$\begin{aligned} f''(t) \approx & f_0 \frac{2((t-x_1)+(t-x_2)+(t-x_3))}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} \\ & + f_1 \frac{2((t-x_0)+(t-x_2)+(t-x_3))}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \\ & + f_2 \frac{2((t-x_0)+(t-x_1)+(t-x_3))}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} \\ & + f_3 \frac{2((t-x_0)+(t-x_1)+(t-x_2))}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}. \end{aligned}$$

Sustituyendo $t = x_0$ y usando que $x_i - x_j = (i-j)h$, nos queda

$$\begin{aligned} f''(x_0) \approx & f_0 \frac{2((x_0-x_1)+(x_0-x_2)+(x_0-x_3))}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} \\ & + f_1 \frac{2((x_0-x_0)+(x_0-x_2)+(x_0-x_3))}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \\ & + f_2 \frac{2((x_0-x_0)+(x_0-x_1)+(x_0-x_3))}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} \\ & + f_3 \frac{2((x_0-x_0)+(x_0-x_1)+(x_0-x_2))}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \\ = & f_0 \frac{2((-h)+(-2h)+(-3h))}{(-h)(-2h)(-3h)} + f_1 \frac{2((0)+(-2h)+(-3h))}{(h)(-h)(-2h)} \\ & + f_2 \frac{2((0)+(-h)+(-3h))}{(2h)(h)(-h)} + f_3 \frac{2((0)+(-h)+(-2h))}{(3h)(2h)(h)} \\ = & f_0 \frac{-12h}{-6h^3} + f_1 \frac{-10h}{2h^3} + f_2 \frac{-8h}{-2h^3} + f_3 \frac{-6h}{6h^3} = \frac{2f_0 - 5f_1 + 4f_2 - f_3}{h^2}, \end{aligned}$$

lo que establece la fórmula. ■

Ejemplo 6.7. Vamos a deducir la fórmula de diferencia progresiva

$$f'''(x_0) \approx \frac{-5f_0 + 18f_1 - 24f_2 + 14f_3 - 3f_4}{2h^3}.$$

Empezamos con el polinomio interpolador de Lagrange de $f(t)$ para los cinco nodos x_0, x_1, x_2, x_3 y x_4 .

$$\begin{aligned} f(t) \approx & f_0 \frac{(t - x_1)(t - x_2)(t - x_3)(t - x_4)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)(x_0 - x_4)} \\ & + f_1 \frac{(t - x_0)(t - x_2)(t - x_3)(t - x_4)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)(x_1 - x_4)} \\ & + f_2 \frac{(t - x_0)(t - x_1)(t - x_3)(t - x_4)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)(x_2 - x_4)} \\ & + f_3 \frac{(t - x_0)(t - x_1)(t - x_2)(t - x_4)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)(x_3 - x_4)} \\ & + f_4 \frac{(t - x_0)(t - x_1)(t - x_2)(t - x_3)}{(x_4 - x_0)(x_4 - x_1)(x_4 - x_2)(x_4 - x_3)}. \end{aligned}$$

Derivando tres veces los numeradores y sustituyendo $x_i - x_j = (i - j)h$ en los denominadores, obtenemos

$$\begin{aligned} f'''(t) \approx & f_0 \frac{6((t - x_1) + (t - x_2) + (t - x_3) + (t - x_4))}{(-h)(-2h)(-3h)(-4h)} \\ & + f_1 \frac{6((t - x_0) + (t - x_2) + (t - x_3) + (t - x_4))}{(h)(-h)(-2h)(-3h)} \\ & + f_2 \frac{6((t - x_0) + (t - x_1) + (t - x_3) + (t - x_4))}{(2h)(h)(-h)(2h)} \\ & + f_3 \frac{6((t - x_0) + (t - x_1) + (t - x_2) + (t - x_4))}{(3h)(2h)(h)(-h)} \\ & + f_4 \frac{6((t - x_0) + (t - x_1) + (t - x_2) + (t - x_3))}{(4h)(3h)(2h)(h)}. \end{aligned}$$

Finalmente, al sustituir $t = x_0$, de manera que $t - x_j = x_0 - x_j = -jh$, nos queda

$$\begin{aligned} f'''(x_0) \approx & f_0 \frac{6((-h) + (-2h) + (-3h) + (-4h))}{24h^4} \\ & + f_1 \frac{6((0) + (-2h) + (-3h) + (-4h))}{-6h^4} \\ & + f_2 \frac{6((0) + (-h) + (-3h) + (-4h))}{4h^4} \\ & + f_3 \frac{6((0) + (-h) + (-2h) + (-4h))}{-6h^4} \\ & + f_4 \frac{6((0) + (-h) + (-2h) + (-3h))}{24h^4}, \\ = & f_0 \frac{-60h}{24h^4} + f_1 \frac{54h}{6h^4} + f_2 \frac{-48h}{4h^4} + f_3 \frac{42h}{6h^4} + f_4 \frac{-36h}{24h^4} \\ = & \frac{-5f_0 + 18f_1 - 24f_2 + 14f_3 - 3f_4}{2h^3}, \end{aligned}$$

lo que establece la fórmula.

Derivada del polinomio interpolador de Newton

En esta sección vamos a mostrar la relación que existe entre las fórmulas de orden $O(h^2)$ para aproximar $f'(x_0)$ y un algoritmo general que nos permite calcular derivadas numéricamente. En la Sección 4.3 vimos que el polinomio interpolador de Newton $P(t)$ de grado $N = 2$ que aproxima a $f(t)$ usando los nodos t_0 , t_1 y t_2 viene dado por

$$(16) \quad P(t) = a_0 + a_1(t - t_0) + a_2(t - t_0)(t - t_1),$$

siendo $a_0 = f(t_0)$, $a_1 = (f(t_1) - f(t_0))/(t_1 - t_0)$ y

$$a_2 = \frac{\frac{f(t_2) - f(t_1)}{t_2 - t_1} - \frac{f(t_1) - f(t_0)}{t_1 - t_0}}{(t_2 - t_0)}.$$

La derivada de $P(t)$ es

$$(17) \quad P'(t) = a_1 + a_2((t - t_0) + (t - t_1))$$

que, cuando se evalúa en $t = t_0$, produce

$$(18) \quad P'(t_0) = a_1 + a_2(t_0 - t_1) \approx f'(t_0).$$

Observemos que en las fórmulas (16), (17) y (18) no hace falta que los nodos $\{t_k\}$ estén equiespaciados. Ordenando los nodos de maneras distintas obtendremos fórmulas de aproximación a $f'(x)$ distintas.

Caso (i): Si $t_0 = x$, $t_1 = x + h$ y $t_2 = x + 2h$, entonces

$$\begin{aligned} a_1 &= \frac{f(x + h) - f(x)}{h}, \\ a_2 &= \frac{f(x) - 2f(x + h) + f(x + 2h)}{2h^2} \end{aligned}$$

y, al sustituir estos valores en (18), obtenemos

$$P'(x) = \frac{f(x + h) - f(x)}{h} + \frac{-f(x) + 2f(x + h) - f(x + 2h)}{2h}.$$

Simplificando un poco, nos queda

$$(19) \quad P'(x) = \frac{-3f(x) + 4f(x + h) - f(x + 2h)}{2h} \approx f'(x),$$

que es la fórmula de diferencias progresivas de segundo orden para $f'(x)$.

Caso (ii): Si $t_0 = x$, $t_1 = x + h$ y $t_2 = x - h$, entonces

$$a_1 = \frac{f(x+h) - f(x)}{h},$$

$$a_2 = \frac{f(x+h) - 2f(x) + f(x-h)}{2h^2}.$$

y, al sustituir estos valores en (18), obtenemos

$$P'(x) = \frac{f(x+h) - f(x)}{h} + \frac{-f(x+h) + 2f(x) - f(x-h)}{2h}.$$

Simplificando un poco, nos queda

$$(20) \quad P'(x) = \frac{f(x+h) - f(x-h)}{2h} \approx f'(x),$$

que es la fórmula de diferencias centradas de segundo orden para $f'(x)$.

Caso (iii): Si $t_0 = x$, $t_1 = x - h$ y $t_2 = x - 2h$, entonces

$$a_1 = \frac{f(x) - f(x-h)}{h},$$

$$a_2 = \frac{f(x) - 2f(x-h) + f(x-2h)}{2h^2}.$$

y, al sustituir estos valores en (18) y simplificar, obtenemos

$$(21) \quad P'(x) = \frac{3f(x) - 4f(x-h) + f(x-2h)}{2h} \approx f'(x),$$

que es la fórmula de diferencias regresivas de segundo orden para $f'(x)$.

El polinomio interpolador de Newton $P(t)$ de grado N que aproxima $f(t)$ usando los nodos t_0, t_1, \dots, t_N viene dado por

$$(22) \quad P(t) = a_0 + a_1(t - t_0) + a_2(t - t_0)(t - t_1) \\ + a_3(t - t_0)(t - t_1)(t - t_2) + \cdots + a_N(t - t_0) \cdots (t - t_{N-1}).$$

La derivada de $P(t)$ es

$$(23) \quad P'(t) = a_1 + a_2((t - t_0) + (t - t_1)) \\ + a_3((t - t_0)(t - t_1) + (t - t_0)(t - t_2) + (t - t_1)(t - t_2)) \\ + \cdots + a_N \sum_{k=0}^{N-1} \prod_{\substack{j=0 \\ j \neq k}}^{N-1} (t - t_j).$$

Cuando evaluamos $P'(t)$ en $t = t_0$, varios de los sumandos son cero, así que podemos escribir $P'(t_0)$ de forma simple como

$$(24) \quad P'(t_0) = a_1 + a_2(t_0 - t_1) + a_3(t_0 - t_1)(t_0 - t_2) + \cdots + a_N(t_0 - t_1)(t_0 - t_2)(t_0 - t_3) \cdots (t_0 - t_{N-1}).$$

La suma parcial k -ésima del miembro derecho de la relación (24) es la derivada del polinomio interpolador de Newton de grado k para los k primeros nodos. Si

$$|t_0 - t_1| \leq |t_0 - t_2| \leq \cdots \leq |t_0 - t_N|$$

y si $\{t_j\}_{j=0}^N$ es un conjunto equiespaciado (quizá reordenándolos) de $N + 1$ nodos, entonces la suma parcial k -ésima es una aproximación a $f'(t_0)$ de orden $O(h^{k-1})$.

Supongamos, por ejemplo, que $N = 5$. Si los cinco nodos son $t_k = x + hk$ para $k = 0, 1, 2, 3$ y 4 , entonces la fórmula (24) es una manera equivalente de calcular la fórmula de diferencias progresivas para aproximar $f'(x)$ de orden $O(h^4)$. Si los cinco nodos $\{t_k\}$ son $t_0 = x, t_1 = x + h, t_2 = x - h, t_3 = x + 2h$ y $t_4 = x - 2h$, entonces (24) es la fórmula de diferencias centradas para aproximar $f'(x)$ de orden $O(h^4)$. Cuando los cinco nodos son $t_k = x - kh$, entonces (24) es la fórmula de diferencias regresivas para aproximar $f'(x)$ de orden $O(h^4)$.

MATLAB

El programa que damos a continuación es una extensión del Programa 4.2 que podemos usar para construir la fórmula (24). Hagamos notar que los nodos no tienen por qué estar equiespaciados y que sólo se calcula la derivada $f'(x_0)$ en un punto.

Programa 6.3 (Derivación basada en $N + 1$ nodos). Construcción del polinomio interpolador de Newton de grado N

$$P(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + a_3(x - x_0)(x - x_1)(x - x_2) + \cdots + a_N(x - x_0) \cdots (x - x_{N-1})$$

para aproximar numéricamente $f'(x)$ usando $f'(x_0) \approx P'(x_0)$ como respuesta final. El método debe usarse sólo con x_0 . Los nodos pueden reordenarse como $\{x_k, x_0, \dots, x_{k-1}, x_{k+1}, \dots, x_N\}$ si se desea calcular $f'(x_k) \approx P'(x_k)$.

```
function [A,df]=diffnew(X,Y)
% Datos
```

```

%      - X es un vector 1 x n que contiene las abscisas
%      - Y es un vector 1 x n que contiene las ordenadas
% Resultados
%      - A es un vector 1 x n que contiene los coeficientes del
%          polinomio de Newton de grado N
%      - df es la derivada aproximada

A=Y;
N=length(X);

for j=2:N
    for k=N:-1:j
        A(k)=(A(k)-A(k-1))/(X(k)-X(k-j+1));
    end
end

x0=X(1);
df=A(2);
prod=1;
n1=length(A)-1;

for k=2:n1
    prod=prod*(x0-X(k));
    df=df+prod*A(k+1);
end

```

Ejercicios

1. Sea $f(x) = \ln(x)$. Trabajando con ocho o nueve cifras decimales:
 - (a) Use la fórmula (6) con $h = 0.05$ para aproximar $f''(5)$.
 - (b) Use la fórmula (6) con $h = 0.01$ para aproximar $f''(5)$.
 - (c) Use la fórmula (12) con $h = 0.1$ para aproximar $f''(5)$.
 - (d) ¿Qué respuesta, (a), (b) o (c), es más precisa?
2. Sea $f(x) = \cos(x)$. Trabajando con ocho o nueve cifras decimales:
 - (a) Use la fórmula (6) con $h = 0.05$ para aproximar $f''(1)$.
 - (b) Use la fórmula (6) con $h = 0.01$ para aproximar $f''(1)$.
 - (c) Use la fórmula (12) con $h = 0.1$ para aproximar $f''(1)$.
 - (d) ¿Qué respuesta, (a), (b) o (c), es más precisa?

3. Considere la siguiente tabulación de la función $f(x) = \ln(x)$ donde los valores están redondeados con cuatro cifras decimales.

x	$f(x) = \ln(x)$
4.90	1.5892
4.95	1.5994
5.00	1.6094
5.05	1.6194
5.10	1.6292

- (a) Use la fórmula (6) con $h = 0.05$ para aproximar $f''(5)$.
(b) Use la fórmula (6) con $h = 0.01$ para aproximar $f''(5)$.
(c) Use la fórmula (12) con $h = 0.05$ para aproximar $f''(5)$.
(d) ¿Qué respuesta, (a), (b) o (c), es más precisa?

4. Considere la siguiente tabulación de la función $f(x) = \cos(x)$ donde los valores están redondeados con cuatro cifras decimales.

x	$f(x) = \cos(x)$
0.90	0.6216
0.95	0.5817
1.00	0.5403
1.05	0.4976
1.10	0.4536

$$f^{(3)}(x) \approx \frac{f(x+2h) - 2f(x+h) + 2f(x-h) - f(x-2h)}{2h^3}.$$

8. Usando la fórmula de Taylor para $f(x+h)$, $f(x-h)$, $f(x+2h)$ y $f(x-2h)$, deduzca la fórmula de diferencias centradas:

$$f^{(4)}(x) \approx \frac{f(x+2h) - 4f(x+h) + 6f(x) - 4f(x-h) + f(x-2h)}{h^4}.$$

9. Determine las aproximaciones de orden $O(h^2)$ a $f'(x_k)$ en cada uno de los cuatro puntos de las siguientes tablas:

(a)

x	$f(x)$
0.0	0.989992
0.1	0.999135
0.2	0.998295
0.3	0.987480

(b)

x	$f(x)$
0.0	0.141120
0.1	0.041581
0.2	-0.058374
0.3	-0.157746

10. Utilice las aproximaciones

$$f'\left(x + \frac{h}{2}\right) \approx \frac{f_1 - f_0}{h} \quad \text{y} \quad f'\left(x - \frac{h}{2}\right) \approx \frac{f_0 - f_{-1}}{h}$$

para deducir la fórmula

$$f''(x) \approx \frac{f_1 - 2f_0 + f_{-1}}{h^2}.$$

11. Use las fórmulas (16), (17) y (18) para deducir una fórmula de aproximación a $f'(x)$ que utilice las abscisas $t_0 = x$, $t_1 = x + h$ y $t_2 = x + 3h$.
12. Use las fórmulas (16), (17) y (18) para deducir una fórmula de aproximación a $f'(x)$ que utilice las abscisas $t_0 = x$, $t_1 = x - h$ y $t_2 = x + 2h$.
13. La resolución numérica de cierta ecuación diferencial requiere una aproximación de orden $O(h^2)$ a $f''(x) + f'(x)$.
- (a) Determine la fórmula de diferencias centradas para aproximar la expresión $f''(x) + f'(x)$ que se obtiene sumando las fórmulas de orden $O(h^2)$ para aproximar $f'(x)$ y $f''(x)$.
- (b) Determine la fórmula de diferencias progresivas para aproximar la expresión $f''(x) + f'(x)$ que se obtiene sumando las fórmulas de orden $O(h^2)$ para aproximar $f'(x)$ y $f''(x)$.
- (c) ¿Qué ocurre si una fórmula de orden $O(h^4)$ para aproximar $f'(x)$ se suma a una fórmula de orden $O(h^2)$ para aproximar $f''(x)$?
14. Critique el siguiente razonamiento: Podemos usar la fórmula de Taylor para obtener las respresentaciones

$$f(x+h) = f(x) + hf'(x) + \frac{h^2 f''(x)}{2} + \frac{h^3 f^{(3)}(c)}{6}$$

y

$$f(x - h) = f(x) - hf'(x) + \frac{h^2 f''(x)}{2} - \frac{h^3 f'''(c)}{6}.$$

Sumando estas relaciones obtenemos

$$f(x + h) + f(x - h) = 2f(x) + h^2 f''(x),$$

de donde podemos despejar $f''(x)$ y obtener, así, una fórmula exacta:

$$f''(x) = \frac{f(x + h) - 2f(x) + f(x - h)}{h^2}.$$

Algoritmos y programas

1. Modifique el Programa 6.3 de manera que sirva para calcular $P'(x_k)$ para $k = 1, 2, \dots, N + 1$.

Integración numérica

La integración numérica es una herramienta esencial que se usa en la ciencia y la ingeniería para obtener valores aproximados de integrales definidas que no pueden calcularse analíticamente. Por ejemplo, en el campo de la termodinámica estadística, el modelo de Debye para calcular la capacidad calórica de un sólido considera la siguiente función

$$\Phi(x) = \int_0^x \frac{t^3}{e^t - 1} dt.$$

Puesto que no hay una expresión analítica para $\Phi(x)$, debemos usar algún método de integración numérica para calcular sus valores. Por ejemplo, el valor

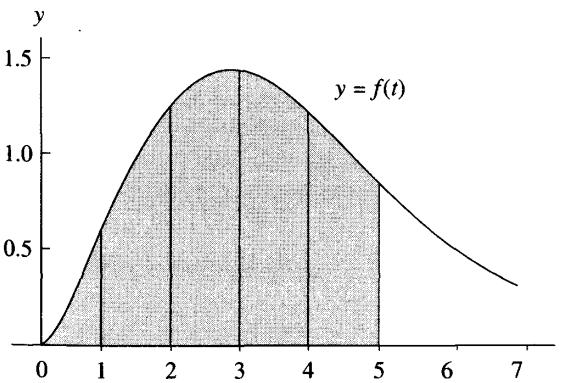


Figura 7.1 El área limitada por la curva $y = f(t)$ para $0 \leq t \leq 5$.

Tabla 7.1 Valores de $\Phi(x)$.

x	$\Phi(x)$
1.0	0.2248052
2.0	1.1763426
3.0	2.5522185
4.0	3.8770542
5.0	4.8998922
6.0	5.5858554
7.0	6.0031690
8.0	6.2396238
9.0	6.3665739
10.0	6.4319219

$\Phi(5)$ es el área bajo la curva $y = f(t) = t^3/(e^t - 1)$ para $0 \leq t \leq 5$ (véase la Figura 7.1). La aproximación numérica a $\Phi(5)$ es

$$\Phi(5) = \int_0^5 \frac{t^3}{e^t - 1} dt \approx 4.8998922.$$

Cualquier otro valor de $\Phi(x)$ debe ser hallado mediante una integración numérica; la Tabla 7.1 muestra algunas de estas aproximaciones en el intervalo $[1, 10]$.

El propósito de este capítulo es desarrollar los principios básicos de la integración numérica. En el Capítulo 9 usaremos las fórmulas de integración numérica para construir los métodos de predicción y corrección que se usan en la resolución numérica de ecuaciones diferenciales.

7.1 Introducción a la integración numérica

Afrontamos ahora el problema de la integración numérica. Nuestro objetivo es aproximar la integral definida de una función $f(x)$ en un intervalo $[a, b]$ evaluando $f(x)$ en un número finito de puntos.

Definición 7.1. Supongamos que $a = x_0 < x_1 < \cdots < x_M = b$. Una fórmula del tipo

$$(1) \quad Q[f] = \sum_{k=0}^M w_k f(x_k) = w_0 f(x_0) + w_1 f(x_1) + \cdots + w_M f(x_M)$$

de manera que

$$(2) \quad \int_a^b f(x) dx = Q[f] + E[f]$$

se llama fórmula de **integración numérica** o de **cuadratura**; el término $E[f]$ se llama **error de truncamiento** de la fórmula; los valores $\{x_k\}_{k=0}^M$ se llaman **nodos de integración** o **nodos de cuadratura** y los valores $\{w_k\}_{k=0}^M$ se llaman **pesos** de la fórmula. ▲

Los nodos $\{x_k\}$ se eligen de diferentes maneras, dependiendo de la situación concreta en la que queramos aplicar una fórmula. Para la regla del trapecio, la regla de Simpson y la regla de Boole, los nodos se toman equiespaciados. Para las fórmulas de Gauss-Legendre, los nodos que se toman son raíces de polinomios de Legendre. Cuando se usa una fórmula de cuadratura para desarrollar una fórmula predictora en la resolución numérica de ecuaciones diferenciales, los nodos elegidos son todos menores que b . Otro aspecto importante en todas las aplicaciones será también el conocimiento del grado de precisión de la solución numérica.

Definición 7.2. El **grado de precisión** de una fórmula de cuadratura es el número natural n que verifica lo siguiente: $E[P_i] = 0$ para todos los polinomios $P_i(x)$ de grado $i \leq n$, y existe un polinomio $P_{n+1}(x)$ de grado $n + 1$ tal que $E[P_{n+1}] \neq 0$. ▲

La forma del término del error $E[P_i]$ puede anticiparse estudiando qué ocurre cuando $f(x)$ es un polinomio: Consideremos un polinomio arbitrario

$$P_i(x) = a_i x^i + a_{i-1} x^{i-1} + \cdots + a_1 x + a_0$$

de grado i . Si $i \leq n$, entonces $P_i^{(n+1)}(x) \equiv 0$ para todo x , mientras que si $i = n + 1$ entonces $P_{n+1}^{(n+1)}(x) = (n + 1)! a_{n+1}$ para todo x . Por tanto, no debe sorprender que la forma general del error de truncamiento sea

$$(3) \quad E[f] = K f^{(n+1)}(c),$$

donde K es una constante adecuada y n es el grado de precisión de la fórmula. La demostración de este resultado de carácter general puede encontrarse en textos avanzados sobre integración numérica.

La deducción de las fórmulas de cuadratura puede hacerse a partir de la interpolación polinomial. Recordemos que existe un único polinomio $P_M(x)$ de grado menor o igual que M que pasa por $M + 1$ puntos dados $\{(x_k, y_k)\}_{k=0}^M$ cuyas abscisas están equiespaciadas. Cuando usamos este polinomio para aproximar la función $f(x)$ en $[a, b]$, de manera que $y_k = f(x_k)$, y luego aproximamos la integral de $f(x)$ por la integral de $P_M(x)$, la fórmula resultante se llama **fórmula de cuadratura de Newton-Cotes** (véase la Figura 7.2). Si el primer nodo es $x_0 = a$ y el último es $x_M = b$, entonces se dice que la fórmula de Newton-Cotes es **cerrada**. En el siguiente resultado se recogen las fórmulas que se obtienen cuando el polinomio de aproximación es de grado $M = 1, 2, 3$ y 4 .

Teorema 7.1 (Fórmulas de cuadratura cerradas de Newton-Cotes). Supongamos que $x_k = x_0 + kh$ ($k = 0, 1, \dots, M$) son nodos equiespaciados y sea $f_k = f(x_k)$ para cada $k = 0, 1, \dots, M$. Las cuatro primeras fórmulas cerradas de Newton-Cotes son

$$(4) \quad \int_{x_0}^{x_1} f(x) dx \approx \frac{h}{2}(f_0 + f_1) \quad (\text{regla del trapecio}),$$

$$(5) \quad \int_{x_0}^{x_2} f(x) dx \approx \frac{h}{3}(f_0 + 4f_1 + f_2) \quad (\text{regla de Simpson}),$$

$$(6) \quad \int_{x_0}^{x_3} f(x) dx \approx \frac{3h}{8}(f_0 + 3f_1 + 3f_2 + f_3) \quad (\text{regla } \frac{3}{8} \text{ de Simpson}),$$

$$(7) \quad \int_{x_0}^{x_4} f(x) dx \approx \frac{2h}{45}(7f_0 + 32f_1 + 12f_2 + 32f_3 + 7f_4)$$

(regla de Boole).

Corolario 7.1 (Precisión de las fórmulas de Newton-Cotes). Supongamos que $f(x)$ es suficientemente derivable; entonces el término del error de truncamiento $E[f]$ de las fórmulas de Newton-Cotes contiene una derivada de orden superior adecuada evaluada en un cierto punto $c \in (a, b)$. La regla del trapecio tiene un grado de precisión de $n = 1$ y si $f \in C^2[a, b]$, entonces

$$(8) \quad \int_{x_0}^{x_1} f(x) dx = \frac{h}{2}(f_0 + f_1) - \frac{h^3}{12}f^{(2)}(c).$$

La regla de Simpson tiene un grado de precisión de $n = 3$ y si $f \in C^4[a, b]$, entonces

$$(9) \quad \int_{x_0}^{x_2} f(x) dx = \frac{h}{3}(f_0 + 4f_1 + f_2) - \frac{h^5}{90}f^{(4)}(c).$$

La regla $\frac{3}{8}$ de Simpson tiene un grado de precisión de $n = 3$ y si $f \in C^4[a, b]$, entonces

$$(10) \quad \int_{x_0}^{x_3} f(x) dx = \frac{3h}{8}(f_0 + 3f_1 + 3f_2 + f_3) - \frac{3h^5}{80}f^{(4)}(c).$$

La regla de Boole tiene un grado de precisión de $n = 5$ y si $f \in C^6[a, b]$, entonces

$$(11) \quad \int_{x_0}^{x_4} f(x) dx = \frac{2h}{45}(7f_0 + 32f_1 + 12f_2 + 32f_3 + 7f_4) - \frac{8h^7}{945}f^{(6)}(c).$$

Demostración del Teorema 7.1. Partimos del polinomio interpolador de Lagrange $P_M(x)$ para los nodos x_0, x_1, \dots, x_M que se usa para aproximar $f(x)$:

$$(12) \quad f(x) \approx P_M(x) = \sum_{k=0}^M f_k L_{M,k}(x),$$

siendo $f_k = f(x_k)$ para $k = 0, 1, \dots, M$. Obtenemos una aproximación a la integral de $f(x)$ reemplazando el integrando por el polinomio $P_M(x)$; este es el procedimiento general para deducir las fórmulas de cuadratura de Newton-Cotes:

$$(13) \quad \begin{aligned} \int_{x_0}^{x_M} f(x) dx &\approx \int_{x_0}^{x_M} P_M(x) dx \\ &= \int_{x_0}^{x_M} \left(\sum_{k=0}^M f_k L_{M,k}(x) \right) dx = \sum_{k=0}^M \left(\int_{x_0}^{x_M} f_k L_{M,k}(x) dx \right) \\ &= \sum_{k=0}^M \left(\int_{x_0}^{x_M} L_{M,k}(x) dx \right) f_k = \sum_{k=0}^M w_k f_k. \end{aligned}$$

Los detalles para determinar los pesos w_k que aparecen en (13) en el caso general son muy tediosos; haremos, a modo de ejemplo, los cálculos para la regla de Simpson, que es el caso $M = 2$. En este caso el polinomio interpolador es

$$(14) \quad P_2(x) = f_0 \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} + f_1 \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} + f_2 \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}.$$

Puesto que f_0, f_1 y f_2 son constantes a la hora de integrar, las relaciones dadas en (13) quedan, en este caso,

$$(15) \quad \begin{aligned} \int_{x_0}^{x_2} f(x) dx &\approx f_0 \int_{x_0}^{x_2} \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} dx + f_1 \int_{x_0}^{x_2} \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} dx \\ &\quad + f_2 \int_{x_0}^{x_2} \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} dx. \end{aligned}$$

Hacemos el cambio de variable $x = x_0 + ht$, de manera que $dx = h dt$, para evaluar las integrales de (15). Los nuevos límites de integración son $t = 0$ y $t = 2$. Como los nodos $x_k = x_0 + kh$ están equiespaciados, podemos escribir $x_k - x_j = (k - j)h$ y $x - x_k = h(t - k)$ y usamos esto para simplificar las

integrales de (15) y escribir

$$\begin{aligned}
 (16) \quad & \int_{x_0}^{x_2} f(x) dx \approx f_0 \int_0^2 \frac{h(t-1)h(t-2)}{(-h)(-2h)} h dt + f_1 \int_0^2 \frac{h(t-0)h(t-2)}{(h)(-h)} h dt \\
 & + f_2 \int_0^2 \frac{h(t-0)h(t-1)}{(2h)(h)} h dt \\
 & = f_0 \frac{h}{2} \int_0^2 (t^2 - 3t + 2) dt - f_1 h \int_0^2 (t^2 - 2t) dt + f_2 \frac{h}{2} \int_0^2 (t^2 - t) dt \\
 & = f_0 \frac{h}{2} \left(\frac{t^3}{3} - \frac{3t^2}{2} + 2t \right) \Big|_{t=0}^{t=2} - f_1 h \left(\frac{t^3}{3} - t^2 \right) \Big|_{t=0}^{t=2} + f_2 \frac{h}{2} \left(\frac{t^3}{3} - \frac{t^2}{2} \right) \Big|_{t=0}^{t=2} \\
 & = f_0 \frac{h}{2} \left(\frac{2}{3} \right) - f_1 h \left(\frac{-4}{3} \right) + f_2 \frac{h}{2} \left(\frac{2}{3} \right) \\
 & = \frac{h}{3}(f_0 + 4f_1 + f_2),
 \end{aligned}$$

lo que completa la demostración. Posponemos una prueba de un caso particular del Corolario 7.1 hasta la Sección 7.2.

Ejemplo 7.1. Consideremos la función $f(x) = 1 + e^{-x} \operatorname{sen}(4x)$, los nodos de cuadratura equiespaciados $x_0 = 0.0$, $x_1 = 0.5$, $x_2 = 1.0$, $x_3 = 1.5$ y $x_4 = 2.0$ y los valores correspondientes de la función $f_0 = 1.00000$, $f_1 = 1.55152$, $f_2 = 0.72159$, $f_3 = 0.93765$ y $f_4 = 1.13390$. Vamos a aplicar las fórmulas de cuadratura (4) a (7).

El incremento es $h = 0.5$ y los cálculos son, entonces,

$$\begin{aligned}
 \int_0^{0.5} f(x) dx & \approx \frac{0.5}{2}(1.00000 + 1.55152) = 0.63788, \\
 \int_0^{1.0} f(x) dx & \approx \frac{0.5}{3}(1.00000 + 4(1.55152) + 0.72159) = 1.32128, \\
 \int_0^{1.5} f(x) dx & \approx \frac{3(0.5)}{8}(1.00000 + 3(1.55152) + 3(0.72159) + 0.93765) \\
 & = 1.64193, \\
 \int_0^{2.0} f(x) dx & \approx \frac{2(0.5)}{45}(7(1.00000) + 32(1.55152) + 12(0.72159) \\
 & + 32(0.93765) + 7(1.13390)) = 2.29444. \quad \blacksquare
 \end{aligned}$$

Es importante tener en cuenta que las fórmulas de cuadratura (4) a (7) aplicadas en este ejemplo proporcionan aproximaciones a integrales definidas en intervalos diferentes. La gráfica de la curva $y = f(x)$ y las áreas limitadas por los polinomios interpoladores de Lagrange $y = P_1(x)$, $y = P_2(x)$, $y = P_3(x)$ e $y = P_4(x)$ se muestran en las Figuras 7.2(a)–(d), respectivamente.

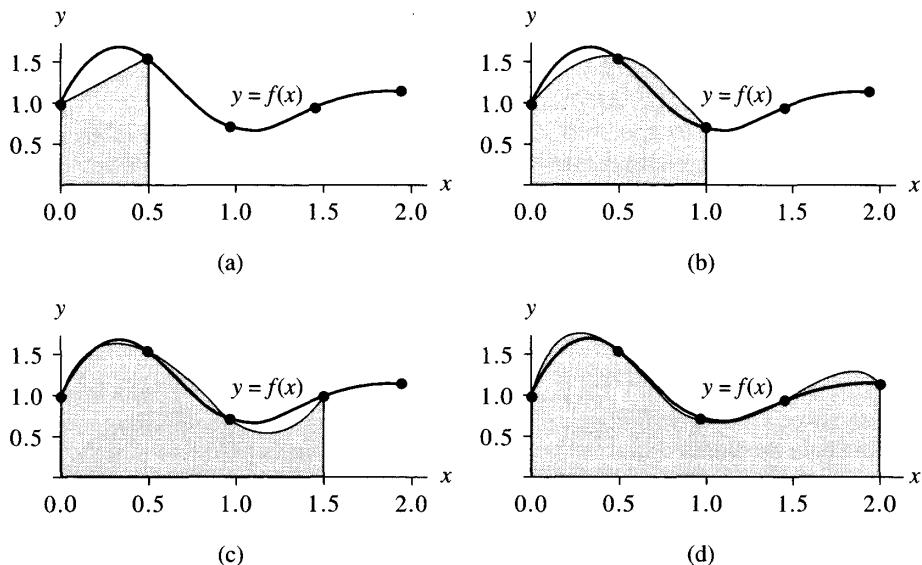


Figura 7.2 (a) La regla del trapecio consiste en integrar $y = P_1(x)$ en $[x_0, x_1] = [0.0, 0.5]$. (b) La regla de Simpson consiste en integrar $y = P_2(x)$ en $[x_0, x_1] = [0.0, 1.0]$. (c) La regla $\frac{3}{8}$ de Simpson consiste en integrar $y = P_3(x)$ en $[x_0, x_3] = [0.0, 1.5]$. (d) La regla de Boole consiste en integrar $y = P_4(x)$ en $[x_0, x_4] = [0.0, 2.0]$.

En el Ejemplo 7.1 hemos aplicado las fórmulas de cuadratura con un incremento fijo $h = 0.5$. Si queremos realizar las aproximaciones sobre un mismo intervalo $[a, b]$, entonces hay que ajustar el incremento a cada caso: Los incrementos son $h = b - a$ para la regla del trapecio, $h = (b - a)/2$ para la regla de Simpson, $h = (b - a)/3$ para la regla $\frac{3}{8}$ de Simpson y $h = (b - a)/4$ para la regla de Boole. El siguiente ejemplo ilustra esta situación.

Ejemplo 7.2. Queremos integrar la función $f(x) = 1 + e^{-x} \operatorname{sen}(4x)$ en el intervalo $[a, b] = [0, 1]$. Para ello vamos a aplicar las fórmulas (4)–(7).

Para la regla del trapecio tenemos $h = 1$ y el resultado es

$$\begin{aligned}\int_0^1 f(x) dx &\approx \frac{1}{2}(f(0) + f(1)) \\ &= \frac{1}{2}(1.00000 + 0.72159) = 0.86079.\end{aligned}$$

Para la regla de Simpson tenemos $h = 1/2$ y el resultado es

$$\begin{aligned}\int_0^1 f(x) dx &\approx \frac{1/2}{3} (f(0) + 4f(\tfrac{1}{2}) + f(1)) \\ &= \frac{1}{6} (1.00000 + 4(1.55152) + 0.72159) = 1.32128.\end{aligned}$$

Para la regla $\frac{3}{8}$ de Simpson tenemos $h = 1/3$ y el resultado es

$$\begin{aligned}\int_0^1 f(x) dx &\approx \frac{3(1/3)}{8} (f(0) + 3f(\tfrac{1}{3}) + 3f(\tfrac{2}{3}) + f(1)) \\ &= \frac{1}{8} (1.00000 + 3(1.69642) + 3(1.23447) + 0.72159) = 1.31440\end{aligned}$$

Para la regla de Boole tenemos $h = 1/4$ y el resultado es

$$\begin{aligned}\int_0^1 f(x) dx &\approx \frac{2(1/4)}{45} (7f(0) + 32f(\tfrac{1}{4}) + 12f(\tfrac{1}{2}) + 32f(\tfrac{3}{4}) + 7f(1)) \\ &= \frac{1}{90} (7(1.00000) + 32(1.65534) + 12(1.55152) \\ &\quad + 32(1.06666) + 7(0.72159)) = 1.30859.\end{aligned}$$

El valor exacto de esta integral definida es

$$\int_0^1 f(x) dx = \frac{21e - 4\cos(4) - \sin(4)}{17e} = 1.3082506046426\dots,$$

así que la aproximación 1.30859 dada por la regla de Boole es la mejor. Las áreas limitadas por cada uno de los polinomios interpoladores de Lagrange $P_1(x)$, $P_2(x)$, $P_3(x)$ y $P_4(x)$ se muestran en las Figuras 7.3(a)–(d), respectivamente.

Para que la comparación entre los métodos de cuadratura sea justa, deberíamos hacer la misma cantidad de evaluaciones de la función integrando. En nuestro siguiente ejemplo, vamos a comparar las aproximaciones que se obtienen cuando se integra sobre un intervalo común $[a, b]$ y se utilizan en cada método exactamente cinco evaluaciones de la función $f_k = f(x_k)$, para $k = 0, 1, \dots, 4$. El proceso de aplicar la regla del trapecio en cada uno de los subintervalos $[x_0, x_1]$, $[x_1, x_2]$, $[x_2, x_3]$ y $[x_3, x_4]$ se llama **regla compuesta del trapecio**:

$$\begin{aligned}(17) \quad \int_{x_0}^{x_4} f(x) dx &= \int_{x_0}^{x_1} f(x) dx + \int_{x_1}^{x_2} f(x) dx + \int_{x_2}^{x_3} f(x) dx + \int_{x_3}^{x_4} f(x) dx \\ &\approx \frac{h}{2} (f_0 + f_1) + \frac{h}{2} (f_1 + f_2) + \frac{h}{2} (f_2 + f_3) + \frac{h}{2} (f_3 + f_4) \\ &= \frac{h}{2} (f_0 + 2f_1 + 2f_2 + 2f_3 + f_4).\end{aligned}$$

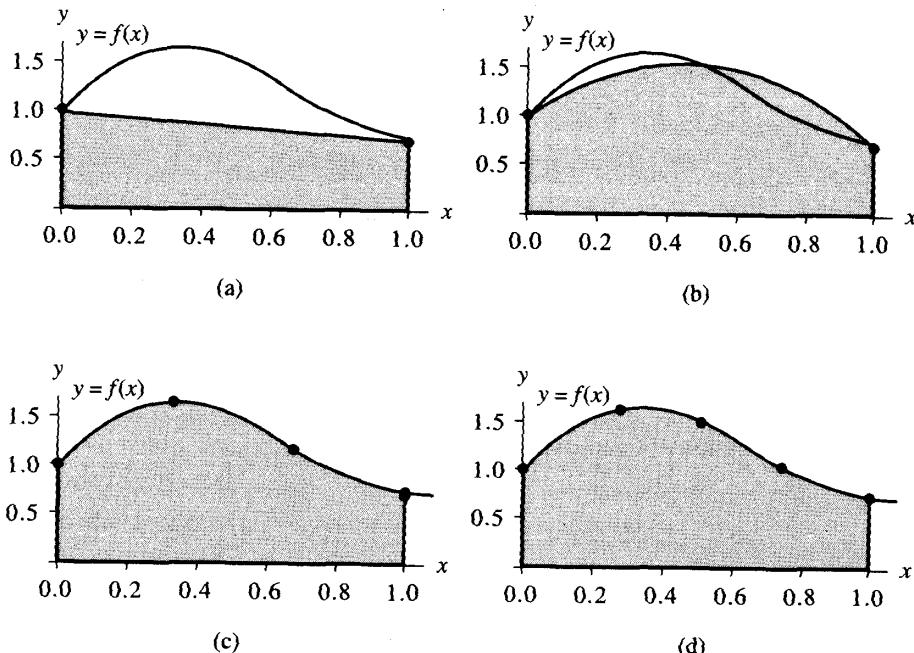


Figura 7.3 (a) La regla del trapezio en $[0, 1]$ proporciona la aproximación 0.86079. (b) La regla de Simpson en $[0, 1]$ proporciona la aproximación 1.32128. (c) La regla $\frac{3}{8}$ de Simpson en $[0, 1]$ proporciona la aproximación 1.31440. (d) La regla de Boole en $[0, 1]$ proporciona la aproximación 1.30859.

La regla de Simpson puede usarse de la misma forma. La aplicación de la regla de Simpson a cada uno de los subintervalos $[x_0, x_2]$ y $[x_2, x_4]$, se llama **regla compuesta de Simpson**:

$$\begin{aligned}
 \int_{x_0}^{x_4} f(x) dx &= \int_{x_0}^{x_2} f(x) dx + \int_{x_2}^{x_4} f(x) dx \\
 (18) \quad &\approx \frac{h}{3}(f_0 + 4f_1 + f_2) + \frac{h}{3}(f_2 + 4f_3 + f_4) \\
 &= \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + f_4).
 \end{aligned}$$

En el siguiente ejemplo se comparan los valores obtenidos con las fórmulas (17), (18) y (7).

Ejemplo 7.3. Queremos integrar la función $f(x) = 1 + e^{-x} \operatorname{sen}(4x)$ en el intervalo $[a, b] = [0, 1]$. Para ello, vamos a aplicar la regla compuesta del trapezio, la regla compuesta de Simpson y la regla de Boole, de manera que cada una de ellas utilice

exactamente cinco evaluaciones de la función. Luego compararemos los resultados que se obtienen.

El incremento común es $h = 1/4$. La regla compuesta del trapecio (17) nos da:

$$\begin{aligned}\int_0^1 f(x) dx &\approx \frac{1/4}{2}(f(0) + 2f(\frac{1}{4}) + 2f(\frac{1}{2}) + 2f(\frac{3}{4}) + f(1)) \\ &= \frac{1}{8}(1.00000 + 2(1.65534) + 2(1.55152) + 2(1.06666) + 0.72159) \\ &= 1.28358.\end{aligned}$$

Con la regla compuesta de Simpson (18) obtenemos:

$$\begin{aligned}\int_0^1 f(x) dx &\approx \frac{1/4}{3}(f(0) + 4f(\frac{1}{4}) + 2f(\frac{1}{2}) + 4f(\frac{3}{4}) + f(1)) \\ &= \frac{1}{12}(1.00000 + 4(1.65534) + 2(1.55152) + 4(1.06666) + 0.72159) \\ &= 1.30938.\end{aligned}$$

El resultado con la regla de Boole ya lo obtuvimos en el Ejemplo 7.2:

$$\begin{aligned}\int_0^1 f(x) dx &\approx \frac{2(1/4)}{45}(7f(0) + 32f(\frac{1}{4}) + 12f(\frac{1}{2}) + 32f(\frac{3}{4}) + 7f(1)) \\ &= 1.30859.\end{aligned}$$

El valor exacto de la integral es

$$\int_0^1 f(x) dx = \frac{21e - 4\cos(4) - \sin(4)}{17e} = 1.3082506046426\dots,$$

así que la aproximación 1.30938 obtenida con la regla de Simpson es mucho mejor que el valor 1.28358 obtenido con la regla del trapecio. De nuevo, es la regla de Boole la que proporciona la mejor aproximación, 1.30859, de las tres. En las Figuras 7.4(a) y (b) se muestran los dibujos de las áreas bajo los trapecios y las paráolas, respectivamente. ■

Ejemplo 7.4. Vamos a determinar el grado de precisión de la regla $\frac{3}{8}$ de Simpson.

Bastará que apliquemos la regla $\frac{3}{8}$ de Simpson en el intervalo $[0, 3]$ con las cinco funciones $f(x) = 1, x, x^2, x^3$ y x^4 . Para las cuatro primeras funciones, la regla $\frac{3}{8}$

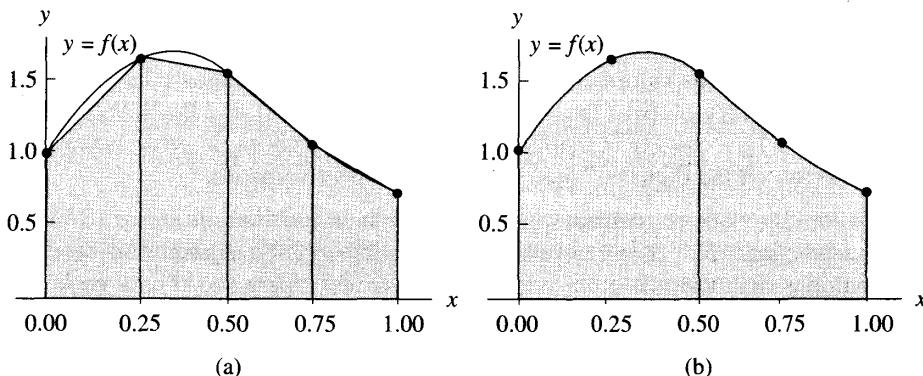


Figura 7.4 (a) La regla compuesta del trapecio proporciona la aproximación 1.28358. (b) La regla compuesta de Simpson proporciona la aproximación 1.30938.

de Simpson es exacta:

$$\begin{aligned}\int_0^3 1 \, dx &= 3 = \frac{3}{8}(1 + 3(1) + 3(1) + 1) \\ \int_0^3 x \, dx &= \frac{9}{2} = \frac{3}{8}(0 + 3(1) + 3(2) + 3) \\ \int_0^3 x^2 \, dx &= 9 = \frac{3}{8}(0 + 3(1) + 3(4) + 9) \\ \int_0^3 x^3 \, dx &= \frac{81}{4} = \frac{3}{8}(0 + 3(1) + 3(8) + 27).\end{aligned}$$

La función $f(x) = x^4$ es la menor potencia de x para la que la regla no es exacta:

$$\int_0^3 x^4 \, dx = \frac{243}{5} \approx \frac{99}{2} = \frac{3}{8}(0 + 3(1) + 3(16) + 81).$$

Por tanto, el grado de aproximación de la regla $\frac{3}{8}$ de Simpson es $n = 3$. ■

Ejercicios

- En los casos que se relacionan a continuación se considera la integración de la función dada $f(x)$ sobre el intervalo fijo $[a, b] = [0, 1]$. Aplique las fórmulas de cuadratura (4)–(7) tomando como incremento: $h = 1$ para la regla del trapecio, $h = \frac{1}{2}$ para la regla de Simpson, $h = \frac{1}{3}$ para la regla $\frac{3}{8}$ de Simpson y $h = \frac{1}{4}$ para la regla de Boole.
 - $f(x) = \sin(\pi x)$

(b) $f(x) = 1 + e^{-x} \cos(4x)$

(c) $f(x) = \operatorname{sen}(\sqrt{x})$

Observación. Los valores exactos de las integrales definidas son: (a) $2/\pi = 0.63661977237\dots$, (b) $(18e - \cos(4) + 4\operatorname{sen}(4))/(17e) = 1.00745963140\dots$ y (c) $2(\operatorname{sen}(1) - \cos(1)) = 0.60233735788\dots$. Las gráficas de estas funciones se muestran en las Figuras 7.5(a), (b) y (c), respectivamente.

2. En los casos que se relacionan a continuación se considera la integración de la función dada $f(x)$ sobre el intervalo fijo $[a, b] = [0, 1]$. Aplique las siguientes fórmulas de cuadratura: la regla compuesta del trapezio (17), la regla compuesta de Simpson (18) y la regla de Boole (7). Utilice cinco evaluaciones de la función en nodos equiespaciados con incremento $h = \frac{1}{4}$.

(a) $f(x) = \operatorname{sen}(\pi x)$

(b) $f(x) = 1 + e^{-x} \cos(4x)$

(c) $f(x) = \operatorname{sen}(\sqrt{x})$

3. Sea $[a, b]$ un intervalo cualquiera. Pruebe que la regla de Simpson proporciona resultados exactos para las funciones $f(x) = x^2$ y $f(x) = x^3$; es decir,

(a) $\int_a^b x^2 dx = \frac{b^3}{3} - \frac{a^3}{3}$

(b) $\int_a^b x^3 dx = \frac{b^4}{4} - \frac{a^4}{4}$

4. Integrando el polinomio de interpolación de Lagrange

$$P_1(x) = f_0 \frac{x - x_1}{x_0 - x_1} + f_1 \frac{x - x_0}{x_1 - x_0}$$

en el intervalo $[x_0, x_1]$, deduzca la regla del trapecio.

5. Determine el grado de precisión de la regla del trapecio. Para ello es suficiente con aplicar la regla en el intervalo $[0, 1]$ a las funciones $f(x) = 1$, x y x^2 .
6. Determine el grado de precisión de la regla de Simpson. Para ello es suficiente con aplicar la regla en el intervalo $[0, 2]$ a las cinco funciones $f(x) = 1$, x , x^2 , x^3 y x^4 . Contraste su resultado con el grado de precisión de la regla $\frac{3}{8}$ de Simpson.

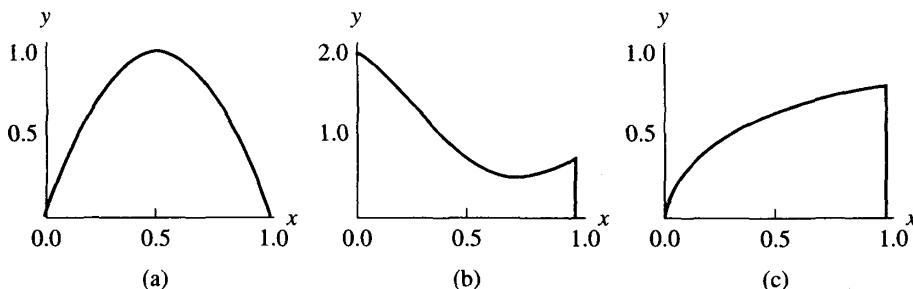


Figura 7.5 (a) $y = \operatorname{sen}(\pi x)$, (b) $y = 1 + e^{-x} \cos(4x)$, (c) $y = \operatorname{sen}(\sqrt{x})$.

7. Determine el grado de precisión de la regla de Boole. Para ello es suficiente con aplicar la regla en el intervalo $[0, 4]$ a las siete funciones $f(x) = 1, x, x^2, x^3, x^4, x^5$ y x^6 .
8. Los intervalos de los Ejercicios 5, 6 y 7 y del Ejemplo 7.4 fueron elegidos para que el cálculo de los nodos de integración fuera sencillo. No obstante, sea cual sea el intervalo de integración $[a, b]$ en el que queramos integrar la función f , cada una de las cuatro fórmulas de cuadratura (4)–(7) tiene el grado de precisión determinado en los Ejercicios 5, 6 y 7 y en el Ejemplo 7.4, respectivamente. Una fórmula de cuadratura en un intervalo $[a, b]$ puede obtenerse a partir de una fórmula de cuadratura en un intervalo $[c, d]$ haciendo el cambio de variables dado por la función lineal

$$x = g(t) = \frac{b-a}{d-c}t + \frac{ad-bc}{d-c}$$

tomando $dx = \frac{b-a}{d-c} dt$.

- (a) Compruebe que $x = g(t)$ es la línea recta que pasa por los puntos (c, a) y (d, b) .
- (b) Compruebe que la regla del trapecio tiene el mismo grado de precisión en cualquier intervalo $[a, b]$ que en el intervalo $[0, 1]$.
- (c) Compruebe que la regla de Simpson tiene el mismo grado de precisión en cualquier intervalo $[a, b]$ que en el intervalo $[0, 2]$.
- (d) Compruebe que la regla de Boole tiene el mismo grado de precisión en cualquier intervalo $[a, b]$ que en el intervalo $[0, 4]$.
9. Deduzca la regla $\frac{3}{8}$ de Simpson a partir del polinomio interpolador de Lagrange.
- Indicación. Después de hacer el cambio de variable se obtienen integrales similares a las de (16):

$$\begin{aligned} & \int_{x_0}^{x_3} f(x) dx \\ & \approx -f_0 \frac{h}{6} \int_0^3 (t-1)(t-2)(t-3) dt + f_1 \frac{h}{2} \int_0^3 (t-0)(t-2)(t-3) dt \\ & \quad - f_2 \frac{h}{2} \int_0^3 (t-0)(t-1)(t-3) dt + f_3 \frac{h}{6} \int_0^3 (t-0)(t-1)(t-2) dt \\ & = f_0 \frac{h}{6} \left(\frac{-t^4}{4} + 2t^3 - \frac{11t^2}{2} + 6t \right) \Big|_{t=0}^{t=3} + f_1 \frac{h}{2} \left(\frac{t^4}{4} - \frac{5t^3}{3} + 3t^2 \right) \Big|_{t=0}^{t=3} \\ & \quad + f_2 \frac{h}{2} \left(\frac{-t^4}{4} + \frac{4t^3}{3} - \frac{3t^2}{2} \right) \Big|_{t=0}^{t=3} + f_3 \frac{h}{6} \left(\frac{t^4}{4} - t^3 + t^2 \right) \Big|_{t=0}^{t=3}. \end{aligned}$$

10. Deduzca la fórmula cerrada de Newton-Cotes usando el polinomio interpolador de Lagrange de grado 5 para los 6 nodos equiespaciados $x_k = x_0 + kh$, con $k = 0, 1, \dots, 5$.

11. En la demostración del Teorema 7.1, la regla de Simpson se dedujo integrando el polinomio interpolador de Lagrange de grado dos para los nodos equiespaciados x_0 , x_1 y x_2 . Deduzca la regla de Simpson integrando el polinomio interpolador de Newton de grado dos para los mismos nodos.

7.2 Las reglas compuestas del trapecio y de Simpson

Un método intuitivo para hallar el área limitada por una curva $y = f(x)$ en un intervalo $[a, b]$ es dar una aproximación a dicha área sumando las áreas de una serie de trapecios construidos sobre los intervalos $\{[x_k, x_{k+1}]\}$ de una partición de $[a, b]$.

Teorema 7.2 (Regla compuesta del trapecio). Supongamos que se divide el intervalo $[a, b]$ en M subintervalos $[x_k, x_{k+1}]$ de anchura común $h = (b-a)/M$ mediante una partición cuyos nodos $x_k = a + kh$, para $k = 0, 1, \dots, M$, están equiespaciados. La **regla compuesta del trapecio con M subintervalos** se puede expresar de cualquiera de las siguientes formas equivalentes:

$$(1a) \quad T(f, h) = \frac{h}{2} \sum_{k=1}^M (f(x_{k-1}) + f(x_k))$$

o bien

$$(1b) \quad T(f, h) = \frac{h}{2}(f_0 + 2f_1 + 2f_2 + 2f_3 + \cdots + 2f_{M-2} + 2f_{M-1} + f_M)$$

o bien

$$(1c) \quad T(f, h) = \frac{h}{2}(f(a) + f(b)) + h \sum_{k=1}^{M-1} f(x_k).$$

Este valor es una aproximación a la integral de $f(x)$ en $[a, b]$, lo que se escribe como

$$(2) \quad \int_a^b f(x) dx \approx T(f, h).$$

Demostración. Aplicando la regla del trapecio sobre cada intervalo $[x_{k-1}, x_k]$ (véase la Figura 7.6) y usando la propiedad de aditividad de la integración, obtenemos:

$$(3) \quad \int_a^b f(x) dx = \sum_{k=1}^M \int_{x_{k-1}}^{x_k} f(x) dx \approx \sum_{k=1}^M \frac{h}{2}(f(x_{k-1}) + f(x_k)).$$

Puesto que $h/2$ es constante, sacándolo como factor común obtenemos la fórmula (1a); la fórmula (1b) no es más que una versión extendida de (1a) y la fórmula (1c) muestra cómo se agrupan los términos intermedios de (1b) que van multiplicados por 2.

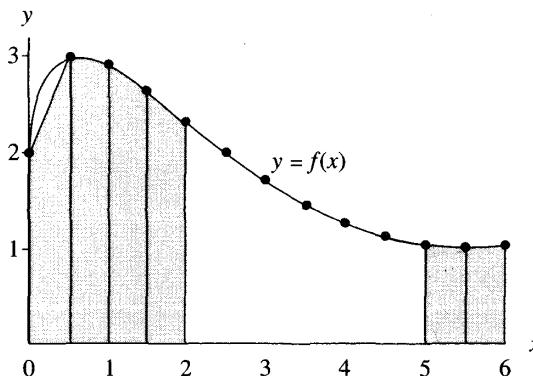


Figura 7.6 Aproximación al área limitada por la curva $y = 2 + \operatorname{sen}(2\sqrt{x})$ mediante la regla compuesta del trapecio.

En la Figura 7.6 puede verse que al aproximar $f(x) = 2 + \operatorname{sen}(2\sqrt{x})$ por un polinomio lineal a trozos, hay sitios en los que la aproximación es buena y sitios en los que no lo es. Para conseguir una buena precisión hay que aplicar la regla compuesta del trapecio con muchos subintervalos. En el siguiente ejemplo mostramos cómo se usa la regla compuesta del trapecio para aproximar la integral en el intervalo $[1, 6]$. Dejamos como ejercicio el analizar qué ocurre si integramos en $[0, 1]$.

Ejemplo 7.5. Consideremos $f(x) = 2 + \operatorname{sen}(2\sqrt{x})$. Vamos a usar la regla compuesta del trapecio con 11 nodos para calcular una aproximación a la integral de $f(x)$ en el intervalo $[1, 6]$.

Para generar los once nodos, tomamos $M = 10$, con lo que $h = (6 - 1)/10 = 1/2$. Usando la fórmula (1c), los cálculos son

$$\begin{aligned}
 T(f, \frac{1}{2}) &= \frac{1/2}{2} (f(1) + f(6)) \\
 &\quad + \frac{1}{2} (f(\frac{3}{2}) + f(2) + f(\frac{5}{2}) + f(3) + f(\frac{7}{2}) + f(4) + f(\frac{9}{2}) + f(5) + f(\frac{11}{2})) \\
 &= \frac{1}{4} (2.90929743 + 1.01735756) \\
 &\quad + \frac{1}{2} (2.63815764 + 2.30807174 + 1.97931647 + 1.68305284 \\
 &\quad + 1.43530410 + 1.24319750 + 1.10831775 + 1.02872220 + 1.00024140) \\
 &= \frac{1}{4} (3.92665499) + \frac{1}{2} (14.42438165) \\
 &= 0.98166375 + 7.21219083 = 8.19385457.
 \end{aligned}$$

Teorema 7.3 (Regla compuesta de Simpson). Supongamos que dividimos $[a, b]$ en $2M$ subintervalos $[x_k, x_{k+1}]$ de la misma anchura $h = (b - a)/(2M)$

mediante una partición de nodos equiespaciados $x_k = a + kh$, para $k = 0, 1, \dots, 2M$. La **regla compuesta de Simpson con $2M$ subintervalos** se puede expresar de cualquiera de las siguientes formas equivalentes:

$$(4a) \quad S(f, h) = \frac{h}{3} \sum_{k=1}^M (f(x_{2k-2}) + 4f(x_{2k-1}) + f(x_{2k}))$$

o bien

$$(4b) \quad S(f, h) = \frac{h}{3} (f_0 + 4f_1 + 2f_2 + 4f_3 \\ + \cdots + 2f_{2M-2} + 4f_{2M-1} + f_{2M})$$

o bien

$$(4c) \quad S(f, h) = \frac{h}{3} (f(a) + f(b)) + \frac{2h}{3} \sum_{k=1}^{M-1} f(x_{2k}) + \frac{4h}{3} \sum_{k=1}^M f(x_{2k-1}).$$

Este valor es una aproximación a la integral de $f(x)$ en $[a, b]$, lo que se escribe como

$$(5) \quad \int_a^b f(x) dx \approx S(f, h).$$

Demostración. Aplicando la regla de Simpson sobre cada $[x_{2k-2}, x_{2k}]$ (véase la Figura 7.7) y usando la propiedad de aditividad de la integración, obtenemos:

$$(6) \quad \begin{aligned} \int_a^b f(x) dx &= \sum_{k=1}^M \int_{x_{2k-2}}^{x_{2k}} f(x) dx \\ &\approx \sum_{k=1}^M \frac{h}{3} (f(x_{2k-2}) + 4f(x_{2k-1}) + f(x_{2k})). \end{aligned}$$

Puesto que $h/3$ es constante, sacándolo como factor común obtenemos la fórmula (4a); la fórmula (4b) no es más que una versión extendida de (4a) y la fórmula (4c) muestra cómo se agrupan los términos intermedios de (4b) que van multiplicados por 2 y los que van multiplicados por 4. •

En la Figura 7.7 puede verse que al aproximar $f(x) = 2 + \operatorname{sen}(2\sqrt{x})$ por un polinomio cuadrático a trozos, hay sitios en los que la aproximación es buena y sitios en los que no lo es. Para conseguir una buena precisión hay que aplicar la regla compuesta de Simpson con varios subintervalos. En el siguiente ejemplo mostramos cómo se usa la regla compuesta de Simpson para aproximar la integral en el intervalo $[1, 6]$. Dejamos como ejercicio el analizar qué ocurre si integramos en $[0, 1]$.

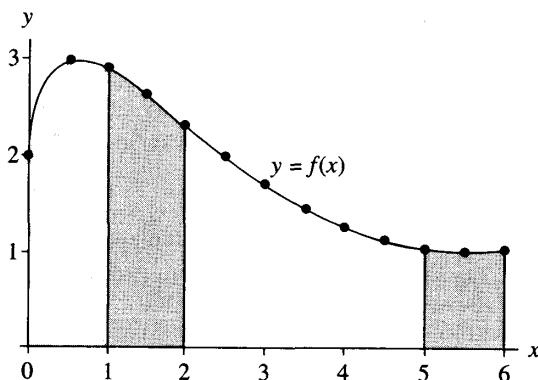


Figura 7.7 Aproximación al área limitada por la curva $y = 2 + \sin(2\sqrt{x})$ mediante la regla compuesta de Simpson.

Ejemplo 7.6. Consideremos $f(x) = 2 + \sin(2\sqrt{x})$. Vamos a usar la regla compuesta de Simpson con 11 nodos para calcular una aproximación a la integral de $f(x)$ en el intervalo $[1, 6]$.

Para generar los once nodos, debemos tomar $M = 5$, con lo cual se tiene $h = (6 - 1)/10 = 1/2$. Usando la fórmula (4c), los cálculos son:

$$\begin{aligned}
 S(f, \frac{1}{2}) &= \frac{1}{6}(f(1) + f(6)) + \frac{1}{3}(f(2) + f(3) + f(4) + f(5)) \\
 &\quad + \frac{2}{3}(f(\frac{3}{2}) + f(\frac{5}{2}) + f(\frac{7}{2}) + f(\frac{9}{2}) + f(\frac{11}{2})) \\
 &= \frac{1}{6}(2.90929743 + 1.01735756) \\
 &\quad + \frac{1}{3}(2.30807174 + 1.68305284 + 1.24319750 + 1.02872220) \\
 &\quad + \frac{2}{3}(2.63815764 + 1.97931647 + 1.43530410 \\
 &\quad \quad + 1.10831775 + 1.00024140) \\
 &= \frac{1}{6}(3.92665499) + \frac{1}{3}(6.26304429) + \frac{2}{3}(8.16133735) \\
 &= 0.65444250 + 2.08768143 + 5.44089157 = 8.18301550. \blacksquare
 \end{aligned}$$

Análisis del error

Lo esencial tras los dos resultados que veremos a continuación es que entendamos que los términos del error $E_T(f, h)$ y $E_S(f, h)$ para las reglas compuestas del trapecio y de Simpson son de orden $O(h^2)$ y $O(h^4)$, respectivamente. Esto prueba que el error de la fórmula de Simpson tiende a cero más rápidamente que el error de la regla del trapecio cuando h tiende a cero. Cuando las derivadas

de $f(x)$ se conocen, las fórmulas

$$E_T(f, h) = \frac{-(b-a)f^{(2)}(c)h^2}{12} \quad \text{y} \quad E_S(f, h) = \frac{-(b-a)f^{(4)}(c)h^4}{180}$$

nos permiten, además, estimar el número de subintervalos necesarios para alcanzar la precisión deseada.

Corolario 7.2 (Análisis del error para la regla compuesta del trapecio). Supongamos que $[a, b]$ se divide en M subintervalos $[x_k, x_{k+1}]$ de tamaño $h = (b-a)/M$. La regla compuesta del trapecio

$$(7) \quad T(f, h) = \frac{h}{2}(f(a) + f(b)) + h \sum_{k=1}^{M-1} f(x_k)$$

es una aproximación a la integral

$$(8) \quad \int_a^b f(x) dx = T(f, h) + E_T(f, h)$$

tal que, si además $f \in C^2[a, b]$, entonces existe un valor c con $a < c < b$ de manera que el término del error $E_T(f, h)$ lo podemos escribir como

$$(9) \quad E_T(f, h) = \frac{-(b-a)f^{(2)}(c)h^2}{12} = O(h^2).$$

Demostración. En primer lugar determinamos el término del error cuando la regla se aplica en el intervalo $[x_0, x_1]$. Para ello, integrando el polinomio interpolador de Lagrange $P_1(x)$ y su término del error, obtenemos

$$(10) \quad \int_{x_0}^{x_1} f(x) dx = \int_{x_0}^{x_1} P_1(x) dx + \int_{x_0}^{x_1} \frac{(x-x_0)(x-x_1)f^{(2)}(c(x))}{2!} dx.$$

Como el término $(x-x_0)(x-x_1)$ no cambia de signo en $[x_0, x_1]$ y $f^{(2)}(c(x))$ es continua, el segundo teorema del valor medio para integrales nos dice que existe un valor c_1 tal que

$$(11) \quad \int_{x_0}^{x_1} f(x) dx = \frac{h}{2}(f_0 + f_1) + f^{(2)}(c_1) \int_{x_0}^{x_1} \frac{(x-x_0)(x-x_1)}{2!} dx$$

y esta integral se calcula fácilmente con el cambio de variable $x = x_0 + ht$:

$$(12) \quad \begin{aligned} \int_{x_0}^{x_1} f(x) dx &= \frac{h}{2}(f_0 + f_1) + \frac{f^{(2)}(c_1)}{2} \int_0^1 h(t-0)h(t-1)h dt \\ &= \frac{h}{2}(f_0 + f_1) + \frac{f^{(2)}(c_1)h^3}{2} \int_0^1 (t^2 - t) dt \\ &= \frac{h}{2}(f_0 + f_1) - \frac{f^{(2)}(c_1)h^3}{12}. \end{aligned}$$

Ahora podemos sumar los términos del error de todos los intervalos $[x_k, x_{k+1}]$:

$$(13) \quad \begin{aligned} \int_a^b f(x) dx &= \sum_{k=1}^M \int_{x_{k-1}}^{x_k} f(x) dx \\ &= \sum_{k=1}^M \frac{h}{2}(f(x_{k-1}) + f(x_k)) - \frac{h^3}{12} \sum_{k=1}^M f^{(2)}(c_k). \end{aligned}$$

El primer sumando de la expresión (13) es la fórmula de la regla compuesta del trapecio $T(f, h)$. En el segundo sumando de (13) reemplazamos uno de los factores h por su valor $h = (b - a)/M$ y nos queda

$$\int_a^b f(x) dx = T(f, h) - \frac{(b - a)h^2}{12} \left(\frac{1}{M} \sum_{k=1}^M f^{(2)}(c_k) \right).$$

Como el término entre paréntesis es una media aritmética de valores de la derivada segunda $f^{(2)}$ y esta función es continua, entonces podemos reemplazarlo por $f^{(2)}(c)$ para algún punto $c \in (a, b)$. En consecuencia, hemos probado que

$$\int_a^b f(x) dx = T(f, h) - \frac{(b - a)f^{(2)}(c)h^2}{12},$$

lo que completa la demostración del Corolario 7.2.

Corolario 7.3 (Análisis del error para la regla compuesta de Simpson). Supongamos que $[a, b]$ se divide en $2M$ subintervalos $[x_k, x_{k+1}]$ de tamaño $h = (b - a)/(2M)$. La regla compuesta de Simpson

$$(14) \quad S(f, h) = \frac{h}{3}(f(a) + f(b)) + \frac{2h}{3} \sum_{k=1}^{M-1} f(x_{2k}) + \frac{4h}{3} \sum_{k=1}^M f(x_{2k-1})$$

es una aproximación a la integral

$$(15) \quad \int_a^b f(x) dx = S(f, h) + E_S(f, h)$$

tal que, si además $f \in C^4[a, b]$, entonces existe un valor c con $a < c < b$ tal que el término del error $E_S(f, h)$ lo podemos escribir como

$$(16) \quad E_S(f, h) = \frac{-(b - a)f^{(4)}(c)h^4}{180} = O(h^4).$$

Tabla 7.2 La regla compuesta del trapecio para $f(x) = 2 + \operatorname{sen}(2\sqrt{x})$ en $[1, 6]$.

M	h	$T(f, h)$	$E_T(f, h) = O(h^2)$
10	0.5	8.19385457	-0.01037540
20	0.25	8.18604926	-0.00257006
40	0.125	8.18412019	-0.00064098
80	0.0625	8.18363936	-0.00016015
160	0.03125	8.18351924	-0.00004003

Ejemplo 7.7. Consideremos $f(x) = 2 + \operatorname{sen}(2\sqrt{x})$ y vamos a analizar el error cuando usamos la regla compuesta del trapecio en el intervalo $[1, 6]$ y el número de subintervalos es 10, 20, 40, 80 y 160.

En la Tabla 7.2 se muestran las aproximaciones $T(f, h)$. Una primitiva de $f(x)$ es

$$F(x) = 2x - \sqrt{x} \cos(2\sqrt{x}) + \frac{\operatorname{sen}(2\sqrt{x})}{2},$$

así que el valor de la integral definida con once cifras significativas es

$$\int_1^6 f(x) dx = F(x) \Big|_{x=1}^{x=6} = 8.1834792077,$$

que es el valor que se usa para calcular los errores $E_T(f, h) = 8.1834792077 - T(f, h)$ que se muestran en la Tabla 7.2. Es importante observar que conforme h disminuye en un factor de $\frac{1}{2}$, los errores sucesivos $E_T(f, h)$ disminuyen en un factor de, aproximadamente, $\frac{1}{4}$; esto confirma que el orden de aproximación es $O(h^2)$. ■

Ejemplo 7.8. Consideremos $f(x) = 2 + \operatorname{sen}(2\sqrt{x})$ y vamos a analizar el error cuando usamos la regla compuesta de Simpson en el intervalo $[1, 6]$ y el número de subintervalos es 10, 20, 40, 80 y 160.

En la Tabla 7.3 se muestran las aproximaciones $S(f, h)$. El valor exacto (con once cifras) de la integral 8.1834792077, calculado antes, es el que usamos para determinar los errores $E_S(f, h) = 8.1834792077 - S(f, h)$ que se muestran en la Tabla 7.3. Es importante observar que conforme h disminuye en un factor de $\frac{1}{2}$, los errores sucesivos $E_S(f, h)$ disminuyen en un factor de, aproximadamente, $\frac{1}{16}$; esto confirma que el orden de aproximación es $O(h^4)$. ■

Ejemplo 7.9. Vamos a determinar el número M de subintervalos y el incremento h de manera que el error $E_T(f, h)$ de la regla compuesta del trapecio en la aproximación $\int_2^7 dx/x \approx T(f, h)$ sea menor que 5×10^{-9} .

Tabla 7.3 La regla compuesta de Simpson para $f(x) = 2 + \operatorname{sen}(2\sqrt{x})$ en $[1, 6]$.

M	h	$S(f, h)$	$E_S(f, h) = O(h^4)$
5	0.5	8.18301549	0.00046371
10	0.25	8.18344750	0.00003171
20	0.125	8.18347717	0.00000204
40	0.0625	8.18347908	0.00000013
80	0.03125	8.18347920	0.00000001

El integrando es $f(x) = 1/x$ y sus dos primeras derivadas son $f'(x) = -1/x^2$ y $f^{(2)}(x) = 2/x^3$; por tanto, el valor máximo de $|f^{(2)}(x)|$ en $[2, 7]$ se alcanza en el punto $x = 2$, lo que nos proporciona la cota $|f^{(2)}(c)| \leq |f^{(2)}(2)| = \frac{1}{4}$ para $2 \leq c \leq 7$. Usando esto con la fórmula (9) obtenemos

$$(17) \quad |E_T(f, h)| = \frac{|-(b-a)f^{(2)}(c)h^2|}{12} \leq \frac{(7-2)\frac{1}{4}h^2}{12} = \frac{5h^2}{48}.$$

El incremento h y el número de subintervalos M verifican la relación $h = 5/M$ que, al usarla en la expresión (17), nos permite obtener

$$(18) \quad |E_T(f, h)| \leq \frac{125}{48M^2} \leq 5 \times 10^{-9}.$$

Vamos a escribir esta relación (18) de manera que podamos despejar M fácilmente:

$$(19) \quad \frac{25}{48} \times 10^9 \leq M^2,$$

con lo cual tenemos que $22821.77 \leq M$. Puesto que M debe ser entero, tomamos $M = 22822$, siendo el incremento correspondiente $h = 5/22822 = 0.000219086846$. Si empleamos la regla compuesta del trapecio, al necesitar un número tan elevado de evaluaciones de la función, existe la posibilidad de que los redondeos en las evaluaciones produzcan un error apreciablemente significativo. Realizando los cálculos con un computador se obtiene

$$T\left(f, \frac{5}{22822}\right) = 1.252762969,$$

que se acerca bastante al valor $\int_2^7 dx/x = \ln(x)|_{x=2}^{x=7} = 1.252762968$. El error es menor que el error predicho porque la cota de $|f^{(2)}(c)|$ que usamos es $\frac{1}{4}$, que resulta ser demasiado alta. Experimentando, se puede observar que el número de evaluaciones de la función necesarias para alcanzar la precisión deseada 5×10^{-9} es de unas 10 000; de hecho, el resultado que se obtiene con $M = 10\,000$ es:

$$T\left(f, \frac{5}{10\,000}\right) = 1.252762973. \quad \blacksquare$$

La regla compuesta del trapecio requiere normalmente de un elevado número de evaluaciones de la función para que se obtenga una respuesta precisa; esto contrasta, como veremos en el ejemplo siguiente, con lo que sucede cuando usamos la regla compuesta de Simpson, que requiere de un número significativamente más bajo de evaluaciones.

Ejemplo 7.10. Vamos a determinar el número M de subintervalos y el incremento h de manera que el error $E_S(f, h)$ de la regla compuesta de Simpson en la aproximación $\int_2^7 dx/x \approx S(f, h)$ sea menor que 5×10^{-9} .

El integrando es $f(x) = 1/x$ y su derivada cuarta es $f^{(4)}(x) = 24/x^5$, cuyo valor máximo en $[2, 7]$ se alcanza en el extremo $x = 2$, lo que nos proporciona la cota $|f^{(4)}(c)| \leq |f^{(4)}(2)| = \frac{3}{4}$ para $2 \leq c \leq 7$. Usando esto con la fórmula (16) obtenemos

$$(20) \quad |E_S(f, h)| = \frac{|-(b-a)f^{(4)}(c)h^4|}{180} \leq \frac{(7-2)\frac{3}{4}h^4}{180} = \frac{h^4}{48}.$$

El incremento h y el número de subintervalos M verifican la relación $h = 5/(2M)$ que, al usarla en (20), nos permite obtener

$$(21) \quad |E_S(f, h)| \leq \frac{625}{768M^4} \leq 5 \times 10^{-9}.$$

Vamos a escribir esta relación (21) de manera que podamos despejar M fácilmente:

$$(22) \quad \frac{125}{768} \times 10^9 \leq M^4,$$

de donde obtenemos $112.95 \leq M$ y, como M debe ser entero, tomamos $M = 113$, siendo el incremento correspondiente $h = 5/226 = 0.02212389381$. Al calcular la aproximación dada por la fórmula compuesta de Simpson, el resultado es

$$S\left(f, \frac{5}{226}\right) = 1.252762969,$$

que coincide con $\int_2^7 dx/x|_{x=2}^{x=7} = \ln(x)|_{x=2}^{x=7} = 1.252762968$ en diez cifras significativas. Experimentando, se puede observar que el número de evaluaciones de la función necesarias para alcanzar la precisión deseada 5×10^{-9} es de unas 129 y si hacemos los cálculos con $M = 64$, el resultado es

$$S\left(f, \frac{5}{128}\right) = 1.252762973. \quad \blacksquare$$

Esto nos muestra que la regla compuesta de Simpson con 229 evaluaciones de $f(x)$ y la regla compuesta del trapecio con 22 823 evaluaciones de $f(x)$, unas cien veces más, proporcionan la misma precisión.

MATLAB

Programa 7.1 (Regla compuesta del trapecio). Construcción de la aproximación a la integral

$$\int_a^b f(x) dx \approx \frac{h}{2}(f(a) + f(b)) + h \sum_{k=1}^{M-1} f(x_k)$$

evaluando $f(x)$ en los $M + 1$ nodos equiespaciados $x_k = a + kh$, para $k = 0, 1, 2, \dots, M$. Nótese que $x_0 = a$ y que $x_M = b$.

```
function s=traprl(f,a,b,M)
% Datos
% - f es el integrando, dado como una
%   cadena de caracteres 'f'
% - a y b son los extremos inferior y superior del
%   intervalo de integración
% - M es el número de subintervalos
% Resultado
% - s es la aproximación obtenida con la
%   regla compuesta del trapecio
h=(b-a)/M;
s=0;
for k=1:(M-1)
    x=a+h*k;
    s=s+feval(f,x);
end
s=h*(feval(f,a)+feval(f,b))/2+h*s;
```

Programa 7.2 (Regla compuesta de Simpson). Construcción de la aproximación a la integral

$$\int_a^b f(x) dx \approx \frac{h}{3}(f(a) + f(b)) + \frac{2h}{3} \sum_{k=1}^{M-1} f(x_{2k}) + \frac{4h}{3} \sum_{k=1}^M f(x_{2k-1})$$

evaluando $f(x)$ en los $2M + 1$ nodos equiespaciados $x_k = a + kh$, para $k = 0, 1, 2, \dots, 2M$. Nótese que $x_0 = a$ y que $x_{2M} = b$.

```
function s=simprl(f,a,b,M)
% Datos
% - f es el integrando, dado como una
%   cadena de caracteres 'f'
```

```

% - a y b son los extremos inferior y superior del
%     intervalo de integración
% - M es el número de subintervalos
% Resultado
% - s es la aproximación obtenida con la
%     regla compuesta de Simpson

h=(b-a)/(2*M);
s1=0;
s2=0;

for k=1:M
    x=a+h*(2*k-1);
    s1=s1+feval(f,x);
end
for k=1:(M-1)
    x=a+h*2*k;
    s2=s2+feval(f,x);
end
s=h*(feval(f,a)+feval(f,b)+4*s1+2*s2)/3;

```

Ejercicios

1. (i) Aproxime cada una de las siguientes integrales usando la regla compuesta del trapecio con $M = 10$.
(ii) Aproxime cada una de las siguientes integrales usando la regla compuesta de Simpson con $M = 5$.

(a) $\int_{-1}^1 (1 + x^2)^{-1} dx$ (b) $\int_0^1 (2 + \operatorname{sen}(2\sqrt{x})) dx$ (c) $\int_{0.25}^4 dx / \sqrt{x}$

(d) $\int_0^4 x^2 e^{-x} dx$ (e) $\int_0^2 2x \cos(x) dx$ (f) $\int_0^\pi \operatorname{sen}(2x) e^{-x} dx$

2. *Longitud de una curva.* La longitud de una curva $y = f(x)$ definida sobre un intervalo $a \leq x \leq b$ es

$$\text{longitud} = \int_a^b \sqrt{1 + (f'(x))^2} dx.$$

- (i) Aproxime la longitud de la curva $y = f(x)$ para cada una de las funciones que se relacionan a continuación usando la regla compuesta del trapecio con $M = 10$.
(ii) Aproxime la longitud de la curva $y = f(x)$ para cada una de las funciones que se relacionan a continuación usando la regla compuesta de Simpson con $M = 5$.

- (a) $f(x) = x^3$ para $0 \leq x \leq 1$
 (b) $f(x) = \operatorname{sen}(x)$ para $0 \leq x \leq \pi/4$
 (c) $f(x) = e^{-x}$ para $0 \leq x \leq 1$

3. *Área de una superficie de revolución.* El área de la superficie del sólido de revolución que se obtiene al girar alrededor del eje OX la región limitada por la curva $y = f(x)$, siendo $a \leq x \leq b$, viene dada por

$$\text{área} = 2\pi \int_a^b f(x) \sqrt{1 + (f'(x))^2} dx.$$

- (i) En cada uno de los casos siguientes, aproxime el área de la superficie de revolución correspondiente usando la regla compuesta del trapecio con $M = 10$.
 (ii) En cada uno de los casos siguientes, aproxime el área de la superficie de revolución correspondiente usando la regla compuesta de Simpson con $M = 5$.
- (a) $f(x) = x^3$ para $0 \leq x \leq 1$
 (b) $f(x) = \operatorname{sen}(x)$ para $0 \leq x \leq \pi/4$
 (c) $f(x) = e^{-x}$ para $0 \leq x \leq 1$
4. (a) Verifique que la regla del trapecio ($M = 1, h = 1$) es exacta para polinomios de grado menor o igual que 1 en $[0, 1]$.
 (b) Utilice como integrando $f(x) = c_2x^2$ para verificar que el término del error de la regla del trapecio ($M = 1, h = 1$) en el intervalo $[0, 1]$ es

$$E_T(f, h) = \frac{-(b-a)f^{(2)}(c)h^2}{12}.$$

5. (a) Verifique que la regla de Simpson ($M = 1, h = 1$) es exacta para polinomios de grado menor o igual que 3 en $[0, 2]$.
 (b) Utilice como integrando $f(x) = c_4x^4$ para verificar que el término del error de la regla de Simpson ($M = 1, h = 1$) en el intervalo $[0, 2]$ es

$$E_S(f, h) = \frac{-(b-a)f^{(4)}(c)h^4}{180}.$$

6. Deduzca la regla del trapecio ($M = 1, h = 1$) usando el método de los coeficientes indeterminados:
- (a) Halle las constantes w_0 y w_1 de manera que $\int_0^1 g(t) dt = w_0g(0) + w_1g(1)$ sea exacta para las funciones $g(t) = 1$ y $g(t) = t$.
 (b) Use la relación $f(x_0 + ht) = g(t)$ y el cambio de variable $x = x_0 + ht$ con $dx = h dt$ para trasladar la regla del trapecio desde $[0, 1]$ hasta el intervalo $[x_0, x_1]$.

Indicación para el apartado (a). Debe obtener un sistema de dos ecuaciones para las incógnitas w_0 y w_1 .

7. Deduzca la regla de Simpson ($M = 1$, $h = 1$) usando el método de los coeficientes indeterminados:

- (a) Determine las constantes w_0 , w_1 y w_2 de manera que $\int_0^2 g(t) dt = w_0g(0) + w_1g(1) + w_2g(2)$ sea exacta para las funciones $g(t) = 1$, $g(t) = t$ y $g(t) = t^2$.
- (b) Use la relación $f(x_0 + ht) = g(t)$ y el cambio de variable $x = x_0 + ht$ con $dx = h dt$ para trasladar la regla de Simpson desde $[0, 2]$ hasta el intervalo $[x_0, x_2]$.

Indicación para el apartado (a). Debe obtener un sistema de tres ecuaciones para las incógnitas w_0 , w_1 y w_2 .

8. Determine, en cada uno de los siguientes casos, el número M y el tamaño de los subintervalos h de manera que la regla del trapecio con M subintervalos nos permita obtener la integral dada con una precisión de 5×10^{-9} .

(a) $\int_{-\pi/6}^{\pi/6} \cos(x) dx$ (b) $\int_2^3 \frac{1}{5-x} dx$ (c) $\int_0^2 xe^{-x} dx$

Indicación para el apartado (c). $f^{(2)}(x) = (x - 2)e^{-x}$.

9. Determine, en cada uno de los siguientes casos, el número M y el tamaño de los subintervalos h de manera que la regla de Simpson con $2M$ subintervalos nos permita obtener la integral dada con una precisión de 5×10^{-9} .

(a) $\int_{-\pi/6}^{\pi/6} \cos(x) dx$ (b) $\int_2^3 \frac{1}{5-x} dx$ (c) $\int_0^2 xe^{-x} dx$

Indicación para el apartado (c). $f^{(4)}(x) = (x - 4)e^{-x}$.

10. Considere la integral $\int_{-0.1}^{0.1} \cos(x) dx = 2 \sin(0.1) = 0.1996668333$. La siguiente tabla proporciona aproximaciones a este valor usando la regla compuesta del trapecio. Calcule $E_T(f, h) = 0.199668 - T(f, h)$ y verifique que es de orden $O(h^2)$.

M	h	$T(f, h)$	$E_T(f, h) = O(h^2)$
1	0.2	0.1990008	
2	0.1	0.1995004	
4	0.05	0.1996252	
8	0.025	0.1996564	
16	0.0125	0.1996642	

11. Considere la integral $\int_{-0.75}^{0.75} \cos(x) dx = 2 \sin(0.75) = 1.363277520$. La siguiente tabla muestra aproximaciones a este valor que se han calculado usando la regla compuesta de Simpson. Calcule $E_S(f, h) = 1.3632775 - S(f, h)$ y

verifique que es de orden $O(h^4)$.

M	h	$S(f, h)$	$E_S(f, h) = O(h^4)$
1	0.75	1.3658444	
2	0.375	1.3634298	
4	0.1875	1.3632869	
8	0.09375	1.3632781	

12. *La regla del punto medio.* La regla del punto medio en un intervalo $[x_0, x_1]$ es

$$\int_{x_0}^{x_1} f(x) dx = hf \left(x_0 + \frac{h}{2} \right) + \frac{h^3}{24} f^{(2)}(c_1),$$

siendo $h = \frac{x_1 - x_0}{2}$.

- (a) Desarrolle $F(x)$, una primitiva de $f(x)$, en serie de Taylor alrededor de $x_0 + h/2$ para establecer la regla del punto medio en $[x_0, x_1]$.
 (b) Use el apartado (a) para probar que la regla compuesta del punto medio $M(f, h)$ para aproximar la integral de $f(x)$ en $[a, b]$ es

$$\int_a^b f(x) dx \approx M(f, h) = h \sum_{k=1}^N f \left(a + \left(k - \frac{1}{2} \right) h \right), \quad \text{siendo } h = \frac{b-a}{N}.$$

- (c) Pruebe que el término del error $E_M(f, h) = \int_a^b f(x) dx - M(f, h)$ de la regla compuesta dada en el apartado (b) es

$$E_M(f, h) = \frac{h^3}{24} \sum_{k=1}^N f^{(2)}(c_k) = \frac{(b-a)f^{(2)}(c)h^2}{24} = O(h^2).$$

13. Use la regla del punto medio con $M = 10$ para aproximar las integrales del Ejercicio 1.

14. Demuestre el Corolario 7.3.

Algoritmos y programas

- (a) Para cada una de las integrales del Ejercicio 1, calcule M y el tamaño de los subintervalos h de manera que se pueda usar la regla compuesta del trapecio para calcular la integral dada con una precisión de nueve cifras decimales. Use el Programa 7.1 para aproximar cada integral.
 (b) Para cada una de las integrales del Ejercicio 1, calcule M y el tamaño de los subintervalos h de manera que se pueda usar la regla compuesta de Simpson para calcular la integral dada con una precisión de nueve cifras decimales. Use el Programa 7.2 para aproximar cada integral.

2. Use el Programa 7.2 para aproximar las integrales definidas del Ejercicio 2 con una precisión de once cifras decimales.
3. Es posible adaptar la regla compuesta del trapecio para integrar una función de la que se conocen sus valores en una cantidad finita de puntos. Adapte el Programa 7.1 para aproximar la integral de una función en un intervalo $[a, b]$ conociendo su valor en M puntos dados. (Nota. Los nodos no necesitan estar equiespaciados.) Use este programa para aproximar la integral de una función que pasa por los puntos $\{(\sqrt{k^2 + 1}, k^{1/3})\}_{k=0}^{13}$.
4. Es posible adaptar la regla compuesta de Simpson para integrar una función de la que se conocen sus valores en una cantidad finita de puntos. Adapte el Programa 7.2 para aproximar la integral de una función en un intervalo $[a, b]$ conociendo su valor en M puntos dados. (Nota. Los nodos no necesitan estar equiespaciados.) Use este programa para aproximar la integral de una función que pasa por los puntos $\{(\sqrt{k^2 + 1}, k^{1/3})\}_{k=0}^{13}$.
5. Modifique el Programa 7.1 de manera que construya la aproximación dada por la regla del punto medio (Ejercicio 12) para aproximar la integral de $f(x)$ en $[a, b]$. Use este programa para aproximar las integrales del Ejercicio 1 con una precisión de once cifras decimales.
6. Obtenga, usando cualquiera de los programas de esta sección, aproximaciones a cada una de las siguientes integrales con una precisión de diez cifras decimales.
- (a) $\int_{1/7\pi}^{1/4\pi} \sin(1/x) dx$
- (b) $\int_{\frac{1}{5\pi}+10^{-5}}^{\frac{1}{4\pi}-10^{-5}} \frac{1}{\sin(1/x)} dx$
7. El siguiente ejemplo muestra como puede usarse la regla de Simpson para aproximar la solución de una ecuación integral. La ecuación es

$$v(x) = x^2 + 0.1 \int_0^1 (x^2 + t)v(t) dt$$

y empezamos usando la regla de Simpson con $h = 0.5$: Sean $t_0 = 0$, $t_1 = 0.5$ y $t_2 = 1$; entonces

$$\int_0^1 (x^2 + t)v(t) dt \approx \frac{0.5}{3}((x_n^2 + 0)v_0 + 4(x_n^2 + 0.5)v_1 + (x_n^2 + 1)v_2).$$

Ahora, dado un punto $x_n \in [0, 1]$, tomamos

$$(1) \quad v(x_n) = x_n^2 + 0.1\left(\frac{1}{6}((x_n^2 + 0)v_0 + 4(x_n^2 + 0.5)v_1 + (x_n^2 + 1)v_2)\right).$$

Al sustituir $x_0 = 0$, $x_1 = 0.5$ y $x_2 = 1$ en la relación (1) obtenemos el sistema

de ecuaciones:

$$(2) \quad \begin{aligned} v_0 &= 0 + \frac{1}{60}((0)v_0 + 2v_1 + v_2) \\ v_1 &= \frac{1}{4} + \frac{1}{60}(0.25v_0 + 3v_1 + 1.25v_2) \\ v_2 &= 1 + \frac{1}{60}(v_0 + 6v_1 + 2v_2). \end{aligned}$$

Sustituyendo la solución del sistema (2) en la relación (1) y simplificando, obtenemos la aproximación

$$(3) \quad v(x) \approx 1.037305x^2 + 0.027297.$$

- (a) A modo de comprobación, sustituya la solución en el miembro derecho de la ecuación integral, integre, simplifique y compare el resultado con la aproximación dada en (3).
- (b) Use la regla compuesta de Simpson con $h = 0.5$ para aproximar la solución de la ecuación integral

$$v(x) = x^2 + 0.1 \int_0^1 (x^2 + t)v(t) dt;$$

luego, proceda como en el apartado (a) para comprobar su solución.

7.3 Reglas recursivas y método de Romberg

En esta sección mostramos cómo se pueden calcular las aproximaciones de la regla de Simpson usando combinaciones lineales especiales de las aproximaciones dadas por la regla del trapecio. Las aproximaciones mejoran su precisión conforme aumenta el número de subintervalos, así que ¿cuántos deberíamos tomar? El proceso secuencial de tomar dos subintervalos, luego cuatro, luego ocho y así hasta que alcancemos la precisión deseada, nos ayudará a responder esta pregunta. Se comienza generando con la regla del trapecio una sucesión de aproximaciones $\{T(J)\}$. Al doblar el número de subintervalos se dobla, prácticamente, el número de evaluaciones de la función ya que hay que evaluar la función en los puntos previos y en los puntos medios de los subintervalos previos (véase la Figura 7.8). El Teorema 7.4 explica cómo podemos eliminar las evaluaciones y las sumas redundantes.

Teorema 7.4 (Reglas del trapecio sucesivas). Supongamos que $J \geq 1$ y que los puntos $\{x_k = a + kh\}$ dividen $[a, b]$ en $2^J = 2M$ subintervalos del mismo tamaño $h = (b - a)/2^J$. Las reglas del trapecio $T(f, h)$ y $T(f, 2h)$ verifican la relación

$$(1) \quad T(f, h) = \frac{T(f, 2h)}{2} + h \sum_{k=1}^M f(x_{2k-1}).$$

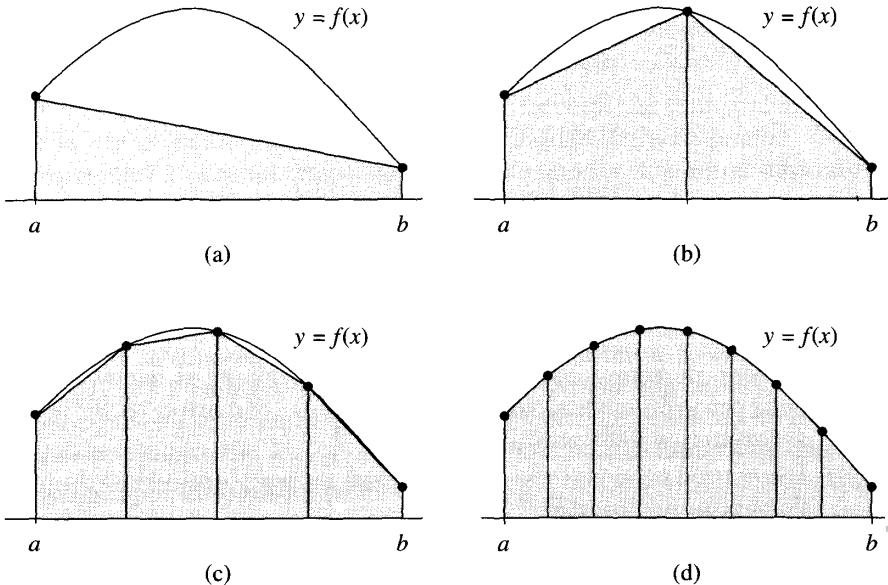


Figura 7.8 (a) $T(0)$ es el área de $2^0 = 1$ trapecio. (b) $T(1)$ es el área de $2^1 = 2$ trapecios. (c) $T(2)$ es el área de $2^2 = 4$ trapecios. (d) $T(3)$ es el área de $2^3 = 8$ trapecios.

Definición 7.3 (Sucesión de aproximaciones con la regla del trapecio). Se define $T(0) = (h/2)(f(a) + f(b))$, que es la regla del trapecio con incremento $h = b - a$ y, para cada $J \geq 1$, se define $T(J) = T(f, h)$, donde $T(f, h)$ es la regla del trapecio con incremento $h = (b - a)/2^J$.

Corolario 7.4 (Regla recursiva del trapecio). A partir del valor inicial $T(0) = (h/2)(f(a) + f(b))$, la sucesión $\{T(J)\}$ de aproximaciones dadas por la regla del trapecio viene generada por la fórmula recursiva

$$(2) \quad T(J) = \frac{T(J-1)}{2} + h \sum_{k=1}^M f(x_{2k-1}) \quad \text{para } J = 1, 2, \dots,$$

siendo $h = (b - a)/2^J$ y $\{x_k = a + kh\}$.

Demostración. En los nodos pares $x_0 < x_2 < \dots < x_{2M-2} < x_{2M}$, usamos la regla del trapecio con tamaño de paso $2h$:

$$(3) \quad T(J-1) = \frac{2h}{2}(f_0 + 2f_2 + 2f_4 + \dots + 2f_{2M-4} + 2f_{2M-2} + f_{2M}).$$

En todos los nodos $x_0 < x_1 < x_2 < \dots < x_{2M-1} < x_{2M}$, usamos la regla del trapecio con tamaño de paso h :

$$(4) \quad T(J) = \frac{h}{2}(f_0 + 2f_1 + 2f_2 + \dots + 2f_{2M-2} + 2f_{2M-1} + f_{2M}).$$

Agrupando los subíndices pares e impares de la expresión (4):

$$(5) \quad T(J) = \frac{h}{2}(f_0 + 2f_2 + \dots + 2f_{2M-2} + f_{2M}) + h \sum_{k=1}^M f_{2k-1}$$

y sustituyendo (3) en (5), obtenemos $T(J) = T(J-1)/2 + h \sum_{k=1}^M f_{2k-1}$, lo que concluye la prueba. •

Ejemplo 7.11. Vamos a usar la regla recursiva del trapecio para calcular las aproximaciones $T(0)$, $T(1)$, $T(2)$ y $T(3)$ a la integral $\int_1^5 dx/x = \ln(5) - \ln(1) = 1.609437912$.

En la Tabla 7.4 se muestran los nueve puntos necesarios para calcular $T(3)$ y los puntos medios necesarios para calcular $T(1)$, $T(2)$ y $T(3)$. Las operaciones detalladas son:

$$\text{Cuando } h = 4: \quad T(0) = \frac{4}{2}(1.000000 + 0.200000) = 2.400000.$$

$$\begin{aligned} \text{Cuando } h = 2: \quad T(1) &= \frac{T(0)}{2} + 2(0.333333) \\ &= 1.200000 + 0.666666 = 1.866666. \end{aligned}$$

$$\begin{aligned} \text{Cuando } h = 1: \quad T(2) &= \frac{T(1)}{2} + 1(0.500000 + 0.250000) \\ &= 0.933333 + 0.750000 = 1.683333. \end{aligned}$$

$$\begin{aligned} \text{Cuando } h = \frac{1}{2}: \quad T(3) &= \frac{T(2)}{2} + \frac{1}{2}(0.666667 + 0.400000 \\ &\quad + 0.285714 + 0.222222) \\ &= 0.841667 + 0.787302 = 1.628968. \end{aligned}$$

Nuestro siguiente resultado establece una relación muy importante entre la regla del trapecio y la regla de Simpson. Cuando calculamos la aproximación dada por la regla del trapecio usando incrementos $2h$ y h , obtenemos $T(f, 2h)$ y $T(f, h)$, respectivamente; pues bien, estos valores pueden combinarse para obtener la aproximación dada por la regla de Simpson:

$$(6) \quad S(f, h) = \frac{4T(f, h) - T(f, 2h)}{3}.$$

Tabla 7.4 Los nueve puntos necesarios para calcular $T(3)$ y los puntos medios necesarios para calcular $T(1)$, $T(2)$ y $T(3)$.

x	$f(x) = \frac{1}{x}$	Extremos para calcular $T(0)$	Puntos medios para calcular $T(1)$	Puntos medios para calcular $T(2)$	Puntos medios para calcular $T(3)$
1.0	1.000000	1.000000			
1.5	0.666667				0.666667
2.0	0.500000			0.500000	0.400000
2.5	0.400000				
3.0	0.333333		0.333333		
3.5	0.285714				0.285714
4.0	0.250000			0.250000	
4.5	0.222222				
5.0	0.200000	0.200000			0.222222

Teorema 7.5 (Regla recursiva de Simpson). Supongamos que $\{T(J)\}$ es la sucesión de aproximaciones obtenidas con la regla del trapecio generada recursivamente como se muestra en el Corolario 7.4. Si $J \geq 1$ y $S(J)$ es la aproximación dada por la regla de Simpson con 2^J subintervalos de $[a, b]$, entonces $S(J)$ y las aproximaciones obtenidas con la regla del trapecio $T(J - 1)$ y $T(J)$ verifican la relación

$$(7) \quad S(J) = \frac{4T(J) - T(J - 1)}{3} \quad \text{para } J = 1, 2, \dots$$

Demostración. La regla del trapecio con incremento h proporciona la aproximación

$$(8) \quad \begin{aligned} \int_a^b f(x) dx &\approx \frac{h}{2}(f_0 + 2f_1 + 2f_2 + \cdots + 2f_{2M-2} + 2f_{2M-1} + f_{2M}) \\ &= T(J) \end{aligned}$$

y la regla del trapecio con incremento $2h$ produce

$$(9) \quad \int_a^b f(x) dx \approx h(f_0 + 2f_2 + \cdots + 2f_{2M-2} + f_{2M}) = T(J - 1).$$

Multiplicando la relación (8) por 4:

$$(10) \quad \begin{aligned} 4 \int_a^b f(x) dx &\approx h(2f_0 + 4f_1 + 4f_2 + \cdots + 4f_{2M-2} + 4f_{2M-1} + 2f_M) \\ &= 4T(J) \end{aligned}$$

y restando (9) de (10) obtenemos

$$(11) \quad 3 \int_a^b f(x) dx \approx h(f_0 + 4f_1 + 2f_2 + \cdots + 2f_{2M-2} + 4f_{2M-1} + f_{2M}) \\ = 4T(J) - T(J-1),$$

expresión que podemos arreglar un poco para deducir

$$(12) \quad \int_a^b f(x) dx \approx \frac{h}{3}(f_0 + 4f_1 + 2f_2 + \cdots + 2f_{2M-2} + 4f_{2M-1} + f_{2M}) \\ = \frac{4T(J) - T(J-1)}{3}.$$

El término central de la relación (12) es la aproximación dada por la regla de Simpson $S(J) = S(f, h)$, lo que concluye la prueba del teorema. •

Ejemplo 7.12. Vamos a usar la regla recursiva de Simpson para calcular las aproximaciones $S(1)$, $S(2)$ y $S(3)$ a la integral del Ejemplo 7.11.

Usando los resultados del Ejemplo 7.11 y la fórmula (7) con $J = 1, 2$ y 3 , obtenemos

$$S(1) = \frac{4T(1) - T(0)}{3} = \frac{4(1.866666) - 2.400000}{3} = 1.688888,$$

$$S(2) = \frac{4T(2) - T(1)}{3} = \frac{4(1.683333) - 1.866666}{3} = 1.622222,$$

$$S(3) = \frac{4T(3) - T(2)}{3} = \frac{4(1.628968) - 1.683333}{3} = 1.610846. \blacksquare$$

En la Sección 7.1 introdujimos la fórmula de aproximación de la regla de Boole en el Teorema 7.1. Para deducirla, lo que hicimos fue integrar el polinomio interpolador de Lagrange de grado cuatro para los nodos x_0, x_1, x_2, x_3 y x_4 . (Un método alternativo para establecer la regla de Boole se menciona en los ejercicios.) La aplicación de la regla de Boole M veces sobre $4M$ subintervalos de $[a, b]$ que tienen todos el mismo tamaño $h = (b - a)/(4M)$, se llama **regla compuesta de Boole**:

$$(13) \quad B(f, h) = \frac{2h}{45} \sum_{k=1}^M (7f_{4k-4} + 32f_{4k-3} + 12f_{4k-2} + 32f_{4k-1} + 7f_{4k}).$$

El siguiente resultado muestra la relación existente entre las reglas recursivas de Boole y Simpson.

Teorema 7.6 (Regla recursiva de Boole). Supongamos que $\{S(J)\}$ es la sucesión recursiva de las aproximaciones dadas por el método de Simpson generadas como se explica en el Teorema 7.5. Si $J \geq 2$ y $B(J)$ es la aproximación

dada por la regla de Boole con 2^J subintervalos de $[a, b]$, entonces $B(J)$ y las aproximaciones obtenidas con la regla del Simpson $S(J - 1)$ y $S(J)$ verifican la relación

$$(14) \quad B(J) = \frac{16S(J) - S(J - 1)}{15} \quad \text{para } J = 2, 3, \dots$$

Demostración. La demostración se deja como ejercicio. •

Ejemplo 7.13. Vamos a usar la regla recursiva de Boole para calcular las aproximaciones $B(2)$ y $B(3)$ a la integral del Ejemplo 7.11.

Usando los resultados del Ejemplo 7.12 y la fórmula (14) con $J = 2$ y 3 , obtenemos

$$B(2) = \frac{16S(2) - S(1)}{15} = \frac{16(1.622222) - 1.688888}{15} = 1.617778,$$

$$B(3) = \frac{16S(3) - S(2)}{15} = \frac{16(1.610846) - 1.622222}{15} = 1.610088.$$

Puede que usted se pregunte hasta dónde queremos llegar. Vamos a probar ahora que las fórmulas (7) y (14) son casos especiales del proceso conocido como método de integración de Romberg; pero antes anunciamos ya que el siguiente nivel de aproximación a la integral del Ejemplo 7.11 es

$$\frac{64B(3) - B(2)}{63} = \frac{64(1.610088) - 1.617778}{63} = 1.609490,$$

respuesta que tiene una precisión de cinco cifras decimales.

Método de integración de Romberg

En la Sección 7.2 vimos que los términos del error $E_T(f, h)$ y $E_S(f, h)$ de las reglas compuestas del trapecio y Simpson son de orden $\mathcal{O}(h^2)$ y $\mathcal{O}(h^4)$, respectivamente. No es difícil probar que el término del error para la regla compuesta de Boole $E_B(f, h)$ es de orden $\mathcal{O}(h^6)$; así que tenemos el patrón

$$(15) \quad \int_a^b f(x) dx = T(f, h) + \mathcal{O}(h^2),$$

$$(16) \quad \int_a^b f(x) dx = S(f, h) + \mathcal{O}(h^4),$$

$$(17) \quad \int_a^b f(x) dx = B(f, h) + \mathcal{O}(h^6).$$

El patrón de los restos en las fórmulas (15), (16) y (17) puede extenderse en el siguiente sentido: Supongamos que usamos una aproximación con incrementos h y $2h$, entonces podemos manipular algebraicamente ambas respuestas para

obtener una respuesta mejorada, de manera que cada nivel de mejora incrementa el orden del término del error de $O(h^{2N})$ a $O(h^{2N+2})$. Este proceso, llamado **método de integración de Romberg**, tiene sus ventajas pero también sus inconvenientes.

Las fórmulas de Newton-Cotes apenas se usan más allá de la regla de Boole, esto se debe a que en las reglas de cuadratura de Newton-Cotes con nueve o más nodos aparecen pesos negativos que podrían producir errores por pérdidas de cifras significativas en los redondeos. El método de Romberg tiene las ventajas de que todos los pesos son positivos y de que las abscisas equiespaciadas son fáciles de calcular.

Una de las debilidades computacionales del método de Romberg es que para reducir el orden del error de $O(h^{2N})$ a $O(h^{2N+2})$ es necesario realizar el doble de evaluaciones de la función. Sin embargo, el uso de las reglas recursivas permite mantener bajo el número de operaciones. El desarrollo del método de Romberg se basa en el hecho teórico de que si $f \in C^N[a, b]$ para todo N , entonces el término del error de la regla del trapecio puede representarse como una serie de potencias de h que sólo contiene potencias pares; es decir,

$$(18) \quad \int_a^b f(x) dx = T(f, h) + E_T(f, h),$$

siendo

$$(19) \quad E_T(f, h) = a_1 h^2 + a_2 h^4 + a_3 h^6 + \dots$$

Una deducción de la fórmula (19) puede encontrarse en la Referencia [153].

Puesto que en el desarrollo dado en (19) sólo aparecen potencias pares de h , podemos usar el método de Richardson, que vimos en el capítulo anterior, para ir eliminando a_1, a_2, a_3, \dots , y generar fórmulas de cuadratura cuyos términos del error tengan órdenes de aproximación $O(h^4)$, $O(h^6)$, $O(h^8)$ y así sucesivamente. Vamos a probar que la primera mejora que se obtiene es la regla de Simpson con $2M$ intervalos. Partiendo de $T(f, 2h)$ y $T(f, h)$ y de las relaciones

$$(20) \quad \int_a^b f(x) dx = T(f, 2h) + a_1 4h^2 + a_2 16h^4 + a_3 64h^6 + \dots$$

e

$$(21) \quad \int_a^b f(x) dx = T(f, h) + a_1 h^2 + a_2 h^4 + a_3 h^6 + \dots,$$

multiplicamos la relación (21) por 4,

$$(22) \quad 4 \int_a^b f(x) dx = 4T(f, h) + a_1 4h^2 + a_2 4h^4 + a_3 4h^6 + \dots,$$

y eliminamos a_1 restando (20) de (22), con lo que obtenemos

$$(23) \quad 3 \int_a^b f(x) dx = 4T(f, h) - T(f, 2h) - a_2 12h^4 - a_3 60h^6 - \dots$$

Ahora dividimos la relación (23) entre 3 y renombramos los coeficientes de la serie para obtener

$$(24) \quad \int_a^b f(x) dx = \frac{4T(f, h) - T(f, 2h)}{3} + b_1 h^4 + b_2 h^6 + \dots$$

Como vimos en (6), el primer término del miembro derecho de (24) es la fórmula de cuadratura de la regla de Simpson $S(f, h)$. La relación (26) prueba, además, que el término $E_S(f, h)$ contiene sólo potencias pares de h :

$$(25) \quad \int_a^b f(x) dx = S(f, h) + b_1 h^4 + b_2 h^6 + b_3 h^8 + \dots$$

Para probar que la mejora que se obtiene en el segundo paso es la regla de Boole, empezamos con la relación (25) y con la fórmula correspondiente al incremento $2h$ que se obtendría para $S(f, 2h)$:

$$(26) \quad \int_a^b f(x) dx = S(f, 2h) + b_1 16h^4 + b_2 64h^6 + b_3 256h^8 + \dots$$

Cuando eliminamos b_1 usando (25) y (26), lo que obtenemos es la regla de Boole

$$(27) \quad \begin{aligned} \int_a^b f(x) dx &= \frac{16S(f, h) - S(f, 2h)}{15} - \frac{b_2 48h^6}{15} - \frac{b_3 240h^8}{15} - \dots \\ &= B(f, h) - \frac{b_2 48h^6}{15} - \frac{b_3 240h^8}{15} - \dots \end{aligned}$$

El esquema general para el método de integración de Romberg se basa en el Lema 7.1.

Lema 7.1 (Esquema de Richardson para el método de integración de Romberg). Dadas dos aproximaciones $R(2h, K - 1)$ y $R(h, K - 1)$ de una cantidad Q que verifican

$$(28) \quad Q = R(h, K - 1) + c_1 h^{2K} + c_2 h^{2K+2} + \dots$$

y

$$(29) \quad Q = R(2h, K - 1) + c_1 4^K h^{2K} + c_2 4^{K+1} h^{2K+2} + \dots,$$

entonces podemos construir una aproximación mejor que viene dada por la fórmula

$$(30) \quad Q = \frac{4^K R(h, K - 1) - R(2h, K - 1)}{4^K - 1} + O(h^{2K+2}).$$

La demostración queda propuesta como ejercicio.

Tabla 7.5 Esquema de integración de Romberg.

J	$R(J, 0)$ Regla del trapezio	$R(J, 1)$ Regla de Simpson	$R(J, 2)$ Regla de Boole	$R(J, 3)$ Tercera mejora	$R(J, 4)$ Cuarta mejora
0	$R(0, 0)$				
1	$R(1, 0)$	$R(1, 1)$			
2	$R(2, 0)$	$R(2, 1)$	$R(2, 2)$		
3	$R(3, 0)$	$R(3, 1)$	$R(3, 2)$	$R(3, 3)$	
4	$R(4, 0)$	$R(4, 1)$	$R(4, 2)$	$R(4, 3)$	$R(4, 4)$

Definición 7.4. Se define la sucesión $\{R(J, K) : J \geq K\}_{J=0}^{\infty}$ de fórmulas de cuadratura para aproximar la integral de $f(x)$ en $[a, b]$ de la siguiente manera

$$(31) \quad \begin{aligned} R(J, 0) &= T(J) && \text{para } J \geq 0, \text{ que es la regla recursiva del trapezio.} \\ R(J, 1) &= S(J) && \text{para } J \geq 1, \text{ que es la regla recursiva de Simpson.} \\ R(J, 2) &= B(J) && \text{para } J \geq 2, \text{ que es la regla recursiva de Boole.} \end{aligned}$$

Las fórmulas $\{R(J, 0)\}$ de partida se usan para generar las primeras mejoras $\{R(J, 1)\}$ que, a su vez, se usan para generar las segundas mejoras $\{R(J, 2)\}$; ya hemos visto como se hace esto:

$$(32) \quad \begin{aligned} R(J, 1) &= \frac{4^1 R(J, 0) - R(J - 1, 0)}{4^1 - 1} && \text{para } J \geq 1 \\ R(J, 2) &= \frac{4^2 R(J, 1) - R(J - 1, 1)}{4^2 - 1} && \text{para } J \geq 2, \end{aligned}$$

que no son más que las relaciones (24) y (27) escritas con la notación dada en (31). La regla general para construir recursivamente las mejoras es

$$(33) \quad R(J, K) = \frac{4^K R(J, K - 1) - R(J - 1, K - 1)}{4^K - 1} \quad \text{para } J \geq K.$$

En la práctica, los valores $R(J, K)$ se disponen en una tabla, que se llama esquema de integración de Romberg, como la Tabla 7.5.

Ejemplo 7.14. Vamos a usar el método de integración de Romberg para calcular aproximaciones a la integral definida

$$\int_0^{\pi/2} (x^2 + x + 1) \cos(x) dx = -2 + \frac{\pi}{2} + \frac{\pi^2}{4} = 2.038197427067\dots$$

Los cálculos se muestran en la Tabla 7.6, donde se puede apreciar que, en cada columna, los valores convergen a $2.038197427067\dots$ y, también, que los valores

Tabla 7.6 Esquema de integración de Romberg para el Ejemplo 7.14.

J	$R(J, 0)$ Regla del trapezio	$R(J, 1)$ Regla de Simpson	$R(J, 2)$ Regla de Boole	$R(J, 3)$ Tercera mejora
0	0.785398163397			
1	1.726812656758	2.040617487878		
2	1.960534166564	2.038441336499	2.038296259740	
3	2.018793948078	2.038213875249	2.038198711166	2.038197162776
4	2.033347341805	2.038198473047	2.038197446234	2.038197426156
5	2.036984954990	2.038197492719	2.038197427363	2.038197427064

Tabla 7.7 Tabla de los errores del esquema de Romberg para el Ejemplo 7.14.

J	h	$E(J, 0) = O(h^2)$	$E(J, 1) = O(h^4)$	$E(J, 2) = O(h^6)$	$E(J, 3) = O(h^8)$
0	$b - a$	-1.252799263670			
1	$\frac{b - a}{2}$	-0.311384770309	0.002420060811		
2	$\frac{b - a}{4}$	-0.077663260503	0.000243909432	0.0000098832673	
3	$\frac{b - a}{8}$	-0.019403478989	0.000016448182	0.000001284099	-0.000000264291
4	$\frac{b - a}{16}$	-0.004850085262	0.000001045980	0.000000019167	-0.000000000912
5	$\frac{b - a}{32}$	-0.001212472077	0.000000065651	0.000000000296	-0.000000000003

dados por la regla de Simpson convergen más rápidamente que los dados por la regla del trapezio; más generalmente, en este ejemplo la convergencia en cada columna es más rápida que en la columna adyacente por la izquierda.

La convergencia de los valores dados en el esquema de Romberg que se muestra en la Tabla 7.6 se aprecia mejor si miramos los errores $E(J, K) = -2 + \pi/2 + \pi^2/4 - R(J, K)$. Sea $h = b - a$ la anchura del intervalo de partida y supongamos que las derivadas de orden superior de $f(x)$ son todas de la misma magnitud. Entonces el error en la columna K -ésima del esquema de Romberg debe ir disminuyendo según, más o menos, el factor $1/2^{2K+2} = 1/4^{K+1}$ conforme descendemos: Los errores $E(J, 0)$ disminuyen según el factor $1/4$, los errores $E(J, 1)$ disminuyen según $1/16$, y así sucesivamente. Esto puede observarse analizando los elementos $\{E(J, K)\}$ que se muestran en la Tabla 7.7. ■

Teorema 7.7 (Precisión del método de integración de Romberg). Supongamos que $f \in C^{2K+2}[a, b]$. Entonces los errores de truncamiento de las aproximaciones generadas con el método de Romberg vienen dadas por

$$(34) \quad \int_a^b f(x) dx = R(J, K) + b_K h^{2K+2} f^{(2K+2)}(c_{J,K}) \\ = R(J, K) + O(h^{2K+2}),$$

donde $h = (b - a)/2^J$, la constante b_K sólo depende de K y $c_{J,K}$ es un punto que está en $[a, b]$ (véase la Referencia [153], página 126).

Ejemplo 7.15. Vamos a aplicar el Teorema 7.7 para probar que

$$\int_0^2 10x^9 dx = 1024 \equiv R(4, 4).$$

El integrando $f(x) = 10x^9$ verifica que $f^{(10)}(x) \equiv 0$; por tanto, para el valor $K = 4$ el término del error debe ser idénticamente cero. Haciendo los cálculos se comprueba que $R(4, 4) = 1024$. ■

MATLAB

Programa 7.3 (Regla recursiva del Trapecio). Construcción de las aproximaciones

$$\int_a^b f(x) dx \approx \frac{h}{2} \sum_{k=1}^{2^J} (f(x_{k-1}) + f(x_k))$$

usando recursivamente la regla del trapecio conforme se incrementa el número de subintervalos de $[a, b]$. En la iteración J -ésima se toman los valores de $f(x)$ en $2^J + 1$ nodos equiespaciados.

```
function T=rctrap(f,a,b,n)
% Datos
%
% - f es el integrando, dado como una
%   cadena de caracteres 'f'
%
% - a y b son los extremos inferior y superior del
%   intervalo de integración
%
% - n es el número de veces que se hace la recursión
%
% Resultado
%
% - T es la lista de las aproximaciones obtenidas con la
%   regla recursiva del trapecio
```

```
M=1;
h=b-a;
T=zeros(1,n+1);
T(1)=h*(feval(f,a)+feval(f,b))/2;
for j=1:n
    M=2*M;
    h=h/2;
    s=0;
    for k=1:M/2
        x=a+h*(2*k-1);
        s=s+feval(f,x);
    end
    T(j+1)=T(j)/2+h*s;
end
```

Programa 7.4 (Método de Integración de Romberg). Construcción de la tabla de aproximaciones $R(J, K)$ (para $J \geq K$) a la integral

$$\int_a^b f(x) dx \approx R(J, K)$$

dando $R(J + 1, J + 1)$ como respuesta final. Las aproximaciones $R(J, K)$ se guardan en una matriz triangular inferior R : los elementos $R(J, 0)$ están en la primera columna de R y son las aproximaciones obtenidas con la regla recursiva del trapecio con 2^{J-1} subintervalos de $[a, b]$; los restantes elementos $R(J, K)$ (que se almacena en $R(J, K+1)$) se calculan usando el método de Romberg de manera que los elementos de la fila J -ésima de R son

$$R(J, K) = R(J, K - 1) + \frac{R(J, K - 1) - R(J - 1, K - 1)}{4^K - 1},$$

para $1 \leq K \leq J$. El criterio de parada del programa es que se termina de iterar en la fila $(J + 1)$ -ésima cuando $|R(J, J) - R(J + 1, J + 1)| < \text{tol}$.

```
function [R,quad,err,h]=romber(f,a,b,n,tol)
% Datos
%     - f es el integrando, dado como una
%       cadena de caracteres 'f'
%     - a y b son los extremos inferior y superior del
%       intervalo de integración
%     - n es el número máximo de filas de la tabla
%     - tol es la tolerancia
% Resultados
%     - R es el esquema de Romberg
```

```

% - quad es la aproximación a la integral
% - err es una estimación del error
% - h es el menor de los incrementos usados

M=1;
h=b-a;
err=1;
J=0;
R=zeros(4,4);
R(1,1)=h*(feval(f,a)+feval(f,b))/2;
while((err>tol)&(J<n))|(J<4)
    J=J+1;
    h=h/2;
    s=0;
    for p=1:M
        x=a+h*(2*p-1);
        s=s+feval(f,x);
    end
    R(J+1,1)=R(J,1)/2+h*s;
    M=2*M;
    for K=1:J
        R(J+1,K+1)=R(J+1,K)+(R(J+1,K)-R(J,K))/(4^K-1);
    end
    err=abs(R(J,J)-R(J+1,K+1));
end
quad=R(J+1,J+1);

```

Ejercicios

1. Para cada una de las siguientes integrales definidas, construya (a mano) el esquema de Romberg (Tabla 7.5) con tres filas.

(a) $\int_0^3 \frac{\sin(2x)}{1+x^2} dx = 0.6717578646\dots$

(b) $\int_0^3 \sin(4x)e^{-2x} dx = 0.1997146621\dots$

(c) $\int_{0.04}^1 \frac{1}{\sqrt{x}} dx = 1.6$

(d) $\int_0^2 \frac{1}{x^2 + \frac{1}{10}} dx = 4.4713993943\dots$

(e) $\int_{1/(2\pi)}^2 \sin\left(\frac{1}{x}\right) dx = 1.1140744942\dots$

$$(f) \int_0^2 \sqrt{4 - x^2} dx = \pi = 3.1415926535\dots$$

2. Supongamos que la regla recursiva del trapecio converge a L (es decir, que $\lim_{J \rightarrow \infty} T(J) = L$).
- Pruebe que la regla recursiva de Simpson también converge a L (o sea, $\lim_{J \rightarrow \infty} S(J) = L$).
 - Pruebe que la regla recursiva de Boole también converge a L (o sea, $\lim_{J \rightarrow \infty} B(J) = L$).
3. (a) Compruebe que la regla de Boole ($M = 1, h = 1$) es exacta para polinomios de grado menor o igual que cinco en $[0, 4]$.
- (b) Use $f(x) = c_6 x^6$ como integrando para comprobar que el término del error para la regla de Boole ($M = 1, h = 1$) en el intervalo $[0, 4]$ es

$$E_B(f, h) = \frac{-2(b-a)f^{(6)}(c)h^6}{945}.$$

4. Deduzca la regla de Boole ($M = 1, h = 1$) usando el método de los coeficientes indeterminados: Determine las constantes w_0, w_1, w_2, w_3 y w_4 de manera que

$$\int_0^4 g(t) dt = w_0 g(0) + w_1 g(1) + w_2 g(2) + w_3 g(3) + w_4 g(4)$$

sea exacta para las funciones $g(t) = 1, t, t^2, t^3$ y t^4 . *Indicación.* Debe llegar al sistema de ecuaciones lineales:

$$\begin{aligned} w_0 + w_1 + w_2 + w_3 + w_4 &= 4 \\ w_1 + 2w_2 + 3w_3 + 4w_4 &= 8 \\ w_1 + 4w_2 + 9w_3 + 16w_4 &= \frac{64}{3} \\ w_1 + 8w_2 + 27w_3 + 64w_4 &= 64 \\ w_1 + 16w_2 + 81w_3 + 256w_4 &= \frac{1024}{5} \end{aligned}$$

5. Establezca la relación $B(J) = (16S(J) - S(J-1))/15$ para el caso $J = 2$ usando los siguientes hechos:

$$S(1) = \frac{2h}{3}(f_0 + 4f_2 + f_4)$$

y

$$S(2) = \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + f_4).$$

6. Regla $\frac{3}{8}$ de Simpson. Consideremos dos reglas del trapecio sobre el intervalo $[x_0, x_3]$: la primera, $T(f, 3h) = (3h/2)(f_0 + f_3)$ con incremento $3h$ y, la segunda, $T(f, h) = (h/2)(f_0 + 2f_1 + 2f_2 + f_3)$ con incremento h . Pruebe que la combinación lineal $(9T(f, h) - T(f, 3h))/8$ coincide con la regla $\frac{3}{8}$ de Simpson.

7. Use las relaciones (25) y (26) para establecer la relación (27).
8. Use las relaciones (28) y (29) para establecer la relación (30).
9. Determine el menor número natural K para el que
 - $\int_0^2 8x^7 dx = 256 \equiv R(K, K)$.
 - $\int_0^2 11x^{10} dx = 2048 \equiv R(K, K)$.
10. Se ha usado el método de integración de Romberg para aproximar las integrales (i) $\int_0^1 \sqrt{x} dx$ y (ii) $\int_0^1 2t^2 dt$ y los resultados obtenidos se muestran en la tabla siguiente:

Aproximaciones a (i)	Aproximaciones a (ii)
$R(0, 0) = 0.5000000$	$R(0, 0) = 1.0000000$
$R(1, 1) = 0.6380712$	$R(1, 1) = 0.6666667$
$R(2, 2) = 0.6577566$	$R(2, 2) = 0.6666667$
$R(3, 3) = 0.6636076$	$R(3, 3) = 0.6666667$
$R(4, 4) = 0.6655929$	$R(4, 4) = 0.6666667$

- (a) Use el cambio de variable $x = t^2$, con lo cual $dx = 2t dt$, para probar que las dos integrales tienen el mismo valor numérico.
- (b) ¿Por qué es más lenta la convergencia de la sucesión generada por el método de Romberg en la integral (i) que en la integral (ii)?
11. *Método de Romberg basado en la regla del punto medio.* La regla compuesta del punto medio es comparable con la regla compuesta del trapecio en lo que respecta al coste computacional y a la velocidad de convergencia. En la regla del punto medio

$$\int_a^b f(x) dx = M(f, h) + E_M(f, h),$$

la fórmula de cuadratura $M(f, h)$ y su término del error $E_M(f, h)$ vienen dados por

$$M(f, h) = h \sum_{k=1}^N f \left(a + \left(k - \frac{1}{2} \right) h \right), \quad \text{siendo } h = \frac{b-a}{N},$$

y

$$E_M(f, h) = a_1 h^2 + a_2 h^4 + a_3 h^6 + \dots$$

- A partir de

$$M(0) = \frac{b-a}{2} f \left(\frac{a+b}{2} \right),$$

desarrolle la regla del punto medio de forma sucesiva:

$$M(J) = M(f, h_J) = h_J \sum_{k=1}^{2^J} f \left(a + \left(k - \frac{1}{2} \right) h_J \right),$$

siendo $h_J = \frac{b-a}{2^J}$.

- (b) Muestre que la regla sucesiva del punto medio puede usarse en lugar de la regla recursiva del trapecio en el método de integración de Romberg.

Algoritmos y programas

1. Use el Programa 7.4 para dar aproximaciones a las integrales del Ejercicio 1 que tengan una precisión de once cifras decimales.
2. Use el Programa 7.4 para dar aproximaciones a las dos integrales definidas siguientes que tengan una precisión de diez cifras decimales. Sabiendo que el valor exacto de cada una de ellas es π , explique las diferencias que puedan aparecer en las velocidades de convergencia de las dos sucesiones generadas por el método de Romberg.
 - $$(a) \int_0^2 \sqrt{4x - x^2} dx$$
 - $$(b) \int_0^1 \frac{4}{1+x^2} dx$$
3. La función de densidad para la distribución de probabilidad normal es $f(t) = (1/\sqrt{2\pi})e^{-t^2/2}$ y la función de distribución correspondiente viene dada por

$$\Phi(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt.$$

Calcule los valores de $\Phi(0.5)$, $\Phi(1.0)$, $\Phi(1.5)$, $\Phi(2.0)$, $\Phi(2.5)$, $\Phi(3.0)$, $\Phi(3.5)$ y $\Phi(4.0)$ con una precisión de ocho cifras decimales.

4. Modifique el Programa 7.3 de manera que el criterio de parada sea el de terminar las iteraciones cuando dos aproximaciones sucesivas $T(K-1)$ y $T(K)$ de la regla recursiva del trapecio difieran en menos que 5×10^{-6} .
5. Modifique el Programa 7.3 de manera que sirva para calcular las aproximaciones dadas por las reglas recursivas de Simpson y Boole.
6. Modifique el Programa 7.4 de manera que sirva para utilizar el método de integración de Romberg generado a partir de la regla del punto medio (use los resultados del Ejercicio 11). Use su programa para aproximar las siguientes integrales con una precisión de diez cifras decimales.

- $$(a) \int_0^1 \frac{\sin(x)}{x} dx$$
- $$(b) \int_{-1}^1 \sqrt{1-x^2} dx$$

7. En el Programa 7.4, las aproximaciones a la integral definida dada se almacenan en la diagonal principal de una matriz triangular inferior. Modifique el Programa 7.4 de manera que las filas sucesivas del esquema de Romberg se almacenen en una matriz columna R de orden $n \times 1$, con el consiguiente ahorro de memoria del computador. Compruebe su programa con las integrales del Ejercicio 1.

7.4 Integración adaptativa

Las reglas compuestas de cuadratura necesitan nodos equiespaciados; típicamente, se usa un incremento pequeño h de manera uniforme en todo el intervalo de integración para garantizar una precisión global. Este proceso no tiene en cuenta el hecho de que en algunas porciones de la curva puedan aparecer oscilaciones más pronunciadas que en otras y, en consecuencia, requieran incrementos más pequeños para conseguir la misma precisión. Sería entonces interesante disponer de un método que vaya ajustando el incremento de manera que sea menor en aquellas porciones de la curva en las que aparezcan oscilaciones más pronunciadas. El método que vamos a presentar, que se basa en la regla de Simpson, se llama **integración adaptativa** o **cuadratura adaptativa**.

La regla de Simpson utiliza dos subintervalos en $[a_k, b_k]$:

$$(1) \quad S(a_k, b_k) = \frac{h}{3}(f(a_k) + 4f(c_k) + f(b_k)),$$

donde $c_k = \frac{1}{2}(a_k + b_k)$ es el centro de $[a_k, b_k]$ y $h = (b_k - a_k)/2$. Es más, si $f \in C^4[a_k, b_k]$, entonces existe un punto $d_1 \in [a_k, b_k]$ tal que

$$(2) \quad \int_{a_k}^{b_k} f(x) dx = S(a_k, b_k) - h^5 \frac{f^{(4)}(d_1)}{90}.$$

Refinamiento

La regla compuesta de Simpson que utiliza cuatro subintervalos de $[a_k, b_k]$ se obtiene dividiendo este intervalo en dos subintervalos del mismo tamaño $[a_{k1}, b_{k1}]$ y $[a_{k2}, b_{k2}]$ y aplicando la fórmula (1) en cada trozo, para lo cual sólo necesitamos dos evaluaciones adicionales de $f(x)$:

$$(3) \quad \begin{aligned} S(a_{k1}, b_{k1}) + S(a_{k2}, b_{k2}) &= \frac{h}{6}(f(a_{k1}) + 4f(c_{k1}) + f(b_{k1})) \\ &\quad + \frac{h}{6}(f(a_{k2}) + 4f(c_{k2}) + f(b_{k2})), \end{aligned}$$

donde $a_{k1} = a_k$, $b_{k1} = a_{k2} = c_k$, $b_{k2} = b_k$, c_{k1} es el punto medio de $[a_{k1}, b_{k1}]$ y c_{k2} es el punto medio de $[a_{k2}, b_{k2}]$. En la fórmula (3) el incremento es $h/2$, razón por la cual aparece $h/6$ en el miembro derecho de la fórmula. Es más, si $f \in C^4[a, b]$, entonces existe un valor $d_2 \in [a_k, b_k]$ tal que

$$(4) \quad \int_{a_k}^{b_k} f(x) dx = S(a_{k1}, b_{k1}) + S(a_{k2}, b_{k2}) - \frac{h^5}{16} \frac{f^{(4)}(d_2)}{90}.$$

Supongamos que $f^{(4)}(d_1) \approx f^{(4)}(d_2)$, entonces podemos usar los miembros derechos de las fórmulas (2) y (4) para obtener la relación

$$(5) \quad S(a_k, b_k) - h^5 \frac{f^{(4)}(d_2)}{90} \approx S(a_{k1}, b_{k1}) + S(a_{k2}, b_{k2}) - \frac{h^5}{16} \frac{f^{(4)}(d_2)}{90}$$

que podemos escribir como

$$(6) \quad -h^5 \frac{f^{(4)}(d_2)}{90} \approx \frac{16}{15} (S(a_{k1}, b_{k1}) + S(a_{k2}, b_{k2}) - S(a_k, b_k)).$$

Sustituyendo (6) en (4) obtenemos la estimación del error:

$$(7) \quad \begin{aligned} & \left| \int_{a_k}^{b_k} f(x) dx - S(a_{k1}, b_{k1}) - S(a_{k2}, b_{k2}) \right| \\ & \approx \frac{1}{15} |S(a_{k1}, b_{k1}) + S(a_{k2}, b_{k2}) - S(a_k, b_k)|. \end{aligned}$$

Para compensar el uso de la hipótesis de que $f^{(4)}(d_1) \approx f^{(4)}(d_2)$, a la hora de desarrollar el método, la fracción $\frac{1}{15}$ se sustituye por la más conservadora $\frac{1}{10}$ en el miembro derecho de (7). Esto justifica plantearse el siguiente criterio de exactitud.

Criterio de exactitud

Supongamos que especificamos una tolerancia $\varepsilon_k > 0$ para la aproximación a la integral en el intervalo $[a_k, b_k]$. Si

$$(8) \quad \frac{1}{10} |S(a_{k1}, b_{k1}) + S(a_{k2}, b_{k2}) - S(a_k, b_k)| < \varepsilon_k,$$

entonces aceptamos que

$$(9) \quad \left| \int_{a_k}^{b_k} f(x) dx - S(a_{k1}, b_{k1}) - S(a_{k2}, b_{k2}) \right| < \varepsilon_k.$$

En consecuencia, usamos la regla compuesta de Simpson (3) para aproximar la integral

$$(10) \quad \int_{a_k}^{b_k} f(x) dx \approx S(a_{k1}, b_{k1}) + S(a_{k2}, b_{k2})$$

y se asume ε_k como cota del error de esta aproximación en $[a_k, b_k]$.

El proceso de integración adaptativa se construye aplicando las reglas de Simpson (1) y (3) de la siguiente manera: Partimos de $\{[a_0, b_0], \varepsilon_0\}$, donde ε_0

es la tolerancia para la cuadratura numérica en $[a_0, b_0]$. Este intervalo se divide en sus dos mitades, que etiquetamos como $[a_{01}, b_{01}]$ y $[a_{02}, b_{02}]$. Si el criterio de exactitud (8) se verifica, entonces la fórmula de cuadratura (3) se aplica en el intervalo $[a_0, b_0]$ y hemos terminado. Por el contrario, si el criterio de exactitud (8) no se verifica, entonces ambos subintervalos se vuelven a etiquetar como $[a_1, b_1]$ y $[a_2, b_2]$, en los que usamos como tolerancias $\varepsilon_1 = \frac{1}{2}\varepsilon_0$ y $\varepsilon_2 = \frac{1}{2}\varepsilon_0$, respectivamente. De esa manera, ahora tenemos dos intervalos con sus tolerancias asociadas a los que tenemos que dividir y aplicar el criterio de exactitud, a saber: $\{[a_1, b_1], \varepsilon_1\}$ y $\{[a_2, b_2], \varepsilon_2\}$, siendo $\varepsilon_1 + \varepsilon_2 = \varepsilon_0$. Vamos, entonces, a continuar el proceso de integración adaptativa con estos dos intervalos y sus tolerancias respectivas.

En el segundo paso, consideramos $\{[a_1, b_1], \varepsilon_1\}$ en primer lugar: dividimos el intervalo $[a_1, b_1]$ en $[a_{11}, b_{11}]$ y $[a_{12}, b_{12}]$ y aplicamos el criterio de exactitud (8). Si el criterio se verifica con la tolerancia ε_1 , entonces aplicamos la fórmula de cuadratura (3) en $[a_1, b_1]$ y aceptamos que sobre este intervalo hemos alcanzado la precisión deseada. Si el criterio (8) no se verifica con la tolerancia ε_1 , entonces hay que dividir y aplicar el criterio de exactitud a cada subintervalo $[a_{11}, b_{11}]$ y $[a_{12}, b_{12}]$ en un tercer paso con una tolerancia reducida $\frac{1}{2}\varepsilon_1$. Pero en este segundo paso hay que estudiar también qué pasa con $\{[a_2, b_2], \varepsilon_2\}$: dividimos $[a_2, b_2]$ en $[a_{21}, b_{21}]$ y $[a_{22}, b_{22}]$ y aplicamos el criterio de exactitud (8). Si el criterio se verifica con la tolerancia ε_2 , entonces aplicamos la fórmula de cuadratura (3) en $[a_2, b_2]$ y aceptamos que sobre este intervalo hemos alcanzado la precisión deseada. Si el criterio (8) no se verifica con la tolerancia ε_2 , entonces hay que dividir y aplicar el criterio de exactitud a cada subintervalo $[a_{21}, b_{21}]$ y $[a_{22}, b_{22}]$ en un tercer paso con una tolerancia reducida $\frac{1}{2}\varepsilon_2$. En resumen, el segundo paso produce dos, tres o cuatro intervalos; produce dos si tanto $[a_1, b_1]$ como $[a_2, b_2]$ pasan el criterio, produce tres si uno de los dos intervalos de partida pasa el criterio pero el otro no y produce cuatro si ninguno de los dos intervalos de partida pasa el criterio. Si sólo hay dos intervalos, entonces el proceso ha terminado. Si hay tres intervalos, entonces los volvemos a etiquetar como $\{\{[a_1, b_1], \varepsilon_1\}, \{[a_2, b_2], \varepsilon_2\} \text{ y } \{[a_3, b_3], \varepsilon_3\}\}$, donde $\varepsilon_1 + \varepsilon_2 + \varepsilon_3 = \varepsilon_0$ y continuamos el proceso de integración adaptativa en un tercer paso con los dos intervalos de menor tamaño. Si hay cuatro intervalos, entonces los volvemos a etiquetar como $\{\{[a_1, b_1], \varepsilon_1\}, \{[a_2, b_2], \varepsilon_2\}, \{[a_3, b_3], \varepsilon_3\} \text{ y } \{[a_4, b_4], \varepsilon_4\}\}$, donde $\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 = \varepsilon_0$ y continuamos el proceso de integración adaptativa en un tercer paso con los cuatro intervalos.

El término del error (4) muestra que conforme hacemos refinamientos, el error se reduce en un factor de $\frac{1}{16}$, por tanto, el proceso debe terminar en un número finito de pasos. Para programar este método, es necesario incluir una variable centinela que nos indique si un cierto subintervalo ha pasado ya el criterio de exactitud. Asimismo, para evitar evaluaciones innecesarias de $f(x)$, los valores de la función pueden almacenarse como una lista de datos correspondientes a cada subintervalo. Estos detalles se muestran en Programa 7.6.

Tabla 7.8 Cálculos del método de integración adaptativa para
 $f(x) = 13(x - x^2)e^{-3x/2}$.

a_k	b_k	$S(a_{k1}, b_{k1}) + S(a_{k2}, b_{k2})$	Cota del error: miembro izquierdo de (8)	Tolerancia ε_k para $[a_k, b_k]$
0.0	0.0625	0.02287184840	0.00000001522	0.00000015625
0.0625	0.125	0.05948686456	0.00000001316	0.00000015625
0.125	0.1875	0.08434213630	0.00000001137	0.00000015625
0.1875	0.25	0.09969871532	0.00000000981	0.00000015625
0.25	0.375	0.21672136781	0.00000025055	0.0000003125
0.375	0.5	0.20646391592	0.00000018402	0.0000003125
0.5	0.625	0.17150617231	0.00000013381	0.0000003125
0.625	0.75	0.12433363793	0.00000009611	0.0000003125
0.75	0.875	0.07324515141	0.00000006799	0.0000003125
0.875	1.0	0.02352883215	0.00000004718	0.0000003125
1.0	1.125	-0.02166038952	0.00000003192	0.0000003125
1.125	1.25	-0.06065079384	0.00000002084	0.0000003125
1.25	1.5	-0.21080823822	0.000000031714	0.000000625
1.5	2.0	-0.60550965007	0.00000003195	0.00000125
2.0	2.25	-0.31985720175	0.00000008106	0.000000625
2.25	2.5	-0.30061749228	0.00000008301	0.000000625
2.5	2.75	-0.27009962412	0.00000007071	0.000000625
2.75	3.0	-0.23474721177	0.00000005447	0.000000625
3.0	3.5	-0.36389799695	0.00000103699	0.00000125
3.5	4.0	-0.24313827772	0.00000041708	0.00000125
Totales		-1.54878823413	0.00000296809	0.00001

Ejemplo 7.16. Vamos a usar el método de integración adaptativa para aproximar numéricamente el valor de la integral definida $\int_0^4 13(x - x^2)e^{-3x/2} dx$ con una tolerancia inicial de $\varepsilon_0 = 0.00001$.

Para llevar a cabo el método fueron necesarios 20 subintervalos y en la Tabla 7.8 se relacionan cada subintervalo $[a_k, b_k]$, la aproximación dada por la regla compuesta de Simpson $S(a_{k1}, b_{k1}) + S(a_{k2}, b_{k2})$, la cota del error aceptada para esta aproximación y la tolerancia asociada ε_k . El valor aproximado de la integral se obtiene sumando las aproximaciones dadas por la regla de Simpson en cada subintervalo y es

$$(11) \quad \int_0^4 13(x - x^2)e^{-3x/2} dx \approx -1.54878823413.$$

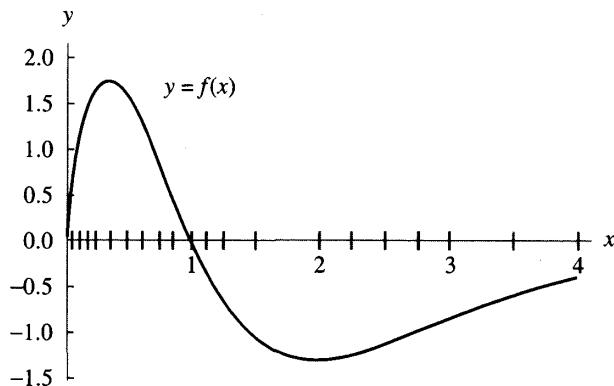


Figura 7.9 Los subintervalos de $[0, 4]$ usados en la cuadratura adaptativa del Ejemplo 7.16.

El valor de la integral es

$$(12) \quad \int_0^4 13(x - x^2)e^{-3x/2} dx = \frac{4108e^{-6} - 52}{27} \\ = -1.5487883725279481333;$$

por tanto, el error de la cuadratura adaptativa es

$$(13) \quad |-1.54878837253 - (-1.54878823413)| = 0.00000013840,$$

que es menor que la tolerancia especificada $\varepsilon_0 = 0.00001$. En este ejemplo, en el método de integración adaptativa se generan 20 subintervalos de $[0, 4]$ y se realizan 81 evaluaciones del integrando; en la Figura 7.9 se muestran la gráfica de $y = f(x)$ y estos 20 subintervalos y en ella podemos observar que los intervalos son menores allí donde la función cambia más rápidamente, es decir, cerca del origen.

En el proceso de división de los intervalos y aplicación del criterio de exactitud, hubo que dividir cada uno de los cuatro primeros intervalos de longitud 0.25 en ocho subintervalos de longitud 0.03125. Si usáramos este incremento en el intervalo total $[0, 4]$ para aplicar la regla compuesta de Simpson, entonces necesitaríamos $M = 128$ subintervalos, con los que obtendríamos la aproximación -1.54878844029 , cuyo error es 0.00000006776. Aunque el error de la aproximación dada por la regla compuesta de Simpson es casi la mitad del error que se produce con la integración adaptativa, hacen falta 176 evaluaciones más; la ganancia en exactitud no compensa el ahorro en el coste computacional que proporciona la integración adaptativa. ■

MATLAB

El Programa 7.5 que damos a continuación, **srule**, es una modificación de la regla de Simpson dada en la Sección 7.1 y nos proporciona un vector Z que

contiene los resultados de la aplicación de la regla de Simpson en el intervalo $[a, b]$. El Programa 7.6 que daremos después utiliza **srule** como subprograma en el que se lleva a cabo la regla de Simpson en cada uno de los subintervalos generados por el proceso de integración adaptativa.

Programa 7.5 (Regla de Simpson). Construcción de la aproximación a la integral

$$\int_a^b f(x) dx \approx \frac{h}{3}(f(a) + 4f(c) + f(b))$$

mediante la regla de Simpson, siendo $c = (a + b)/2$.

```
function Z=srule(f,a,b,tol)

% Datos
% - f es el integrando, dado como una
%   cadena de caracteres 'f'
% - a y b son los extremos inferior y superior del
%   intervalo de integración
% - tol es la tolerancia
% Resultado
% - Z es un vector de orden 1 x 6: [a b S S2 err tol1]

h=(b-a)/2;
C=zeros(1,3);
C=feval(f,[a (a+b)/2 b]);
S=h*(C(1)+4*C(2)+C(3))/3;
S2=S;
tol1=tol;
err=tol;
Z=[a b S S2 err tol1];
```

El Programa 7.6 produce: una matriz **SRmat**, la aproximación **quad** a la integral mediante cuadratura adaptativa y una cota **err** del error de la aproximación. Las filas de **SRmat** contienen los extremos del subintervalo correspondiente, la aproximación dada por la regla de Simpson sobre dicho subintervalo y las cotas del error generadas por el proceso de integración adaptativa.

Programa 7.6 (Método de integración adaptativa con la regla de Simpson). Construcción de la aproximación a la integral

$$\int_a^b f(x) dx \approx \sum_{k=1}^{M/4} (f(x_{4k-4}) + 4f(x_{4k-3}) + 2f(x_{4k-2}) \\ + 4f(x_{4k-1}) + f(x_{4k}))$$

mediante la aplicación de la regla compuesta de Simpson aplicada con $4M$ subintervalos $[x_{4k-4}, x_{4k}]$, donde $[a, b] = [x_0, x_{4M}]$ y $x_{4k-4+j} = x_{4k-4} + jh_k$, para cada $k = 1, \dots, M$ y $j = 1, \dots, 4$.

```
function [SRmat,quad,err]=adapt(f,a,b,tol)
% Datos
% - f es el integrando, dado como
%   una cadena de caracteres 'f'
% - a y b son los extremos inferior y superior del
%   intervalo de integración
% - tol es la tolerancia
% Resultados
% - SRmat es la tabla de valores
% - quad es la aproximación a la integral
% - err el error estimado
% Inicialización de los valores
SRmat = zeros(30,6);
iterating=0;
done=1;
SRvec=zeros(1,6);
SRvec=srule(f,a,b,tol);
SRmat(1,1:6)=SRvec;
m=1;
state=iterating;
while(state==iterating)
    n=m;
    for j=n:-1:1
        p=j;
        SR0vec=SRmat(p,:);
        err=SR0vec(5);
        tol=SR0vec(6);
        if (tol<=err)
            % Se divide el intervalo, se aplica la regla
            % de Simpson y se determina el error
            state=done;
```

```

SR1vec=SR0vec;
SR2vec=SR0vec;
a=SR0vec(1);
b=SR0vec(2);
c=(a+b)/2;
err=SR0vec(5);
tol=SR0vec(6);
tol2=tol/2;
SR1vec=srule(f,a,c,tol2);
SR2vec=srule(f,c,b,tol2);
err=abs(SR0vec(3)-SR1vec(3)-SR2vec(3))/10;

% Criterio de exactitud
if (err<tol)
    SRmat(p,:)=SR0vec;
    SRmat(p,4)=SR1vec(3)+SR2vec(3);
    SRmat(p,5)=err;
else
    SRmat(p+1:m+1,:)=SRmat(p:m,:);
    m=m+1;
    SRmat(p,:)=SR1vec;
    SRmat(p+1,:)=SR2vec;
    state=iterating;
end
end
end
quad=sum(SRmat(:,4));
err=sum(abs(SRmat(:,5)));
SRmat=SRmat(1:m,1:6);

```

Algoritmos y programas

1. Use el Programa 7.6 para aproximar el valor de cada una de las integrales definidas que se relacionan a continuación; tome $\varepsilon_0 = 0.00001$ como tolerancia inicial.

$$\begin{array}{lll}
 \text{(a)} \int_0^3 \frac{\sin(2x)}{1+x^5} dx & \text{(b)} \int_0^3 \sin(4x)e^{-2x} dx & \text{(c)} \int_{0.04}^1 \frac{1}{\sqrt{x}} dx \\
 \text{(d)} \int_0^2 \frac{1}{x^2 + \frac{1}{10}} dx & \text{(e)} \int_{1/(2\pi)}^2 \sin\left(\frac{1}{x}\right) dx & \text{(f)} \int_0^2 \sqrt{4x - x^2} dx
 \end{array}$$

2. Construya una gráfica análoga a la de la Figura 7.9 para cada una de las integrales del Problema 1. *Indicación.* La primera columna de `SRmat` contiene los extremos de los subintervalos (excepto b) del proceso adaptativo; entonces, con `T=SRmat(:,1)` y `Z=zeros(length(T))'`, la instrucción `plot(T,Z,'.'`) dibujará los subintervalos (excepto el extremo derecho b).
3. Modifique el Programa 7.6 de manera que utilice la regla de Boole en cada subintervalo $[a_k, b_k]$.
4. Use su programa modificado como se sugiere en el Problema 3 para calcular aproximaciones y dibujar gráficas análogas a las de la Figura 7.9 para las integrales definidas del Problema 1.

7.5 El método de integración de Gauss-Legendre (opcional)

Queremos hallar el área limitada por la curva

$$y = f(x), \quad -1 \leq x \leq 1,$$

¿que método proporciona la mejor respuesta si sólo pueden hacerse dos evaluaciones de la función? Ya hemos visto que la regla del trapecio es un método para aproximar el área limitada por una curva que sólo realiza dos evaluaciones de la función en los extremos del intervalo $(-1, f(-1))$ y $(1, f(1))$. Sin embargo, si la curva $y = f(x)$ es cóncava, entonces el error de la aproximación es el área de toda la región comprendida entre la curva y el segmento rectilíneo que une sus extremos (otro caso en el que pueden aparecer errores grandes se muestra en la Figura 7.10(a)).

Si usamos dos nodos distintos x_1 y x_2 interiores al intervalo $[-1, 1]$, entonces la línea recta que pasa por $(x_1, f(x_1))$ y $(x_2, f(x_2))$ corta a la curva y el área limitada por la recta es una aproximación mejor al área limitada por la curva, como se aprecia en la Figura 7.10(b). La ecuación de esta línea recta es

$$(1) \quad y = f(x_1) + \frac{(x - x_1)(f(x_2) - f(x_1))}{x_2 - x_1}$$

y el área del trapecio limitado por dicha recta es

$$(2) \quad A_{\text{trap}} = \frac{2x_2}{x_2 - x_1} f(x_1) - \frac{2x_1}{x_2 - x_1} f(x_2).$$

Hagamos notar que la regla del trapecio es un caso especial de la fórmula (2), ya que si elegimos $x_1 = -1$, $x_2 = 1$ y $h = 2$, entonces

$$T(f, h) = \frac{2}{2} f(x_1) - \frac{-2}{2} f(x_2) = f(x_1) + f(x_2).$$

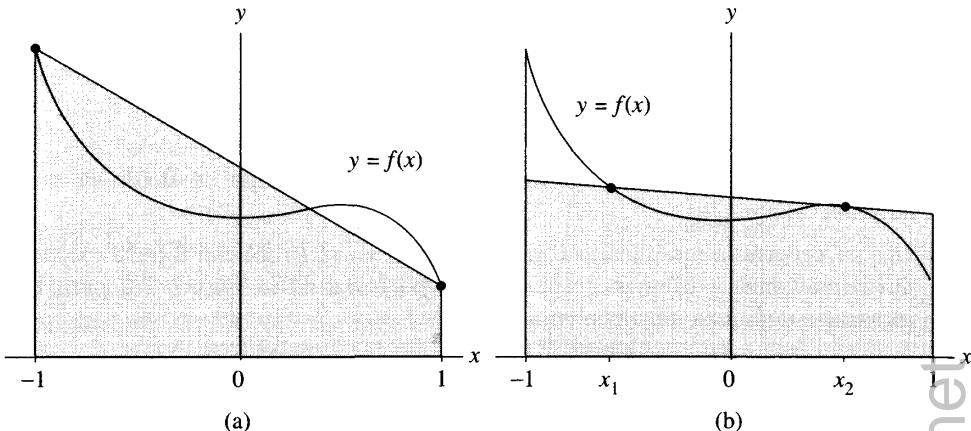


Figura 7.10 (a) Aproximación trapezoidal con abscisas -1 y 1 . (b) Aproximación trapezoidal con abscisas x_1 y x_2 .

Vamos a usar el método de los coeficientes indeterminados para hallar dos abscisas x_1 , x_2 y dos pesos w_1 , w_2 de manera que la fórmula

$$(3) \quad \int_{-1}^1 f(x) dx \approx w_1 f(x_1) + w_2 f(x_2)$$

sea exacta para polinomios cúbicos de la forma $f(x) = a_3x^3 + a_2x^2 + a_1x + a_0$. Puesto que hay que determinar cuatro números: w_1 , w_2 , x_1 y x_2 en la ecuación (3), podemos seleccionar cuatro condiciones que deben cumplirse. Usando que la integración es aditiva, será suficiente con exigir que la fórmula (3) sea exacta para las cuatro funciones $f(x) = 1$, x , x^2 , x^3 . Las cuatro condiciones de integración son, entonces,

$$(4) \quad \begin{aligned} f(x) = 1: \quad & \int_{-1}^1 1 dx = 2 = w_1 + w_2, \\ f(x) = x: \quad & \int_{-1}^1 x dx = 0 = w_1 x_1 + w_2 x_2, \\ f(x) = x^2: \quad & \int_{-1}^1 x^2 dx = \frac{2}{3} = w_1 x_1^2 + w_2 x_2^2, \\ f(x) = x^3: \quad & \int_{-1}^1 x^3 dx = 0 = w_1 x_1^3 + w_2 x_2^3, \end{aligned}$$

y tenemos que resolver el sistema de ecuaciones no lineales

$$(5) \quad w_1 + w_2 = 2,$$

$$(6) \quad w_1 x_1 = -w_2 x_2,$$

$$(7) \quad w_1 x_1^2 + w_2 x_2^2 = \frac{2}{3},$$

$$(8) \quad w_1 x_1^3 = -w_2 x_2^3.$$

Dividiendo (8) entre (6) y teniendo en cuenta que $x_1 \neq x_2$, nos queda

$$(9) \quad x_1^2 = x_2^2 \quad \text{así que} \quad x_1 = -x_2.$$

Ahora, usando (9) y dividiendo (6) entre x_1 por la derecha y $-x_2$ por la izquierda, obtenemos

$$(10) \quad w_1 = w_2.$$

Sustituyendo (10) en (5) resulta $2w_1 = 2$ y, por tanto,

$$(11) \quad w_1 = w_2 = 1.$$

Usando (11) y (9) en (7), tenemos

$$(12) \quad w_1 x_1^2 + w_2 x_2^2 = x_2^2 + x_2^2 = \frac{2}{3} \quad \text{con lo cual} \quad x_2^2 = \frac{1}{3}.$$

Finalmente, de (12) y (9) se deduce que los nodos son

$$-x_1 = x_2 = 1/3^{1/2} \approx 0.5773502692.$$

Hemos encontrado los nodos y los pesos con los que se construye la regla de Gauss-Legendre con dos nodos. Puesto que la fórmula es exacta para polinomios de grado tres, el término del error incluirá la derivada cuarta (la deducción de la forma de dicho término del error puede consultarse en la Referencia [41]).

Teorema 7.8 (Regla de Gauss-Legendre con dos nodos). Si f es continua en $[-1, 1]$, entonces

$$(13) \quad \int_{-1}^1 f(x) dx \approx G_2(f) = f\left(\frac{-1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

La regla de Gauss-Legendre con dos nodos $G_2(f)$ tiene grado de precisión $n = 3$ y si $f \in C^4[-1, 1]$, entonces

$$(14) \quad \int_{-1}^1 f(x) dx = f\left(\frac{-1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) + E_2(f),$$

siendo

$$(15) \quad E_2(f) = \frac{f^{(4)}(c)}{135}$$

para algún punto $c \in [-1, 1]$.

Ejemplo 7.17. Vamos a usar la regla de Gauss-Legendre con dos nodos para aproximar

$$\int_{-1}^1 \frac{dx}{x+2} = \ln(3) - \ln(1) \approx 1.09861$$

y compararemos el resultado que se obtiene con las aproximaciones dadas por la regla del trapecio $T(f, h)$ para $h = 2$ y por la regla de Simpson $S(f, h)$ para $h = 1$.

Sea $G_2(f)$ la aproximación que proporciona la regla de Gauss-Legendre con dos nodos, entonces

$$\begin{aligned} G_2(f) &= f(-0.57735) + f(0.57735) \\ &= 0.70291 + 0.38800 = 1.09091, \end{aligned}$$

$$\begin{aligned} T(f, 2) &= f(-1.00000) + f(1.00000) \\ &= 1.00000 + 0.33333 = 1.33333, \end{aligned}$$

$$S(f, 1) = \frac{f(-1) + 4f(0) + f(1)}{3} = \frac{1 + 2 + \frac{1}{3}}{3} = 1.11111.$$

Los errores son 0.00770, -0.23472 y -0.01250, respectivamente, de manera que la regla de Gauss-Legendre es la que proporciona la mejor aproximación. Hagamos notar, además, que en la regla de Gauss-Legendre sólo se hicieron dos evaluaciones de la función, por tres en la regla de Simpson. En este ejemplo, el tamaño del error de $G_2(f)$ es un 61% del tamaño del error de $S(f, 1)$.

La regla general de Gauss-Legendre con N nodos es exacta para funciones polinomiales de grado menor o igual que $2N - 1$ y su fórmula de cuadratura es

$$(16) \quad G_N(f) = w_{N,1}f(x_{N,1}) + w_{N,2}f(x_{N,2}) + \cdots + w_{N,N}f(x_{N,N}).$$

Los nodos $x_{N,k}$ y los pesos $w_{N,k}$ que hay que usar están tabulados y pueden conseguirse fácilmente; en la Tabla 7.9 se relacionan los valores correspondientes para las reglas de Gauss-Legendre con hasta ocho nodos, así como la forma de los términos del error $E_N(f)$ correspondientes a las aproximaciones $G_N(f)$; estos términos pueden usarse para estimar la precisión del método de integración de Gauss-Legendre.

Los valores reflejados en la Tabla 7.9 no tienen, en general, una representación sencilla, por eso el método no es muy atractivo si hay que realizar los cálculos a mano. Sin embargo, una vez que los tenemos almacenados en la memoria del computador, no tiene ninguna dificultad usarlos en nuestros cálculos. Los nodos son, de hecho, las raíces de los polinomios de Legendre y los pesos correspondientes se obtienen resolviendo un sistema de ecuaciones lineales. Para la regla de Gauss-Legendre con tres nodos, los nodos son $-(0.6)^{1/2}$, 0 y $(0.6)^{1/2}$, y los pesos correspondientes son $5/9$, $8/9$ y $5/9$.

Tabla 7.9 Nodos y pesos para el método de Gauss-Legendre.

N	Nodos, $x_{N,k}$	Pesos, $w_{N,k}$	Error, $E_N(f)$
2	-0.5773502692	1.0000000000	$\frac{f^{(4)}(c)}{135}$
	0.5773502692	1.0000000000	
3	± 0.7745966692	0.5555555556	$\frac{f^{(6)}(c)}{15\ 750}$
	0.0000000000	0.8888888888	
4	± 0.8611363116	0.3478548451	$\frac{f^{(8)}(c)}{3\ 472\ 875}$
	± 0.3399810436	0.6521451549	
5	± 0.9061798459	0.2369268851	$\frac{f^{(10)}(c)}{1\ 237\ 732\ 650}$
	± 0.5384693101	0.4786286705	
	0.0000000000	0.5688888888	
6	± 0.9324695142	0.1713244924	$\frac{f^{(12)}(c)2^{13}(6!)^4}{(12!)^313!}$
	± 0.6612093865	0.3607615730	
	0.2386191861	0.4679139346	
7	± 0.9491079123	0.1294849662	$\frac{f^{(14)}(c)2^{15}(7!)^4}{(14!)^315!}$
	± 0.7415311856	0.2797053915	
	± 0.4058451514	0.3818300505	
	0.0000000000	0.4179591837	
8	± 0.9602898565	0.1012285363	$\frac{f^{(16)}(c)2^{17}(8!)^4}{(16!)^317!}$
	± 0.7966664774	0.2223810345	
	± 0.5255324099	0.3137066459	
	0.1834346425	0.3626837834	

Teorema 7.9 (Regla de Gauss-Legendre con tres nodos). Si f es continua en $[-1, 1]$, entonces

$$(17) \quad \int_{-1}^1 f(x) dx \approx G_3(f) = \frac{5f(-\sqrt{3/5}) + 8f(0) + 5f(\sqrt{3/5})}{9}.$$

La regla de Gauss-Legendre con tres nodos $G_3(f)$ tiene grado de precisión $n = 5$. Si, además, $f \in C^6[-1, 1]$, entonces

$$(18) \quad \int_{-1}^1 f(x) dx = \frac{5f(-\sqrt{3/5}) + 8f(0) + 5f(\sqrt{3/5})}{9} + E_3(f)$$

y existe algún punto $c \in [-1, 1]$ tal que

$$(19) \quad E_3(f) = \frac{f^{(6)}(c)}{15\ 750}.$$

Ejemplo 7.18. Vamos a probar que la regla de Gauss-Legendre con tres puntos es exacta para el polinomio $5x^4$; o sea,

$$\int_{-1}^1 5x^4 dx = 2 = G_3(f).$$

Puesto que el integrando es $f(x) = 5x^4$ y $f^{(6)}(x) = 0$, la expresión (19) nos dice directamente que $E_3(f) = 0$. Pero también es instructivo usar la fórmula (17) y realizar los cálculos correspondientes:

$$G_3(f) = \frac{5(5)(0.6)^2 + 0 + 5(5)(0.6)^2}{9} = \frac{18}{9} = 2.$$

El siguiente resultado muestra cómo debemos cambiar la variable de integración para que podamos aplicar las reglas de Gauss-Legendre en un intervalo cualquiera $[a, b]$.

Teorema 7.10 (Traslación del método de Gauss-Legendre). Supongamos que tenemos los nodos $\{x_{N,k}\}_{k=1}^N$ y los pesos $\{w_{N,k}\}_{k=1}^N$ necesarios para aplicar la regla de Gauss-Legendre con N nodos en $[-1, 1]$. Entonces, para aplicar el método de Gauss-Legendre en un intervalo $[a, b]$, se puede usar el cambio de variable

$$(20) \quad t = \frac{a+b}{2} + \frac{b-a}{2}x \quad \text{con} \quad dt = \frac{b-a}{2} dx$$

y la relación

$$(21) \quad \int_a^b f(t) dt = \int_{-1}^1 f\left(\frac{a+b}{2} + \frac{b-a}{2}x\right) \frac{b-a}{2} dx$$

proporciona la fórmula de cuadratura

$$(22) \quad \int_a^b f(t) dt = \frac{b-a}{2} \sum_{k=1}^N w_{N,k} f\left(\frac{a+b}{2} + \frac{b-a}{2}x_{N,k}\right).$$

Ejemplo 7.19. Vamos a usar el método de Gauss-Legendre con tres nodos para aproximar

$$\int_1^5 \frac{dt}{t} = \ln(5) - \ln(1) \approx 1.609438$$

y compararemos el resultado con la aproximación dada por la regla de Boole $B(2)$ con $h = 1$.

En este caso $a = 1$ y $b = 5$, así que la fórmula (23) queda

$$\begin{aligned} G_3(f) &= (2) \frac{5f(3 - 2(0.6)^{1/2}) + 8f(3 + 0) + 5f(3 + 2(0.6)^{1/2})}{9} \\ &= (2) \frac{3.446359 + 2.666667 + 1.099096}{9} = 1.602694. \end{aligned}$$

En el Ejemplo 7.13 vimos que la regla de Boole proporciona $B(2) = 1.617778$. Los errores son 0.006744 y -0.008340 , respectivamente, de manera que la regla de Gauss-Legendre es ligeramente mejor en este caso, el error es prácticamente el mismo, a pesar de que sólo necesita tres evaluaciones, frente a las cinco necesarias para la regla de Boole.

Las fórmulas de integración de Gauss-Legendre tienen una precisión muy alta y deben ser tenidas en cuenta si hay que realizar muchas integrales de funciones parecidas sobre un mismo intervalo. En este caso se procede de la siguiente manera: Se toman algunas integrales representativas, incluyendo las que se sospechen que puedan comportar mayores errores en su evaluación numérica. Después, se determina el número N de nodos necesarios para obtener dichas integrales con la precisión deseada y se usa la regla de Gauss-Legendre con N nodos para calcular todas las integrales.

MATLAB

En el Programa 7.7 que damos a continuación, donde el número de nodos N está fijo, hay que proporcionar como datos los nodos y los pesos de la Tabla 7.9, almacenados en sendas matrices de orden $1 \times N$ llamadas A y W , respectivamente. Esto puede hacerse directamente en la ventana de trabajo del programa MATLAB o en matrices previamente guardadas en los archivos $A.m$ y $W.m$. Una posibilidad más expeditiva es almacenar la Tabla 7.9 como una matriz G de orden 35×2 cuya primera columna contenga los nodos y cuya segunda columna contenga los pesos; así, para cada valor de N las matrices A y W serían submatrices de G . Por ejemplo, si $N = 3$, entonces $A=G(3:5,1)'$ y $W=G(3:5,2)'$.

Programa 7.7 (Método de integración de Gauss-Legendre). Construcción de la aproximación a la integral

$$\int_a^b f(x) dx \approx \frac{b-a}{2} \sum_{k=1}^N w_{N,k} f(t_{N,k})$$

que se obtiene evaluando $f(x)$ en N nodos (desigualmente espaciados) $\{t_{N,k}\}_{k=1}^N$ dados por el cambio de variable

$$t = \frac{a+b}{2} + \frac{b-a}{2}x, \quad \text{con} \quad dt = \frac{b-a}{2} dx,$$

a partir de los nodos $\{x_{N,k}\}_{k=1}^N$ que, así como los pesos correspondientes $\{w_{N,k}\}_{k=1}^N$, deben ser previamente obtenidos de una tabla de valores.

```

function quad=gauss(f,a,b,A,W)
% Datos
%   - f es el integrando, dado como
%     una cadena de caracteres 'f'
%   - a y b son los extremos inferior y superior del
%     intervalo de integración
%   - A es el vector 1 x N de nodos de la Tabla 7.9
%   - W es el vector 1 x N de pesos de la Tabla 7.9
% Resultado
%   - quad es la aproximación al valor de la integral
N=length(A);
T=zeros(1,N);
T=((a+b)/2)+((b-a)/2)*A;
quad=((b-a)/2)*sum(W.*feval(f,T));

```

Ejercicios

En los Ejercicios 1 a 5, pruebe que las dos integrales son iguales y calcule la aproximación $G_2(f)$.

$$1. \int_0^2 6t^5 dt = \int_{-1}^1 6(x+1)^5 dx \qquad \qquad 2. \int_0^2 \sin(t) dt = \int_{-1}^1 \sin(x+1) dx$$

$$3. \int_0^1 \frac{\sin(t)}{t} dt = \int_{-1}^1 \frac{\sin((x+1)/2)}{x+1} dx$$

$$4. \frac{1}{\sqrt{2\pi}} \int_0^1 e^{-t^2/2} dt = \frac{1}{\sqrt{2\pi}} \int_{-1}^1 \frac{e^{-(x+1)^2/8}}{2} dx$$

$$5. \frac{1}{\pi} \int_0^\pi \cos(0.6 \sin(t)) dt = 0.5 \int_{-1}^1 \cos\left(0.6 \sin\left((x+1)\frac{\pi}{2}\right)\right) dx$$

6. Use el término del error $E_N(f)$ que se muestra en la Tabla 7.9 y realice el cambio de variable dado en el Teorema 7.10 para hallar el menor número natural N tal que $E_N(f) = 0$ para los casos

(a) $\int_0^2 8x^7 dx = 256 = G_N(f)$.

(b) $\int_0^2 11x^{10} dx = 2048 = G_N(f)$.

7. Determine las raíces de los siguientes polinomios de Legendre y compare sus resultados con los nodos dados en la Tabla 7.9.

(a) $P_2(x) = (3x^2 - 1)/2$

(b) $P_3(x) = (5x^3 - 3x)/2$

(c) $P_4(x) = (35x^4 - 30x^2 + 3)/8$

8. El error de truncamiento de la regla de Gauss-Legendre con dos nodos en el intervalo $[-1, 1]$ es $f^{(4)}(c_1)/135$. El error de truncamiento de la regla de Simpson en $[a, b]$ es $-h^5 f^{(4)}(c_2)/90$. Compare ambos términos del error cuando $[a, b] = [-1, 1]$. ¿Qué método piensa que es mejor? ¿Por qué?

9. La regla de Gauss-Legendre con tres nodos es

$$\int_{-1}^1 f(x) dx \approx \frac{5f(-(0.6)^{1/2}) + 8f(0) + 5f((0.6)^{1/2})}{9}.$$

Pruebe que la fórmula es exacta para $f(x) = 1, x, x^2, x^3, x^4, x^5$. *Indicación.* Si f es una función impar (o sea, $f(-x) = -f(x)$), entonces la integral de f en $[-1, 1]$ es cero.

10. El error de truncamiento de la regla de Gauss-Legendre con tres puntos en el intervalo $[-1, 1]$ es $f^{(6)}(c_1)/15\,750$. El error de truncamiento de la regla de Boole en $[a, b]$ es $-8h^7 f^{(6)}(c_2)/945$. Compare ambos términos del error cuando $[a, b] = [-1, 1]$. ¿Qué método piensa que es mejor? ¿Por qué?
11. Deduzca la regla de Gauss-Legendre con tres puntos dando los siguientes pasos y utilizando el hecho de que los nodos son las raíces del polinomio de Legendre de grado 3:

$$x_1 = -(0.6)^{1/2}, \quad x_2 = 0, \quad x_3 = (0.6)^{1/2}.$$

Determine los pesos w_1, w_2 y w_3 de manera que la relación

$$\int_{-1}^1 f(x) dx \approx w_1 f(-(0.6)^{1/2}) + w_2 f(0) + w_3 f((0.6)^{1/2})$$

sea exacta para las funciones $f(x) = 1, x$ y x^2 . *Indicación.* Obtenga y resuelva el sistema de ecuaciones lineales

$$\begin{aligned} w_1 + w_2 + w_3 &= 2 \\ -(0.6)^{1/2}w_1 + (0.6)^{1/2}w_3 &= 0 \\ 0.6w_1 + 0.6w_3 &= \frac{2}{3}. \end{aligned}$$

12. En la práctica, cuando hay que calcular muchas integrales del mismo tipo, se suele hacer un análisis preliminar para determinar el número de evaluaciones de la función que hay que realizar para obtener la precisión deseada. Supongamos que deben realizarse 17 evaluaciones del integrando. Compare la aproximación $R(4, 4)$ dada por el método de Romberg con la aproximación $G_{17}(f)$ dada por la regla de Gauss-Legendre.

Algoritmos y programas

1. Use el Programa 7.7 para calcular, hallando $G_6(f)$, $G_7(f)$ y $G_8(f)$, cada una de las integrales de los Ejercicios 1 a 5.

432 CAP. 7 INTEGRACIÓN NUMÉRICA

- 2. (a)** Modifique el Programa 7.7 de manera que vaya calculando $G_1(f)$, $G_2(f)$, ..., $G_8(f)$ y detenga los cálculos cuando el error relativo entre dos aproximaciones consecutivas $G_{N-1}(f)$ y $G_N(f)$ sea menor que el valor fijado para la tolerancia `tol`, es decir

$$\frac{2|G_{N-1}(f) - G_N(f)|}{|G_{N-1}(f) + G_N(f)|} < \text{tol}.$$

Indicación. Como se observó al final de la sección, almacene la Tabla 7.9 en un archivo `G.m` como una matriz de orden 35×2 .

- (b)** Use su programa del apartado (a) para aproximar las integrales de los Ejercicios 1 a 5 con una precisión de cinco cifras decimales.
- 3. (a)** Use la regla de Gauss-Legendre con seis puntos para aproximar la solución de la ecuación integral

$$v(x) = x^2 + 0.1 \int_0^3 (x^2 + t)v(t) dt.$$

Para ello, sustituya su solución aproximada en el miembro derecho de la ecuación integral y simplifique.

- (b)** Repita el apartado (a) usando la regla de Gauss-Legendre con ocho puntos.

Optimización numérica

La ecuación de ondas bidimensional se usa en ingeniería mecánica para modelar las vibraciones de una placa rectangular. Si las placas están ancladas por sus cuatro esquinas, las vibraciones sinusoidales pueden describirse mediante una serie de Fourier doble. Supongamos que en un cierto instante de tiempo

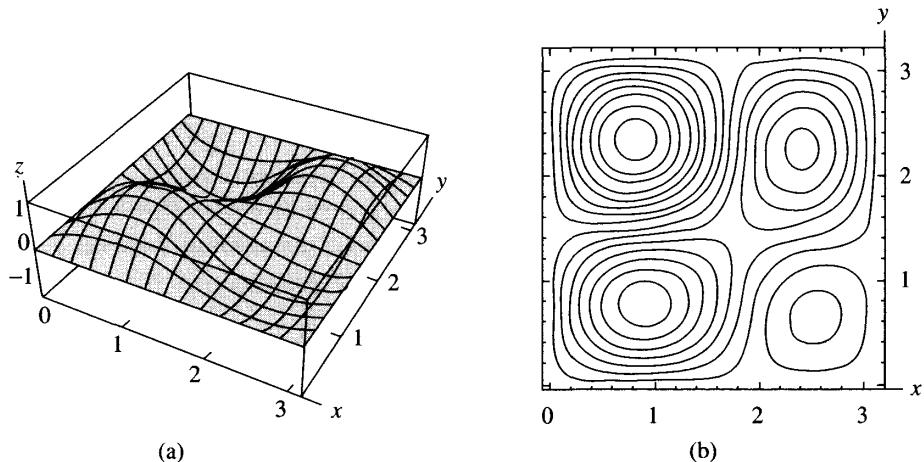


Figura 8.1 (a) El desplazamiento $z = f(x, y)$ de una placa vibrante. (b) Las curvas de nivel $f(x, y) = C$ para dicha placa.

la altura del desplazamiento $z = f(x, y)$ sobre el punto (x, y) viene dada por la función

$$\begin{aligned} z = f(x, y) = & 0.02 \operatorname{sen}(x) \operatorname{sen}(y) - 0.03 \operatorname{sen}(2x) \operatorname{sen}(y) \\ & + 0.04 \operatorname{sen}(x) \operatorname{sen}(2y) + 0.08 \operatorname{sen}(2x) \operatorname{sen}(2y). \end{aligned}$$

¿Dónde se localizan los puntos en los que el desplazamiento es mayor? Mirando la gráfica tridimensional y las curvas de nivel correspondientes que se muestran en las Figuras 8.1(a) y (b), respectivamente, vemos que hay dos mínimos locales y dos máximos locales en el cuadrado $0 \leq x \leq \pi$, $0 \leq y \leq \pi$. Los métodos numéricos nos permiten determinar aproximadamente su localización:

$$f(0.8278, 2.3322) = -0.1200 \quad \text{y} \quad f(2.5351, 0.6298) = -0.0264$$

son los mínimos locales;

$$f(0.9241, 0.7640) = 0.0998 \quad \text{y} \quad f(2.3979, 2.2287) = 0.0853$$

son los máximos locales.

En este capítulo damos una breve introducción a algunos de los métodos básicos para localizar extremos de funciones de una o varias variables.

8.1 Minimización de una función

Definición 8.1 (Extremo local). Se dice que una función f tiene, o alcanza, un **mínimo local** en $x = p$ si existe un intervalo abierto I tal que $p \in I$ y $f(p) \leq f(x)$ para todo $x \in I$, en ese caso también se dice que $f(p)$ es un **valor mínimo local** de $f(x)$.

De manera análoga, se dice que f tiene un **máximo local** en $x = p$ si $f(x) \leq f(p)$ para todo $x \in I$, en ese caso también se dice que $f(p)$ es un **valor máximo local** de $f(x)$. Si f tiene un máximo o un mínimo local en $x = p$, entonces se dice que tiene un **extremo local** en $x = p$. ▲

Definición 8.2 (Creciente y decreciente). Supongamos que f está definida en un intervalo I .

- (i) Se dice que f es **creciente** en I si para cualesquiera $x_1, x_2 \in I$ con $x_1 < x_2$, se tiene que $f(x_1) < f(x_2)$.
- (ii) Se dice que f es **decreciente** en I si para cualesquiera $x_1, x_2 \in I$ con $x_1 < x_2$, se tiene que $f(x_1) > f(x_2)$. ▲

Teorema 8.1. Supongamos que $f(x)$ es continua en $I = [a, b]$ y derivable en (a, b) .

- (i) Si $f'(x) > 0$ para todo $x \in (a, b)$, entonces $f(x)$ es creciente en I .
- (ii) Si $f'(x) < 0$ para todo $x \in (a, b)$, entonces $f(x)$ es decreciente en I .

Teorema 8.2. Supongamos que f está definida en un intervalo $I = [a, b]$ y que tiene un extremo local en un punto $p \in (a, b)$. Si $f(x)$ es derivable en $x = p$, entonces $f'(p) = 0$.

Teorema 8.3 (Criterio de la derivada primera). Supongamos que $f(x)$ es continua en $[a, b]$ y derivable en todo punto $x \in (a, b)$, salvo quizás en $x = p$.

- (i) Si $f'(x) < 0$ en (a, p) y $f'(x) > 0$ en (p, b) , entonces $f(x)$ tiene un mínimo local en p .
- (ii) Si $f'(x) > 0$ en (a, p) y $f'(x) < 0$ en (p, b) , entonces $f(x)$ tiene un máximo local en p .

Teorema 8.4 (Criterio de la derivada segunda). Supongamos que $f(x)$ es continua en $[a, b]$ y dos veces derivable en (a, b) . Supongamos también que $p \in (a, b)$ es un punto crítico de $f(x)$, o sea, que $f'(p) = 0$. Entonces se verifica:

- (i) Si $f''(p) > 0$, entonces $f(x)$ tiene un mínimo local en p .
- (ii) Si $f''(p) < 0$, entonces $f(x)$ tiene un máximo local en p .
- (iii) Si $f''(p) = 0$, entonces no se puede afirmar nada.

Ejemplo 8.1. Vamos a usar el criterio de la derivada segunda para clasificar los extremos locales de $f(x) = x^3 + x^2 - x + 1$ en el intervalo $[-2, 2]$.

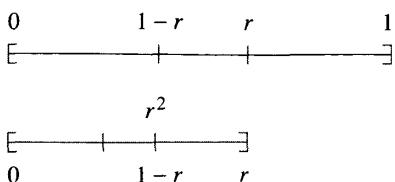
La derivada primera es $f'(x) = 3x^2 + 2x - 1 = (3x - 1)(x + 1)$ y la derivada segunda $f''(x) = 6x + 2$. Hay dos puntos en los que $f'(x) = 0$, a saber, $x = 1/3$ y $x = -1$.

Caso (i): En $x = 1/3$ se tiene que $f'(1/3) = 0$ y que $f''(1/3) = 4 > 0$, así que $f(x)$ tiene un mínimo local en $x = 1/3$.

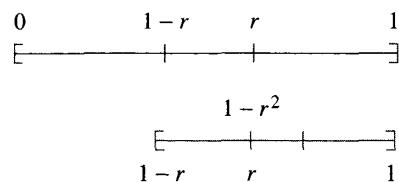
Caso (ii): En $x = -1$ se tiene que $f'(-1) = 0$ y que $f''(-1) = -4 < 0$, así que $f(x)$ tiene un máximo local en $x = -1$. ■

Métodos de búsqueda

Otro método para hallar el mínimo de $f(x)$ es evaluar la función en muchos puntos y buscar un mínimo local entre ellos. Para reducir el número de evaluaciones de la función es importante tener una buena estrategia que determine dónde tenemos que evaluar $f(x)$. Uno de los métodos más eficaces se conoce como el **método de búsqueda de la sección áurea** y se llama así porque la estrategia para ir seleccionando los puntos en los que evaluar la función $f(x)$ depende de lo que se conoce como la proporción áurea.



Se recorta por la derecha y el nuevo intervalo es $[0, r]$.



Se recorta por la izquierda y el nuevo intervalo es $[1-r, 1]$.

Figura 8.2 Los intervalos involucrados en el método de búsqueda de la sección áurea.

El Método de búsqueda de la razón áurea

Supongamos que el intervalo inicial es $[0, 1]$. Tomamos un número r tal que $0.5 < r < 1$, con lo cual $0 < 1 - r < 0.5$, y dividimos el intervalo inicial en los tres subintervalos $[0, 1 - r]$, $[1 - r, r]$ y $[r, 1]$. Ahora hay que decidir entre eliminar el intervalo de la derecha y quedarse con $[0, r]$ o eliminar el intervalo de la izquierda y quedarse con $[1 - r, 1]$. Después, dividimos el nuevo intervalo en tres subintervalos según la misma proporción que teníamos en la división del intervalo original $[0, 1]$.

Con objeto de reducir el número de evaluaciones, lo que se hace es elegir r de manera que el extremo de los subintervalos de la primera división que está en el interior del intervalo con el que nos hemos quedado (es decir, $1 - r$ si nos quedamos con $[0, r]$ y r si nos quedamos con $[1 - r, 1]$) coincida con uno de los extremos de los subintervalos de la nueva división, como se muestra en la Figura 8.2. Para que esto ocurra, las razones de división del segundo caso $(1 - r) : r$ y del primero $r : 1$ deben coincidir. Por tanto r debe verificar la ecuación de segundo grado $1 - r = r^2$ o, lo que es lo mismo, $r^2 + r - 1 = 0$. La solución r que verifica $0.5 < r < 1$ resulta ser $r = (\sqrt{5} - 1)/2$, que fue llamada proporción áurea por los matemáticos de la Grecia clásica.

Si queremos usar el método de búsqueda de la sección áurea para hallar el mínimo de $f(x)$, hay una condición que nos asegura que existe sólo un mínimo y que el método converge realmente a dicho mínimo.

Definición 8.3 (Función unimodal). Se dice que una función $f(x)$ es unimodal en $I = [a, b]$ si existe un único número $p \in I$ tal que

$$(1) \quad f(x) \text{ es decreciente en } [a, p]$$

$$(2) \quad f(x) \text{ es creciente en } [p, b],$$

lo que implica, en particular, que f alcanza su mínimo global en p . ▲

Si se sabe que $f(x)$ es unimodal en $[a, b]$, entonces es posible sustituir el intervalo inicial por un subintervalo en el que $f(x)$ alcanza su mínimo. En

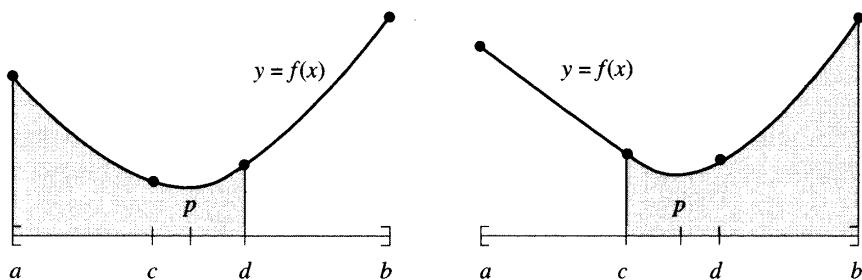


Figura 8.3 El proceso de decisión en el método de búsqueda de la sección áurea. Izquierdo: Si $f(c) \leq f(d)$, entonces se recorta por la derecha y se usa $[a, d]$. Derecha: Si $f(d) < f(c)$, entonces se recorta por la izquierda y se usa $[c, b]$.

el método de búsqueda de la sección áurea se utilizan dos puntos interiores $c = a + (1 - r)(b - a)$ y $d = a + r(b - a)$, siendo r la proporción áurea mencionada antes, de manera que $a < c < d < b$. La condición de que $f(x)$ es unimodal garantiza que los valores $f(c)$ y $f(d)$ son ambos menores que $\max\{f(a), f(b)\}$. Ahora hay que considerar dos casos (véase la Figura 8.3).

Si $f(c) \leq f(d)$, entonces el mínimo debe estar en el subintervalo $[a, d]$, así que reemplazamos b por d y continuamos la búsqueda en el nuevo subintervalo.

Si $f(d) < f(c)$, entonces el mínimo debe estar en el subintervalo $[c, b]$, así que reemplazamos a por c y continuamos la búsqueda.

En el siguiente ejemplo comparamos el método de búsqueda de la sección áurea con el método de hallar las raíces de la derivada primera.

Ejemplo 8.2. Vamos a determinar el mínimo de la función unimodal dada por $f(x) = x^2 - \operatorname{sen}(x)$ en el intervalo $[0, 1]$.

Resolviendo $f'(x) = 0$. Usamos un método de cálculo de raíces para determinar los puntos en los que se anula la derivada primera, $f'(x) = 2x - \cos(x)$. Puesto que $f'(0) = -1$ y $f'(1) = 1.4596977$, existe una raíz de $f'(x)$ en el intervalo $[0, 1]$. Empezando con $p_0 = 0$ y $p_1 = 1$, la Tabla 8.1 muestra las iteraciones realizadas con el método de la secante.

Tras aplicar el método de la secante, concluimos que $f'(0.4501836) = 0$. La segunda derivada es $f''(x) = 2 + \operatorname{sen}(x)$, con lo cual $f''(0.4501836) = 2.435131 > 0$. Así que el valor mínimo de f se alcanza en $p = 0.4501836$ y es $f(p) = -0.2324656$.

Usando el método de la sección áurea. En cada paso, los valores de la función $f(c)$ y $f(d)$ se comparan y se decide si la búsqueda continúa en $[a, d]$ o bien continúa en $[c, b]$. En la Tabla 8.2 se muestran algunos de los cálculos.

En la iteración vigésimo tercera, el intervalo se ha estrechado hasta ser $[a_{23}, b_{23}] = [0.4501827, 0.4501983]$. Este intervalo tiene una anchura de 0.0000156; sin embargo, los valores de la función calculados en sus extremos coinciden en ocho cifras deci-

Tabla 8.1 Método de la secante para resolver
 $f'(x) = 2x - \cos(x) = 0$.

k	p_k	$2p_k - \cos(p_k)$
0	0.0000000	-1.0000000
1	1.0000000	1.45969769
2	0.4065540	-0.10538092
3	0.4465123	-0.00893398
4	0.4502137	0.00007329
5	0.4501836	-0.00000005

Tabla 8.2 Método de la sección áurea para hallar el mínimo de $f(x) = x^2 - \operatorname{sen}(x)$.

k	a_k	c_k	d_k	b_k	$f(c_k)$	$f(d_k)$
0	0.0000000	0.3819660	0.6180340	1	-0.22684748	-0.19746793
1	0.0000000	<u>0.2360680</u>	0.3819660	0.6180340	<u>-0.17815339</u>	-0.22684748
2	0.2360680	<u>0.3819660</u>	0.4721360	0.6180340	<u>-0.22684748</u>	-0.23187724
3	0.3819660	0.4721360	<u>0.5278640</u>	0.6180340	-0.23187724	-0.22504882
4	0.3819660	0.4376941	<u>0.4721360</u>	0.5278640	-0.23227594	<u>-0.23187724</u>
5	0.3819660	<u>0.4164079</u>	0.4376941	0.4721360	<u>-0.23108238</u>	-0.23227594
6	0.4164079	<u>0.4376941</u>	0.4508497	0.4721360	<u>-0.23227594</u>	-0.23246503
:	:	:	:	:	:	:
21	0.4501574	<u>0.4501730</u>	0.4501827	0.4501983	-0.23246558	-0.23246558
22	0.4501730	<u>0.4501827</u>	0.4501886	0.4501983	-0.23246558	-0.23246558
23	0.4501827	<u>0.4501886</u>	0.4501923	0.4501983	-0.23246558	-0.23246558

males: $f(a_{23}) \approx -0.23246558 \approx f(b_{23})$ y terminamos aquí las iteraciones. Uno de los problemas de los métodos de búsqueda es que la función suele ser bastante plana cerca del mínimo y esto limita la exactitud que podemos obtener; en este ejemplo, el método de la secante nos proporciona una respuesta más exacta $p_5 = 0.4501836$.

Aunque el método de búsqueda de la sección áurea es más lento en este ejemplo, tiene un aspecto deseable y es que puede usarse cuando $f(x)$ no es derivable o cuando su derivada no puede calcularse fácilmente.

Cálculo de los extremos de $f(x, y)$

La Definición 8.1 se extiende fácilmente a funciones de varias variables. Supongamos que $f(x, y)$ está definida en la región circular

$$(3) \quad R = \{(x, y) : (x - p)^2 + (y - q)^2 < r^2\}.$$

Se dice que $f(x, y)$ tiene un mínimo local en el punto (p, q) cuando

$$(4) \quad f(p, q) \leq f(x, y) \quad \text{para cada punto } (x, y) \in R.$$

Se dice que $f(x, y)$ tiene un máximo local en el punto (p, q) cuando

$$(5) \quad f(x, y) \leq f(p, q) \quad \text{para cada punto } (x, y) \in R.$$

El criterio de la derivada segunda para determinar los extremos locales de una función de dos variables es una extensión del Teorema 8.4.

Teorema 8.5 (Criterio de la derivada segunda). Supongamos que $f(x, y)$ así como sus derivadas parciales primeras y segundas son continuas en la región R . Supongamos que $(p, q) \in R$ es un punto crítico, o sea, que $f_x(p, q) = 0$ y $f_y(p, q) = 0$. Entonces podemos usar las derivadas de orden superior para determinar la naturaleza de este punto crítico:

- (i) Si $f_{xx}(p, q)f_{yy}(p, q) - f_{xy}^2(p, q) > 0$ y $f_{xx}(p, q) > 0$, entonces f tiene un mínimo local en (p, q) .
- (ii) Si $f_{xx}(p, q)f_{yy}(p, q) - f_{xy}^2(p, q) > 0$ y $f_{xx}(p, q) < 0$, entonces f tiene un máximo local en (p, q) .
- (iii) Si $f_{xx}(p, q)f_{yy}(p, q) - f_{xy}^2(p, q) < 0$, entonces f no tiene un extremo local en (p, q) ; este punto es un punto de silla.
- (iv) Si $f_{xx}(p, q)f_{yy}(p, q) - f_{xy}^2(p, q) = 0$, entonces no se puede afirmar nada.

Ejemplo 8.3. Vamos a calcular el mínimo de $f(x, y) = x^2 - 4x + y^2 - y - xy$.

Las derivadas parciales primeras son

$$(6) \quad f_x(x, y) = 2x - 4 - y \quad \text{y} \quad f_y(x, y) = 2y - 1 - x.$$

Al igualar estas derivadas parciales a cero, obtenemos el sistema de ecuaciones lineales

$$(7) \quad \begin{aligned} 2x - y &= 4 \\ -x + 2y &= 1, \end{aligned}$$

cuya solución es $(x, y) = (3, 2)$. Ahora, las derivadas parciales segundas de $f(x, y)$ son

$$f_{xx}(x, y) = 2, \quad f_{yy}(x, y) = 2, \quad \text{y} \quad f_{xy}(x, y) = -1.$$

Es fácil ver que estamos en el caso (i) del Teorema 8.5, esto es,

$$f_{xx}(3, 2)f_{yy}(3, 2) - f_{xy}^2(3, 2) = 3 > 0 \quad \text{and} \quad f_{xx}(3, 2) = 2 > 0$$

así que, $f(x, y)$ alcanza en el punto $(3, 2)$ un mínimo local cuyo valor es $f(3, 2) = -7$

El método de Nelder-Mead

Nelder y Mead han desarrollado un método de búsqueda para hallar un mínimo local de una función de varias variables. Este método utiliza un tipo de cuerpo geométrico llamado simplex que en el caso del plano es un triángulo y en el caso del espacio tridimensional es un tetraedro. En el caso de dos variables, cuando tenemos un triángulo, el método consiste en comparar los valores de la función en los vértices y sustituir el peor vértice, aquel en el que $f(x, y)$ es mayor, por un vértice nuevo. De esa manera, se forma un nuevo triángulo y la búsqueda continúa. En el proceso se genera una sucesión de triángulos (que pueden tener formas diferentes), en los que los valores de la función van decreciendo. El tamaño de estos triángulos se reduce y, cuando los vértices están suficientemente juntos, hemos encontrado el mínimo local.

El algoritmo general que presentaremos al final nos permitirá calcular mínimos de funciones de N variables. Es un algoritmo efectivo y computacionalmente compacto.

El triángulo inicial OBP

Sea $f(x, y)$ la función que queremos minimizar. Partimos de un triángulo inicial cuyos vértices son $\mathbf{V}_k = (x_k, y_k)$, $k = 1, 2, 3$. Entonces evaluamos la función $f(x, y)$ en cada uno de los vértices y obtenemos $z_k = f(x_k, y_k)$ para $k = 1, 2, 3$. Ahora ordenamos los subíndices de manera que $z_1 \leq z_2 \leq z_3$ e introducimos la notación

$$(8) \quad \mathbf{O} = (x_1, y_1), \quad \mathbf{B} = (x_2, y_2) \quad \text{y} \quad \mathbf{P} = (x_3, y_3)$$

que nos debe ayudar a recordar que \mathbf{O} es el vértice óptimo, \mathbf{B} es el vértice bueno (el siguiente al óptimo) y \mathbf{P} es el vértice peor.

El punto medio del lado bueno

El siguiente paso en el proceso de construcción es calcular el punto medio del segmento que une \mathbf{O} con \mathbf{B} , para ello hallamos la media de las coordenadas:

$$(9) \quad \mathbf{M} = \frac{\mathbf{O} + \mathbf{B}}{2} = \left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2} \right).$$

Reflexión usando el punto R

Puesto que la función decrece al movernos desde \mathbf{P} hasta \mathbf{O} a lo largo de ese lado del triángulo inicial y también al movernos desde \mathbf{P} hasta \mathbf{B} , es de esperar que $f(x, y)$ tome valores menores en puntos alejados del peor vértice \mathbf{P} que estén situados al otro lado del segmento que une \mathbf{O} con \mathbf{B} . Lo que hacemos es tomar un punto de prueba: el punto \mathbf{R} que se obtiene “reflejando” el triángulo

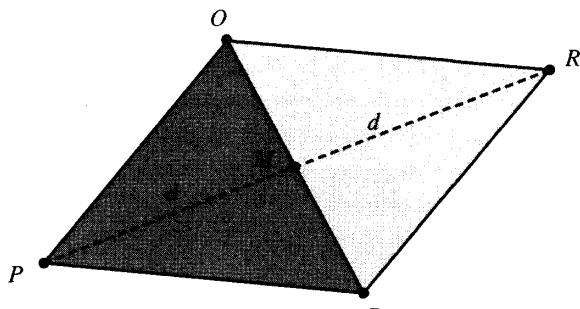


Figura 8.4 El triángulo OBP , el punto medio M y el punto reflejado R en el método de Nelder-Mead.

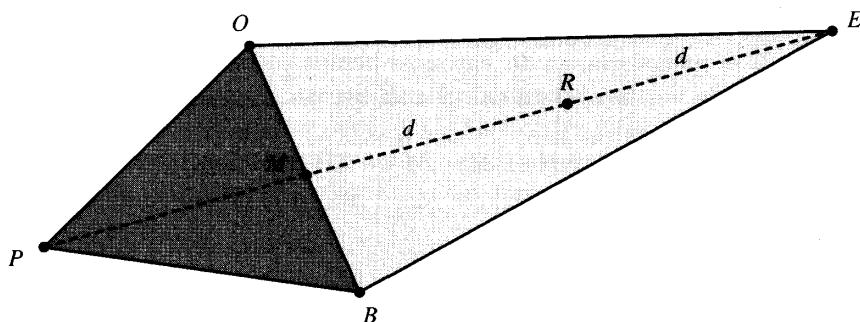


Figura 8.5 El triángulo OBP y el punto extendido E .

a través del lado \overline{OB} . Para determinar \mathbf{R} , usamos el punto medio \mathbf{M} del lado \overline{OB} . Si dibujamos el segmento rectilíneo que une \mathbf{P} con \mathbf{M} y denotamos por d su longitud, entonces extendemos este segmento una distancia d al otro lado de \mathbf{M} que nos lleva al punto \mathbf{R} (véase la Figura 8.4); en otras palabras, \mathbf{R} es el punto simétrico de \mathbf{P} respecto de \mathbf{M} . El nuevo triángulo es $OB\mathbf{R}$ y la fórmula vectorial para hallar \mathbf{R} es

$$(10) \quad \mathbf{R} = \mathbf{M} + (\mathbf{M} - \mathbf{P}) = 2\mathbf{M} - \mathbf{P}.$$

Extensión usando el punto E

Si el valor de la función en el punto \mathbf{R} es menor que el valor en el punto \mathbf{P} , entonces nos hemos movido en la dirección correcta hacia el mínimo; quizás el mínimo está algo más allá del punto \mathbf{R} , así que extendemos una distancia adicional d el segmento que une \mathbf{M} y \mathbf{R} hasta un punto \mathbf{E} y, de esa manera, formamos un triángulo extendido OBE (véase la Figura 8.5). Si el valor de la

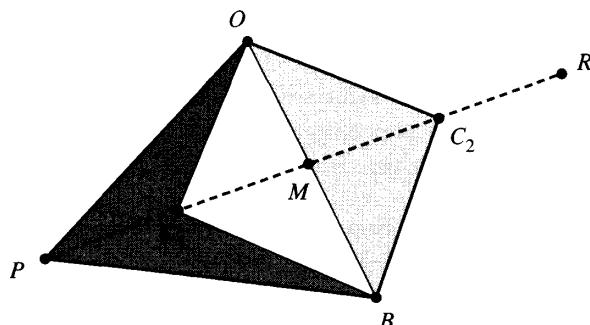


Figura 8.6 El punto de contracción C_1 o bien C_2 en el método de Nelder-Mead.

función en E es menor que en R , entonces hemos encontrado un vértice mejor que R ; la fórmula vectorial para calcular E es

$$(11) \quad E = R + (R - M) = 2R - M.$$

Contracción usando el punto C

Si los valores de la función en R y P son iguales, o en P es menor que en R , entonces hay que probar otro punto. Quizás la función es menor en el punto M , pero no podemos reemplazar P por M porque nos quedaríamos sin triángulo. Consideraremos los puntos medios C_1 y C_2 de los segmentos rectilíneos \overline{PM} y \overline{MR} , respectivamente (véase la Figura 8.6); el punto en el que la función tome un valor menor lo llamamos C y el nuevo triángulo es ahora OCB . Nota: la elección entre C_1 y C_2 es importante en dimensiones superiores.

Encogimiento hacia O

Si el valor de la función en C no es menor que el valor en P , entonces tenemos que encoger el triángulo en la dirección de O (véase la Figura 8.7): El punto B se reemplaza por M y el punto P se reemplaza por S que es el punto medio del segmento que une O con P .

Decisiones lógicas en cada paso

Un algoritmo que sea computacionalmente eficiente debería realizar una evaluación de la función sólo si es necesario. En cada paso se determina un nuevo vértice que se usa para reemplazar el vértice peor P . Una vez que se encuentra este nuevo vértice, no hace falta realizar más evaluaciones; los detalles del proceso de decisión lógica en el caso bidimensional se explican en la Tabla 8.3.

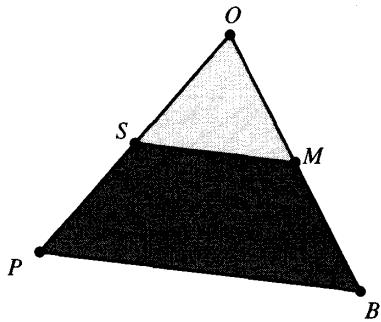


Figura 8.7 Encogimiento del triángulo hacia O .

Tabla 8.3 Decisiones lógicas en el algoritmo de Nelder-Mead.

SI $f(R) < f(B)$, ENTONCES se hace el caso (i) {reflejar o extender}	
SI NO, se hace el caso (ii) {contraer o encoger}	
COMIENZO {Caso (i)}	COMIENZO {Caso (ii)}
SI $f(O) < f(R)$ ENTONCES	SI $f(R) < f(P)$ ENTONCES
se reemplaza P por R	se reemplaza P por R
SI NO	SI NO
entonces	se calculan $C = (P + M)/2$
se calculan E y $f(E)$	o bien $C = (M + R)/2$ y $f(C)$
SI $f(E) < f(O)$ ENTONCES	SI $f(C) < f(P)$ ENTONCES
se reemplaza P por E	se reemplaza P por C
SI NO	SI NO
se reemplaza P por R	se calculan S y $f(S)$
FIN del SI	se reemplaza P por S
FIN del SI	se reemplaza B por M
FIN {Caso (i)}	FIN del SI
	FIN del SI
	FIN {Caso (ii)}

Ejemplo 8.4. Vamos a usar el algoritmo de Nelder-Mead para hallar el mínimo de $f(x, y) = x^2 - 4x + y^2 - y - xy$. Empezamos con los vértices

$$\mathbf{V}_1 = (0, 0), \quad \mathbf{V}_2 = (1.2, 0.0), \quad \mathbf{V}_3 = (0.0, 0.8).$$

La función $f(x, y)$ toma los valores

$$f(0, 0) = 0.0, \quad f(1.2, 0.0) = -3.36, \quad f(0.0, 0.8) = -0.16.$$

Comparando los valores, determinamos \mathbf{O} , \mathbf{B} y \mathbf{P} :

$$\mathbf{O} = (1.2, 0.0), \quad \mathbf{B} = (0.0, 0.8), \quad \mathbf{P} = (0, 0).$$

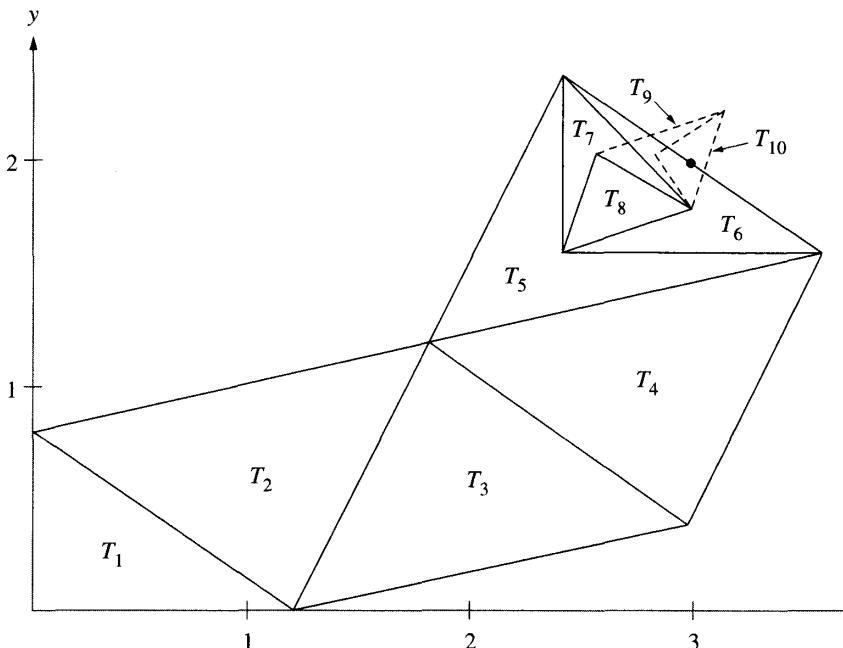


Figura 8.8 La sucesión de triángulos $\{T_k\}$ que converge al punto $(3, 2)$ en el método de Nelder-Mead.

Hay que reemplazar el vértice $\mathbf{P} = (0, 0)$. Los puntos \mathbf{M} y \mathbf{R} son

$$\mathbf{M} = \frac{\mathbf{O} + \mathbf{B}}{2} = (0.6, 0.4) \quad \text{y} \quad \mathbf{R} = 2\mathbf{M} - \mathbf{P} = (1.2, 0.8).$$

El valor de la función en \mathbf{R} es $f(\mathbf{R}) = f(1.2, 0.8) = -4.48$ que es menor que $f(\mathbf{B})$, así que estamos en el caso (i). Puesto que $f(\mathbf{R}) \leq f(\mathbf{O})$, nos hemos movido en la dirección adecuada y construimos el vértice \mathbf{E} :

$$\mathbf{E} = 2\mathbf{R} - \mathbf{M} = 2(1.2, 0.8) - (0.6, 0.4) = (1.8, 1.2).$$

El valor $f(\mathbf{E}) = f(1.8, 1.2) = -5.88$ es menor que $f(\mathbf{O})$, así que los vértices del nuevo triángulo son

$$\mathbf{V}_1 = (1.8, 1.2), \quad \mathbf{V}_2 = (1.2, 0.0), \quad \mathbf{V}_3 = (0.0, 0.8).$$

El proceso continúa y genera una sucesión de triángulos que converge al punto solución $(3, 2)$ (véase la Figura 8.8). En la Tabla 8.4 se muestran los valores de la función en los vértices del triángulo en algunos pasos de la iteración. El programa para el computador llega hasta el paso trigésimo tercero, paso en el que se obtiene como vértice óptimo el punto $\mathbf{O} = (2.99996456, 1.99983839)$ con

Tabla 8.4 Valores de la función en los triángulos sucesivos del Ejemplo 8.4.

k	Punto óptimo	Punto bueno	Punto peor
1	$f(1.2, 0.0) = -3.36$	$f(0.0, 0.8) = -0.16$	$f(0.0, 0.0) = 0.00$
2	$f(1.8, 1.2) = -5.88$	$f(1.2, 0.0) = -3.36$	$f(0.0, 0.8) = -0.16$
3	$f(1.8, 1.2) = -5.88$	$f(3.0, 0.4) = -4.44$	$f(1.2, 0.0) = -3.36$
4	$f(3.6, 1.6) = -6.24$	$f(1.8, 1.2) = -5.88$	$f(3.0, 0.4) = -4.44$
5	$f(3.6, 1.6) = -6.24$	$f(2.4, 2.4) = -6.24$	$f(1.8, 1.2) = -5.88$
6	$f(2.4, 1.6) = -6.72$	$f(3.6, 1.6) = -6.24$	$f(2.4, 2.4) = -6.24$
7	$f(3.0, 1.8) = -6.96$	$f(2.4, 1.6) = -6.72$	$f(2.4, 2.4) = -6.24$
8	$f(3.0, 1.8) = -6.96$	$f(2.55, 2.05) = -6.7725$	$f(2.4, 1.6) = -6.72$
9	$f(3.0, 1.8) = -6.96$	$f(3.15, 2.25) = -6.9525$	$f(2.55, 2.05) = -6.7725$
10	$f(3.0, 1.8) = -6.96$	$f(2.8125, 2.0375) = -6.95640625$	$f(3.15, 2.25) = -6.9525$

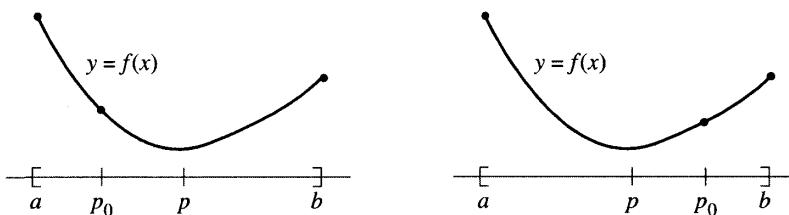


Figura 8.9 Cómo se usa $f'(x)$ para hallar el mínimo de una función unimodal $f(x)$ en el intervalo $[a, b]$. Izquierda: Si $f'(p_0) < 0$, entonces p está en $[p_0, b]$. Derecha: Si $f'(p_0) > 0$, entonces p está en $[a, p_0]$.

$f(\mathbf{O}) = -6.99999998$. Estos valores son aproximaciones a la solución $f(3, 2) = -7$ encontrada en el Ejemplo 8.3. La razón por la cual la iteración se detiene antes de llegar al punto $(3, 2)$ es que la función es muy plana cerca del mínimo: los valores $f(\mathbf{O})$, $f(\mathbf{B})$ y $f(\mathbf{P})$ son iguales (este es un ejemplo típico de error de redondeo), así que el algoritmo no puede seguir. ■

Métodos de minimización mediante derivadas

Supongamos que $f(x)$ es unimodal en $[a, b]$ y que su único mínimo se alcanza en $x = p$. Supongamos también que $f(x)$ es derivable en (a, b) . Tomemos un valor inicial p_0 en (a, b) . Si $f'(p_0) < 0$, entonces el mínimo p está a la derecha de p_0 ; mientras que si $f'(p_0) > 0$, entonces p está a la izquierda de p_0 (véase la Figura 8.9).

Localización del mínimo

Nuestro primer objetivo es obtener tres valores de partida

$$(12) \quad p_0, \quad p_1 = p_0 + h, \quad \text{y} \quad p_2 = p_0 + 2h,$$

tales que

$$(13) \quad f(p_0) > f(p_1) \quad \text{y} \quad f(p_1) < f(p_2).$$

Supongamos que $f'(p_0) < 0$, entonces $p_0 < p$ y elegimos un incremento h positivo. No es difícil encontrar h de manera que los tres puntos de (12) verifiquen (13): Empezamos con $h = 1$ en la fórmula (12) (supuesto que $a + 1 < b$, si esto no ocurre se toma $h = 1/2$ y así sucesivamente).

Caso (i): Si se verifica (13), entonces hemos terminado.

Caso (ii): Si $f(p_0) > f(p_1)$ y $f(p_1) > f(p_2)$, entonces $p_2 < p$ y tenemos que buscar más a la derecha: doblamos el incremento y repetimos el proceso.

Caso (iii): Si $f(p_0) \leq f(p_1)$, entonces hemos ido demasiado lejos a la derecha de p y h es demasiado grande, necesitamos valores más cercanos a p_0 : reducimos el incremento a la mitad y repetimos el proceso.

Cuando $f'(p_0) > 0$, entonces tomamos un incremento h negativo y se repite un proceso análogo al de los casos (i) a (iii) anteriores.

Aproximación cuadrática para calcular p

Supongamos que ya tenemos tres puntos como los dados en (12) que verifican las relaciones dadas en (13), vamos a usar interpolación cuadrática para hallar una aproximación a p que denotamos por p_{\min} . El polinomio interpolador de Lagrange para los nodos dados en (12) es

$$(14) \quad Q(x) = \frac{y_0(x - p_1)(x - p_2)}{2h^2} - \frac{y_1(x - p_0)(x - p_2)}{h^2} + \frac{y_2(x - p_0)(x - p_1)}{2h^2},$$

siendo $y_i = f(p_i)$ para $i = 0, 1, 2$. La derivada de $Q(x)$ es

$$(15) \quad Q'(x) = \frac{y_0(2x - p_1 - p_2)}{2h^2} - \frac{y_1(2x - p_0 - p_2)}{h^2} + \frac{y_2(2x - p_0 - p_1)}{2h^2}.$$

Ahora resolvemos $Q'(x) = 0$ pero escribiendo esto como $Q'(p_0 + h_{\min}) = 0$, con lo cual tenemos

$$(16) \quad 0 = \frac{y_0(2(p_0 + h_{\min}) - p_1 - p_2)}{2h^2} - \frac{y_1(4(p_0 + h_{\min}) - 2p_0 - 2p_2)}{2h^2} \\ + \frac{y_2(2(p_0 + h_{\min}) - p_0 - p_1)}{2h^2}.$$

Multiplicando cada término de (16) por $2h^2$ y agrupando los términos que contienen el factor común h_{\min} , obtenemos

$$\begin{aligned}-h_{\min}(2y_0 - 4y_1 + 2y_2) &= y_0(2p_0 - p_1 - p_2) \\&\quad - y_1(4p_0 - 2p_0 - 2p_2) + y_2(2p_0 - p_0 - p_1) \\&= y_0(-3h) - y_1(-4h) + y_2(-h),\end{aligned}$$

de donde despejamos fácilmente h_{\min} :

$$(17) \quad h_{\min} = \frac{h(4y_1 - 3y_0 - y_2)}{4y_1 - 2y_0 - 2y_2}.$$

El valor $p_{\min} = p_0 + h_{\min}$ es una aproximación a p mejor que p_0 . Ahora reemplazamos p_0 por p_{\min} y repetimos los dos procesos descritos antes, obteniendo un nuevo incremento h y un nuevo h_{\min} . La iteración continúa hasta que se obtiene el mínimo con la precisión deseada; los detalles se recogen en el Programa 8.3.

Método del gradiente o del descenso por la máxima pendiente

Volvamos al problema de minimizar una función $f(\mathbf{X})$ de N variables, donde $\mathbf{X} = (x_1, x_2, \dots, x_N)$. El gradiente de $f(\mathbf{X})$ es la función vectorial definida por

$$(18) \quad \text{grad } f(\mathbf{X}) = (f_1, f_2, \dots, f_N),$$

siendo $f_k = \partial f / \partial x_k$ las derivadas parciales evaluadas en \mathbf{X} .

Recordemos que el vector gradiente (18) señala en cada punto la dirección en la que la velocidad de crecimiento de $f(\mathbf{X})$ es mayor; por tanto, $-\text{grad } f(\mathbf{X})$ señala en cada punto la dirección en la que la velocidad de decrecimiento de la función es mayor. Esto sugiere la siguiente estrategia: Empezamos en un punto \mathbf{P}_0 y buscamos una mejora en la semirrecta que parte de \mathbf{P}_0 en la dirección señalada por el vector $\mathbf{S}_0 = -\mathbf{G} / \|\mathbf{G}\|$, siendo $\mathbf{G} = \text{grad } f(\mathbf{P}_0)$. Este es un problema en una variable (el parámetro de la semirrecta) que podemos resolver por cualquiera de los métodos señalados antes. De esta manera llegamos a un punto \mathbf{P}_1 , en el que hay un mínimo local de la función $f(\mathbf{X})$ restringida a los puntos \mathbf{X} de la semirrecta $\mathbf{X} = \mathbf{P}_0 + t\mathbf{S}_0$ con $t \geq 0$.

Ahora, calculamos $\mathbf{G} = \text{grad } f(\mathbf{P}_1)$ y usamos como nueva dirección de búsqueda la señalada por el vector $\mathbf{S}_1 = -\mathbf{G} / \|\mathbf{G}\|$; llegamos a un punto \mathbf{P}_2 , en el que hay un mínimo local de la función $f(\mathbf{X})$ restringida a los puntos \mathbf{X} de la semirrecta $\mathbf{X} = \mathbf{P}_1 + t\mathbf{S}_1$ con $t \geq 0$. Este proceso iterativo produce una sucesión $\{\mathbf{P}_k\}$ de puntos que tienen la siguiente propiedad:

$$f(\mathbf{P}_0) > f(\mathbf{P}_1) > \cdots > f(\mathbf{P}_k) > \cdots.$$

Si $\lim_{k \rightarrow \infty} \mathbf{P}_k = \mathbf{P}$, entonces $f(\mathbf{P})$ es un mínimo local de $f(\mathbf{X})$.

Descripción del método del gradiente

Supongamos que hemos obtenido el punto \mathbf{P}_k

Paso 1. Evaluamos el vector gradiente $\mathbf{G} = \text{grad } f(\mathbf{P}_k)$.

Paso 2. Calculamos la dirección de búsqueda $\mathbf{S} = -\mathbf{G}/\|\mathbf{G}\|$.

Paso 3. Utilizamos un método de minimización en una variable con la función $\Psi(t) = f(\mathbf{P}_k + t\mathbf{S})$ en intervalo $[0, b]$, donde b es grande. Esto producirá un valor $t = h_{\min}$ en el que $\Psi(t)$ tiene un mínimo local; entonces, la relación $\Psi(h_{\min}) = f(\mathbf{P}_k + h_{\min}\mathbf{S})$ nos dice que este valor es un mínimo local de $f(\mathbf{X})$ restringida a la semirecta de búsqueda $\mathbf{X} = \mathbf{P}_k + h_{\min}\mathbf{S}$.

Paso 4. Construimos el nuevo punto $\mathbf{P}_{k+1} = \mathbf{P}_k + h_{\min}\mathbf{S}$.

Paso 5. Aplicamos el criterio de parada: ¿están los valores de la función $f(\mathbf{P}_k)$ y $f(\mathbf{P}_{k+1})$ suficientemente próximos y es la distancia entre los puntos $\|\mathbf{P}_{k+1} - \mathbf{P}_k\|$ suficientemente pequeña? Si no, repetimos el proceso.

MATLAB

Programa 8.1 (Método de búsqueda de la sección áurea para minimizar). Construcción de la aproximación numérica a un mínimo de la función $f(x)$ en el intervalo $[a, b]$ usando el método de la sección áurea. Este método sólo debe usarse si la función $f(x)$ es unimodal en el intervalo $[a, b]$.

```
function[S,E,G]=golden(f,a,b,delta,epsilon)
% Datos
%   - f es la función objetivo, introducida como una
%     cadena de caracteres 'f'
%   - a y b son los extremos del intervalo
%   - delta es la tolerancia para las abscisas
%   - epsilon es la tolerancia para las ordenadas
% Resultados
%   - S=(p,yp) contiene la abscisa p y
%     la ordenada yp del mínimo
%   - E=(dp,dy) contiene las estimaciones de los errores
%     para p e yp
%   - G es una matriz de orden n x 4 cuya fila k-ésima
%     [ak ck dk bk] contiene los valores a, c, d y b
%     de la k-ésima iteración
r1=(sqrt(5)-1)/2;
r2=r1^2;
h=b-a;
```

```
ya=feval(f,a);
yb=feval(f,b);
c=a+r2*h;
d=a+r1*h;
yc=feval(f,c);
yd=feval(f,d);
k=1;
A(k)=a;B(k)=b;C(k)=c;D(k)=d;
while(abs(yb-ya)>epsilon) | (h>delta)
    k=k+1;
    if(yc<yd)
        b=d;
        yb=yd;
        d=c;
        yc=yd;
        h=b-a;
        c=a+r2*h;
        yc=feval(f,c);
    else
        a=c;
        ya=yc;
        c=d;
        yc=yd;
        h=b-a;
        d=a+r1*h;
        yd=feval(f,d);
    end
    A(k)=a;B(k)=b;C(k)=c;D(k)=d;
end
dp=abs(b-a);
dy=abs(yb-ya);
p=a;
yp=ya;
if(yb<ya)
    p=b;
    yp=yb;
end
G=[A' C' D' B'];
S=[p yp];
E=[dp dy];
```

Los Programas 8.2 y 8.4 que damos a continuación necesitan que la función objetivo F sea almacenada como un archivo F.m, cuyo argumento debe ser una

matriz de orden $1 \times n$. Veamos, a modo de ilustración, cómo se almacena la función del Ejemplo 8.3 en el archivo F.m:

```
function z=F(V)
z=0; x=V(1); y=V(2);
z=x.^2-4x+y.^2-y.*x;
```

Programa 8.2 (Método de Nelder-Mead). Aproximación a un mínimo local de $f(x_1, x_2, \dots, x_N)$, siendo f una función continua de N variables reales y $\mathbf{V}_k = (v_{k,1}, \dots, v_{k,N})$ (para $k = 0, 1, \dots, N$) los $N + 1$ vectores que forman el simplex inicial.

```
function[V0,y0,dV,dy,P,Q]=nelder(F,V,min1,max1,epsilon,show)
% Datos
% - F es la función objetivo, introducida como una
%   cadena de caracteres 'F'
% - V es una matriz de orden 3 x n
%   que contiene el simplex inicial
% - min1 y max1 son los números mínimo y
%   máximo de iteraciones
% - epsilon es la tolerancia
% - show == 1 va mostrando las iteraciones (P y Q)
% Resultados
% - V0 es el vértice en el que se alcanza el mínimo
% - y0 es el valor mínimo de la función F(V0)
% - dV es una estimación del volumen del simplex final
% - dy es la estimación del error para el valor mínimo
% - P es la matriz que contiene los vértices
%   de cada iteración
% - Q es la matriz que da los valores de la función F(P)
if nargin==5,
    show=0;
end
[mm n]=size(V);
% Ordenación de los vértices iniciales
for j=1:n+1
    Z=V(j,1:n);
    Y(j)=feval(F,Z);
end
[mm lo]=min(Y);
[mm hi]=max(Y);
li=hi;
ho=lo;
```

```
for j=1:n+1
    if(j~=lo&j~=hi&Y(j)<=Y(li))
        li=j;
    end
    if(j~=hi&j~=lo&Y(j)>=Y(ho))
        ho=j;
    end
end
cnt=0;
% Comienzo del algoritmo de Nelder-Mead
while(Y(hi)>Y(lo)+epsilon&cnt<max1)|cnt<min1
    S=zeros(1,1:n);
    for j=1:n+1
        S=S+V(j,1:n);
    end
    M=(S-V(hi,1:n))/n;
    R=2*M-V(hi,1:n);
    yR=feval(F,R);
    if(yR<Y(ho))
        if(Y(li)<yR)
            V(hi,1:n)=R;
            Y(hi)=yR;
        else
            E=2*R-M;
            yE=feval(F,E);
            if(yE<Y(li))
                V(hi,1:n)=E;
                Y(hi)=yE;
            else
                V(hi,1:n)=R;
                Y(hi)=yR;
            end
        end
    end
    else
        if(yR<Y(hi))
            V(hi,1:n)=R;
            Y(hi)=yR;
        end
    end
    C=(V(hi,1:n)+M)/2;
    yC=feval(F,C);
    C2=(M+R)/2;
    yC2=feval(F,C2);
    if(yC2<yC)
```

```
C=C2;
yC=yC2;
end
if(yC<Y(hi))
    V(hi,1:n)=C;
    Y(hi)=yC;
else
    for j=1:n+1
        if(j~=lo)
            V(j,1:n)=(V(j,1:n)+V(lo,1:n))/2;
            Z=V(j,1:n);
            Y(j)=feval(F,Z);
        end
    end
end
[mm lo]=min(Y);
[mm hi]=max(Y);
li=hi;
ho=lo;
for j=1:n+1
    if(j~=lo&j~=hi&Y(j)<=Y(li))
        li=j;
    end
    if(j~=hi&j~=lo&Y(j)>=Y(ho))
        ho=j;
    end
end
cnt=cnt+1;
P(cnt,:)=V(lo,:);
Q(cnt)=Y(lo);
end
% Fin del algoritmo de Nelder-Mead
% Estimación del volúmen del simplex
snorm=0;
for j=1:n+1
    s=norm(V(j)-V(lo));
    if(s>=snorm)
        snorm=s;
    end
end
Q=Q';
V0=V(lo,1:n);
```

```
y0=Y(lo);
dV=snorm;
dy=abs(Y(hi)-Y(lo));
if (show==1)
    disp(P);
    disp(Q);
end
```

Programa 8.3 (Búsqueda de un mínimo local mediante interpolación cuadrática). Búsqueda de un mínimo local de la función $f(x)$ en el intervalo $[a, b]$, realizada a partir de una aproximación inicial p_0 .

```
function[p,yp,dp,dy,P]=quadmin(f,a,b,delta,epsilon)

% Datos
% - f es la función objetivo, introducida como una
%   cadena de caracteres 'f'
% - a y b son los extremos del intervalo
% - delta es la tolerancia para las abscisas
% - epsilon es la tolerancia para las ordenadas
% Resultados
% - p es la abscisa del mínimo
% - yp es la ordenada del mínimo
% - dp es la estimación del error de p
% - dy es la estimación del error de yp
% - P es el vector de las iteraciones

p0=a;
maxj=20;
maxk=30;
big=1e6;
err=1;
k=1;
P(k)=p0;
cond=0;
h=1;
if (abs(p0)>1e4),h=abs(p0)/1e4;end
while(k<maxk&err>epsilon&cond~=5)
    f1=(feval(f,p0+0.00001)-feval(f,p0-0.00001))/0.00002;
    if(f1>0),h=-abs(h);end
    p1=p0+h;
    p2=p0+2*h;
    pmin=p0;
    y0=feval(f,p0);
    y1=feval(f,p1);
```

```
y2=feval(f,p2);
ymin=y0;
cond=0;
j=0;
% Cálculo de h para que y1<y0&y1<y2
while(j<maxj&abs(h)>delta&cond==0)
    if (y0<=y1),
        p2=p1;
        y2=y1;
        h=h/2;
        p1=p0+h;
        y1=feval(f,p1);
    else
        if(y2<y1),
            p1=p2;
            y1=y2;
            h=2*h;
            p2=p0+2*h;
            y2=feval(f,p2);
        else
            cond=-1;
        end
    end
j=j+1;
if(abs(h)>big|abs(p0)>big),cond=5;end
end
if(cond==5),
    pmin=p1;
    ymin=feval(f,p1);
else
    % Interpolación cuadrática para hallar yp
    d=4*y1-2*y0-2*y2;
    if(d<0),
        hmin=h*(4*y1-3*y0-y2)/d;
    else
        hmin=h/3;
        cond=4;
    end
    pmin=p0+hmin;
    ymin=feval(f,pmin);
    h=abs(h);
    h0=abs(hmin);
    h1=abs(hmin-h);
```

```

h2=abs(hmin-2*h);

% Cálculo del siguiente incremento h
if(h0<h),h=h0;end
if(h1<h),h=h1;end
if(h2<h),h=h2;end
if(h==0),h=hmin;end
if(h<delta),cond=1;end
if (abs(h)>big|abs(pmin)>big),cond=5;end

% Criterio de parada de la minimización
e0=abs(y0-ymin);
e1=abs(y1-ymin);
e2=abs(y2-ymin);
if(e0~=0 & e0<err),err=e0;end
if(e1~=0 & e1<err),err=e1;end
if(e2~=0 & 2<err),err=e2;end
if(e0~=0 & e1==0 & e2==0),error=0;end
if(err<epsilon),cond=2;end
p0=pmin;
k=k+1;
P(k)=p0;
end
if(cond==2&h<delta),cond=3;end
end
p=p0;
dp=h;
yp=feval(f,p);
dy=err;

```

El Programa 8.4 necesita que la función objetivo f se almacene como un archivo `f.m` y lo mismo con la dirección de búsqueda $-\text{grad } f / \|\text{grad } f\|$. Veamos, a modo de ilustración, cómo se almacena el vector gradiente $(2x-4-y, 2y-1-x)$ de la función f del Ejemplo 8.3 en el archivo `G.m`:

```

function z=G(V)
z=zeros(1,2);
x=V(1);y=V(2);
g=[2x-4-y 2*y-1-x];
z=-(1/norm(g))*g;

```

Programa 8.4 (Método del descenso por la máxima pendiente o método del gradiente). Aproximación a un mínimo local de una función diferenciable $f(\mathbf{X})$ de N variables reales $\mathbf{X} = (x_1, x_2, \dots, x_N)$, usando el método del gradiente a partir de un punto inicial \mathbf{P}_0 .

```

function[P0,y0,err]=grads(F,G,P0,max1,delta,epsilon,show)
% Datos
%   - F es la función objetivo, introducida como una
%     cadena de caracteres 'F'
%   - G ==-(1/norm(grad F))*grad F es la dirección
%     de búsqueda, introducida como una cadena de
%     caracteres 'G'
%   - P0 es el punto inicial
%   - max1 es el número máximo de iteraciones
%   - delta es la tolerancia para hmin en el proceso
%     de minimización unidimensional en la
%     dirección de búsqueda
%   - epsilon es la tolerancia para el error en y0
%   - show; si show==1, entonces las iteraciones
%     se van mostrando en la pantalla
% Resultados
%   - P0 es la aproximación del punto donde
%     se alcanza el mínimo
%   - y0 es la aproximación al valor mínimo F(P0)
%   - err es la estimación del error para y0
%   - P es el vector que contiene las iteraciones

if nargin==5,show=0;end
[m n]=size(P0);
maxj=10; big=1e8; h=1;
P=zeros(maxj,n+1);
len=norm(P0);
y0=feval(F,P0);
if (len>e4),h=len/1e4;end
err=1;cnt=0;cond=0;
P(cnt+1,:)=[P0 y0];
while(cnt<max1&cond~=5&(h>delta|err>epsilon))
    % Cálculo de la dirección de búsqueda
    S=feval(G,P0);

    % Comienzo de la minimización unidimensional mediante
    % interpolación cuadrática
    P1=P0+h*S;
    P2=P0+2*h*S;
    y1=feval(F,P1);
    y2=feval(F,P2);
    cond=0;j=0;
    while(j<maxj&cond==0)
        len=norm(P0);

```

```
if (y0<y1)
P2=P1;
y2=y1;
h=h/2;
P1=P0+h*S;
y1=feval(F,P1);
else
    if(y2<y1)
        P1=P2;
        y1=y2;
        h=2*h;
        P2=P0+2*h*S;
        y2=feval(F,P2);
    else
        cond=-1;
    end
end
j=j+1;
if(h<delta),cond=1;end
if(abs(h)>big|len>big),cond=5;end
end
if(cond==5)
    Pmin=P1;
    ymin=y1;
else
    d=4*y1-2*y0-2*y2;
    if(d<0)
        hmin=h*(4*y1-3*y0-y2)/d;
    else
        cond=4;
        hmin=h/3;
    end
% Construcción del próximo punto
Pmin=P0+hmin*S;
ymin=feval(F,Pmin);

% Determinación de la magnitud del próximo incremento h
h0=abs(hmin);
h1=abs(hmin-h);
h2=abs(hmin-2*h);
if(h0<h),h=h0;end
if(h1<h),h=h1;end
if(h2<h),h=h2;end
```

```
if(h==0),h=hmin;end
if(h<delta),cond=1;end

% Criterio de parada de la minimización
e0=abs(y0-ymin);
e1=abs(y1-ymin);
e2=abs(y2-ymin);
if(e0~=0&e0<err),err=e0;end
if(e1~=0&e1<err),err=e1;end
if(e2~=0&e2<err),err=e2;end
if(e0==0&e1==0&e2==0),err=0;end
if(err<epsilon),cond=2;end
if(cond==2&h<delta),cond=3;end
end
cnt=cnt+1;
P(cnt,:)=[Pmin ymin];
P0=Pmin;
y0=ymin;
end
if(show==1)
    disp(P);
end
```

Ejercicios

1. Use el Teorema 8.1 para determinar dónde son crecientes y dónde son decrecientes las siguientes funciones.
 - (a) $f(x) = 2x^3 - 9x^2 + 12x - 5$
 - (b) $f(x) = x/(x + 1)$
 - (c) $f(x) = (x + 1)/x$
 - (d) $f(x) = x^x$
2. Use la Definición 8.3 para probar que cada una de las siguientes funciones es unimodal en el intervalo que se indica.
 - (a) $f(x) = x^2 - 2x + 1$; $[0, 4]$
 - (b) $f(x) = \cos(x)$; $[0, 3]$
 - (c) $f(x) = x^x$; $[1, 10]$
 - (d) $f(x) = -x(3 - x)^{5/3}$; $[0, 3]$
3. Use los Teoremas 8.3 y 8.4 para, si es posible, hallar todos los máximos y mínimos locales de cada una de las siguientes funciones en el intervalo que se indica.
 - (a) $f(x) = 4x^3 - 8x^2 - 11x + 5$; $[0, 2]$

- (b) $f(x) = x + 3/x^2$; $[0.5, 3]$
(c) $f(x) = (x + 2.5)/(4 - x^2)$; $[-1.9, 1.9]$
(d) $f(x) = e^x/x^2$; $[0.5, 3]$
(e) $f(x) = -\operatorname{sen}(x) - \operatorname{sen}(3x)/3$; $[0, 2]$
(f) $f(x) = -2\operatorname{sen}(x) + \operatorname{sen}(2x) - 2\operatorname{sen}(3x)/3$; $[1, 3]$
4. Determine el punto de la parábola $y = x^2$ que está más cerca del punto $(3, 1)$.
5. Determine el punto de la curva $y = \operatorname{sen}(x)$ que está más cerca del punto $(2, 1)$.
6. Determine el punto o los puntos de la circunferencia $x^2 + y^2 = 25$ que están más alejados de la cuerda AB siendo $A = (3, 4)$ y $B = (-1, \sqrt{24})$.
7. Use el Teorema 8.5 para hallar el (o los) mínimo local de cada una de las siguientes funciones:
(a) $f(x, y) = x^3 + y^3 - 3x - 3y + 5$
(b) $f(x, y) = x^2 + y^2 + x - 2y - xy + 1$
(c) $f(x, y) = x^2y + xy^2 - 3xy$
(d) $f(x, y) = (x - y)/(x^2 + y^2 + 2)$
(e) $f(x, y) = 100(y - x^2)^2 + (1 - x)^2$
(El valle parabólico de la función de Rosenbrock, circa 1960)
8. Sean $\mathbf{O} = (2, -3)$, $\mathbf{B} = (1, 1)$ y $\mathbf{P} = (5, 2)$. Calcule los puntos \mathbf{M} , \mathbf{R} y \mathbf{E} y esboce los triángulos involucrados en el método de minimización de Nelder-Mead.
9. Sean $\mathbf{O} = (-1, 2)$, $\mathbf{B} = (-2, -5)$ y $\mathbf{P} = (3, 1)$. Calcule los puntos \mathbf{M} , \mathbf{R} y \mathbf{E} y esboce los triángulos involucrados en el método de minimización de Nelder-Mead.
10. Dé una demostración vectorial de que $\mathbf{M} = (\mathbf{O} + \mathbf{B})/2$ es el punto medio del segmento rectilíneo que une \mathbf{O} con \mathbf{B} .
11. Dé una demostración vectorial de la igualdad (10).
12. Dé una demostración vectorial de la igualdad (11).
13. Demuestre vectorialmente que las medianas de un triángulo se cortan en un punto que está a dos tercios de la distancia desde cada vértice al punto medio del lado opuesto.
14. Sean $\mathbf{O} = (0, 0, 0)$, $\mathbf{B} = (1, 1, 0)$, $\mathbf{A} = (0, 0, 1)$ y $\mathbf{P} = (1, 0, 0)$.
(a) Esboce el tetraedro de vértices \mathbf{O} , \mathbf{B} , \mathbf{A} y \mathbf{P} .
(b) Calcule $\mathbf{M} = (\mathbf{O} + \mathbf{B} + \mathbf{A})/3$.
(c) Calcule $\mathbf{R} = 2\mathbf{M} - \mathbf{W}$ y esboce el tetraedro de vértices \mathbf{O} , \mathbf{B} , \mathbf{A} y \mathbf{R} .
(d) Calcule $\mathbf{E} = 2\mathbf{R} - \mathbf{M}$ y esboce el tetraedro de vértices \mathbf{O} , \mathbf{B} , \mathbf{A} y \mathbf{E} .
15. Sean $\mathbf{O} = (0, 0, 0)$, $\mathbf{B} = (0, 2, 0)$, $\mathbf{A} = (0, 1, 1)$ y $\mathbf{P} = (2, 1, 0)$. Haga lo mismo que en el Ejercicio 14.

Algoritmos y programas

1. Use el Programa 8.1 para hallar el mínimo local de cada una de las funciones del Ejercicio 3 con una precisión de ocho cifras decimales.
2. Use el Programa 8.3 para hallar el mínimo local de cada una de las funciones del Ejercicio 3 con una precisión de ocho cifras decimales. Empiece con el punto medio de cada intervalo.
3. Use el Programa 8.2 para hallar el mínimo de cada una de las funciones del Ejercicio 7 con una precisión de ocho cifras decimales. Utilice los siguientes vértices de partida, respectivamente.
 - (a) $(1, 2), (2, 0)$ y $(2, 2)$
 - (b) $(0, 0), (2, 0)$ y $(2, 1)$
 - (c) $(0, 0), (2, 0)$ y $(2, 1)$
 - (d) $(0, 0), (0, 1)$ y $(1, 1)$
 - (e) $(0, 0), (1, 0)$ y $(0, 2)$
4. Use el Programa 8.4 para hallar el mínimo de cada una de las funciones del Ejercicio 7 con una precisión de ocho cifras decimales. Utilice el siguiente vértice de partida, respectivamente.

(a) $(1, 2)$	(b) $(0, 0, 3)$	(c) $(0.1, 0.1)$
(d) $(0.5, 0.11)$	(e) $(0, 0)$	
5. En el Programa 8.4 las coordenadas x e y de las iteraciones se almacenan en las dos primeras columnas de la matriz P , respectivamente. Modifique el Programa 8.4 de manera que proporcione el dibujo de las coordenadas x e y de las iteraciones sobre un mismo gráfico. Use su programa con las funciones del Ejercicio 7. *Indicación.* Incorpore la instrucción `plot(P(:,1),P(:,2),'.')` a su programa.
6. Use el Programa 8.2 para hallar un mínimo local de cada una de las siguientes funciones con una precisión de ocho cifras decimales.
 - (a) $f(x, y, z) = 2x^2 + 2y^2 + z^2 - 2xy + yz - 7y - 4z$ con el tetraedro de partida: $(1, 1, 1), (0, 1, 0), (1, 0, 1)$ y $(0, 0, 1)$.
 - (b) $f(x, y, z, u) = 2(x^2 + y^2 + z^2 + u^2) - x(y + z - u) + yz - 3x - 8y - 5z - 9u$. Comience la búsqueda cerca del punto $(1, 1, 1, 1)$.
 - (c) $f(x, y, z, u) = xyzu + \frac{1}{x} + \frac{1}{y} + \frac{1}{z} + \frac{1}{u}$ Comience la búsqueda cerca del punto $(0.7, 0.7, 0.7, 0.7)$.
7. Use el Programa 8.4 para hallar el mínimo local de cada una de las funciones del Problema 6. Utilice un punto inicial que esté cerca de alguno de los vértices dados.

8. Use los Programas 8.1 y 8.3 para calcular máximos y mínimos locales de la siguiente función en el intervalo $[0, 2]$.

$$f(x) = \frac{x^3 + x^2 - 12x - 12}{2x^6 - 3x^5 - 4x^4 + 9x^2 + 12x - 18}$$

9. Determine el punto de la superficie $z = x^2 + y^2$ que está más cerca del punto $(2, 3, 1)$.
10. Una compañía tiene cinco fábricas A, B, C, D y E, localizadas en los puntos $(10, 10)$, $(30, 50)$, $(16.667, 29)$, $(0.555, 29.888)$ y $(22.2221, 49.988)$, respectivamente, de un plano cartesiano. Supongamos que la distancia entre dos puntos representa la distancia que hay por carretera, en kilómetros, entre las fábricas correspondientes. La compañía planea construir un almacén para componentes en algún punto del plano y se sabe de antemano que, en una semana media, habrá que enviar 10, 18, 20, 14 y 25 remesas a cada una de las fábricas A, B, C, D y E, respectivamente. Si se quiere minimizar la distancia semanal recorrida por el total de vehículos que participan en los envíos, ¿en qué lugar hay que situar, idealmente, el almacén?
11. En el Problema 10, ¿dónde hay que situar el almacén si, debido a restricciones sobre la zona de localización, debe estar sobre la curva $y = x^2$?

Ecuaciones diferenciales ordinarias

Las ecuaciones diferenciales se usan habitualmente para construir modelos matemáticos de problemas de la ciencia y la ingeniería. A menudo se da el caso de que no hay una solución analítica conocida, por lo que necesitamos aproximaciones numéricas. A modo de ejemplo consideremos, en el contexto de la dinámica de poblaciones, un sistema no lineal que es una modificación de las ecuaciones de Lotka-Volterra:

$$x' = f(t, x, y) = x - xy - \frac{1}{10}x^2 \quad \text{e} \quad y' = g(t, x, y) = xy - y - \frac{1}{20}y^2,$$

para $0 \leq t \leq 30$ con la condición inicial $x(0) = 2$ e $y(0) = 1$. La solución

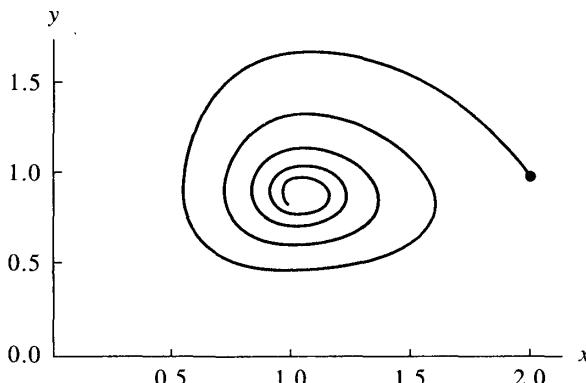


Figura 9.1 Trayectoria de un sistema de ecuaciones diferenciales no lineales $x' = f(t, x, y)$ e $y' = g(t, x, y)$.

numérica es una mera tabla de números, así que resulta más ilustrativo dibujar la trayectoria poligonal, mostrada en la Figura 9.1, que une los puntos de la solución aproximada $\{(x_k, y_k)\}$. En este capítulo presentamos los métodos habituales de resolución numérica de ecuaciones diferenciales ordinarias, sistemas de ecuaciones diferenciales y problemas de contorno.

9.1 Introducción a las ecuaciones diferenciales

Consideremos la ecuación

$$(1) \quad \frac{dy}{dt} = 1 - e^{-t}.$$

Esta es una ecuación diferencial porque en ella aparece la derivada dy/dt de la “función desconocida” $y = y(t)$. En el miembro derecho de la ecuación (1) sólo aparece la variable independiente t , así que las soluciones son las primitivas de $1 - e^{-t}$. Usando las técnicas del cálculo de primitivas podemos hallar $y(t)$:

$$(2) \quad y(t) = t + e^{-t} + C,$$

donde C es la constante de integración. Todas las funciones de la forma (2) son soluciones de la ecuación (1) porque verifican que $y'(t) = 1 - e^{-t}$; sus gráficas forman la familia de curvas que se muestran en la Figura 9.2.

Las técnicas del cálculo de primitivas nos han permitido hallar la fórmula explícita de las soluciones que aparece en (2); observamos en ella que hay un grado de libertad en la elección de la solución, la constante de integración C , lo que se pone de manifiesto en la Figura 9.2: variando el valor de C “movemos”

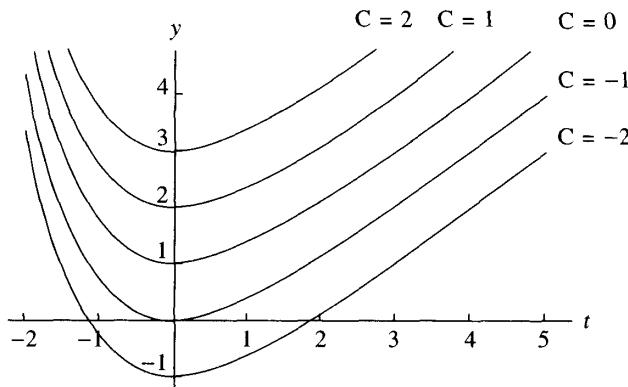


Figura 9.2 La familia de curvas $y(t) = t + e^{-t} + C$.

la curva solución” hacia arriba o hacia abajo, de manera que siempre podemos encontrar una curva particular que pase por un punto dado de antemano. Los secretos del mundo no se suelen esconder detrás de fórmulas explícitas; en vez de eso, lo que normalmente somos capaces de medir es cómo los cambios de una variable afectan a otra variable. Cuando traducimos esto en un modelo matemático, el resultado es una ecuación diferencial que involucra la velocidad de cambio de la función desconocida y , en la mayoría de las ocasiones, las variables dependiente e independiente.

Consideremos la temperatura $y(t)$ de un objeto que se enfriá. Podríamos conjeturar que la velocidad de cambio de la temperatura del cuerpo está relacionada con la diferencia entre su temperatura y la del medio que lo rodea; los experimentos confirman esta conjetura y la ley del enfriamiento de Newton establece que dicha velocidad de cambio es directamente proporcional a la diferencia de estas temperaturas. Si denotamos por A la temperatura del medio que lo rodea y por $y(t)$ la temperatura del cuerpo en el instante t , entonces

$$(3) \quad \frac{dy}{dt} = -k(y - A),$$

donde k es una constante positiva; hace falta incluir el signo negativo porque dy/dt será negativa (la temperatura decrece) siempre que la temperatura del cuerpo sea mayor que la del medio.

Si conocemos la temperatura y_0 del cuerpo en el instante $t = 0$, entonces incluimos esta información, que se denomina condición inicial, en el enunciado del problema, de manera que lo que queremos resolver es

$$(4) \quad \frac{dy}{dt} = -k(y - A) \quad \text{con} \quad y(0) = y_0.$$

La solución la calculamos usando la técnica de separación de variables, obteniendo

$$(5) \quad y = A + (y_0 - A)e^{-kt}.$$

Cada elección de y_0 nos proporciona una solución distinta; es como si el valor inicial fuera el punto de anclaje de la curva correspondiente a la solución, de forma que no podemos “saltar” de una curva a otra. En la Figura 9.3 se muestran varias soluciones del problema y en ella podemos observar que, conforme t crece, la temperatura del cuerpo se aproxima a la temperatura ambiente y , también, que si $y_0 < A$, entonces el cuerpo no se enfriá sino que se calienta.

Problemas de valor inicial

Definición 9.1. Una *solución del problema de valor inicial*

$$(6) \quad y' = f(t, y) \quad \text{con} \quad y(t_0) = y_0$$

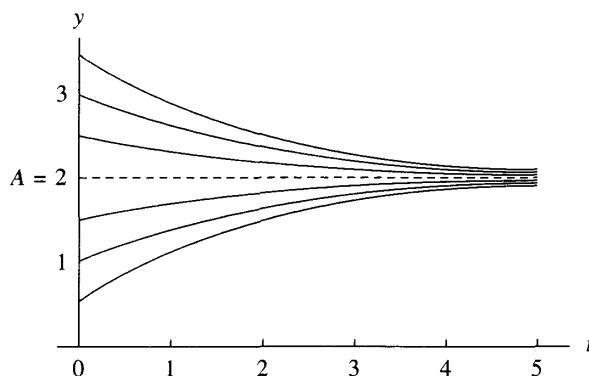


Figura 9.3 Gráficas de las curvas $y = A + (y_0 - A)e^{-kt}$, soluciones de la ecuación del enfriamiento (y calentamiento) de Newton.

en un intervalo $[t_0, t_1]$ es una función derivable $y = y(t)$ tal que

$$(7) \quad y(t_0) = y_0 \quad \text{e} \quad y'(t) = f(t, y(t)) \quad \text{para todo } t \in [t_0, t_1].$$

Hagamos notar que la gráfica de la solución $y = y(t)$ debe pasar por el punto inicial (t_0, y_0) .

Interpretación geométrica

En cada punto (t, y) del rectángulo $R = \{(t, y) : a \leq t \leq b, c \leq y \leq d\}$ la pendiente m de la solución $y = y(t)$ puede hallarse mediante la fórmula implícita $m = f(t, y(t))$. Por tanto, cada valor $m_{i,j} = f(t_i, y_j)$, calculado para distintos puntos del rectángulo, representa la pendiente de la recta tangente a la solución que pasa por (t_i, y_j) .

Un campo de direcciones, o campo de pendientes, es una gráfica en la que se representan las pendientes $\{m_{i,j}\}$ en una colección de puntos del rectángulo y puede usarse para ver cómo se va ajustando una solución a la pendiente dada. Para movernos a lo largo de la solución, debemos ponernos en el punto inicial y calcular la pendiente $f(t_0, y_0)$ para determinar en qué dirección debemos movernos. Ahora damos un pasito horizontal desde t_0 hasta $t_0 + h$ y luego nos desplazamos verticalmente una distancia apropiada $hf(t_0, y_0)$, llegando a un punto que denotamos por (t_1, y_1) , de manera que el desplazamiento total que resulta tenga la inclinación requerida. Una vez en el punto (t_1, y_1) , repetimos el proceso y continuamos nuestro viaje a lo largo de la solución. Puesto que sólo podemos dar un número finito de pasos, este método producirá una aproximación a la solución.

Ejemplo 9.1. En la Figura 9.4 se muestran el campo de direcciones de la ecuación $y' = (t - y)/2$ en el rectángulo $R = \{(t, y) : 0 \leq t \leq 5, 0 \leq y \leq 4\}$ y las soluciones correspondientes a los siguientes valores iniciales:

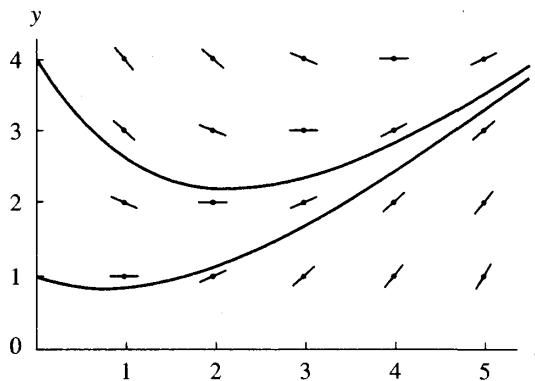


Figura 9.4 El campo de direcciones de la ecuación diferencial $y' = f(x, y) = (t - y)/2$.

1. Para $y(0) = 1$, la solución es $y(t) = 3e^{-t/2} - 2 + t$.
2. Para $y(0) = 4$, la solución es $y(t) = 6e^{-t/2} - 2 + t$. ■

Definición 9.2. Dado el rectángulo $R = \{(t, y) : a \leq t \leq b, c \leq y \leq d\}$, supongamos que $f(t, y)$ es continua en R . Se dice que la función f verifica una **condición de Lipschitz** con respecto a su variable y en R si existe una constante $L > 0$ tal que

$$(8) \quad |f(t, y_1) - f(t, y_2)| \leq L |y_1 - y_2|$$

para cualesquiera $(t, y_1), (t, y_2) \in R$. La constante L se llama **constante de Lipschitz** de f . ▲

Teorema 9.1. Supongamos que $f(t, y)$ está definida en un rectángulo R . Si existe una constante $L > 0$ tal que

$$(9) \quad |f_y(t, y)| \leq L \quad \text{para todo } (t, y) \in R,$$

entonces f verifica una condición de Lipschitz con respecto a su variable y en R , siendo L su constante de Lipschitz.

Demuestração. Fijando t y usando el teorema del valor medio, obtenemos c_1 , con $y_1 < c_1 < y_2$, tal que

$$\begin{aligned} |f(t, y_1) - f(t, y_2)| &= |f_y(t, c_1)(y_1 - y_2)| \\ &= |f_y(t, c_1)||y_1 - y_2| \leq L|y_1 - y_2|. \end{aligned}$$

Teorema 9.2 (Existencia y unicidad de soluciones). Supongamos que $f(t, y)$ es continua en el rectángulo $R = \{(t, y) : t_0 \leq t \leq t_1, c \leq y \leq d\}$. Si f verifica una condición de Lipschitz con respecto a su variable y en R y $(t_0, y_0) \in R$, entonces el problema de valor inicial (6), $y' = f(t, y)$ con $y(t_0) = y_0$, tiene solución única $y = y(t)$ en algún subintervalo $t_0 \leq t \leq t_0 + \delta$.

Demostración. Puede hallarse la demostración en cualquier texto de ecuaciones diferenciales como, por ejemplo, la Referencia [38]. •

Aplicemos los Teoremas 9.1 y 9.2 a la función $f(t, y) = (t - y)/2$: La derivada parcial es $f_y(t, y) = -1/2$, por tanto $|f_y(t, y)| \leq \frac{1}{2}$ y, de acuerdo con el Teorema 9.1, la constante de Lipschitz es $L = \frac{1}{2}$. En consecuencia, por el Teorema 9.2, el problema de valor inicial tiene solución única. Podemos hacer un esbozo del campo de direcciones y de las soluciones usando las instrucciones `meshgrid` y `quiver` del paquete de programas MATLAB. El siguiente archivo permite generar una gráfica similar a la de la Figura 9.4. En general, hay que tener la precaución de evitar los puntos (t, y) en los que y' no está definida.

```
[t,y]=meshgrid(1:5,4:-1:1);
dt=ones(5,4);
dy=(t-y)/2;
quiver(t,y,dt,dy);
hold on
x=0:.01:5;
z1=3*exp(-x/2)-2+x;
z2=6*exp(-x/2)-2+x;
plot(x,z1,x,z2)
hold off
```

Ejercicios

En los Ejercicios 1 a 5:

- (a) Pruebe que $y(t)$ es la solución de la ecuación diferencial sustituyendo $y(t)$ e $y'(t)$ en la ecuación $y'(t) = f(t, y(t))$.

- (b) Use el Teorema 9.1 para hallar una constante de Lipschitz L de f en el rectángulo $R = \{(t, y) : 0 \leq t \leq 3, 0 \leq y \leq 5\}$.

1. $y' = t^2 - y, y(t) = Ce^{-t} + t^2 - 2t + 2$

2. $y' = 3y + 3t, y(t) = Ce^{3t} - t - \frac{1}{3}$

3. $y' = -ty, y(t) = Ce^{-t^2/2}$

4. $y' = e^{-2t} - 2y, y(t) = Ce^{-2t} + te^{-2t}$

5. $y' = 2ty^2, y(t) = 1/(C - t^2)$

En los Ejercicios 6 a 9, construya una gráfica del campo de direcciones de la ecuación diferencial, $m_{i,j} = f(t_i, y_j)$, y de las soluciones indicadas en el rectángulo $R = \{(t, y) : 0 < t \leq 4, 0 < y \leq 4\}$.

6. $y' = -t/y$, $y(t) = (C - t^2)^{1/2}$ para $C = 1, 2, 4, 9$
7. $y' = t/y$, $y(t) = (C + t^2)^{1/2}$ para $C = -4, -1, 1, 4$
8. $y' = 1/y$, $y(t) = (C + 2t)^{1/2}$ para $C = -4, -2, 0, 2$
9. $y' = y^2$, $y(t) = 1/(C - t)$ para $C = 1, 2, 3, 4$
10. He aquí un ejemplo de un problema de valor inicial que tiene “dos soluciones”: $y' = \frac{3}{2}y^{1/3}$ con $y(0) = 0$.
 - (a) Compruebe que $y(t) = 0$ para $t \geq 0$ es una solución.
 - (b) Compruebe que $y(t) = t^{3/2}$ para $t \geq 0$ es una solución.
 - (c) ¿Viola este ejemplo el Teorema 9.2? ¿Por qué?
11. Consideremos el problema de valor inicial

$$y' = (1 - y^2)^{1/2} \quad y(0) = 0.$$

- (a) Compruebe que $y(t) = \sin(t)$ es una solución en $[0, \pi/4]$.
- (b) Determine el intervalo más grande en el que la solución existe.
12. Pruebe que la integral definida $\int_a^b f(t) dt$ puede calcularse resolviendo el problema de valor inicial

$$y' = f(t) \quad \text{para } a \leq t \leq b \quad \text{con} \quad y(a) = 0.$$

En los Ejercicios 13 a 15, halle la solución del problema de valor inicial.

13. $y' = 3t^2 + \sin(t)$, $y(0) = 2$.
14. $y' = \frac{1}{1+t^2}$, $y(0) = 0$.
15. $y' = e^{-t^2/2}$, $y(0) = 0$. *Indicación.* La solución debe expresarse mediante una cierta integral.
16. Consideremos la ecuación diferencial lineal de primer orden

$$y'(t) + p(t)y(t) = q(t).$$

Pruebe que se puede hallar la solución general $y(t)$ usando dos integrales especiales: En primer lugar, defina $F(t)$ mediante

$$F(t) = e^{\int p(t) dt}.$$

Luego, defina $y(t)$ mediante

$$y(t) = \frac{1}{F(t)} \left(\int F(t)q(t) dt + C \right).$$

Indicación. Derive el producto $F(t)y(t)$.

- 17.** Consideremos la desintegración de una sustancia radiactiva. Si $y(t)$ representa la cantidad de sustancia presente en un instante t , entonces $y(t)$ decrece y los experimentos sugieren que la velocidad de decrecimiento de $y(t)$ es proporcional a la cantidad presente. Por tanto, el problema de valor inicial para la desintegración de una sustancia radiactiva es

$$y' = -ky \quad \text{con} \quad y(0) = y_0.$$

- (a) Pruebe que la solución es $y(t) = y_0 e^{-kt}$.
- (b) La vida media de una sustancia radiactiva se define como el tiempo necesario para que se desintegre la mitad de la cantidad presente al principio; por ejemplo, la vida media del isótopo ^{14}C es 5730 años. Halle la fórmula $y(t)$ que proporciona la cantidad de ^{14}C presente en un instante t . *Indicación.* Determine k de manera que $y(5730) = 0.5y_0$.
- (c) Tras analizar un trozo de madera, se establece que la cantidad de ^{14}C presente es 0.712 veces la cantidad que había cuando el árbol estaba vivo. ¿Cuántos años tiene ese trozo de madera?
- (d) En un cierto instante hay 10 mg de una sustancia radiactiva y 23 segundos después sólo queda 1 mg. ¿Cuál es la vida media de esa sustancia?

En los Ejercicios 18 y 19, establezca el problema de valor inicial y halle la solución.

- 18.** Se proyecta una nueva liga europea de fútbol de manera que la venta anual de entradas crezca a una velocidad proporcional a la diferencia entre las ventas en un instante t y una cota superior de 300 millones de euros. Supongamos que en el instante inicial no se ha vendido ninguna entrada y que después de tres años deben haberse alcanzado unas ventas de 40 millones de euros (si no, la competición se suprime). Basándose en esta hipótesis, ¿cuánto tiempo hace falta para que la venta anual de entradas alcance los 220 millones de euros?
- 19.** El interior de una biblioteca tiene un volumen de 5 millones de metros cúbicos. El sistema de ventilación introduce aire fresco en la biblioteca a una velocidad de 45 000 metros cúbicos por segundo. Antes de encender el sistema de ventilación, los porcentajes de dióxido de carbono presentes en el interior de la biblioteca y en el exterior son, respectivamente, el 0.4% y el 0.5%. Determine el porcentaje de dióxido de carbono presente en la biblioteca dos horas después de encender el sistema de ventilación.

9.2 El método de Euler

No todos los problemas de valor inicial pueden resolverse explícitamente; con frecuencia es imposible hallar una fórmula que represente la solución $y(t)$. Por ejemplo, no existe una “expresión cerrada” para la solución del problema de valor inicial $y' = t^3 + y^2$ con $y(0) = 0$. En consecuencia, es necesario disponer de métodos que aproximen la solución de problemas que aparecen en la ciencia

y la ingeniería. Si, además, se requiere que la aproximación a la solución tenga muchas cifras decimales de precisión, entonces necesitaremos usar métodos sofisticados y el esfuerzo computacional será grande.

El primer método que veremos es el método de Euler y nos servirá para ilustrar una serie de conceptos que juegan un papel importante en métodos más avanzados. El método de Euler no se suele utilizar en la práctica debido a que la solución que proporciona acumula errores apreciables a lo largo del proceso; sin embargo, es importante estudiarlo porque es más fácil llevar a cabo el análisis del error de este método que el de otros más exactos pero más complejos.

Sea $[a, b]$ el intervalo en el que queremos hallar la solución de un problema de valor inicial $y' = f(t, y)$ con $y(a) = y_0$ que está bien planteado (en el sentido de que f satisface una condición de Lipschitz). Hay que advertir que, de hecho, no vamos a encontrar una función derivable que sea solución del problema de valor inicial; en vez de eso, lo que se construye es un conjunto finito de puntos $\{(t_k, y_k)\}$ que son aproximaciones de la solución (o sea, $y(t_k) \approx y_k$). ¿Cómo podemos construir un “conjunto finito de puntos” que “verifiquen aproximadamente una ecuación diferencial”? En primer lugar, elegimos las abscisas de los puntos. Por comodidad, dividimos el intervalo $[a, b]$ en M subintervalos del mismo tamaño usando la partición dada por los siguientes puntos:

$$(1) \quad t_k = a + kh \quad \text{para } k = 0, 1, \dots, M, \text{ siendo } h = \frac{b - a}{M}.$$

El valor del incremento h se llama **tamaño de paso**. Procedemos ahora a resolver aproximadamente

$$(2) \quad y' = f(t, y) \quad \text{en} \quad [t_0, t_M] \quad \text{con} \quad y(t_0) = y_0.$$

Suponiendo que $y(t)$, $y'(t)$ e $y''(t)$ son continuas y usando el teorema de Taylor para desarrollar $y(t)$ alrededor de $t = t_0$, para cada punto t existe un punto c_1 entre t_0 y t tal que

$$(3) \quad y(t) = y(t_0) + y'(t_0)(t - t_0) + \frac{y''(c_1)(t - t_0)^2}{2}.$$

Al sustituir $y'(t_0) = f(t_0, y(t_0))$ y $h = t_1 - t_0$ en la ecuación (3), el resultado es una expresión para el valor $y(t_1)$:

$$(4) \quad y(t_1) = y(t_0) + hf(t_0, y(t_0)) + y''(c_1) \frac{h^2}{2}.$$

Si el tamaño de paso h es suficientemente pequeño, entonces podemos despreciar el término que contiene h^2 y obtener

$$(5) \quad y(t_1) \approx y_1 = y_0 + hf(t_0, y_0),$$

que se llama **aproximación de Euler**.

Repetiendo el proceso generamos una sucesión de puntos que se aproximan a la gráfica de la solución $y = y(t)$. El paso general del método de Euler es

$$(6) \quad t_{k+1} = t_k + h, \quad y_{k+1} = y_k + hf(t_k, y_k) \quad \text{para } k = 0, 1, \dots, M - 1.$$

Ejemplo 9.2. Vamos a usar el método de Euler para hallar una solución aproximada del problema de valor inicial

$$(7) \quad y' = Ry \quad \text{en } [0, 1] \text{ con } y(0) = y_0 \text{ y } R \text{ constante.}$$

Debemos elegir un tamaño de paso y , luego, usar la segunda fórmula de (6) para calcular las ordenadas. A veces, esta segunda fórmula se llama ecuación en diferencias y, en nuestro caso, es

$$(8) \quad y_{k+1} = y_k(1 + hR) \quad \text{para } k = 0, 1, \dots, M - 1.$$

Si vamos escribiendo estos valores recursivamente, obtenemos

$$(9) \quad \begin{aligned} y_1 &= y_0(1 + hR) \\ y_2 &= y_1(1 + hR) = y_0(1 + hR)^2 \\ &\vdots \\ y_M &= y_{M-1}(1 + hR) = y_0(1 + hR)^M. \end{aligned}$$

En la mayoría de los problemas no se puede hallar una fórmula explícita para determinar las aproximaciones, de manera que cada punto debe calcularse sucesivamente a partir del punto anterior. Sin embargo, con el problema de valor inicial (7) hemos tenido suerte; el método de Euler proporciona una solución explícita:

$$(10) \quad t_k = kh \quad y_k = y_0(1 + hR)^k \quad \text{para } k = 0, 1, \dots, M.$$

La fórmula (10) puede entenderse como una fórmula para calcular el interés compuesto y la aproximación de Euler proporciona el capital acumulado a partir de un depósito inicial. ■

Ejemplo 9.3. Supongamos que se depositan 1000 euros durante cinco años a un interés compuesto continuamente del 10%. ¿Cuál es el capital acumulado al cabo de esos cinco años?

Vamos a utilizar las aproximaciones de Euler con tamaños de paso $h = 1, \frac{1}{12}$ y $\frac{1}{360}$ para aproximar $y(5)$ en el problema de valor inicial

$$y' = 0.1y \quad \text{en } [0, 5] \text{ con } y(0) = 1000.$$

Los resultados de aplicar la fórmula (10) con $R = 0.1$ se recogen en la Tabla 9.1. ■

Tabla 9.1 El interés compuesto en el Ejemplo 9.3.

Tamaño de paso, h	Número de iteraciones, M	Aproximación y_M a $y(5)$
1	5	$1000 \left(1 + \frac{0.1}{1}\right)^5 = 1610.51$
$\frac{1}{12}$	60	$1000 \left(1 + \frac{0.1}{12}\right)^{60} = 1645.31$
$\frac{1}{360}$	1800	$1000 \left(1 + \frac{0.1}{360}\right)^{1800} = 1648.61$

Pensemos en los valores y_5 , y_{60} e y_{1800} que aproximan el capital acumulado a los cinco años. Estos valores se han obtenido usando tamaños de paso diferentes y realizando un esfuerzo computacional también diferente. La solución del problema de valor inicial es $y(5) = 1000e^{0.5} = 1648.72$. Si no usáramos la fórmula explícita (10), entonces hubieramos necesitado realizar 1800 iteraciones en el método de Euler para obtener y_{1800} y, aun así, ¡sólo obtenemos cinco cifras significativas de precisión en nuestra respuesta!

Cuando se tiene que aproximar la solución del problema de valor inicial (7) en un banco, se utiliza el método de Euler porque se dispone de la fórmula explícita (10). Otros métodos de aproximación a la solución más sofisticados no tienen una fórmula explícita para calcular y_k , pero necesitan un esfuerzo computacional menor.

Descripción geométrica

Si partimos del punto (t_0, y_0) , calculamos el valor de la pendiente $m_0 = f(t_0, y_0)$, nos movemos horizontalmente una distancia h y verticalmente una distancia $hf(t_0, y_0)$, entonces lo que hacemos es desplazarnos a lo largo de la recta tangente a la curva $y(t)$ terminando en el punto (t_1, y_1) (véase la Figura 9.5). Hagamos notar que (t_1, y_1) no es un punto de la curva deseada!, aunque sea la aproximación que se genera. Ahora debemos usar (t_1, y_1) , como si fuera un punto correcto, para calcular la pendiente $m_1 = f(t_1, y_1)$ y usar este valor para obtener el siguiente desplazamiento vertical $hf(t_1, y_1)$, que nos lleva al punto (t_2, y_2) , y así sucesivamente.

Tamaño de paso frente a error

Los métodos de aproximación a la solución de un problema de valor inicial que vamos a presentar se llaman **métodos de diferencia** o **métodos de variable**

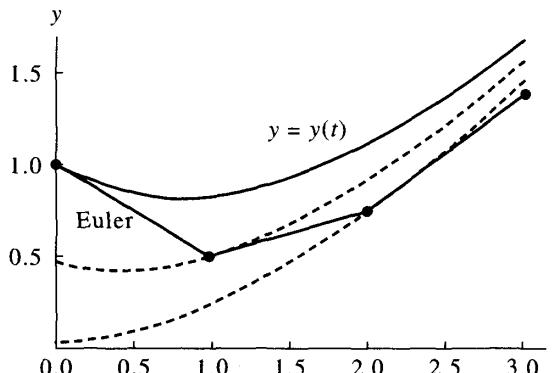


Figura 9.5 Aproximaciones de Euler $y_{k+1} = y_k + h f(t_k, y_k)$.

discreta. En este tipo de métodos, la solución se aproxima en un conjunto finito de puntos que llamaremos nodos. Un método elemental de la forma $y_{k+1} = y_k + h\Phi(t_k, y_k)$, para cierta función Φ llamada **función incremental**, se dice que es de paso simple, o de un solo paso, porque en el cálculo del nuevo punto sólo interviene de manera explícita el punto inmediatamente anterior.

Cuando usamos un método de variable discreta para resolver de manera aproximada un problema de valor inicial, existen dos fuentes de error: la discretización y el redondeo.

Definición 9.3 (Error de discretización). Supongamos que $\{(t_k, y_k)\}_{k=0}^M$ es un conjunto finito de aproximaciones a la única solución $y = y(t)$ de un problema de valor inicial.

El **error de truncamiento global** o **error de discretización global** e_k se define como

$$(11) \quad e_k = y(t_k) - y_k \quad \text{para } k = 0, 1, \dots, M.$$

Este error es la diferencia entre la solución exacta y la calculada con el método en el nodo correspondiente.

El **error de truncamiento local** o **error de discretización local** ε_{k+1} se define como

$$(12) \quad \varepsilon_{k+1} = y(t_{k+1}) - y_k - h\Phi(t_k, y_k) \quad \text{para } k = 0, 1, \dots, M-1.$$

Este error es el que se comete en un solo paso, el que nos lleva desde el nodo t_k hasta el nodo t_{k+1} . ▲

Cuando se obtiene la ecuación (6) en la presentación del método de Euler, el término que se desprecia en cada paso es $y''(c_k)(h^2/2)$. Si este fuera el único

error que se comete en cada paso, entonces en el otro extremo del intervalo, una vez dados los M pasos, el error acumulado sería

$$\sum_{k=1}^M y''(c_k) \frac{h^2}{2} \approx My''(c) \frac{h^2}{2} = \frac{hM}{2} y''(c)h = \frac{(b-a)y''(c)}{2} h = O(h^1).$$

Podría haber otros errores, pero esta estimación es la que predomina. Una discusión detallada de este tema puede encontrarse en textos avanzados de métodos numéricos para ecuaciones diferenciales (Referencia [75]).

Teorema 9.3 (Precisión del método de Euler). Sea $y(t)$ la solución del problema de valor inicial (2). Si $y(t) \in C^2[t_0, b]$ y $\{(t_k, y_k)\}_{k=0}^M$ es la sucesión de aproximaciones generada por el método de Euler, entonces

$$(13) \quad \begin{aligned} |e_k| &= |y(t_k) - y_k| = O(h), \\ |\varepsilon_{k+1}| &= |y(t_{k+1}) - y_k - hf(t_k, y_k)| = O(h^2). \end{aligned}$$

El error al final del intervalo se llama **error global final** y viene dado por

$$(14) \quad E(y(b), h) = |y(b) - y_M| = O(h).$$

Observación. El error global final $E(y(b), h)$ se usa para estudiar el comportamiento del error para tamaños de paso diferentes y nos permite tener una idea del esfuerzo computacional que hay que realizar para obtener aproximaciones con la precisión deseada.

Los Ejemplos 9.4 y 9.5 ilustran los conceptos que aparecen en el Teorema 9.3. Si calculásemos las aproximaciones usando como tamaños de paso h y $h/2$, deberíamos tener

$$(15) \quad E(y(b), h) \approx Ch$$

para el tamaño de paso más grande y

$$(16) \quad E\left(y(b), \frac{h}{2}\right) \approx C \frac{h}{2} = \frac{1}{2} Ch \approx \frac{1}{2} E(y(b), h).$$

Por tanto, el Teorema 9.3 nos dice que si reducimos a la mitad el tamaño de paso en el método de Euler, entonces cabe esperar que el error final global se reduzca también a la mitad.

Ejemplo 9.4. Vamos a usar el método de Euler para resolver el problema de valor inicial

$$y' = \frac{t-y}{2} \quad \text{en } [0, 3] \text{ con } y(0) = 1$$

y a comparar las soluciones que se obtienen con $h = 1, \frac{1}{2}, \frac{1}{4}$ y $\frac{1}{8}$.

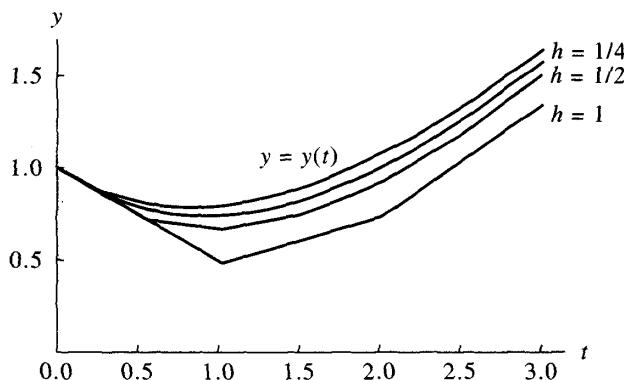


Figura 9.6 Comparación de las aproximaciones de Euler obtenidas con tamaños de paso diferentes para $y' = (t - y)/2$ en $[0, 3]$ con la condición $y(0) = 1$.

La Figura 9.6 muestra las gráficas de las cuatro soluciones obtenidas por el método de Euler y la gráfica de la solución exacta $y(t) = 3e^{-t/2} - 2 + t$. En la Tabla 9.2 se muestran los valores de las cuatro aproximaciones en algunos nodos. Para el tamaño de paso $h = 0.25$, los cálculos son

$$\begin{aligned}y_1 &= 1.0 + 0.25 \left(\frac{0.0 - 1.0}{2} \right) = 0.875, \\y_2 &= 0.875 + 0.25 \left(\frac{0.25 - 0.875}{2} \right) = 0.796875, \quad \text{etc.}\end{aligned}$$

La iteración continúa hasta que llegamos al otro extremo del intervalo

$$y(3) \approx y_{12} = 1.440573 + 0.25 \left(\frac{2.75 - 1.440573}{2} \right) = 1.604252.$$

Ejemplo 9.5. Vamos a comparar los errores globales finales cuando se usa el método de Euler para resolver el problema de valor inicial

$$y' = \frac{t - y}{2} \quad \text{en } [0, 3] \text{ con } y(0) = 1,$$

usando como tamaños de paso $1, \frac{1}{2}, \dots, \frac{1}{64}$.

La Tabla 9.3 muestra los errores globales finales para los diferentes tamaños de paso; en ella vemos que el error en la aproximación a $y(3)$ decrece en un factor de $\frac{1}{2}$ cuando el tamaño de paso se reduce a la mitad. Para los tamaños de paso más pequeños, también podemos comprobar la conclusión del Teorema 9.3

$$E(y(3), h) = y(3) - y_M = O(h^1) \approx Ch, \quad \text{siendo } C = 0.256. \quad \blacksquare$$

Tabla 9.2 Comparación de las aproximaciones de Euler obtenidas con tamaños de paso diferentes para $y' = (t - y)/2$ en $[0, 3]$ con $y(0) = 1$.

t_k	y_k				$y(t_k)$ Exacto
	$h = 1$	$h = \frac{1}{2}$	$h = \frac{1}{4}$	$h = \frac{1}{8}$	
0	1.0	1.0	1.0	1.0	1.0
0.125			0.9375	0.943239	
0.25			0.886719	0.897491	
0.375			0.846924	0.862087	
0.50		0.75	0.796875	0.817429	0.836402
0.75			0.759766	0.786802	0.811868
1.00	0.5	0.6875	0.758545	0.790158	0.819592
1.50		0.765625	0.846386	0.882855	0.917100
2.00	0.75	0.949219	1.030827	1.068222	1.103638
2.50		1.211914	1.289227	1.325176	1.359514
3.00	1.375	1.533936	1.604252	1.637429	1.669390

Tabla 9.3 Relación entre el tamaño de paso y el error global final para las aproximaciones de Euler a la solución de $y' = (t - y)/2$ en $[0, 3]$ con $y(0) = 1$.

Tamaño de paso, h	Número de pasos, M	Aproximación y_M a $y(3)$	Error global final, $y(3) - y_M$	$O(h) \approx Ch$ con $C = 0.256$
1	3	1.375	0.294390	0.256
$\frac{1}{2}$	6	1.533936	0.135454	0.128
$\frac{1}{4}$	12	1.604252	0.065138	0.064
$\frac{1}{8}$	24	1.637429	0.031961	0.032
$\frac{1}{16}$	48	1.653557	0.015833	0.016
$\frac{1}{32}$	96	1.661510	0.007880	0.008
$\frac{1}{64}$	192	1.665459	0.003931	0.004

MATLAB

Programa 9.1 (Método de Euler). Construcción de las aproximaciones a la solución del problema inicial $y' = f(t, y)$ con $y(a) = y_0$ en $[a, b]$ dadas por

$$y_{k+1} = y_k + h f(t_k, y_k) \quad \text{para } k = 0, 1, \dots, M - 1.$$

```
function E=euler(f,a,b,ya,M)
% Datos
%     - f es la función, almacenada
%         como una cadena de caracteres 'f'
%     - a y b son los extremos derecho e izquierdo
%         del intervalo
%     - ya es la condición inicial y(a)
%     - M es el número de pasos
% Resultado
%     - E=[T' Y'] siendo T el vector de las abscisas e
%         Y el vector de las ordenadas

h=(b-a)/M;
T=zeros(1,M+1);
Y=zeros(1,M+1);
T=a:h:b;
Y(1)=ya;
for j=1:M
    Y(j+1)=Y(j)+h*feval(f,T(j),Y(j));
end
E=[T' Y'];
```

Ejercicios

En los Ejercicios 1 a 5 resuelva la ecuación diferencial usando el método de Euler.

- (a) Tome $h = 0.2$ y dé dos pasos calculando los valores a mano. Luego tome $h = 0.1$ y dé cuatro pasos calculando los valores a mano.
- (b) Compare la solución exacta $y(0.4)$ con las dos aproximaciones calculadas en el apartado (a).
- (c) ¿Se comporta el error global final de las aproximaciones obtenidas en el apartado (a) como se espera cuando h se divide entre dos?

1. $y' = t^2 - y$ con $y(0) = 1$, $y(t) = -e^{-t} + t^2 - 2t + 2$

2. $y' = 3y + 3t$ con $y(0) = 1$, $y(t) = \frac{4}{3}e^{3t} - t - \frac{1}{3}$

3. $y' = -ty$ con $y(0) = 1$, $y(t) = e^{-t^2/2}$
4. $y' = e^{-2t} - 2y$ con $y(0) = \frac{1}{10}$, $y(t) = \frac{1}{10}e^{-2t} + te^{-2t}$
5. $y' = 2ty^2$ con $y(0) = 1$, $y(t) = 1/(1-t^2)$
6. Crecimiento logístico de una población. Se supone que la curva de población $P(t)$ para los Estados Unidos de América en el siglo XX obedece la ecuación diferencial logística $P' = aP - bP^2$ siendo $a = 0.02$ y $b = 0.00004$. Denotemos por t los años transcurridos desde 1900 y tomemos como tamaño de paso $h = 10$. Determine, haciendo las operaciones a mano o con una calculadora, las aproximaciones de Euler a $P(t)$ (redondeando cada valor P_k a la décima) y rellene la siguiente tabla

Año	t_k	$P(t_k)$ real	P_k , aproximación de Euler
1900	0.0	76.1	76.1
1910	10.0	92.4	89.0
1920	20.0	106.5	_____
1930	30.0	123.1	_____
1940	40.0	132.6	138.2
1950	50.0	152.3	_____
1960	60.0	180.7	_____
1970	70.0	204.9	202.8
1980	80.0	226.5	_____

7. Pruebe que si se utiliza el método de Euler para resolver el problema de valor inicial

$$y' = f(t) \quad \text{en } [a, b] \quad \text{con } y(a) = y_0 = 0$$

el resultado es

$$y(b) \approx \sum_{k=0}^{M-1} f(t_k)h,$$

que es la suma de Riemann que aproxima la integral definida de $f(t)$ en el intervalo $[a, b]$.

8. Pruebe que el método de Euler falla cuando queremos aproximar la solución $y(t) = t^{3/2}$ del problema de valor inicial

$$y' = f(t, y) = 1.5y^{1/3} \quad \text{con } y(0) = 0.$$

Justifique su respuesta. ¿Cuál es el problema?

9. ¿Puede usarse el método de Euler para resolver el problema de valor inicial

$$y' = 1 + y^2 \quad \text{en } [0, 3] \quad \text{con } y(0) = 0?$$

Indicación. La solución exacta es $y(t) = \tan(t)$.

Algoritmos y programas

En los Problemas 1 a 5, resuelva la ecuación diferencial usando el método de Euler.

- Tome $h = 0.1$ y dé 20 pasos con el Programa 9.1. Luego tome $h = 0.05$ y dé 40 pasos con el Programa 9.1.
 - Compare la solución exacta $y(2)$ con las dos aproximaciones obtenidas en el apartado (a).
 - ¿Se comporta el error global final de las aproximaciones obtenidas en el apartado (a) como se espera cuando h se divide entre dos?
 - Dibuje las aproximaciones y la solución exacta en una misma gráfica. *Indicación.* La matriz E que se obtiene como resultado en el Programa 9.1 contiene las coordenadas x e y de las aproximaciones y la instrucción del paquete MATLAB `plot(E(:,1),E(:,2))` producirá una dibujo análogo al de la Figura 9.6.
- $y' = t^2 - y$ con $y(0) = 1$, $y(t) = -e^{-t} + t^2 - 2t + 2$
 - $y' = 3y + 3t$ con $y(0) = 1$, $y(t) = \frac{4}{3}e^{3t} - t - \frac{1}{3}$
 - $y' = -ty$ con $y(0) = 1$, $y(t) = e^{-t^2/2}$
 - $y' = e^{-2t} - 2y$ con $y(0) = \frac{1}{10}$, $y(t) = \frac{1}{10}e^{-2t} + te^{-2t}$
 - $y' = 2ty^2$ con $y(0) = 1$, $y(t) = 1/(1 - t^2)$
 - Considere el problema de valor inicial $y' = 0.12y$ en $[0, 5]$ con $y(0) = 1000$.
 - Aplique la fórmula (10) para hallar la aproximación de Euler a $y(5)$ tomando como tamaños de paso $h = 1$, $\frac{1}{12}$ y $\frac{1}{360}$.
 - ¿Cuál es límite en el apartado (a) cuando h tiende a cero?
 - Crecimiento exponencial de una población.* La población de ciertas especies crece a una velocidad que es proporcional a la población presente y que responde a un problema de valor inicial como el siguiente

$$y' = 0.02y \quad \text{en } [0, 5] \quad \text{con } y(0) = 5000.$$

- Aplique la fórmula (10) para calcular la aproximación de Euler a $y(5)$ usando los tamaños de paso $h = 1$, $\frac{1}{12}$ y $\frac{1}{360}$.
- ¿Cuál es límite en el apartado (a) cuando h tiende a cero?
- Un paracaidista salta desde un avión. Hasta el momento en que abre el paracaídas, la resistencia del aire es proporcional a $v^{3/2}$ (v representa la velocidad). Supongamos que el intervalo temporal es $[0, 6]$ y que la ecuación diferencial para la velocidad de descenso es

$$v' = 10 - 0.01v^{3/2} \quad \text{en } [0, 6] \quad \text{con } v(0) = 0.$$

Use el método de Euler con $h = 0.05$ para estimar $v(6)$.

- 9. Modelo de una epidemia.** Vamos a describir a continuación un modelo matemático para la extensión de una epidemia. Supongamos que tenemos una comunidad de L personas que contiene inicialmente P personas contagiadas y Q sin contagiar. Sea $y(t)$ el número de personas contagiadas en un instante t . Si la enfermedad no es muy grave, como el resfriado común, todo el mundo continúa en activo y la epidemia se extiende. Puesto que hay PQ posibles contactos entre personas de uno y otro grupo, la velocidad de cambio de $y(t)$ es proporcional a PQ , así que el problema puede modelarse mediante el problema de valor inicial

$$y' = ky(L - y) \quad \text{con} \quad y(0) = y_0.$$

- (a) Tomando $L = 25\,000$, $k = 0.00003$ y $h = 0.2$ con la condición inicial $y(0) = 250$, use el Programa 9.1 para calcular la aproximación de Euler en el intervalo $[0, 60]$.
- (b) Dibuje la gráfica de la solución aproximada del apartado (a).
- (c) Estime el número medio de personas contagiadas calculando la media aritmética de las ordenadas obtenidas en el apartado (a) con el método de Euler.
- (d) Estime el número medio de personas contagiadas ajustando una curva a los datos del apartado (a) y usando el Teorema 1.10 (teorema del valor medio para integrales).
10. Consideremos la ecuación íntegro-diferencial de primer orden

$$y' = 1.3y - 0.25y^2 - 0.0001y \int_0^t y(\tau) d\tau.$$

- (a) Use el método de Euler con $h = 0.2$ e $y(0) = 250$ en el intervalo $[0, 20]$ (aproxime las integrales mediante la regla del trapecio). *Indicación.* El paso general (6) para el método de Euler es

$$y_{k+1} = y_k + h \left(1.3y_k - 0.25y_k^2 - 0.0001y_k \int_0^{t_k} y(\tau) d\tau \right).$$

Si usamos la regla del trapecio para aproximar la integral, entonces esta expresión se transforma en

$$y_{k+1} = y_k + h \left(1.3y_k - 0.25y_k^2 - 0.0001y_k T_k(h) \right),$$

siendo $T_0(h) = 0$ y

$$T_k(h) = T_{k-1}(h) + \frac{h}{2} (y_{k-1} + y_k) \quad \text{para } k = 0, 1, \dots, 99.$$

- (b) Repita el apartado (a) con los valores iniciales $y(0) = 200$ e $y(0) = 300$.
- (c) Dibuje las soluciones aproximadas obtenidas en los apartados (a) y (b) sobre un mismo gráfico.

9.3 El método de Heun

La siguiente técnica que presentamos, el método de Heun, introduce una idea nueva en la construcción de un algoritmo para resolver el problema de valor inicial

$$(1) \quad y'(t) = f(t, y(t)) \quad \text{en} \quad [a, b] \quad \text{con} \quad y(t_0) = y_0.$$

Para obtener el punto (t_1, y_1) , podemos usar el teorema fundamental del cálculo e integrar $y'(t)$ en $[t_0, t_1]$ de manera que

$$(2) \quad \int_{t_0}^{t_1} f(t, y(t)) dt = \int_{t_0}^{t_1} y'(t) dt = y(t_1) - y(t_0),$$

donde hemos usado como primitiva de $y'(t)$ la función deseada $y(t)$. Despejando $y(t_1)$ en la igualdad (2), nos queda

$$(3) \quad y(t_1) = y(t_0) + \int_{t_0}^{t_1} f(t, y(t)) dt.$$

Ahora podríamos usar un método de integración para aproximar la integral definida en la expresión (3). Si usamos la regla del trapecio con incremento $h = t_1 - t_0$, entonces el resultado es

$$(4) \quad y(t_1) \approx y(t_0) + \frac{h}{2}(f(t_0, y(t_0)) + f(t_1, y(t_1))).$$

Hagamos notar que en la fórmula del miembro derecho de (4) aparece el valor $y(t_1)$ que queremos determinar; lo que hacemos es usar una estimación de $y(t_1)$ y, para nuestro propósito, la aproximación de Euler es suficiente. Al sustituir ésta en (4), obtenemos una fórmula de aproximación a $y(t_1)$ que se llama **método de Heun**:

$$(5) \quad y_1 = y(t_0) + \frac{h}{2}(f(t_0, y_0) + f(t_1, y_0 + hf(t_0, y_0))).$$

Repetiendo el proceso se genera una sucesión de puntos que aproximan la solución $y = y(t)$. En cada paso, la aproximación dada por el método de Euler se usa como una predicción del valor que queremos calcular y luego la regla del trapecio se usa para hacer una corrección y obtener el valor definitivo. El paso general del método de Heun es

$$(6) \quad \begin{aligned} p_{k+1} &= y_k + hf(t_k, y_k), & t_{k+1} &= t_k + h, \\ y_{k+1} &= y_k + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, p_{k+1})). \end{aligned}$$

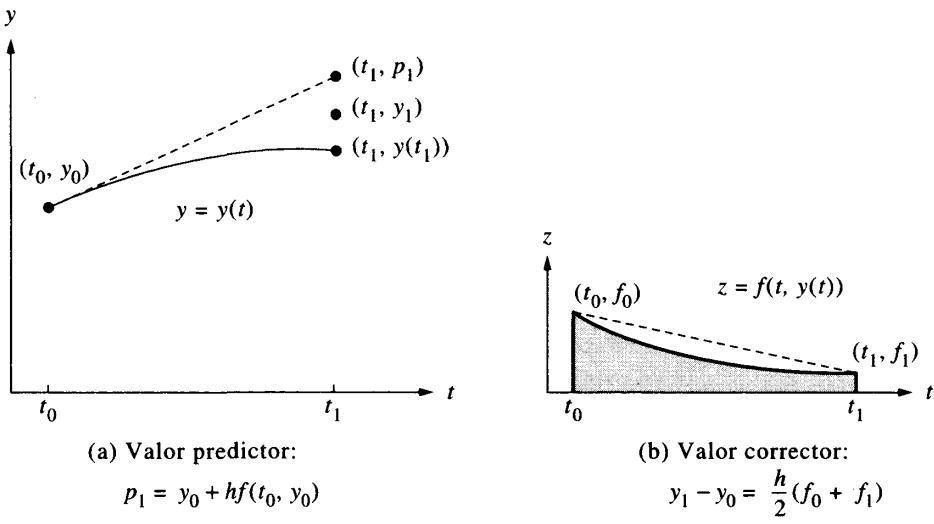


Figura 9.7 Las gráficas $y = y(t)$ y $z = f(t, y(t))$ que se usan en la construcción del método de Heun.

Veamos el papel que juegan la derivación y la integración en el método de Heun: Dibujamos la recta tangente a la gráfica de la solución $y = y(t)$ en el punto (t_0, y_0) y la usamos para construir el punto (t_1, p_1) . Ahora miramos la gráfica de $z = f(t, y(t))$ y consideraremos los puntos (t_0, f_0) y (t_1, f_1) , donde $f_0 = f(t_0, y_0)$ y $f_1 = f(t_1, p_1)$. El área del trapecio con vértices (t_0, f_0) y (t_1, f_1) se usa como aproximación de la integral que aparece en (3) que, a su vez, se usa para obtener el valor final dado por la expresión (5). Las gráficas correspondientes se muestran en la Figura 9.7.

Tamaño de paso frente a error

El término del error de la regla del trapecio que hemos usado para aproximar la integral de la expresión (3) es

$$(7) \quad -y^{(2)}(c_k) \frac{h^3}{12}.$$

Si el único error que se cometiera en cada paso fuera el que se da en la expresión (7), entonces después de M pasos del método de Heun el error acumulado sería

$$(8) \quad -\sum_{k=1}^M y^{(2)}(c_k) \frac{h^3}{12} \approx \frac{b-a}{12} y^{(2)}(c) h^2 = O(h^2).$$

El siguiente teorema es importante porque establece la relación entre el error global final y el tamaño de paso y puede usarse para darnos una idea del esfuerzo computacional requerido por el método de Heun si queremos obtener la solución con una cierta precisión fijada de antemano.

Teorema 9.4 (Precisión del método de Heun). Supongamos que $y(t)$ es una solución del problema de valor inicial (1). Si $y(t) \in C^3[t_0, b]$ y $\{(t_k, y_k)\}_{k=0}^M$ es la sucesión de aproximaciones dadas por el método de Heun, entonces

$$(9) \quad \begin{aligned} |e_k| &= |y(t_k) - y_k| = O(h^2), \\ |\varepsilon_{k+1}| &= |y(t_{k+1}) - y_k - h\Phi(t_k, y_k)| = O(h^3), \end{aligned}$$

donde $\Phi(t_k, y_k) = y_k + (h/2)(f(t_k, y_k) + f(t_{k+1}, y_k + hf(t_k, y_k)))$.

En particular, el error global final en el extremo derecho del intervalo verifica

$$(10) \quad E(y(b), h) = |y(b) - y_M| = O(h^2).$$

Los Ejemplos 9.6 y 9.7 ilustran el Teorema 9.4: Si calculásemos las aproximaciones usando como tamaños de paso h y $h/2$, entonces deberíamos tener

$$(11) \quad E(y(b), h) \approx Ch^2$$

para el tamaño de paso más grande y

$$(12) \quad E\left(y(b), \frac{h}{2}\right) \approx C \frac{h^2}{4} = \frac{1}{4} Ch^2 \approx \frac{1}{4} E(y(b), h).$$

En consecuencia, el Teorema 9.4 nos dice que si el tamaño de paso en el método de Heun se reduce a la mitad, entonces cabe esperar que el error global final se reduzca a su cuarta parte.

Ejemplo 9.6. Vamos a usar el método de Heun para resolver el problema

$$y' = \frac{t-y}{2} \quad \text{en } [0, 3] \text{ con } y(0) = 1$$

y a comparar las soluciones obtenidas con $h = 1, \frac{1}{2}, \frac{1}{4}$ y $\frac{1}{8}$.

En la Figura 9.8 se muestran las gráficas de las dos primeras soluciones dadas por el método de Heun y la gráfica de la solución exacta $y(t) = 3e^{-t/2} - 2 + t$. En la Tabla 9.4 se recogen los valores de las cuatro soluciones en algunos nodos. Un cálculo típico, que hacemos con el tamaño de paso $h = 0.25$, sería

$$f(t_0, y_0) = \frac{0-1}{2} = -0.5, \quad p_1 = 1.0 + 0.25(-0.5) = 0.875,$$

$$f(t_1, p_1) = \frac{0.25-0.875}{2} = -0.3125,$$

$$y_1 = 1.0 + 0.125(-0.5 - 0.3125) = 0.8984375.$$

La iteración continúa hasta que llegamos al último paso

$$y(3) \approx y_{12} = 1.511508 + 0.125(0.619246 + 0.666840) = 1.672269.$$

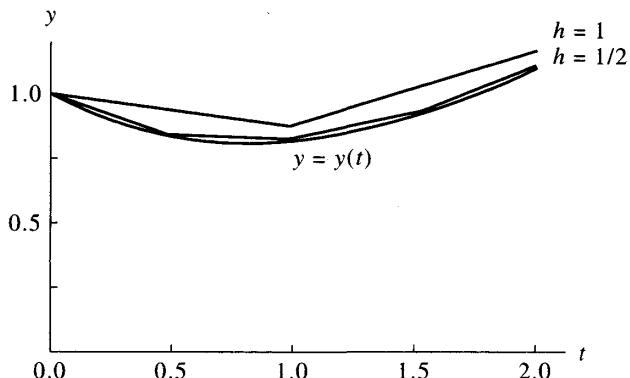


Figura 9.8 Comparación de las soluciones obtenidas con el método de Heun con diferentes tamaños de paso para $y' = (t - y)/2$ en $[0, 2]$ con la condición inicial $y(0) = 1$.

Tabla 9.4 Comparación de las soluciones obtenidas con el método de Heun con diferentes tamaños de paso para $y' = (t - y)/2$ en $[0, 2]$ con la condición $y(0) = 1$.

t_k	y_k				$y(t_k)$ Exacto
	$h = 1$	$h = \frac{1}{2}$	$h = \frac{1}{4}$	$h = \frac{1}{8}$	
0	1.0	1.0	1.0	1.0	1.0
0.125				0.943359	0.943239
0.25			0.898438	0.897717	0.897491
0.375				0.862406	0.862087
0.50		0.84375	0.838074	0.836801	0.836402
0.75			0.814081	0.812395	0.811868
1.00	0.875	0.831055	0.822196	0.820213	0.819592
1.50		0.930511	0.920143	0.917825	0.917100
2.00	1.171875	1.117587	1.106800	1.104392	1.103638
2.50		1.373115	1.362593	1.360248	1.359514
3.00	1.732422	1.682121	1.672269	1.670076	1.669390

Ejemplo 9.7. Vamos a comparar los errores globales finales cuando usamos el método de Heun para resolver el problema de valor inicial

$$y' = \frac{t - y}{2} \quad \text{en } [0, 3] \text{ con } y(0) = 1,$$

usando los tamaños de paso $1, \frac{1}{2}, \dots, \frac{1}{64}$.

En la Tabla 9.5 se recogen los errores globales finales y en ella podemos observar que el error de la aproximación a $y(3)$ decrece a su cuarta parte $\frac{1}{4}$ cuando reducimos

Tabla 9.5 Relación entre el tamaño de paso y el error global final de las soluciones obtenidas con el método de Heun para $y' = (t - y)/2$ en $[0, 2]$ con la condición inicial $y(0) = 1$.

Tamaño de paso, h	Número de pasos, M	Aproximación y_M a $y(3)$	Error global final, $y(3) - y_M$	$O(h^2) \approx Ch^2$ con $C = -0.0432$
1	3	1.732422	-0.063032	-0.043200
$\frac{1}{2}$	6	1.682121	-0.012731	-0.010800
$\frac{1}{4}$	12	1.672269	-0.002879	-0.002700
$\frac{1}{8}$	24	1.670076	-0.000686	-0.000675
$\frac{1}{16}$	48	1.669558	-0.000168	-0.000169
$\frac{1}{32}$	96	1.669432	-0.000042	-0.000042
$\frac{1}{64}$	192	1.669401	-0.000011	-0.000011

el tamaño de paso a la mitad $\frac{1}{2}$:

$$E(y(3), h) = y(3) - y_M = O(h^2) \approx Ch^2, \quad \text{con } C = -0.0432.$$

MATLAB

Programa 9.2 (Método de Heun). Construcción de las aproximaciones a la solución del problema inicial $y' = f(t, y)$ con $y(a) = y_0$ en $[a, b]$ dadas por

$$y_{k+1} = y_k + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, y_k + f(t_k, y_k)))$$

para $k = 0, 1, \dots, M - 1$.

```
function H=heun(f,a,b,ya,M)
% Datos
%   - f es la función, almacenada como
%     una cadena de caracteres 'f'
%   - a y b son los extremos derecho e izquierdo
%     del intervalo
```

```

% - ya es la condición inicial y(a)
% - M es el número de pasos
% Resultado
% - H=[T' Y'] siendo T el vector de las abscisas e
%   Y el vector de las ordenadas
h=(b-a)/M;
T=zeros(1,M+1);
Y=zeros(1,M+1);
T=a:h:b;
Y(1)=ya;
for j=1:M
    k1=feval(f,T(j),Y(j));
    k2=feval(f,T(j+1),Y(j)+h*k1);
    Y(j+1)=Y(j)+(h/2)*(k1+k2);
end
H=[T'Y'];

```

Ejercicios

En los Ejercicios 1 a 5 resuelva la ecuación diferencial usando el método de Heun.

- (a) Tome $h = 0.2$ y dé dos pasos calculando los valores a mano. Luego tome $h = 0.1$ y dé cuatro pasos calculando los valores a mano.
 - (b) Compare la solución exacta $y(0.4)$ con las dos aproximaciones calculadas en el apartado (a).
 - (c) ¿Se comporta el error global final de las aproximaciones obtenidas en el apartado (a) como se espera cuando h se divide entre dos?
1. $y' = t^2 - y$ con $y(0) = 1$, $y(t) = -e^{-t} + t^2 - 2t + 2$
 2. $y' = 3y + 3t$ con $y(0) = 1$, $y(t) = \frac{4}{3}e^{3t} - t - \frac{1}{3}$
 3. $y' = -ty$ con $y(0) = 1$, $y(t) = e^{-t^2/2}$
 4. $y' = e^{-2t} - 2y$ con $y(0) = \frac{1}{10}$, $y(t) = \frac{1}{10}e^{-2t} + te^{-2t}$
 5. $y' = 2ty^2$ con $y(0) = 1$, $y(t) = 1/(1-t^2)$
Note que el método de Heun genera una aproximación de $y(1)$ aunque la solución no está definida en $t = 1$.
 6. Pruebe que cuando se usa el método de Heun para resolver el problema de valor inicial $y' = f(t)$ en $[a, b]$ con $y(a) = y_0 = 0$ el resultado es

$$y(b) = \frac{h}{2} \sum_{k=0}^{M-1} (f(t_k) + f(t_{k+1})),$$

que es la aproximación dada por la regla compuesta del trapecio para aproximar la integral definida de $f(t)$ en el intervalo $[a, b]$.

7. El método de Richardson para mejorar las aproximaciones, que ya presentamos en el Lema 7.1 (Sección 7.3), puede usarse en conjunción con el método de Heun. Si usamos el método de Heun con tamaño de paso h , entonces tenemos

$$y(b) \approx y_h + Ch^2,$$

si ahora lo usamos con tamaño de paso $2h$, entonces tenemos

$$y(b) \approx y_{2h} + 4Ch^2.$$

Los términos que contienen Ch^2 pueden eliminarse para obtener una aproximación mejorada a $y(b)$:

$$y(b) \approx \frac{4y_h - y_{2h}}{3}.$$

Este esquema de mejora puede usarse con los valores mostrados en el Ejemplo 9.7 para obtener una aproximación mejor a $y(3)$. Calcule las entradas que faltan en la tabla siguiente:

h	y_h	$(4y_h - y_{2h})/3$
1	1.732422	_____
1/2	1.682121	1.665354
1/4	1.672269	_____
1/8	1.670076	_____
1/16	1.669558	1.669385
1/32	1.669432	_____
1/64	1.669401	_____

8. Pruebe que el método de Heun falla cuando queremos aproximar la solución $y(t) = t^{3/2}$ del problema de valor inicial

$$y' = f(t, y) = 1.5y^{1/3} \quad \text{con} \quad y(0) = 0.$$

Justifique su respuesta. ¿Cuál es el problema?

Algoritmos y programas

En los Problemas 1 a 5, resuelva la ecuación diferencial usando el método de Heun.

- (a) Tome $h = 0.1$ y dé 20 pasos con el Programa 9.2. Luego tome $h = 0.05$ y dé 40 pasos con el Programa 9.2.

- (b) Compare la solución exacta $y(2)$ con las dos aproximaciones obtenidas en el apartado (a).
- (c) ¿Se comporta el error global final de las aproximaciones obtenidas en el apartado (a) como se espera cuando h se divide entre dos?
- (d) Dibuje las aproximaciones y la solución exacta en una misma gráfica. *Indicación.* La matriz H que se obtiene como resultado en el Programa 9.2 contiene las coordenadas x e y de las aproximaciones y la instrucción del paquete MATLAB `plot(H(:,1),H(:,2))` producirá una dibujo análogo al de la Figura 9.8.
1. $y' = t^2 - y$ con $y(0) = 1$, $y(t) = -e^{-t} + t^2 - 2t + 2$
 2. $y' = 3y + 3t$ con $y(0) = 1$, $y(t) = \frac{4}{3}e^{3t} - t - \frac{1}{3}$
 3. $y' = -ty$ con $y(0) = 1$, $y(t) = e^{-t^2/2}$
 4. $y' = e^{-2t} - 2y$ con $y(0) = \frac{1}{10}$, $y(t) = \frac{1}{10}e^{-2t} + te^{-2t}$
 5. $y' = 2ty^2$ con $y(0) = 1$, $y(t) = 1/(1 - t^2)$
 6. Consideremos un proyectil que se dispara hacia arriba y luego cae siguiendo una trayectoria rectilínea. Si la resistencia del aire es proporcional a la velocidad, entonces el problema de valor inicial para la velocidad $v(t)$ es

$$v' = -10 - \frac{K}{M}v \quad \text{con} \quad v(0) = v_0,$$

siendo v_0 la velocidad inicial, M la masa y K el coeficiente de resistencia del aire. Supongamos que $v_0 = 40$ m/s y $K/M = 0.1$. Use el método de Heun con $h = 0.5$ para resolver el problema de valor inicial

$$v' = -10 - 0.1v \quad \text{en } [0, 4] \text{ con } v(0) = 40.$$

Dibuje su solución y la solución exacta $v(t) = 140e^{-t/10} - 100$ en una misma gráfica. (Observe que la velocidad límite es -100 m/s.)

7. En psicología, la ley de estímulo-respuesta de Wever-Fechner establece que la tasa de variación dR/dE de la reacción R ante un estímulo E es inversamente proporcional al estímulo. Si llamamos valor umbral al mínimo nivel de estímulo S_0 que es posible detectar, entonces el problema de valor inicial que modela esta situación es

$$R' = \frac{k}{S} \quad \text{con} \quad R(S_0) = 0.$$

Supongamos que $S_0 = 0.1$ y que $R(0.1) = 0$. Use el método de Heun con $h = 0.1$ para resolver

$$R' = \frac{1}{S} \quad \text{en } [0.1, 5.1] \text{ con } R(0.1) = 0.$$

8. (a) Escriba un programa que sirva para llevar a cabo el método de mejora de Richardson descrito en el Ejercicio 7.
- (b) Use su programa para aproximar $y(2)$ en cada una de las ecuaciones de los Problemas 1–5 tomando $h = 0.05$ como tamaño de paso inicial. El

programa debería terminar cuando el valor absoluto de la diferencia entre dos mejoras consecutivas sea menor que 10^{-6} .

9.4 El método de la serie de Taylor

El método de la serie de Taylor es de aplicabilidad general y es el método estándar con el que se compara la precisión de otros métodos numéricos para resolver problemas de valor inicial, ya que puede ser construido de manera que tenga un grado de exactitud fijado de antemano. Empezamos reformulando el Teorema de Taylor de manera adecuada para la resolución de una ecuación diferencial.

Teorema 9.5 (Teorema de Taylor). Supongamos que $y(t) \in C^{N+1}[t_0, b]$ y que $y(t)$ tiene el desarrollo de Taylor de orden N alrededor de un punto $t = t_k \in [t_0, b]$ dado por

$$(1) \quad y(t_k + h) = y(t_k) + hT_N(t_k, y(t_k)) + O(h^{N+1}),$$

donde

$$(2) \quad T_N(t_k, y(t_k)) = \sum_{j=1}^N \frac{y^{(j)}(t_k)}{j!} h^{j-1}$$

en donde $y^{(j)}(t) = f^{(j-1)}(t, y(t))$ denota la derivada $(j-1)$ -ésima de la función $f(t, y(t))$ con respecto a t . Las fórmulas de estas derivadas pueden calcularse recursivamente usando la regla de la cadena:

$$\begin{aligned} (3) \quad y'(t) &= f \\ y''(t) &= f_t + f_y y' = f_t + f_y f \\ y^{(3)}(t) &= f_{tt} + 2f_{ty}y' + f_{yy}y'' + f_{yy}(y')^2 \\ &= f_{tt} + 2f_{ty}f + f_{yy}f^2 + f_y(f_t + f_y f) \\ y^{(4)}(t) &= f_{ttt} + 3f_{tty}y' + 3f_{tyy}y'' + 3f_{tyy}(y')^2 + f_{yy}y''' + 3f_{yy}y'y'' + f_{yyy}(y')^3 \\ &= (f_{ttt} + 3f_{tty}f + 3f_{tyy}f^2 + f_{yyy}f^3) + f_y(f_{tt} + 2f_{ty}f + f_{yy}f^2) \\ &\quad + 3(f_t + f_y f)(f_{ty} + f_{yy}f) + f_y^2(f_t + f_y f) \end{aligned}$$

y, en general,

$$(4) \quad y^{(N)}(t) = P^{(N-1)}f(t, y(t)),$$

donde P es el operador de derivación

$$P = \left(\frac{\partial}{\partial t} + f \frac{\partial}{\partial y} \right).$$

El valor numérico aproximado de la solución del problema de valor inicial $y'(t) = f(t, y)$ en $[t_0, t_M]$ se calcula usando la fórmula (1) en cada subintervalo $[t_k, t_{k+1}]$, de manera que el paso general del método de Taylor de orden N es

$$(5) \quad y_{k+1} = y_k + d_1 h + \frac{d_2 h^2}{2!} + \frac{d_3 h^3}{3!} + \cdots + \frac{d_N h^N}{N!},$$

siendo $d_j = y^{(j)}(t_k)$ para $j = 1, 2, \dots, N$ en cada paso $k = 0, 1, \dots, M - 1$.

El método de Taylor de orden N tiene la propiedad de que el error global final es de orden $\mathcal{O}(h^{N+1})$; por tanto, se puede elegir N de manera que este error sea tan pequeño como queramos. Si fijamos el orden N entonces es teóricamente posible fijar el tamaño de paso h de forma que el error global final sea tan pequeño como queramos; sin embargo, en la práctica lo que se hace es construir dos conjuntos de aproximaciones usando tamaños de paso h y $h/2$ y comparar los resultados.

Teorema 9.6 (Precisión del método de Taylor de orden N). Supongamos que $y(t)$ es la solución del problema de valor inicial. Si $y(t) \in C^{N+1}[t_0, b]$ y $\{(t_k, y_k)\}_{k=0}^M$ es la sucesión de aproximaciones generadas por el método de Taylor de orden N , entonces

$$(6) \quad \begin{aligned} |e_k| &= |y(t_k) - y_k| = \mathcal{O}(h^{N+1}), \\ |\varepsilon_{k+1}| &= |y(t_{k+1}) - y_k - hT_N(t_k, y_k)| = \mathcal{O}(h^N). \end{aligned}$$

En particular, el error global final en el extremo derecho del intervalo es

$$(7) \quad E(y(b), h) = |y(b) - y_M| = \mathcal{O}(h^N).$$

La demostración puede hallarse en la Referencia [78].

Los Ejemplos 9.8 y 9.9 ilustran el Teorema 9.6 para el caso $N = 4$. Si calculásemos las aproximaciones tomando como tamaños de paso h y $h/2$, entonces deberíamos tener

$$(8) \quad E(y(b), h) \approx Ch^4$$

para el tamaño de paso más grande y

$$(9) \quad E\left(y(b), \frac{h}{2}\right) \approx C \frac{h^4}{16} = \frac{1}{16} Ch^4 \approx \frac{1}{16} E(y(b), h).$$

Por tanto, el Teorema 9.6 nos dice que si el tamaño de paso se reduce a la mitad, entonces el error global final debería reducirse en un factor del orden de $\frac{1}{16}$.

Ejemplo 9.8. Vamos a usar el método de Taylor de orden $N = 4$ para resolver $y' = (t - y)/2$ en $[0, 3]$ con $y(0) = 1$ y a comparar las soluciones obtenidas con $h = 1$, $\frac{1}{2}$, $\frac{1}{4}$ y $\frac{1}{8}$.

Primero hay que determinar las derivadas de $y(t)$. Recordando que la solución $y(t)$ es función de t y derivando la fórmula $y'(t) = f(t, y(t))$ con respecto a t obtenemos $y^{(2)}(t)$. Después, continuamos el proceso para hallar las derivadas de orden superior:

$$\begin{aligned}y'(t) &= \frac{t - y}{2}, \\y^{(2)}(t) &= \frac{d}{dt} \left(\frac{t - y}{2} \right) = \frac{1 - y'}{2} = \frac{1 - (t - y)/2}{2} = \frac{2 - t + y}{4}, \\y^{(3)}(t) &= \frac{d}{dt} \left(\frac{2 - t + y}{4} \right) = \frac{0 - 1 + y'}{4} = \frac{-1 + (t - y)/2}{4} = \frac{-2 + t - y}{8}, \\y^{(4)}(t) &= \frac{d}{dt} \left(\frac{-2 + t - y}{8} \right) = \frac{-0 + 1 - y'}{8} = \frac{1 - (t - y)/2}{8} = \frac{2 - t + y}{16}.\end{aligned}$$

Para hallar y_1 , debemos evaluar las derivadas que acabamos de calcular en el punto $(t_0, y_0) = (0, 1)$, obteniéndose

$$\begin{aligned}d_1 &= y'(0) = \frac{0.0 - 1.0}{2} = -0.5, \\d_2 &= y^{(2)}(0) = \frac{2.0 - 0.0 + 1.0}{4} = 0.75, \\d_3 &= y^{(3)}(0) = \frac{-2.0 + 0.0 - 1.0}{8} = -0.375, \\d_4 &= y^{(4)}(0) = \frac{2.0 - 0.0 + 1.0}{16} = 0.1875.\end{aligned}$$

Ahora sustituimos las derivadas $\{d_j\}$ en la expresión (5) con $h = 0.25$ y usamos el esquema de Horner-Ruffini para calcular el valor y_1 :

$$\begin{aligned}y_1 &= 1.0 + 0.25 \left(-0.5 + 0.25 \left(\frac{0.75}{2} + 0.25 \left(\frac{-0.375}{6} + 0.25 \left(\frac{0.1875}{24} \right) \right) \right) \right) \\&= 0.8974915.\end{aligned}$$

El punto calculado es pues $(t_1, y_1) = (0.25, 0.8974915)$.

Para determinar y_2 , hay que evaluar las derivadas $\{d_j\}$ en el punto $(t_1, y_1) = (0.25, 0.8974915)$. Los cálculos empiezan a requerir un esfuerzo computacional con-

Tabla 9.6 Comparación de las soluciones obtenidas con el método de Taylor de orden $N = 4$ para el problema de valor inicial $y' = (t - y)/2$ en $[0, 3]$ con $y(0) = 1$.

t_k	y_k				Exacto
	$h = 1$	$h = \frac{1}{2}$	$h = \frac{1}{4}$	$h = \frac{1}{8}$	
0	1.0	1.0	1.0	1.0	1.0
0.125			0.8974915	0.9432392	0.9432392
0.25				0.8974908	0.8974917
0.375				0.8620874	0.8620874
0.50		0.8364258	0.8364037	0.8364024	0.8364023
0.75			0.8118696	0.8118679	0.8118678
1.00	0.8203125	0.8196285	0.8195940	0.8195921	0.8195920
1.50		0.9171423	0.9171021	0.9170998	0.9170997
2.00	1.1045125	1.1036826	1.1036408	1.1036385	1.1036383
2.50		1.3595575	1.3595168	1.3595145	1.3595144
3.00	1.6701860	1.6694308	1.6693928	1.6693906	1.6693905

siderable y resulta muy tedioso el hacerlos a mano:

$$d_1 = y'(0.25) = \frac{0.25 - 0.8974915}{2} = -0.3237458,$$

$$d_2 = y^{(2)}(0.25) = \frac{2.0 - 0.25 + 0.8974915}{4} = 0.6618729,$$

$$d_3 = y^{(3)}(0.25) = \frac{-2.0 + 0.25 - 0.8974915}{8} = -0.3309364,$$

$$d_4 = y^{(4)}(0.25) = \frac{2.0 - 0.25 + 0.8974915}{16} = 0.1654682.$$

Ahora sustituimos estas derivadas en la expresión (5) con $h = 0.25$ y usamos el esquema de Horner-Ruffini para calcular y_2 :

$$\begin{aligned} y_2 &= 0.8974915 + 0.25 \left(-0.3237458 \right. \\ &\quad \left. + 0.25 \left(\frac{0.6618729}{2} + 0.25 \left(\frac{-0.3309364}{6} + 0.25 \left(\frac{0.1654682}{24} \right) \right) \right) \right) \\ &= 0.8364037. \end{aligned}$$

Por tanto, el nuevo punto es $(t_2, y_2) = (0.50, 0.8364037)$. En la Tabla 9.6 se muestran las aproximaciones obtenidas en algunas abscisas con tamaños de paso diferentes. ■

Ejemplo 9.9. Vamos a comparar el error global final de las soluciones obtenidas con el método de Taylor en el Ejemplo 9.8 para el problema de valor inicial $y' = (t - y)/2$ en $[0, 3]$ con $y(0) = 1$.

Tabla 9.7 Relación entre el tamaño de paso y el error global final para las soluciones obtenidas con el método de Taylor para el problema $y' = (t - y)/2$ en $[0, 3]$ con $y(0) = 1$.

Tamaño de paso, h	Número de pasos, M	Aproximación y_M a $y(3)$	Error global final $y(3) - y_M$	$O(h^2) \approx Ch^4$ con $C = -0.000614$
1	3	1.6701860	-0.0007955	-0.0006140
$\frac{1}{2}$	6	1.6694308	-0.0000403	-0.0000384
$\frac{1}{4}$	12	1.6693928	-0.0000023	-0.0000024
$\frac{1}{8}$	24	1.6693906	-0.0000001	-0.0000001

En la Tabla 9.7 se recogen los errores globales finales para los correspondientes tamaños de paso y en ella podemos ver que las aproximaciones a $y(3)$ decrecen en un factor del orden de $\frac{1}{16}$ cuando el tamaño de paso se reduce a la mitad:

$$E(y(3), h) = y(3) - y_M = O(h^4) \approx Ch^4, \quad \text{con} \quad C = -0.000614.$$

MATLAB

Para usar el programa que damos a continuación, hace falta que las derivadas y' , y'' , y''' e y'''' estén almacenadas en un archivo llamado, por ejemplo, `df.m`; así, el siguiente archivo serviría para almacenar las derivadas del Ejemplo 9.8 según el formato requerido por el Programa 9.3.

```
function z=df(t,y)
z=[(t-y)/2, (2-t+y)/4, (-2+t-y)/8, (2-t+y)/16];
```

Programa 9.3 (Método de Taylor de Orden 4). Construcción de las aproximaciones a la solución de $y' = f(t, y)$ con $y(a) = y_0$ en $[a, b]$ evaluando y'', y''' e y'''' y usando el polinomio de Taylor en cada paso.

```
function T4=taylor(df,a,b,ya,M)
% Datos
% - df=[y' y'' y''' y'''''] almacenada como una
%   cadena de caracteres 'df', siendo y'=f(t,y)
% - a y b son los extremos derecho e izquierdo
%   del intervalo
```

% - ya es la condición inicial $y(a)$
% - M es el número de pasos
% Resultado
% - $T4=[T' Y']$ siendo T el vector de las abscisas e
% Y el vector de las ordenadas

```
h=(b-a)/M;
T=zeros(1,M+1);
Y=zeros(1,M+1);
T=a:h:b;
Y(1)=ya;
for j=1:M
    D=feval(df,T(j),Y(j));
    Y(j+1)=Y(j)+h*(D(1)+h*(D(2)/2+h*(D(3)/6+h*D(4)/24)));
end
T4=[T' Y'];
```

Ejercicios

En los Ejercicios 1 a 5 resuelva la ecuación diferencial usando el método de Taylor de orden $N = 4$.

- (a) Tome $h = 0.2$ y dé dos pasos calculando los valores a mano. Luego tome $h = 0.1$ y dé cuatro pasos calculando los valores a mano.
- (b) Compare la solución exacta $y(0.4)$ con las dos aproximaciones calculadas en el apartado (a).
- (c) ¿Se comporta el error global final de las aproximaciones obtenidas en el apartado (a) como se espera cuando h se divide entre dos?
1. $y' = t^2 - y$ con $y(0) = 1$, $y(t) = -e^{-t} + t^2 - 2t + 2$
 2. $y' = 3y + 3t$ con $y(0) = 1$, $y(t) = \frac{4}{3}e^{3t} - t - \frac{1}{3}$
 3. $y' = -ty$ con $y(0) = 1$, $y(t) = e^{-t^2/2}$
 4. $y' = e^{-2t} - 2y$ con $y(0) = \frac{1}{10}$, $y(t) = \frac{1}{10}e^{-2t} + te^{-2t}$
 5. $y' = 2ty^2$ con $y(0) = 1$, $y(t) = 1/(1 - t^2)$
6. El método de mejora de Richardson, presentado en el Lema 7.1 (Sección 7.3) puede usarse en conjunción con el método de Taylor. Si en el método de Taylor de orden $N = 4$ se usa un tamaño de paso h , entonces $y(b) \approx y_h + Ch^4$, pero si usamos un tamaño de paso $2h$, entonces $y(b) \approx y_{2h} + 16Ch^4$. Los términos que contienen Ch^4 pueden eliminarse para obtener una aproximación mejorada:

$$y(b) \approx \frac{16y_h - y_{2h}}{15}.$$

Este esquema de mejora puede usarse con los valores mostrados en el Ejemplo 9.9 para obtener una aproximación mejor a $y(3)$. Calcule las entradas que faltan en la tabla siguiente:

h	y_h	$(16y_h - y_{2h})/15$
1.0	1.6701860	_____
0.5	1.6694308	_____
0.25	1.6693928	_____
0.125	1.6693906	_____

7. Pruebe que cuando se utiliza el método de Taylor de orden N con tamaños de paso h y $h/2$, entonces el error global final se reduce, aproximadamente, en un factor de 2^{-N} .
8. Pruebe que el método de Taylor falla cuando queremos aproximar la solución $y(t) = t^{3/2}$ del problema de valor inicial

$$y' = f(t, y) = 1.5y^{1/3} \quad \text{con} \quad y(0) = 0.$$

Justifique su respuesta. ¿Cuál es el problema?

9. (a) Verifique que la solución del problema de valor inicial $y' = y^2$, $y(0) = 1$ en el intervalo $[0, 1]$ es $y(t) = 1/(1-t)$.
(b) Verifique que la solución del problema de valor inicial $y' = 1+y^2$, $y(0) = 1$ en el intervalo $[0, \pi/4]$ es $y(t) = \tan(t + \pi/4)$.
(c) Use los resultados de los apartados (a) y (b) para deducir que la solución del problema de valor inicial $y' = t^2 + y^2$, $y(0) = 1$ tiene una asíntota vertical entre $\pi/4$ y 1 (localizada cerca de $t = 0.96981$).
10. Consideraremos el problema de valor inicial $y' = 1+y^2$, $y(0) = 1$.
 - (a) Determine las expresiones de $y^{(2)}(t)$, $y^{(3)}(t)$ e $y^{(4)}(t)$.
 - (b) Evalúe las derivadas en $t = 0$ y úselas para calcular los cinco primeros términos del desarrollo de Maclaurin de $\tan(t)$.

Algoritmos y programas

En los Problemas 1 a 5, resuelva la ecuación diferencial usando el método de Taylor de orden $N = 4$.

- (a) Tome $h = 0.1$ y dé 20 pasos con el Programa 9.3. Luego tome $h = 0.05$ y dé 40 pasos con el Programa 9.3.
- (b) Compare la solución exacta $y(2)$ con las dos aproximaciones obtenidas en el apartado (a).

- (c) ¿Se comporta el error global final de las aproximaciones obtenidas en el apartado (a) como se espera cuando h se divide entre dos?
- (d) Dibuje las aproximaciones y la solución exacta en una misma gráfica. *Indicación.* La matriz $T4$ que se obtiene como resultado en el Programa 9.3 contiene las coordenadas x e y de las aproximaciones y la instrucción del paquete MATLAB `plot(T4(:,1),T4(:,2))` producirá un dibujo análogo al de la Figura 9.6.
1. $y' = t^2 - y$ con $y(0) = 1$, $y(t) = -e^{-t} + t^2 - 2t + 2$
 2. $y' = 3y + 3t$ con $y(0) = 1$, $y(t) = \frac{4}{3}e^{3t} - t - \frac{1}{3}$
 3. $y' = -ty$ con $y(0) = 1$, $y(t) = e^{-t^2/2}$
 4. $y' = e^{-2t} - 2y$ con $y(0) = \frac{1}{10}$, $y(t) = \frac{1}{10}e^{-2t} + te^{-2t}$
 5. $y' = 2ty^2$ con $y(0) = 1$, $y(t) = 1/(1-t^2)$
6. (a) Escriba un programa que sirva para llevar a cabo el método de mejora de Richardson descrito en el Ejercicio 6.
(b) Use su programa del apartado (a) para hallar una aproximación a $y(0.8)$ para el problema de valor inicial $y' = t^2 + y^2$, $y(0) = 1$ en el intervalo $[0, 0.8]$. Se sabe que la solución en $t = 0.8$ vale (con ocho cifras significativas) $y(0.8) = 5.8486168$. Empiece con un tamaño de paso $h = 0.05$ y haga que el programa se detenga cuando el valor absoluto de la diferencia entre dos mejoras consecutivas sea menor que 10^{-6} .
7. (a) Modifique el Programa 9.3 de manera que sirva para llevar a cabo el método de Taylor de orden $N = 3$.
(b) Use su programa del apartado (a) para resolver el problema de valor inicial $y' = t^2 + y^2$, $y(0) = 1$ en el intervalo $[0, 0.8]$. Determine las aproximaciones correspondientes a los tamaños de paso $h = 0.05, 0.025, 0.0125$ y 0.00625 y dibuje dichas aproximaciones en una misma gráfica.

9.5 Los métodos de Runge-Kutta

Los métodos de Taylor de la sección precedente tienen la característica deseable de que el error global final es de orden $O(h^N)$, de manera que podemos escoger N tan grande como queramos para que el error sea tan pequeño como deseemos. Sin embargo, los métodos de Taylor presentan dos inconvenientes: la necesidad de determinar N a priori y el cálculo de las derivadas de orden superior, que puede ser bastante complicado. Los métodos de Runge-Kutta se construyen a partir de un método de Taylor, digamos de orden N , de tal manera que el error global final sea del mismo orden $O(h^N)$ pero se evite la evaluación de las derivadas parciales; esto puede conseguirse a cambio de evaluar, en cada paso, la función en varios puntos. Aunque estos métodos se pueden construir para cualquier orden N , nosotros nos centraremos en el método de Runge-Kutta de

orden $N = 4$, que es el más popular y es también, para propósitos generales, una buena elección ya que es bastante preciso, estable y fácil de programar. Aún más, la mayoría de los expertos dicen que no es necesario trabajar con métodos de orden superior porque el aumento del coste computacional no compensa la mayor exactitud; si es necesario obtener mayor precisión en la solución aproximada, entonces es mejor usar un tamaño de paso menor o un método adaptativo.

El método de cuarto orden de Runge-Kutta (que llamaremos RK4) simula la precisión del método de la serie de Taylor de orden $N = 4$ y consiste en calcular la aproximación y_{k+1} de la siguiente manera:

$$(1) \quad y_{k+1} = y_k + w_1 k_1 + w_2 k_2 + w_3 k_3 + w_4 k_4,$$

donde k_1, k_2, k_3 y k_4 son de la forma

$$(2) \quad \begin{aligned} k_1 &= hf(t_k, y_k), & k_3 &= hf(t_k + a_2 h, y_k + b_2 k_1 + b_3 k_2), \\ k_2 &= hf(t_k + a_1 h, y_k + b_1 k_1), & k_4 &= hf(t_k + a_3 h, y_k + b_4 k_1 + b_5 k_2 + b_6 k_3). \end{aligned}$$

Emparejando estos coeficientes con los del método de la serie de Taylor de orden $N = 4$ de manera que el error de truncamiento local sea de orden $O(h^5)$ (nosotros veremos luego cómo se hace esto para el método de orden $N = 2$), Runge y Kutta fueron capaces de obtener el siguiente sistema de ecuaciones:

$$(3) \quad \begin{aligned} b_1 &= a_1, \\ b_2 + b_3 &= a_2, \\ b_4 + b_5 + b_6 &= a_3, \\ w_1 + w_2 + w_3 + w_4 &= 1, \\ w_2 a_1 + w_3 a_2 + w_4 a_3 &= \frac{1}{2}, \\ w_2 a_1^2 + w_3 a_2^2 + w_4 a_3^2 &= \frac{1}{3}, \\ w_2 a_1^3 + w_3 a_2^3 + w_4 a_3^3 &= \frac{1}{4}, \\ w_3 a_1 b_3 + w_4 (a_1 b_5 + a_2 b_6) &= \frac{1}{6}, \\ w_3 a_1 a_2 b_3 + w_4 a_3 (a_1 b_5 + a_2 b_6) &= \frac{1}{8}, \\ w_3 a_1^2 b_3 + w_4 (a_1^2 b_5 + a_2^2 b_6) &= \frac{1}{12}, \\ w_4 a_1 b_3 b_6 &= \frac{1}{24}. \end{aligned}$$

Este sistema tiene 11 ecuaciones con 13 incógnitas, así que debemos añadir dos

condiciones adicionales para resolverlo. La elección más útil resulta ser

$$(4) \quad a_1 = \frac{1}{2} \quad \text{y} \quad b_2 = 0.$$

Entonces los valores de la solución para las demás variables son

$$(5) \quad \begin{aligned} a_2 &= \frac{1}{2}, \quad a_3 = 1, \quad b_1 = \frac{1}{2}, \quad b_3 = \frac{1}{2}, \quad b_4 = 0, \quad b_5 = 0, \quad b_6 = 1, \\ w_1 &= \frac{1}{6}, \quad w_2 = \frac{1}{3}, \quad w_3 = \frac{1}{3}, \quad w_4 = \frac{1}{6}. \end{aligned}$$

Sustituyendo en las expresiones (1) y (2) los valores dados en (4) y (5), obtenemos la fórmula para el método de Runge-Kutta de orden $N = 4$ estándar: A partir del punto inicial (t_0, y_0) se genera la sucesión de aproximaciones usando la fórmula recursiva

$$(6) \quad y_{k+1} = y_k + \frac{h(f_1 + 2f_2 + 2f_3 + f_4)}{6},$$

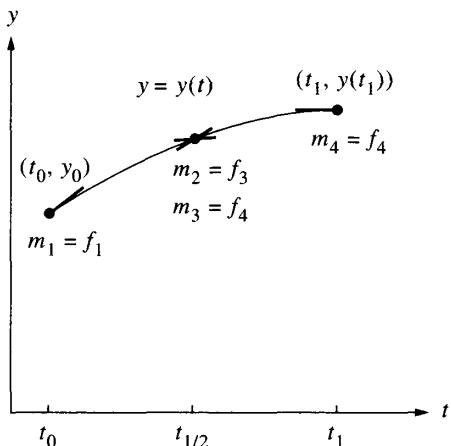
donde

$$(7) \quad \begin{aligned} f_1 &= f(t_k, y_k), \\ f_2 &= f\left(t_k + \frac{h}{2}, y_k + \frac{h}{2}f_1\right), \\ f_3 &= f\left(t_k + \frac{h}{2}, y_k + \frac{h}{2}f_2\right), \\ f_4 &= f(t_k + h, y_k + hf_3). \end{aligned}$$

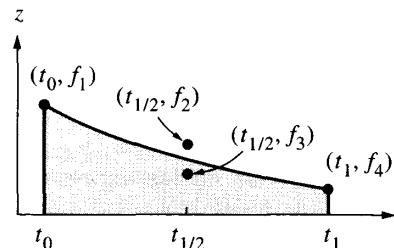
Algunos aspectos relevantes del método

El desarrollo completo del método hasta que se obtienen las expresiones dadas en (7) está más allá de los objetivos de este libro y puede encontrarse en textos más avanzados; no obstante, sí podemos profundizar en algunos aspectos de interés. Consideremos la gráfica de la solución $y = y(t)$ en el primer subintervalo $[t_0, t_1]$. Los valores de la función que se recogen en (7) son aproximaciones a valores de la derivada en algunos puntos de dicha gráfica: El valor f_1 es la derivada en el extremo derecho del subintervalo, los valores f_2 y f_3 son estimaciones de la derivada en el punto medio y f_4 es la derivada en el extremo derecho (véase la Figura 9.9(a)). El nuevo punto (t_1, y_1) se obtiene integrando la función derivada

$$(8) \quad y(t_1) - y(t_0) = \int_{t_0}^{t_1} f(t, y(t)) dt.$$



(a) Aproximaciones m_j a la derivada de la solución $y = y(t)$



(b) Aproximación integral:

$$y(t_1) - y_0 = \frac{h}{6}(f_1 + 2f_2 + 2f_3 + f_4)$$

Figura 9.9 Las gráficas $y = y(t)$ y $z = f(t, y(t))$ que aparecen en el desarrollo del método de Runge-Kutta de orden $N = 4$.

Si aplicamos aquí la regla de Simpson con incremento $h/2$, entonces la aproximación a la integral de (8) que obtenemos es

$$(9) \quad \int_{t_0}^{t_1} f(t, y(t)) dt \approx \frac{h}{6}(f(t_0, y(t_0)) + 4f(t_{1/2}, y(t_{1/2})) + f(t_1, y(t_1))),$$

siendo $t_{1/2}$ el punto medio del subintervalo. Necesitamos, entonces, tres valores de la función f , así que hacemos las elecciones obvias $f(t_0, y(t_0)) = f_1$ y $f(t_1, y(t_1)) \approx f_4$, mientras que para el valor en el punto medio hacemos la media de f_2 y f_3 :

$$f(t_{1/2}, y(t_{1/2})) \approx \frac{f_2 + f_3}{2}.$$

Sustituyendo estos valores en (9) y dicha aproximación a la integral en (8) obtenemos y_1 :

$$(10) \quad y_1 = y_0 + \frac{h}{6} \left(f_1 + \frac{4(f_2 + f_3)}{2} + f_4 \right).$$

Simplificando esta fórmula, obtenemos la ecuación (6) con $k = 0$. En la Figura 9.9(b) se muestra la gráfica de la integral que aparece en (9).

Tamaño de paso frente a error

El término del error de la regla de Simpson con incremento $h/2$ es

$$(11) \quad -y^{(4)}(c_1) \frac{h^5}{2880}.$$

Si el único error que apareciera en cada paso fuera el dado en la expresión (11), entonces, después de M pasos, el error acumulado al llevar a cabo el método RK4 sería

$$(12) \quad -\sum_{k=1}^M y^{(4)}(c_k) \frac{h^5}{2880} \approx \frac{b-a}{5760} y^{(4)}(c) h^4 \approx O(h^4).$$

El siguiente teorema establece la relación que hay entre el error global final y el tamaño de paso y nos sirve para darnos una idea de cuál es el esfuerzo computacional que se realiza al utilizar el método RK4.

Teorema 9.7 (Precisión del método de Runge-Kutta). Supongamos que $y(t)$ es la solución del problema de valor inicial. Si $y(t) \in C^5[t_0, b]$ y $\{(t_k, y_k)\}_{k=0}^M$ es la sucesión de aproximaciones generada por el método de Runge-Kutta de orden 4, entonces

$$(13) \quad \begin{aligned} |e_k| &= |y(t_k) - y_k| = O(h^4), \\ |\varepsilon_{k+1}| &= |y(t_{k+1}) - y_k - hT_N(t_k, y_k)| = O(h^5). \end{aligned}$$

En particular, el error global final verifica

$$(14) \quad E(y(b), h) = |y(b) - y_M| = O(h^4).$$

Los Ejemplos 9.10 y 9.11 ilustran el Teorema 9.7. Si calculásemos las aproximaciones usando como tamaños de paso h y $h/2$, entonces deberíamos tener

$$(15) \quad E(y(b), h) \approx Ch^4$$

para el tamaño de paso más grande y

$$(16) \quad E\left(y(b), \frac{h}{2}\right) \approx C \frac{h^4}{16} = \frac{1}{16} Ch^4 \approx \frac{1}{16} E(y(b), h).$$

Luego el Teorema 9.7 nos dice que si el tamaño de paso en el método RK4 se reduce a la mitad, entonces debemos esperar que el error global final se reduzca, aproximadamente, en un factor de $\frac{1}{16}$.

Tabla 9.8 Comparación de las soluciones de $y' = (t - y)/2$ en $[0, 3]$ con $y(0) = 1$ obtenidas con el método RK4 para diferentes tamaños de paso.

t_k	y_k				$y(t_k)$ Exacto
	$h = 1$	$h = \frac{1}{2}$	$h = \frac{1}{4}$	$h = \frac{1}{8}$	
0	1.0	1.0	1.0	1.0	1.0
0.125			0.8974915	0.9432392	0.9432392
0.25			0.8974908	0.8974917	0.8974917
0.375			0.8620874	0.8620874	0.8620874
0.50		0.8364258	0.8364037	0.8364024	0.8364023
0.75			0.8118696	0.8118679	0.8118678
1.00	0.8203125	0.8196285	0.8195940	0.8195921	0.8195920
1.50		0.9171423	0.9171021	0.9170998	0.9170997
2.00	1.1045125	1.1036826	1.1036408	1.1036385	1.1036383
2.50		1.3595575	1.3595168	1.3595145	1.3595144
3.00	1.6701860	1.6694308	1.6693928	1.6693906	1.6693905

Ejemplo 9.10. Vamos a usar el método RK4 para resolver el problema de valor inicial $y' = (t - y)/2$ en $[0, 3]$ con $y(0) = 1$ y a comparar las soluciones obtenidas para $h = 1, \frac{1}{2}, \frac{1}{4}$ y $\frac{1}{8}$.

En la Tabla 9.8 se muestran los valores de la solución en algunas abscisas escogidas. Para el tamaño de paso $h = 0.25$, un cálculo típico es el siguiente:

$$f_1 = \frac{0.0 - 1.0}{2} = -0.5,$$

$$f_2 = \frac{0.125 - (1 + 0.25(0.5)(-0.5))}{2} = -0.40625,$$

$$f_3 = \frac{0.125 - (1 + 0.25(0.5)(-0.40625))}{2} = -0.4121094,$$

$$f_4 = \frac{0.25 - (1 + 0.25(-0.4121094))}{2} = -0.3234863,$$

$$y_1 = 1.0 + 0.25 \left(\frac{-0.5 + 2(-0.40625) + 2(-0.4121094) - 0.3234863}{6} \right) \\ = 0.8974915.$$

Ejemplo 9.11. Vamos a comparar el error global final cuando se usa el método RK4 para resolver $y' = (t - y)/2$ en $[0, 3]$ con $y(0) = 1$ con tamaños de paso $h = 1, \frac{1}{2}, \frac{1}{4}$ y $\frac{1}{8}$.

En la Tabla 9.9 se muestran los errores globales finales para los distintos tamaños de paso; en ella podemos ver que el error de la aproximación a $y(3)$ decrece, más o menos, en un factor $\frac{1}{16}$ cuando el tamaño de paso se reduce a la mitad:

$$E(y(3), h) = y(3) - y_M = O(h^4) \approx Ch^4 \quad \text{con} \quad C = -0.000614.$$

Tabla 9.9 Relación entre el error global final y el tamaño de paso en las soluciones de $y' = (t - y)/2$ en $[0, 3]$ con $y(0) = 1$ obtenidas con el método RK4.

Tamaño de paso, h	Número de pasos, M	Aproximación y_M a $y(3)$	Error global final, $y(3) - y_M$	$O(h^4) \approx Ch^4$ con $C = -0.000614$
1	3	1.6701860	-0.0007955	-0.0006140
$\frac{1}{2}$	6	1.6694308	-0.0000403	-0.0000384
$\frac{1}{4}$	12	1.6693928	-0.0000023	-0.0000024
$\frac{1}{8}$	24	1.6693906	-0.0000001	-0.0000001

Comparando los resultados obtenidos en los Ejemplos 9.10 y 9.11 y en los Ejemplos 9.8 y 9.9 vemos lo que se quiere decir en la frase escrita antes “El método de cuarto orden de Runge-Kutta (RK4) simula la precisión del método de la serie de Taylor de orden $N = 4$ ”. En estos ejemplos, ambos métodos generan soluciones idénticas $\{(t_k, y_k)\}$ en el intervalo dado. La ventaja del método RK4 es obvia: no hay que calcular las fórmulas de las derivadas de orden superior ni hay que incluirlas en el programa.

No es fácil el determinar la precisión con la que se calcula una solución obtenida por el método de Runge-Kutta; podríamos estimar la magnitud de $y^{(4)}(c)$ y usar la fórmula (12), o bien repetir el algoritmo con un tamaño de paso menor y comparar los resultados. Una tercera vía es la determinación adaptativa del tamaño de paso, como se hace en el Programa 9.5. En la Sección 9.6 veremos cómo se realiza el cambio de tamaño de paso en los métodos multipaso.

Los métodos de Runge-Kutta de orden $N = 2$

El método de Runge-Kutta de segundo orden (que llamaremos RK2) simula la precisión del método de la serie de Taylor de orden $N = 2$. Aunque no es un método tan bueno como el RK4, los razonamientos que nos conducen a su desarrollo son más fáciles de entender y sirven para ilustrar las ideas involucradas en los métodos de Runge-Kutta. Empezamos escribiendo el desarrollo en serie de Taylor para $y(t + h)$:

$$(17) \quad y(t + h) = y(t) + hy'(t) + \frac{1}{2}h^2y''(t) + C_T h^3 + \dots,$$

donde C_T es una constante que incluye la derivada tercera de $y(t)$ y los demás términos corresponden a las potencias h^j para $j > 3$.

Vamos a expresar las derivadas $y'(t)$ e $y''(t)$ de la ecuación (17) en términos de $f(t, y)$ y de sus derivadas parciales: recordando que

$$(18) \quad y'(t) = f(t, y)$$

y derivando la igualdad (18) con respecto a t usando la regla de la cadena para funciones de dos variables, obtenemos

$$y''(t) = f_t(t, y) + f_y(t, y)y'(t).$$

Esto podemos escribirlo, usando la igualdad (18), como

$$(19) \quad y''(t) = f_t(t, y) + f_y(t, y)f(t, y).$$

Sustituyendo las derivadas (18) y (19) en la expresión (17), obtenemos una nueva expresión del desarrollo de Taylor para $y(t + h)$:

$$(20) \quad \begin{aligned} y(t + h) &= y(t) + hf(t, y) + \frac{1}{2}h^2 f_t(t, y) \\ &\quad + \frac{1}{2}h^2 f_y(t, y)f(t, y) + C_T h^3 + \dots \end{aligned}$$

En el método de Runge-Kutta de orden $N = 2$ se utiliza una combinación lineal de dos funciones que nos permita expresar $y(t + h)$:

$$(21) \quad y(t + h) = y(t) + Ahf_0 + Bhf_1,$$

donde

$$(22) \quad \begin{aligned} f_0 &= f(t, y), \\ f_1 &= f(t + Ph, y + Qhf_0). \end{aligned}$$

Usando la fórmula de Taylor para una función de dos variables (véanse los Ejercicios 8 y 9), podemos aproximar $f(t, y)$ obteniendo la siguiente representación de f_1 :

$$(23) \quad f_1 = f(t, y) + Phf_t(t, y) + Qhf_y(t, y)f(t, y) + C_P h^2 + \dots,$$

en la cual el coeficiente C_P incluye las derivadas parciales segundas de $f(t, y)$. Sustituyendo la expresión (23) en (21) obtenemos la representación de $y(t + h)$ que se usa en el método RK2:

$$(24) \quad \begin{aligned} y(t + h) &= y(t) + (A + B)hf(t, y) + BPh^2 f_t(t, y) \\ &\quad + BQh^2 f_y(t, y)f(t, y) + BC_P h^3 + \dots \end{aligned}$$

Comparando los términos correspondientes en las expresiones (20) y (24) llegamos a las siguientes conclusiones:

$$\begin{aligned} hf(t, y) &= (A + B)hf(t, y) && \text{implica que } 1 = A + B, \\ \frac{1}{2}h^2 f_t(t, y) &= BPh^2 f_t(t, y) && \text{implica que } \frac{1}{2} = BP, \\ \frac{1}{2}h^2 f_y(t, y)f(t, y) &= BQh^2 f_y(t, y)f(t, y) && \text{implica que } \frac{1}{2} = BQ. \end{aligned}$$

Por consiguiente, si exigimos que A , B , P y Q verifiquen las relaciones

$$(25) \quad A + B = 1 \quad BP = \frac{1}{2} \quad BQ = \frac{1}{2},$$

entonces el método RK2 dado en (24) tendrá el mismo orden de precisión que el método de Taylor dado en (20).

Puesto que hay sólo tres ecuaciones para cuatro incógnitas, el sistema de ecuaciones (25) está subdeterminado, lo que nos permite elegir libremente uno de los coeficientes; vamos a presentar dos de entre las varias elecciones posibles que aparecen en la literatura.

Caso (i): Elegimos $A = \frac{1}{2}$, lo que nos lleva a $B = \frac{1}{2}$, $P = 1$ y $Q = 1$. Escribiendo la ecuación (21) con estos parámetros, nos queda la fórmula

$$(26) \quad y(t+h) = y(t) + \frac{h}{2}(f(t, y) + f(t+h, y+hf(t, y)))$$

y cuando usamos este esquema para generar la sucesión $\{(t_k, y_k)\}$, lo que obtenemos es el método de Heun.

Caso (ii): Elegimos $A = 0$, lo que nos lleva a $B = 1$, $P = \frac{1}{2}$ y $Q = \frac{1}{2}$. Escribiendo la ecuación (21) con estos parámetros, nos queda la fórmula

$$(27) \quad y(t+h) = y(t) + hf\left(t + \frac{h}{2}, y + \frac{h}{2}f(t, y)\right)$$

y cuando usamos este esquema para generar la sucesión $\{(t_k, y_k)\}$, lo que obtenemos se conoce como **método de Euler modificado** o **método de Cauchy**.

El método de Runge-Kutta-Fehlberg (RKF45)

Una forma de garantizar la exactitud de la solución aproximada de un problema de valor inicial es resolver el problema dos veces, con tamaños de paso h y $h/2$, y comparar las respuestas en los nodos correspondientes al tamaño de paso más grande. Sin embargo, la segunda resolución supone un coste computacional significativo y, además, hay que hacer una tercera si se establece que la precisión no es aún la adecuada.

El método de Runge-Kutta-Fehlberg (que denotaremos RKF45) es una forma de resolver este problema ya que incluye un criterio para determinar en cada momento si estamos utilizando el tamaño de paso h apropiado. En cada paso se calculan dos aproximaciones distintas de la solución y se comparan: si los dos valores son suficientemente parecidos, entonces se acepta la aproximación y, además, se aumenta el tamaño de paso si coinciden en más cifras significativas que las requeridas; mientras que si los valores no coinciden con la precisión especificada, entonces se reduce el tamaño de paso.

En cada paso se calculan los siguientes seis valores:

$$(28) \quad \begin{aligned} k_1 &= hf(t_k, y_k), \\ k_2 &= hf\left(t_k + \frac{1}{4}h, y_k + \frac{1}{4}k_1\right), \\ k_3 &= hf\left(t_k + \frac{3}{8}h, y_k + \frac{3}{32}k_1 + \frac{9}{32}k_2\right), \\ k_4 &= hf\left(t_k + \frac{12}{13}h, y_k + \frac{1932}{2197}k_1 - \frac{7200}{2197}k_2 + \frac{7296}{2197}k_3\right), \\ k_5 &= hf\left(t_k + h, y_k + \frac{439}{216}k_1 - 8k_2 + \frac{3680}{513}k_3 - \frac{845}{4104}k_4\right), \\ k_6 &= hf\left(t_k + \frac{1}{2}h, y_k - \frac{8}{27}k_1 + 2k_2 - \frac{3544}{2565}k_3 + \frac{1859}{4104}k_4 - \frac{11}{40}k_5\right). \end{aligned}$$

Entonces se calcula una aproximación a la solución del problema de valor inicial usando un método de Runge-Kutta de orden 4:

$$(29) \quad y_{k+1} = y_k + \frac{25}{216}k_1 + \frac{1408}{2565}k_3 + \frac{2197}{4101}k_4 - \frac{1}{5}k_5,$$

donde se utilizan cuatro valores de la función: f_1 , f_3 , f_4 y f_5 ; hagamos notar que el valor f_2 no se emplea en la fórmula (29). Ahora se calcula una aproximación mejor que (29) usando un método de Runge-Kutta de orden 5:

$$(30) \quad z_{k+1} = y_k + \frac{16}{135}k_1 + \frac{6656}{12,825}k_3 + \frac{28,561}{56,430}k_4 - \frac{9}{50}k_5 + \frac{2}{55}k_6.$$

El tamaño de paso óptimo sh se determina, entonces, multiplicando el tamaño de paso en uso h por un escalar s dado por

$$(31) \quad s = \left(\frac{\tau h}{2|z_{k+1} - y_{k+1}|} \right)^{1/4} \approx 0.84 \left(\frac{\tau h}{|z_{k+1} - y_{k+1}|} \right)^{1/4}$$

siendo τ la tolerancia especificada para controlar el error.

La deducción de la fórmula (31) puede hallarse en libros de texto avanzados de cálculo numérico. Es importante aprender que el uso de un tamaño de paso

Tabla 9.10 Solución de $y' = 1 + y^2$, $y(0) = 0$ obtenida por el método RKF45.

k	t_k	Aproximación RK45 y_k	$y(t_k)$ exacta, $\tan(t_k)$	Error $y(t_k) - y_k$
0	0.0	0.0000000	0.0000000	0.0000000
1	0.2	0.2027100	0.2027100	0.0000000
2	0.4	0.4227933	0.4227931	-0.0000002
3	0.6	0.6841376	0.6841368	-0.0000008
4	0.8	1.0296434	1.0296386	-0.0000048
5	1.0	1.5574398	1.5774077	-0.0000321
6	1.1	1.9648085	1.9647597	-0.0000488
7	1.2	2.5722408	2.5721516	-0.0000892
8	1.3	3.6023295	3.6021024	-0.0002271
9	1.35	4.4555714	4.4552218	-0.0003496
10	1.4	5.7985045	5.7978837	-0.0006208

fijo no es la mejor estrategia aunque proporcione mejor apariencia a una tabla de valores; si es necesario conocer el valor de la solución en un punto que no está en la tabla, entonces se puede usar la técnica de interpolación polinomial.

Ejemplo 9.12. Vamos a comparar las soluciones obtenidas con los métodos RKF45 y RK4 para el problema de valor inicial

$$y' = 1 + y^2 \quad \text{con} \quad y(0) = 0 \quad \text{en} \quad [0, 1.4].$$

Usando un programa para el método RKF45, con el valor $\tau = 2 \times 10^{-5}$ para la tolerancia, se generan aproximaciones en 10 puntos que se muestran en la Tabla 9.10, donde podemos ver que el tamaño de paso ha ido cambiando automáticamente. Usando un programa para el método RK4, con tamaño de paso fijo $h = 0.1$, se generan 14 aproximaciones en nodos equiespaciados que se muestran en la Tabla 9.11. Las aproximaciones en el extremo derecho del intervalo para los métodos RKF45 y RK4 son, respectivamente,

$$y(1.4) \approx y_{10} = 5.7985045 \quad \text{e} \quad y(1.4) \approx y_{14} = 5.7919748$$

con errores

$$E_{10} = -0.0006208 \quad \text{y} \quad E_{14} = 0.0059089.$$

Puede observarse que el método RKF45 proporciona una aproximación mejor. ■

Tabla 9.11 Solución de $y' = 1 + y^2$, $y(0) = 0$ obtenida por el método RK4.

k	t_k	Aproximación RK4 y_k	$y(t_k)$ exacta, $\tan(t_k)$	Error $y(t_k) - y_k$
0	0.0	0.0000000	0.0000000	0.0000000
1	0.1	0.1003346	0.1003347	0.0000001
2	0.2	0.2027099	0.2027100	0.0000001
3	0.3	0.3093360	0.3093362	0.0000002
4	0.4	0.4227930	0.4227932	0.0000002
5	0.5	0.5463023	0.5463025	0.0000002
6	0.6	0.6841368	0.6841368	0.0000000
7	0.7	0.8422886	0.8422884	-0.0000002
8	0.8	1.0296391	1.0296386	-0.0000005
9	0.9	1.2601588	1.2601582	-0.0000006
10	1.0	1.5574064	1.5574077	0.0000013
11	1.1	1.9647466	1.9647597	0.0000131
12	1.2	2.5720718	2.5721516	0.0000798
13	1.3	3.6015634	3.6021024	0.0005390
14	1.4	5.7919748	5.7978837	0.0059089

MATLAB

Programa 9.4 (Método de Runge-Kutta de orden 4). Aproximación a la solución del problema de valor inicial $y' = f(t, y)$ con $y(a) = y_0$ en $[a, b]$ usando la fórmula

$$y_{k+1} = y_k + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4).$$

```
function R=rk4(f,a,b,ya,M)
% Datos
% - f es la función, almacenada como una
%   cadena de caracteres 'f'
% - a y b son los extremos derecho e izquierdo
%   del intervalo
% - ya es la condición inicial y(a)
% - M es el número de pasos
% Resultado
% - R=[T' Y'] siendo T el vector de las abscisas e
%   Y el vector de las ordenadas
h=(b-a)/M;
T=zeros(1,M+1);
```

```

Y=zeros(1,M+1);
T=a:h:b;
Y(1)=ya;
for j=1:M
    k1=h*feval(f,T(j),Y(j));
    k2=h*feval(f,T(j)+h/2,Y(j)+k1/2);
    k3=h*feval(f,T(j)+h/2,Y(j)+k2/2);
    k4=h*feval(f,T(j)+h,Y(j)+k3);
    Y(j+1)=Y(j)+(k1+2*k2+2*k3+k4)/6;
end
R=[T' Y'];

```

En el siguiente programa se implementa el método de Runge-Kutta-Fehlberg (RKF45) descrito en las expresiones (28) a (31).

Programa 9.5 (Método de Runge-Kutta-Fehlberg). Aproximación a la solución del problema de valor inicial $y' = f(t, y)$ con $y(a) = y_0$ en $[a, b]$ controlando el error para estimar el tamaño de paso.

```

function R=rkf45(f,a,b,ya,M,tol)
% Datos
%     - f es la función, almacenada como
%         una cadena de caracteres 'f'
%     - a y b son los extremos derecho e izquierdo
%         del intervalo
%     - ya es la condición inicial y(a)
%     - M es el número de pasos
%     - tol es la tolerancia
% Resultado
%     - R=[T' Y'] siendo T el vector de las abscisas e
%         Y el vector de las ordenadas
% Coeficientes necesarios para calcular las
% aproximaciones (28) y (29)
a2=1/4;b2=1/4;a3=3/8;b3=3/32;c3=9/32;a4=12/13;
b4=1932/2197;c4=-7200/2197;d4=7296/2197;a5=1;
b5=439/216;c5=-8;d5=3680/513;e5=-845/4104;a6=1/2;
b6=-8/27;c6=2;d6=-3544/2565;e6=1859/4104;
f6=-11/40;r1=1/360;r3=-128/4275;r4=-2197/75240;r5=1/50;
r6=2/55;n1=25/216;n3=1408/2565;n4=2197/4104;n5=-1/5;
big=1e15;
h=(b-a)/M;
hmin=h/64;
hmax=64*h;

```

```

max1=200;
Y(1)=ya;
T(1)=a;
j=1;
br=b-0.00001*abs(b);

while (T(j)<b)
    if ((T(j)+h)>br)
        h=b-T(j);
    end

% Cálculo de las aproximaciones (28) y (29)
k1=h*feval(f,T(j),Y(j));
y2=Y(j)+b2*k1;
if big<abs(y2)break,end
k2=h*feval(f,T(j)+a2*h,y2);
y3=Y(j)+b3*k1+c3*k2;
if big<abs(y3)break,end
k3=h*feval(f,T(j)+a3*h,y3);
y4=Y(j)+b4*k1+c4*k2+d4*k3;
if big<abs(y4)break,end
k4=h*feval(f,Y(j)+a4*h,y4);
y5=Y(j)+b5*k1+c5*k2+d5*k3+e5*k4;
if big<abs(y5)break,end
k5=h*feval(f,T(j)+a5*h,y5);
y6=Y(j)+b6*k1+c6*k2+d6*k3+e6*k4+f6*k5;
if big<abs(y6)break,end
k6=h*feval(f,Y(j)+a6*h,y6);

err=abs(r1*k1+r3*k3+r4*k4+r5*k5+r6*k6);
ynew=Y(j)+n1*k1+n3*k3+n4*k4+n5*k5;

% Control del error y del incremento
if((err<tol)|(h<2*hmin))
    Y(j+1)=ynew;
    if((T(j)+h)>br)
        T(j+1)=b;
    else
        T(j+1)=T(j)+h;
    end
    j=j+1;
end
if (err==0)
    s=0;
else
    s=0.84*(tol*h/err)^(0.25);

```

```

end
if((s<0.75)&(h>2*hmin))
    h=h/2;
end
if((s>1.50)&(2*h<hmax))
h=2*h;
end
if((big<abs(Y(j))|(max1==j)),break,end
M=j;
if (b>T(j))
    M=j+1;
else
    M=j;
end
end
R=[T' Y'];

```

Ejercicios

En los Ejercicios 1 a 5 resuelva la ecuación diferencial usando el método de Runge-Kutta de orden $N = 4$.

- Tome $h = 0.2$ y dé dos pasos calculando los valores a mano. Luego, tome $h = 0.1$ y dé cuatro pasos calculando los valores a mano.
 - Compare la solución exacta $y(0.4)$ con las dos aproximaciones calculadas en el apartado (a).
 - ¿Se comporta el error global final de las aproximaciones obtenidas en el apartado (a) como se espera cuando h se divide entre dos?
- $y' = t^2 - y$ con $y(0) = 1$, $y(t) = -e^{-t} + t^2 - 2t + 2$
 - $y' = 3y + 3t$ con $y(0) = 1$, $y(t) = \frac{4}{3}e^{3t} - t - \frac{1}{3}$
 - $y' = -ty$ con $y(0) = 1$, $y(t) = e^{-t^2/2}$
 - $y' = e^{-2t} - 2y$ con $y(0) = \frac{1}{10}$, $y(t) = \frac{1}{10}e^{-2t} + te^{-2t}$
 - $y' = 2ty^2$ con $y(0) = 1$, $y(t) = 1/(1 - t^2)$
 - Pruebe que cuando se usa el método de Runge-Kutta de orden $N = 4$ para resolver el problema de valor inicial $y' = f(t)$ en $[a, b]$ con $y(a) = 0$ el resultado es

$$y(b) \approx \frac{h}{6} \sum_{k=0}^{M-1} (f(t_k) + 4f(t_{k+1/2}) + f(t_{k+1})),$$

donde $h = (b - a)/M$, $t_k = a + kh$ y $t_{k+1/2} = a + (k + \frac{1}{2})h$, que es la aproximación dada por la regla de Simpson (con incremento $h/2$) a la integral definida de $f(t)$ en el intervalo $[a, b]$.

7. El método de mejora de Richardson, presentado en el Lema 7.1 (Sección 7.3) puede usarse en conjunción con el método de Runge-Kutta. Si en el método de Runge-Kutta de orden $N = 4$ se usa un tamaño de paso h , entonces

$$y(b) \approx y_h + Ch^4,$$

pero si usamos un tamaño de paso $2h$, entonces

$$y(b) \approx y_{2h} + 16Ch^4.$$

Los términos que contienen Ch^4 pueden eliminarse para obtener una aproximación mejorada a $y(b)$:

$$y(b) \approx \frac{16y_h - y_{2h}}{15}.$$

Este esquema de mejora puede usarse con los valores mostrados en el Ejemplo 9.11 para obtener una aproximación mejor a $y(3)$. Calcule las entradas que faltan en la tabla siguiente:

h	y_h	$(16y_h - y_{2h})/15$
1	1.6701860	_____
$\frac{1}{2}$	1.6694308	_____
$\frac{1}{4}$	1.6693928	_____
$\frac{1}{8}$	1.6693906	_____

Para los Ejercicios 8 y 9: el polinomio de Taylor de grado $N = 2$ de una función $f(t, y)$ de dos variables t e y desarrollado alrededor de (a, b) es

$$\begin{aligned} P_2(t, y) = & f(a, b) + f_t(a, b)(t - a) + f_y(a, b)(y - b) \\ & + \frac{f_{tt}(a, b)(t - a)^2}{2} + f_{ty}(a, b)(t - a)(y - b) + \frac{f_{yy}(a, b)(y - b)^2}{2}. \end{aligned}$$

8. (a) Halle el polinomio de Taylor de grado $N = 2$ de la función $f(t, y) = y/t$ desarrollado alrededor de $(1, 1)$.
 (b) Calcule $P_2(1.05, 1.1)$ y compare su valor con $f(1.05, 1.1)$.
9. (a) Determine el polinomio de Taylor de grado $N = 2$ de la función $f(t, y) = (1 + t - y)^{1/2}$ desarrollado alrededor de $(0, 0)$.
 (b) Calcule $P_2(0.04, 0.08)$ y compárelo con $f(0.04, 0.08)$.

Algoritmos y programas

En los Problemas 1 a 5, resuelva la ecuación diferencial usando el método de Runge-Kutta de orden $N = 4$.

- Tome $h = 0.1$ y dé 20 pasos con el Programa 9.4. Luego, tome $h = 0.05$ y dé 40 pasos con el Programa 9.4.
- Compare la solución exacta $y(2)$ con las dos aproximaciones obtenidas en el apartado (a).
- ¿Se comporta el error global final de las aproximaciones obtenidas en el apartado (a) como se espera cuando h se divide entre dos?
- Dibuje las aproximaciones y la solución exacta en una misma gráfica.

Indicación. La matriz R que se obtiene como resultado en el Programa 9.4 contiene las coordenadas x e y de las aproximaciones y la instrucción del paquete MATLAB `plot(R(:,1),R(:,2))` producirá un dibujo análogo al de la Figura 9.6.

- $y' = t^2 - y$ con $y(0) = 1$, $y(t) = -e^{-t} + t^2 - 2t + 2$
- $y' = 3y + 3t$ con $y(0) = 1$, $y(t) = \frac{4}{3}e^{3t} - t - \frac{1}{3}$
- $y' = -ty$ con $y(0) = 1$, $y(t) = e^{-t^2/2}$
- $y' = e^{-2t} - 2y$ con $y(0) = \frac{1}{10}$, $y(t) = \frac{1}{10}e^{-2t} + te^{-2t}$
- $y' = 2ty^2$ con $y(0) = 1$, $y(t) = 1/(1 - t^2)$

En los Problemas 6 y 7, resuelva la ecuación usando el método de Runge-Kutta-Fehlberg.

- Use el Programa 9.5 con un tamaño de paso inicial $h = 0.1$ y con una tolerancia $\tau = 10^{-7}$.
 - Compare la solución exacta $y(b)$ con la aproximación obtenida.
 - Dibuje las aproximaciones y la solución en una misma gráfica.
- $y' = 9te^{3t}$, $y(0) = 0$ en $[0, 3]$, $y(t) = 3te^{3t} - e^{3t} + 1$
 - $y' = 2 \tan^{-1}(t)$, $y(0) = 0$ en $[0, 1]$, $y(t) = 2t \tan^{-1}(t) - \ln(1 + t^2)$
 - En una reacción química, una molécula de una sustancia A se combina con una molécula de una sustancia B para formar una molécula de una sustancia C. Se sabe que la concentración $y(t)$ de la sustancia C en el instante t es la solución del problema de valor inicial

$$y' = k(a - y)(b - y) \quad \text{con} \quad y(0) = 0,$$

donde k es una constante positiva y a y b son las concentraciones iniciales de las sustancias A y B, respectivamente. Supongamos que $k = 0.01$, $a = 70$ milímoles/litro y $b = 50$ milímoles/litro. Use el método de Runge-Kutta de orden $N = 4$ con $h = 0.5$ para hallar la solución en el intervalo $[0, 20]$.
Observación. Compare la solución dada por el computador con la solución

exacta $y(t) = 350(1 - e^{-0.2t})/(7 - 5e^{-0.2t})$. Observe que el valor límite cuando $t \rightarrow +\infty$ es 50.

9. Resolviendo un problema de valor inicial adecuado, haga una tabla de valores de la función de distribución normal $f(t)$ definida por la integral:

$$f(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2} dt \quad \text{para } 0 \leq x \leq 3.$$

Utilice en sus cálculos el método de Runge-Kutta de orden $N = 4$ con $h = 0.1$. Su solución debería coincidir con los valores que se muestran en la siguiente tabla. *Observación.* Esta es una manera bastante aceptable de generar una tabla de valores de la función de distribución normal estándar.

x	$f(x)$
0.0	0.5
0.5	0.6914625
1.0	0.8413448
1.5	0.9331928
2.0	0.9772499
2.5	0.9937903
3.0	0.9986501

10. (a) Escriba un programa para el método de mejora de Richardson descrito en el Ejercicio 7.
 (b) Use su programa del apartado (a) para aproximar $y(0.8)$ en el problema de valor inicial $y' = t^2 + y^2$, $y(0) = 1$ en $[0, 0.8]$. Se sabe que el valor de la solución exacta en $t = 0.8$ es $y(0.8) = 5.8486168$. Empiece con un tamaño de paso $h = 0.05$ y utilice como criterio de parada el que el valor absoluto de la diferencia entre dos mejoras de Richardson consecutivas sea menor que 10^{-7} .

11. Considere la ecuación íntegro-diferencial de primer orden:

$$y' = 1.3y - 0.25y^2 - 0.0001y \int_0^t y(\tau) d\tau.$$

- (a) Use el método de Runge-Kutta de orden $N = 4$ con $h = 0.2$, el valor inicial $y(0) = 250$ en el intervalo $[0, 20]$ y la regla del trapecio para calcular una solución aproximada de la ecuación (vea el Problema 10 de la subsección “Algoritmos y programas” de la Sección 9.2).
 (b) Repita el apartado (a) usando como valores iniciales $y(0) = 200$ e $y(0) = 300$.
 (c) Dibuje las soluciones aproximadas calculadas en los apartados (a) y (b) sobre una misma gráfica.

9.6 Métodos de predicción y corrección

Los métodos de Euler, Heun, Taylor y Runge-Kutta se llaman **métodos de paso simple**, o **métodos de un sólo paso** porque en el cálculo de cada punto sólo utilizan la información del punto previo; esto es, únicamente se usa el punto inicial (t_0, y_0) para calcular (t_1, y_1) y, en general, sólo se necesita conocer y_k para calcular y_{k+1} . Sin embargo, una vez calculados varios puntos, sería posible usar varios de los puntos que ya tenemos en el cálculo del punto siguiente. Para ilustrar esta técnica vamos a desarrollar el método de cuatro pasos de Adams-Bashforth, en el que se necesitan los valores y_{k-3} , y_{k-2} , y_{k-1} e y_k para calcular y_{k+1} . Naturalmente, al comienzo necesitaremos conocer por adelantado cuatro puntos (t_0, y_0) , (t_1, y_1) , (t_2, y_2) y (t_3, y_3) para poder generar la sucesión $\{(t_k, y_k) : k \geq 4\}$ (esto puede hacerse dando tres pasos con alguno de los métodos de paso simple).

Un aspecto atractivo de los métodos multipaso es que se puede estimar el valor del error de truncamiento local y luego incluir un término de corrección, lo cual mejora la precisión de la respuesta en cada paso. Asimismo, es posible determinar si el tamaño de paso es suficientemente pequeño como para obtener una buena aproximación de y_{k+1} y, a la vez, suficientemente grande como para evitar cálculos innecesarios que consuman tiempo de computación. El uso de la combinación de un valor predictor con otro corrector sólo requiere que se realicen dos evaluaciones de la función $f(t, y)$ en cada paso.

El método de Adams-Bashforth-Moulton

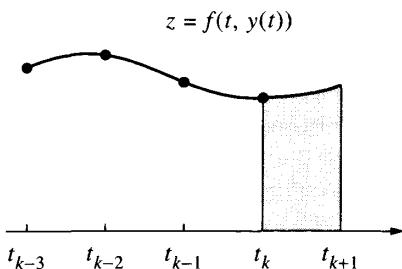
El método de predicción y corrección (o predictor-corrector) de Adams-Bashforth-Moulton es un método multipaso que se genera a partir del teorema fundamental del cálculo:

$$(1) \quad y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(t, y(t)) dt.$$

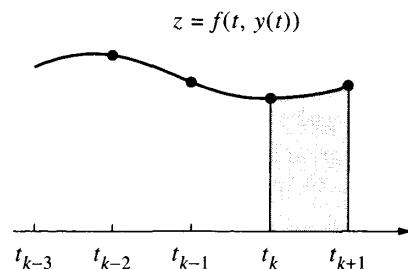
Para calcular el valor predictor se utiliza el polinomio de interpolación de Lagrange de $f(t, y(t))$ que pasa por los puntos (t_{k-3}, f_{k-3}) , (t_{k-2}, f_{k-2}) , (t_{k-1}, f_{k-1}) y (t_k, f_k) . Integrando este polinomio en el intervalo $[t_k, t_{k+1}]$ y sustituyendo el resultado en la expresión (1) obtenemos lo que se conoce como valor predictor de Adams-Bashforth:

$$(2) \quad p_{k+1} = y_k + \frac{h}{24}(-9f_{k-3} + 37f_{k-2} - 59f_{k-1} + 55f_k).$$

El valor corrector se calcula de manera parecida, sólo que ahora usamos el predictor p_{k+1} que acabamos de calcular: Construimos un segundo polinomio



(a) Los cuatro nodos para el valor predictor de Adams-Basforth (se extrae).



(a) Los cuatro nodos para el valor corrector de Adams-Moulton (se interpola).

Figura 9.10 Integraciones en el intervalo $[t_k, t_{k+1}]$ en el método de Adams-Basforth-Moulton.

interpolador de Lagrange de $f(t, y(t))$, el que pasa por (t_{k-2}, f_{k-2}) , (t_{k-1}, f_{k-1}) , (t_k, f_k) y también por el nuevo punto $(t_{k+1}, f_{k+1}) = (t_{k+1}, f(t_{k+1}, p_{k+1}))$. Integrando este polinomio en el intervalo $[t_k, t_{k+1}]$ y sustituyendo el resultado en (1) obtenemos lo que se conoce como valor corrector de Adams-Moulton:

$$(3) \quad y_{k+1} = y_k + \frac{h}{24}(f_{k-2} - 5f_{k-1} + 19f_k + 9f_{k+1}).$$

La Figura 9.10 muestra los nodos de interpolación para los polinomios de Lagrange que se usan en el desarrollo de las fórmulas (2) y (3), respectivamente.

Estimación del error y corrección

Los términos del error de las fórmulas de integración numérica que se usan para obtener los valores predictor y corrector son de orden $O(h^5)$. Los errores de truncamiento local de las fórmulas (2) y (3) son, de hecho,

$$(4) \quad y(t_{k+1}) - p_{k+1} = \frac{251}{720}y^{(5)}(c_{k+1})h^5 \quad (\text{para el valor predictor}),$$

$$(5) \quad y(t_{k+1}) - y_{k+1} = \frac{-19}{720}y^{(5)}(d_{k+1})h^5 \quad (\text{para el valor corrector}).$$

Supongamos que h es pequeño y que $y^{(5)}(t)$ es prácticamente constante en el intervalo; entonces podemos eliminar los términos de las fórmulas (4) y (5) que contienen la derivada quinta para obtener

$$(6) \quad y(t_{k+1}) - y_{k+1} \approx \frac{-19}{270}(y_{k+1} - p_{k+1}).$$

La razón de la importancia del método de predicción y corrección se ve ahora claramente: La fórmula (6) proporciona una estimación aproximada del error que se basa en dos valores calculados p_{k+1} e y_{k+1} y que no requiere conocer la derivada quinta $y^{(5)}(t)$.

Consideraciones prácticas

Para calcular y_{k+1} en la fórmula del valor corrector (3) se usa la estimación $f_{k+1} \approx f(t_{k+1}, p_{k+1})$. Puesto que y_{k+1} es también una aproximación de $y(t_{k+1})$, podríamos emplear ahora este valor en la fórmula del corrector (3) para generar una nueva aproximación f_{k+1} que, a su vez, generaría un nuevo valor y_{k+1} . Sin embargo, cuando se repite esta iteración sobre el valor corrector, lo que ocurre es que los valores calculados convergen a un punto fijo de (3) más que a la solución de la ecuación diferencial. Si es necesario conseguir una precisión mayor, entonces es más eficiente reducir el tamaño de paso.

También podemos usar la fórmula (6) para determinar cuándo debe cambiar el tamaño de paso. Aunque hay otros métodos más elaborados, vamos a mostrar cómo decidir si dejamos el tamaño de paso como está, lo reducimos a la mitad $h/2$ o lo aumentamos al doble $2h$. Por ejemplo, sea $\varepsilon = 5 \times 10^{-6}$ nuestra tolerancia para el error relativo y tomemos $\rho = 10^{-5}$:

$$(7) \quad \text{Si } \frac{19}{270} \frac{|y_{k+1} - p_{k+1}|}{|y_{k+1}| + \rho} > \varepsilon, \quad \text{entonces se toma } h = \frac{h}{2}.$$

$$(8) \quad \text{Si } \frac{19}{270} \frac{|y_{k+1} - p_{k+1}|}{|y_{k+1}| + \rho} < \frac{\varepsilon}{100}, \quad \text{entonces se toma } h = 2h.$$

Cuando los valores predictor y corrector no coinciden en al menos cinco cifras significativas, entonces el criterio (7) nos dice que debemos reducir el tamaño de paso; mientras que si coinciden en siete o más cifras significativas, entonces el criterio (8) nos dice que debemos aumentar el tamaño de paso. (Le aconsejamos que realice un ajuste fino de estos parámetros ε y ρ para adecuarlos a su computador.)

Al reducir el tamaño de paso, hacen falta cuatro valores de partida nuevos. Para ello, interpolamos $f(t, y(t))$ mediante un polinomio de cuarto grado que nos aproxima los valores, que no teníamos antes, de dicha función en los puntos medios de los intervalos $[t_{k-2}, t_{k-1}]$ y $[t_{k-1}, t_k]$. Los cuatro nodos $t_{k-3/2}$, t_{k-1} , $t_{k-1/2}$ y t_k que se usan en los cálculos posteriores se muestran en la Figura 9.11. Estos valores interpolados necesarios para disponer de cuatro valores de partida separados por un tamaño de paso $h/2$ son

$$(9) \quad f_{k-1/2} = \frac{-5f_{k-4} + 28f_{k-3} - 70f_{k-2} + 140f_{k-1} + 35f_k}{128},$$

$$f_{k-3/2} = \frac{3f_{k-4} - 20f_{k-3} + 90f_{k-2} + 60f_{k-1} - 5f_k}{128}.$$

Nodos nuevos →

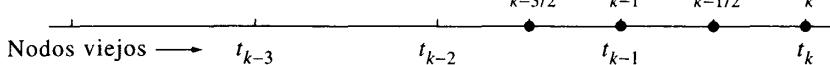


Figura 9.11 Reducción del tamaño de paso a $h/2$ en un método adaptativo.

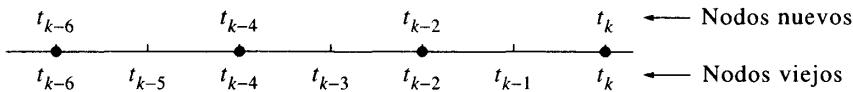


Figura 9.12 Aumento del tamaño de paso a $2h$ en un método adaptativo.

Aumentar el tamaño de paso al doble no presenta tantas complicaciones: tomando los siete puntos anteriores obtenemos los cuatro necesarios omitiendo uno de cada dos, como se muestra en la Figura 9.12.

El método de Milne-Simpson

Otro esquema popular de predicción y corrección es el llamado método de Milne-Simpson. Su valor predictor se construye integrando numéricamente $f(t, y(t))$ en el intervalo $[t_{k-3}, t_{k+1}]$:

$$(10) \quad y(t_{k+1}) = y(t_{k-3}) + \int_{t_{k-3}}^{t_{k+1}} f(t, y(t)) dt;$$

para ello se utiliza el polinomio de interpolación de Lagrange de $f(t, y(t))$ que pasa por los puntos (t_{k-3}, f_{k-3}) , (t_{k-2}, f_{k-2}) , (t_{k-1}, f_{k-1}) y (t_k, f_k) que, al integrarlo en $[t_{k-3}, t_{k+1}]$, proporciona la fórmula para lo que se conoce como predictor de Milne:

$$(11) \quad p_{k+1} = y_{k-3} + \frac{4h}{3}(2f_{k-2} - f_{k-1} + 2f_k).$$

El valor corrector se calcula de manera parecida, sólo que ahora usamos también el predictor p_{k+1} que acabamos de calcular: Construimos un segundo polinomio interpolador de Lagrange de $f(t, y(t))$, el que pasa por (t_{k-1}, f_{k-1}) , (t_k, f_k) y también por el nuevo punto $(t_{k+1}, f_{k+1}) = (t_{k+1}, f(t_{k+1}, p_{k+1}))$. Integrando este polinomio en el intervalo $[t_{k-1}, t_{k+1}]$ obtenemos el valor corrector mediante la ya familiar regla de Simpson:

$$(12) \quad y_{k+1} = y_{k-1} + \frac{h}{3}(f_{k-1} + 4f_k + f_{k+1}).$$

Estimación del error y corrección

Los términos del error de las fórmulas de integración numérica que se usan para obtener los valores predictor y corrector en el método de Milne-Simpson son de orden $O(h^5)$. Los errores de truncamiento local de las fórmulas (11) y (12) son, de hecho,

$$(13) \quad y(t_{k+1}) - p_{k+1} = \frac{28}{90}y^{(5)}(c_{k+1})h^5 \quad (\text{para el valor predictor}),$$

$$(14) \quad y(t_{k+1}) - y_{k+1} = \frac{-1}{90}y^{(5)}(d_{k+1})h^5 \quad (\text{para el valor corrector}).$$

Supongamos que h es pequeño y que $y^{(5)}(t)$ es prácticamente constante en el intervalo $[t_{k-3}, t_{k+1}]$; entonces podemos eliminar los términos de las fórmulas (13) y (14) que contienen la derivada quinta para obtener

$$(15) \quad y(t_{k+1}) - p_{k+1} \approx \frac{28}{29}(y_{k+1} - p_{k+1}).$$

La fórmula (15) proporciona una estimación aproximada del error que se basa en dos valores calculados p_{k+1} e y_{k+1} y que no requiere conocer la derivada quinta $y^{(5)}(t)$. Podemos usar esta fórmula para mejorar el valor del predictor: Si suponemos que la diferencia entre los valores predictores y correctores cambia lentamente de un paso a otro, entonces hacemos jugar a p_k e y_k el papel de p_{k+1} e y_{k+1} en (15) para obtener la siguiente fórmula modificada:

$$(16) \quad m_{k+1} = p_{k+1} + 28\frac{y_k - p_k}{29}.$$

Ahora usamos este valor modificado en la fórmula de corrección en p_{k+1} , con lo que la ecuación (12) queda como

$$(17) \quad y_{k+1} = y_{k-1} + \frac{h}{3}(f_{k-1} + 4f_k + f(t_{k+1}, m_{k+1})).$$

Podemos resumir, entonces, el método de Milne-Simpson mejorado (o modificado) en los siguientes cálculos:

$$p_{k+1} = y_{k-3} + \frac{4h}{3}(2f_{k-2} - f_{k-1} + 2f_k) \quad (\text{valor predictor})$$

$$(18) \quad m_{k+1} = p_{k+1} + 28\frac{y_k - p_k}{29} \quad (\text{valor modificador})$$

$$f_{k+1} = f(t_{k+1}, m_{k+1})$$

$$y_{k+1} = y_{k-1} + \frac{h}{3}(f_{k-1} + 4f_k + f_{k+1}) \quad (\text{valor corrector}).$$

Otro procedimiento de predicción y corrección muy utilizado es el método de Hamming. Aunque no vamos a mostrar su desarrollo, al final de la sección

Tabla 9.12 Comparación de los métodos de Adams-Bashforth-Moulton, Milne-Simpson y Hamming para resolver $y' = (t - y)/2$, $y(0) = 1$.

<i>k</i>	Adams-Bashforth-Moulton	Error	Milne-Simpson	Error	Método de Hamming	Error
0.0	1.00000000	0E-8	1.00000000	0E-8	1.00000000	0E-8
0.5	0.83640227	8E-8	0.83640231	4E-8	0.83640234	1E-8
0.625	0.81984673	16E-8	0.81984687	2E-8	0.81984688	1E-8
0.75	0.81186762	22E-8	0.81186778	6E-8	0.81186783	1E-8
0.875	0.81194530	28E-8	0.81194555	3E-8	0.81194558	0E-8
1.0	0.81959166	32E-8	0.81959190	8E-8	0.81959198	0E-8
1.5	0.91709920	46E-8	0.91709957	9E-8	0.91709967	-1E-8
2.0	1.10363781	51E-8	1.10363822	10E-8	1.10363834	-2E-8
2.5	1.35951387	52E-8	1.35951429	10E-8	1.35951441	-2E-8
2.625	1.43243853	52E-8	1.43243899	6E-8	1.43243907	-2E-8
2.75	1.50851827	52E-8	1.50851869	10E-8	1.50851881	-2E-8
2.875	1.58756195	51E-8	1.58756240	6E-8	1.58756248	-2E-8
3.0	1.66938998	50E-8	1.66939038	10E-8	1.66939050	-2E-8

sí daremos un programa para poder utilizarlo con el paquete MATLAB. Como advertencia final, mencionemos que todos los métodos de predicción y corrección presentan problemas de estabilidad. El estudio de la estabilidad de los métodos numéricos es un tema avanzado que, no obstante, no debe descuidar ninguna persona que desee profundizar seriamente en el análisis numérico.

Ejemplo 9.13. Vamos a usar los métodos de Adams-Bashforth-Moulton, Milne-Simpson y Hamming con $h = \frac{1}{8}$ para calcular aproximaciones a la solución del problema de valor inicial

$$y' = \frac{t - y}{2}, \quad y(0) = 1 \quad \text{en } [0, 3].$$

Usando uno de los métodos de Runge-Kutta, calculamos los valores iniciales necesarios

$$y_1 = 0.94323919, \quad y_2 = 0.89749071 \quad \text{e} \quad y_3 = 0.86208736.$$

Usando los Programas 9.6 a 9.8 con un computador obtenemos los valores que se recogen en la Tabla 9.12. En esta tabla los errores de cada aproximación se dan como múltiplos de 10^{-8} y puede observarse que todas las aproximaciones tienen una precisión de seis cifras decimales, por lo menos, y también que el método de Hamming es el que produce las mejores aproximaciones en este caso. ■

El tamaño de paso adecuado

Hay varias razones que justifican la selección de los métodos que hemos presentado: en primer lugar, porque su desarrollo es accesible para estudiantes de un primer curso; en segundo lugar, porque métodos más avanzados se desarrollan de manera similar y en tercer lugar, porque la mayoría de los problemas de aplicación que pueden presentarse en otras asignaturas de las carreras científico-técnicas pueden resolverse con alguno de estos métodos. Sin embargo, cuando se usa un método de predicción y corrección para resolver un problema de valor inicial $y' = f(t, y)$, con $y(t_0) = y_0$, en un intervalo grande, entonces pueden aparecer ciertos problemas.

Si $f_y(t, y) < 0$ y el tamaño de paso es demasiado grande, entonces los métodos de predicción y corrección pueden ser inestables. Como norma general, la estabilidad se tiene cuando un error pequeño se propaga de manera decreciente, mientras que la inestabilidad se produce cuando un error pequeño se propaga de manera creciente. Cuando se utiliza un tamaño de paso demasiado grande en un intervalo grande, suele producirse inestabilidad, que se manifiesta usualmente mediante la aparición de oscilaciones en la solución calculada. Este fenómeno puede atenuarse cambiando el tamaño de paso por uno menor, siguiendo el procedimiento sugerido por las fórmulas (7) a (9); para incluir el control del tamaño de paso en un algoritmo, las estimaciones del error correspondientes son:

$$(19) \quad y(t_k) - y_k \approx 19 \frac{p_k - y_k}{270} \quad (\text{Adams-Bashforth-Moulton}),$$

$$(20) \quad y(t_k) - y_k \approx \frac{p_k - y_k}{29} \quad (\text{Milne-Simpson}),$$

$$(21) \quad y(t_k) - y_k \approx 9 \frac{p_k - y_k}{121} \quad (\text{Hamming}).$$

En todos los métodos, el valor corrector se calcula dando un sólo paso en un esquema de iteración punto fijo y puede probarse que el método converge si el tamaño de paso h verifica, en cada caso, la condición:

$$(22) \quad h \ll \frac{2.66667}{|f_y(t, y)|} \quad (\text{Adams-Bashforth-Moulton}),$$

$$(23) \quad h \ll \frac{3.00000}{|f_y(t, y)|} \quad (\text{Milne-Simpson}),$$

$$(24) \quad h \ll \frac{2.66667}{|f_y(t, y)|} \quad (\text{Hamming}),$$

donde la notación \ll en las fórmulas (22) a (24) significa “mucho más pequeño que”. El siguiente ejemplo muestra que hay que usar desigualdades más restric-

tivas, a saber:

$$(25) \quad h < \frac{0.75}{|f_y(t, y)|} \quad (\text{Adams-Bashforth-Moulton}),$$

$$(26) \quad h < \frac{0.45}{|f_y(t, y)|} \quad (\text{Milne-Simpson}),$$

$$(27) \quad h < \frac{0.69}{|f_y(t, y)|} \quad (\text{Hamming}).$$

Las desigualdades (25)–(27) pueden encontrarse en textos más avanzados de cálculo numérico, veremos su utilidad en el ejemplo.

Ejemplo 9.14. Vamos a usar los métodos de Adams-Bashforth-Moulton, Milne-Simpson y Hamming para calcular aproximaciones a la solución de

$$y' = 30 - 5y, \quad y(0) = 1 \quad \text{en el intervalo } [0, 10].$$

Los tres métodos son de orden $O(h^4)$ y cuando se usan $N = 120$ pasos, el error máximo en cada método se produce en puntos distintos

$$y(0.41666667) - y_5 \approx -0.00277037 \quad (\text{Adams-Bashforth-Moulton}),$$

$$y(0.33333333) - y_4 \approx -0.00139255 \quad (\text{Milne-Simpson}),$$

$$y(0.33333333) - y_4 \approx -0.00104982 \quad (\text{Hamming}).$$

En el extremo derecho $t = 10$, el error es

$$y(10) - y_{120} \approx 0.00000000 \quad (\text{Adams-Bashforth-Moulton}),$$

$$y(10) - y_{120} \approx 0.00001015 \quad (\text{Milne-Simpson}),$$

$$y(10) - y_{120} \approx 0.00000000 \quad (\text{Hamming}).$$

Los métodos de Adams-Bashforth-Moulton y de Hamming producen soluciones aproximadas con ocho cifras significativas correctas en el extremo derecho.

Resulta muy instructivo observar que cuando el tamaño de paso es demasiado grande, la solución calculada presenta oscilaciones en torno a la solución exacta; este fenómeno se muestra en las Figuras 9.13 (a) y (b). El número de pasos más bajo fue determinado experimentalmente para que las oscilaciones resultaran ser de la misma magnitud, mientras que el número mayor de pasos que es necesario dar para atenuar las oscilaciones se obtuvo aplicando las fórmulas (25)–(27).

MATLAB

Cada uno de los tres programas que damos a continuación requiere que las cuatro primeras coordenadas de T e Y sean valores iniciales calculados con otro

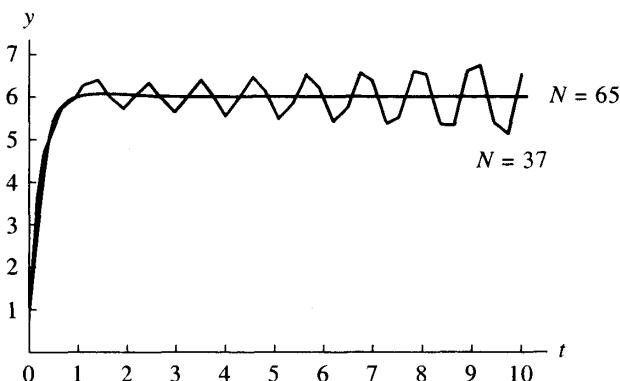


Figura 9.13 (a) La solución de $y' = 30 - 5y$ obtenida mediante el método de Adams-Bashforth-Moulton con $N = 37$ pasos presenta oscilaciones; se estabiliza cuando $N = 65$ porque $h = 10/65 = 0.1538 \approx 0.15 = 0.75/5 = 0.75/|f_y(t, y)|$.

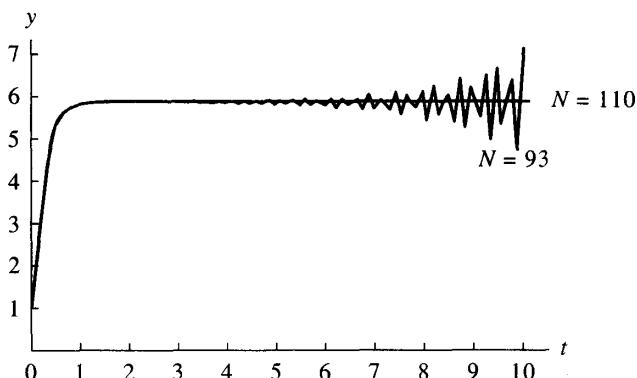


Figura 9.13 (b) La solución de $y' = 30 - 5y$ obtenida mediante el método de Milne-Simpson con $N = 93$ pasos presenta oscilaciones; se estabiliza cuando $N = 110$ porque $h = 10/110 = 0.0909 \approx 0.09 = 0.45/5 = 0.45/|f_y(t, y)|$.

método. Si consideramos el Ejemplo 9.13, donde el tamaño de paso es $h = \frac{1}{8}$ y el intervalo $[0, 3]$, entonces la siguiente secuencia de instrucciones del paquete MATLAB produce los vectores iniciales T e Y adecuados.

```
>>T=zeros(1,25); Y=zeros(1,25);
>>T=0:1/8:3; Y(1:4)=[1 0.94323919 0.89749071 0.86208736];
```

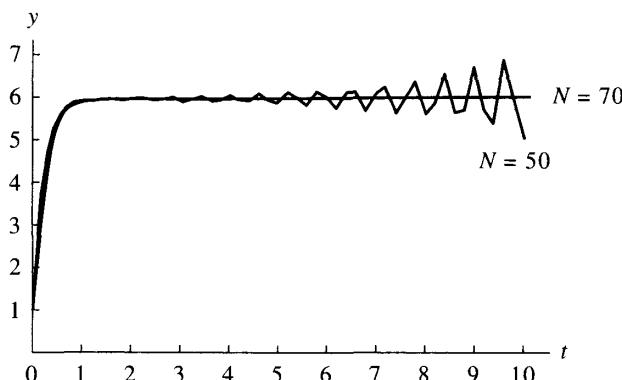


Figura 9.13 (c) La solución de $y' = 30 - 5y$ obtenida mediante el método de Hamming con $N = 50$ pasos presenta oscilaciones; se estabiliza cuando $N = 70$ porque $h = 10/70 = 0.1428 \approx 0.138 = 0.69/5 = 0.69/|f_y(t, y)|$.

Programa 9.6 (Método de Adams-Bashforth-Moulton). Construcción de las aproximaciones a la solución del problema de valor inicial $y' = f(t, y)$ con $y(a) = y_0$ en $[a, b]$ usando el valor predictor

$$p_{k+1} = y_k + \frac{h}{24}(-9f_{k-3} + 37f_{k-2} - 59f_{k-1} + 55f_k)$$

y el valor corrector

$$y_{k+1} = y_k + \frac{h}{24}(f_{k-2} - 5f_{k-1} + 19f_k + 9f_{k+1}).$$

```

function A=abm(f,T,Y)
% Datos
%      - f es la función, almacenada como
%          una cadena de caracteres 'f'
%      - T es el vector de las abscisas; su dimensión
%          es el número de pasos
%      - Y es el vector de las ordenadas
% Observación
%      Las cuatro primeras coordenadas de T e Y
%      deben contener valores iniciales
%      calculados con el método RK4
% Resultado
%      - A=[T' Y'] siendo T el vector de las abscisas e

```

```

% Y el vector de las ordenadas
n=length(T);
if n<5, break, end;
F=zeros(1,4);
F=feval(f,T(1:4),Y(1:4));
h=T(2)-T(1);
for k=4:n-1
    % Predictor
    p=Y(k)+(h/24)*(F*[-9 37 -59 55]');
    T(k+1)=T(1)+h*k;
    F=[F(2) F(3) F(4) feval(f,T(k+1),p)];
    % Corrector
    Y(k+1)=Y(k)+(h/24)*(F*[1 -5 19 9]');
    F(4)=feval(f,T(k+1),Y(k+1));
end
A=[T' Y'];

```

Programa 9.7 (Método de Milne-Simpson modificado). Construcción de las aproximaciones a la solución del problema de valor inicial $y' = f(t, y)$ con $y(a) = y_0$ en $[a, b]$ usando el valor predictor

$$p_{k+1} = y_{k-3} + \frac{4h}{3}(2f_{k-2} - f_{k-1} + 2f_k),$$

la modificación

$$m_{k+1} = p_{k+1} + 28 \frac{y_k - p_k}{29} \quad f_{k+1} = f(t_{k+1}, m_{k+1})$$

y el valor corrector

$$y_{k+1} = y_{k-1} + \frac{h}{3}(f_{k-1} + 4f_k + f_{k+1}).$$

```

function M=milne(f,T,Y)
% Datos
% - f es la función, almacenada como
% una cadena de caracteres 'f'
% - T es el vector de las abscisas; su dimensión
% es el número de pasos
% - Y es el vector de las ordenadas
% Observación
% Las cuatro primeras coordenadas de T e Y
% deben contener valores iniciales
% calculados con el método RK4

```

```
% Resultado
% - M=[T' Y'] siendo T el vector de las abscisas e
%     Y el vector de las ordenadas
n=length(T);
if n<5, break, end;
F=zeros(1,4);
F=feval(f,T(1:4),Y(1:4));
h=T(2)-T(1);
pold=0;
yold=0;
for k=4:n-1
    % Predictor
    pnew=Y(k-3)+(4*h/3)*(F(2:4)*[2 -1 2]');
    % Modificador
    pmod=pnew+28*(yold-pold)/29;
    T(k+1)=T(1)+h*k;
    F=[F(2) F(3) F(4) feval(f,T(k+1),pmod)];
    % Corrector
    Y(k+1)=Y(k-1)+(h/3)*(F(2:4)*[1 4 1]');
    pold=pnew;
    yold=Y(k+1);
    F(4)=feval(f,T(k+1),Y(k+1));
end
M=[T' Y'];

```

Programa 9.8 (Método de Hamming). Construcción de las aproximaciones a la solución del problema de valor inicial $y' = f(t, y)$ con $y(a) = y_0$ en $[a, b]$ usando el valor predictor

$$p_{k+1} = y_{k-3} + \frac{4h}{3}(2f_{k-2} - f_{k-1} + 2f_k),$$

una modificación y el valor corrector

$$y_{k+1} = \frac{-y_{k-2} + 9y_k}{8} + \frac{3h}{8}(-f_{k-1} + 2f_k + f_{k+1}).$$

```
function H=hamming(f,T,Y)
% Datos
% - f es la función, almacenada como
%     una cadena de caracteres 'f'
% - T es el vector de las abscisas; su dimensión
%     es el número de pasos
% - Y es el vector de las ordenadas
```

```
% Observación
% Las cuatro primeras coordenadas de T e Y
% deben contener valores iniciales
% calculados con el método RK4
% Resultado
% - H=[T' Y'] siendo T el vector de las abscisas e
% Y el vector de las ordenadas
n=length(T);
if n<5, break, end;
F=zeros(1,4);
F=feval(f,T(1:4),Y(1:4));
h=T(2)-T(1);
pold=0;
cold=0;
for k=4:n-1
    % Predictor
    pnew=Y(k-3)+(4*h/3)*(F(2:4)*[2 -1 2]');
    % Modificador
    pmod=pnew+112*(cold-pold)/121;
    T(k+1)=T(1)+h*k;
    F=[F(2) F(3) F(4) feval(f,T(k+1),pmod)];
    % Corrector
    cnew=(9*Y(k)-Y(k-2)+3*h*(F(2:4)*[-1 2 1]'))/8;
    Y(k+1)=cnew+9*(pnew-cnew)/121;
    pold=pnew;
    cold=cnew;
    F(4)=feval(f,T(k+1),Y(k+1));
end
H=[T' Y'];
```

Ejercicios

En los Ejercicios 1 a 3, use el método de Adams-Bashforth-Moulton, con valores iniciales y_1 , y_2 e y_3 y tamaño de paso $h = 0.05$, para calcular a mano los valores y_4 e y_5 en el problema de valor inicial dado. Compare su solución con la solución exacta $y(t)$ que se proporciona.

1. $y' = t^2 - y$, $y(0) = 1$ en $[0, 5]$, $y(t) = -e^{-t} + t^2 - 2t + 2$
 $y(0.05) = 0.95127058$, $y(0.10) = 0.90516258$, $y(0.15) = 0.86179202$
2. $y' = y + 3t - t^2$, $y(0) = 1$ en $[0, 5]$, $y(t) = 2e^t + t^2 - t - 1$
 $y(0.05) = 1.0550422$, $y(0.10) = 1.1203418$, $y(0.15) = 1.1961685$

3. $y' = -t/y$, $y(1) = 1$ en $[1, 1.4]$, $y(t) = (2 - t^2)^{1/2}$

$$y(1.05) = 0.94736477, y(1.10) = 0.88881944, y(1.15) = 0.82310388$$

En los Ejercicios 4 a 6, use el método de Milne-Simpson, con valores iniciales y_1 , y_2 e y_3 y tamaño de paso $h = 0.05$, para calcular a mano los valores y_4 e y_5 en el problema de valor inicial dado. Compare su solución con la solución exacta $y(t)$ que se proporciona.

4. $y' = e^{-t} - y$, $y(0) = 1$ en $[0, 5]$, $y(t) = te^{-t} + e^{-t}$

$$y(0.05) = 0.99879090, y(0.10) = 0.99532116, y(0.15) = 0.98981417$$

5. $y' = 2ty^2$, $y(0) = 1$ en $[0, 0.95]$, $y(t) = 1/(1 - t^2)$

$$y(0.05) = 1.0025063, y(0.10) = 1.0101010, y(0.15) = 1.0230179$$

6. $y' = 1 + y^2$, $y(0) = 1$ en $[0, 0.75]$, $y(t) = \tan(t + \pi/4)$

$$y(0.05) = 1.1053556, y(0.10) = 1.2230489, y(0.15) = 1.3560879$$

En los Ejercicios 7 a 9, use el método de Hamming, con valores iniciales y_1 , y_2 e y_3 y tamaño de paso $h = 0.05$, para calcular a mano los valores y_4 e y_5 en el problema de valor inicial dado. Compare su solución con la solución exacta $y(t)$ que se proporciona.

7. $y' = 2y - y^2$, $y(0) = 1$ en $[0, 5]$, $y(t) = 1 + \tanh(t)$

$$y(0.05) = 1.0499584, y(0.10) = 1.0996680, y(0.15) = 1.1488850$$

8. $y' = (1 - y^2)^{1/2}$, $y(0) = 0$ en $[0, 1.55]$, $y(t) = \sin(t)$

$$y(0.05) = 0.049979169, y(0.10) = 0.099833417, y(0.15) = 0.14943813$$

9. $y' = y^2 \sin(t)$, $y(0) = 1$ en $[0, 1.55]$, $y(t) = \sec(t)$

$$y(0.05) = 1.0012513, y(0.10) = 1.0050209, y(0.15) = 1.0113564$$

Algoritmos y programas

- 1. (a)** Use el Programa 9.6 para resolver las ecuaciones diferenciales de los Ejercicios 1 a 3.
(b) Dibuje su aproximación y la solución exacta en una misma gráfica.
- 2. (a)** Use el Programa 9.7 para resolver las ecuaciones diferenciales de los Ejercicios 4 a 6.
(b) Dibuje su aproximación y la solución exacta en una misma gráfica.
- 3. (a)** Use el Programa 9.8 para resolver las ecuaciones diferenciales de los Ejercicios 7 a 9.
(b) Dibuje su aproximación y la solución exacta en una misma gráfica.

4. Realice una gráfica análoga a la de la Figura 9.13 usando el Programa 9.6 con $N = 37$ y $N = 65$ para resolver el problema de valor inicial

$$y' = 30 - 5y, \quad y(0) = 1 \quad \text{en } [0, 10].$$

5. Para el problema de valor inicial $y' = 45 - 9y$, $y(1) = 0$ en $[1, 20]$:
- (a) Use la desigualdad (22) para determinar los valores del tamaño de paso para los que el método de Adams-Bashforth-Moulton podría ser inestable.
 - (b) Basándose en los resultados del apartado (a), seleccione dos tamaños de paso h_e y h_i para los que el método de Adams-Bashforth-Moulton debe ser estable e inestable, respectivamente. Use un método de Runge-Kutta para generar tres valores iniciales y_1 , y_2 e y_3 para cada uno de los tamaños de paso seleccionados.
 - (c) Use el Programa 9.6 para generar dos vectores de aproximaciones, uno para cada tamaño de paso, a la solución del problema de valor inicial.
 - (d) Use los resultados del apartado (c) para dibujar una gráfica similar a la Figura 9.13; puede que sea necesario experimentar con varios tamaños de paso.

9.7 Sistemas de ecuaciones diferenciales

Esta sección es una introducción a los sistemas de ecuaciones diferenciales. Para ilustrar los conceptos, consideremos el problema de valor inicial

$$(1) \quad \begin{aligned} \frac{dx}{dt} &= f(t, x, y) \\ \frac{dy}{dt} &= g(t, x, y) \end{aligned} \quad \text{con} \quad \begin{cases} x(t_0) = x_0, \\ y(t_0) = y_0. \end{cases}$$

Una solución del sistema (1) es un par de funciones derivables $x(t)$ e $y(t)$ tales que cuando t , $x(t)$ e $y(t)$ se sustituyen en $f(t, x, y)$ y $g(t, x, y)$, el resultado es igual a la derivada $x'(t)$ e $y'(t)$, respectivamente; es decir

$$(2) \quad \begin{aligned} x'(t) &= f(t, x(t), y(t)) \\ y'(t) &= g(t, x(t), y(t)) \end{aligned} \quad \text{con} \quad \begin{cases} x(t_0) = x_0, \\ y(t_0) = y_0. \end{cases}$$

Por ejemplo, consideremos el sistema de ecuaciones diferenciales

$$(3) \quad \begin{aligned} \frac{dx}{dt} &= x + 2y \\ \frac{dy}{dt} &= 3x + 2y \end{aligned} \quad \text{con} \quad \begin{cases} x(0) = 6, \\ y(0) = 4. \end{cases}$$

La solución del problema de valor inicial (3) es

$$(4) \quad \begin{aligned} x(t) &= 4e^{4t} + 2e^{-t}, \\ y(t) &= 6e^{4t} - 2e^{-t}. \end{aligned}$$

Esto puede verificarse sustituyendo directamente $x(t)$ e $y(t)$ en el miembro derecho de (3), calculando las derivadas en (4) y sustituyéndolas en el miembro izquierdo de (3), con lo que se obtiene:

$$\begin{aligned} 16e^{4t} - 2e^{-t} &= (4e^{4t} + 2e^{-t}) + 2(6e^{4t} - 2e^{-t}), \\ 24e^{4t} + 2e^{-t} &= 3(4e^{4t} + 2e^{-t}) + 2(6e^{4t} - 2e^{-t}). \end{aligned}$$

Resolución numérica

Podemos encontrar una solución numérica del sistema (1) en un intervalo dado $a \leq t \leq b$ considerando los diferenciales

$$(5) \quad dx = f(t, x, y) dt \quad \text{y} \quad dy = g(t, x, y) dt.$$

El método de Euler para resolver este problema es fácil de formular: Sustituyendo en (5) los diferenciales por incrementos $dt = t_{k+1} - t_k$, $dx = x_{k+1} - x_k$ y $dy = y_{k+1} - y_k$ obtenemos

$$(6) \quad \begin{aligned} x_{k+1} - x_k &\approx f(t_k, x_k, y_k)(t_{k+1} - t_k), \\ y_{k+1} - y_k &\approx g(t_k, x_k, y_k)(t_{k+1} - t_k). \end{aligned}$$

Dividiendo el intervalo en M subintervalos de anchura $h = (b - a)/M$ y usando en (6) los puntos $t_{k+1} = t_k + h$ como nodos, obtenemos las fórmulas recursivas del método de Euler

$$(7) \quad \begin{aligned} t_{k+1} &= t_k + h, \\ x_{k+1} &= x_k + hf(t_k, x_k, y_k), \\ y_{k+1} &= y_k + hg(t_k, x_k, y_k), \quad \text{para } k = 0, 1, \dots, M - 1. \end{aligned}$$

Para conseguir un grado de precisión razonable, es necesario utilizar un método de orden mayor. Por ejemplo, las fórmulas para el método de Runge-Kutta de orden 4 son

$$(8) \quad \begin{aligned} x_{k+1} &= x_k + \frac{h}{6}(f_1 + 2f_2 + 2f_3 + f_4), \\ y_{k+1} &= y_k + \frac{h}{6}(g_1 + 2g_2 + 2g_3 + g_4), \end{aligned}$$

donde

$$\begin{aligned} f_1 &= f(t_k, x_k, y_k), & g_1 &= g(t_k, x_k, y_k), \\ f_2 &= f\left(t_k + \frac{h}{2}, x_k + \frac{h}{2}f_1, y_k + \frac{h}{2}g_1\right), & g_2 &= g\left(t_k + \frac{h}{2}, x_k + \frac{h}{2}f_1, y_k + \frac{h}{2}g_1\right), \\ f_3 &= f\left(t_k + \frac{h}{2}, x_k + \frac{h}{2}f_2, y_k + \frac{h}{2}g_2\right), & g_3 &= g\left(t_k + \frac{h}{2}, x_k + \frac{h}{2}f_2, y_k + \frac{h}{2}g_2\right), \\ f_4 &= f(t_k + h, x_k + hf_3, y_k + hg_3), & g_4 &= g(t_k + h, x_k + hf_3, y_k + hg_3). \end{aligned}$$

Ejemplo 9.15. Vamos a usar el método de Runge-Kutta dado en (8) para calcular una solución numérica del sistema (3) en el intervalo $[0.0, 0.2]$ tomando diez subintervalos con tamaño de paso $h = 0.02$.

Para el primer punto tenemos $t_1 = 0.02$ y las operaciones intermedias necesarias para obtener x_1 e y_1 son

$$f_1 = f(0.00, 6.0, 4.0) = 14.0$$

$$g_1 = g(0.00, 6.0, 4.0) = 26.0$$

$$x_0 + \frac{h}{2}f_1 = 6.14$$

$$y_0 + \frac{h}{2}g_1 = 4.26$$

$$f_2 = f(0.01, 6.14, 4.26) = 14.66$$

$$g_2 = g(0.01, 6.14, 4.26) = 26.94$$

$$x_0 + \frac{h}{2}f_2 = 6.1466$$

$$y_0 + \frac{h}{2}g_2 = 4.2694$$

$$f_3 = f(0.01, 6.1466, 4.2694) = 14.6854$$

$$g_3 = f(0.01, 6.1466, 4.2694) = 26.9786$$

$$x_0 + hf_3 = 6.293708 \quad y_0 + hg_3 = 4.539572$$

$$f_4 = f(0.02, 6.293708, 4.539572) = 15.372852$$

$$g_4 = f(0.02, 6.293708, 4.539572) = 27.960268$$

Usando estos valores en la fórmula (8) nos queda:

$$x_1 = 6 + \frac{0.02}{6}(14.0 + 2(14.66) + 2(14.6854) + 15.372852) = 6.29354551,$$

$$y_1 = 4 + \frac{0.02}{6}(26.0 + 2(26.94) + 2(26.9786) + 27.960268) = 4.53932490.$$

Los cálculos en los demás nodos se recogen en la Tabla 9.13. ■

Las soluciones así calculadas presentan errores que se van acumulando paso a paso. En el ejemplo anterior, el error crece y alcanza su máximo en el extremo derecho $t = 0.2$ del intervalo:

$$x(0.2) - x_{10} = 10.5396252 - 10.5396230 = 0.0000022,$$

$$y(0.2) - y_{10} = 11.7157841 - 11.7157807 = 0.0000034.$$

Ecuaciones diferenciales de orden superior

Las ecuaciones diferenciales de orden superior son las que involucran las derivadas de orden superior $x''(t)$, $x'''(t)$ y así sucesivamente. Este tipo de ecuaciones aparecen en modelos matemáticos de problemas de la física y la ingeniería. Por ejemplo,

$$mx''(t) + cx'(t) + kx(t) = g(t)$$

representa un sistema mecánico en el que un muelle, cuya constante de recuperación es k y que está atado a una masa m , ha sido separado de su posición

Tabla 9.13 Aproximación a la solución de $x'(t) = x + 2y$, $y'(t) = 3x + 2y$ con valores iniciales $x(0) = 6$ e $y(0) = 4$ mediante el método de Runge-Kutta.

k	t_k	x_k	y_k
0	0.00	6.00000000	4.00000000
1	0.02	6.29354551	4.53932490
2	0.04	6.61562213	5.11948599
3	0.06	6.96852528	5.74396525
4	0.08	7.35474319	6.41653305
5	0.10	7.77697287	7.14127221
6	0.12	8.23813750	7.92260406
7	0.14	8.74140523	8.76531667
8	0.16	9.29020955	9.67459538
9	0.18	9.88827138	10.6560560
10	0.20	10.5396230	11.7157807

de equilibrio, a la que tiende a volver. Se supone que la amortiguación debida al rozamiento es proporcional a la velocidad, que existe una fuerza externa $g(t)$ y, como ocurre con frecuencia, que se conocen la posición $x(t_0)$ y la velocidad $x'(t_0)$ en un cierto instante t_0 .

Despejando la derivada segunda, podemos escribir el problema de valor inicial de segundo orden como

$$(9) \quad x''(t) = f(t, x(t), x'(t)) \quad \text{con } x(t_0) = x_0 \text{ y } x'(t_0) = y_0.$$

Esta ecuación diferencial de segundo orden puede reformularse como un sistema con dos ecuaciones de primer orden usando la sustitución

$$(10) \quad x'(t) = y(t).$$

Entonces $x''(t) = y'(t)$ y la ecuación diferencial (9) se convierte en el sistema

$$(11) \quad \begin{aligned} \frac{dx}{dt} &= y \\ \frac{dy}{dt} &= f(t, x, y) \end{aligned} \quad \text{con} \quad \begin{cases} x(t_0) = x_0, \\ y(t_0) = y_0. \end{cases}$$

Al resolver el sistema (11) con un método numérico como el de Runge-Kutta, se generan dos sucesiones $\{x_k\}$ e $\{y_k\}$, siendo $\{x_k\}$ la solución de (9). El siguiente ejemplo puede interpretarse como un movimiento armónico amortiguado.

Ejemplo 9.16. Consideremos el problema de valor inicial de segundo orden

$$x''(t) + 4x'(t) + 5x(t) = 0 \quad \text{con } x(0) = 3 \text{ y } x'(0) = -5.$$

Tabla 9.14 Solución numérica de $x''(t) + 4x'(t) + 5x(t) = 0$ con las condiciones iniciales $x(0) = 3$ y $x'(0) = -5$ obtenida con el método de Runge-Kutta.

k	t_k	x_k	$x(t_k)$
0	0.0	3.00000000	3.00000000
1	0.1	2.52564583	2.52565822
2	0.2	2.10402783	2.10404686
3	0.3	1.73506269	1.73508427
4	0.4	1.41653369	1.41655509
5	0.5	1.14488509	1.14490455
10	1.0	0.33324302	0.33324661
20	2.0	-0.00620684	-0.00621162
30	3.0	-0.00701079	-0.00701204
40	4.0	-0.00091163	-0.00091170
48	4.8	-0.00004972	-0.00004969
49	4.9	-0.00002348	-0.00002345
50	5.0	-0.00000493	-0.00000490

- (a) Vamos a escribir un sistema con dos ecuaciones de primer orden que sea equivalente.
- (b) Vamos a resolver el problema reformulado usando el método de Runge-Kutta en el intervalo $[0, 5]$ con $M = 50$ intervalos de anchura $h = 0.1$.
- (c) Vamos a comparar la solución numérica con la exacta:

$$x(t) = 3e^{-2t} \cos(t) + e^{-2t} \sin(t).$$

La ecuación diferencial la escribimos como

$$x''(t) = f(t, x(t), x'(t)) = -4x'(t) - 5x(t).$$

Usando el cambio dado en (10), el problema reformulado queda

$$\begin{aligned} \frac{dx}{dt} &= y \\ \frac{dy}{dt} &= -5x - 4y \end{aligned} \quad \text{con} \quad \begin{cases} x(0) = 3, \\ y(0) = -5. \end{cases}$$

En la Tabla 9.14 se recogen algunas de las aproximaciones numéricas que se obtienen. Los valores $\{y_k\}$ no nos interesan, así que no se muestran; sí se muestran, en cambio, los valores exactos $\{x(t_k)\}$ para que podamos hacer la correspondiente comparación. ■

Ejercicios

En los Ejercicios 1 a 4, tome $h = 0.05$ y use

- (a) el método de Euler (7) para calcular a mano (x_1, y_1) y (x_2, y_2) ,
- (b) el método de Runge-Kutta (8) para calcular a mano (x_1, y_1) .

1. Resuelva el sistema $x' = 2x + 3y$, $y' = 2x + y$ con la condición inicial $x(0) = -2.7$ e $y(0) = 2.8$ en el intervalo $0 \leq t \leq 1.0$. La curva poligonal formada por las coordenadas de la solución numérica obtenida se muestra en la Figura 9.14 y puede compararse con la solución exacta

$$x(t) = -\frac{69}{25}e^{-t} + \frac{3}{50}e^{4t} \quad \text{e} \quad y(t) = \frac{69}{25}e^{-t} + \frac{1}{25}e^{4t}.$$

2. Resuelva el sistema $x' = 3x - y$, $y' = 4x - y$ con la condición inicial $x(0) = 0.2$ e $y(0) = 0.5$ en el intervalo $0 \leq t \leq 2$. La curva poligonal formada por las coordenadas de la solución numérica obtenida se muestra en la Figura 9.15 y puede compararse con la solución exacta

$$x(t) = \frac{1}{5}e^t - \frac{1}{10}te^t \quad \text{e} \quad y(t) = \frac{1}{2}e^t - \frac{1}{5}te^t.$$

3. Resuelva el sistema $x' = x - 4y$, $y' = x + y$ con la condición inicial $x(0) = 2$ e $y(0) = 3$ en el intervalo $0 \leq t \leq 2$. La curva poligonal formada por las coordenadas de la solución numérica obtenida se muestra en la Figura 9.16 y puede compararse con la solución exacta

$$x(t) = -2e^t + 4e^t \cos^2(t) - 12e^t \cos(t) \sin(t),$$

$$y(t) = -3e^t + 6e^t \cos^2(t) + 2e^t \cos(t) \sin(t).$$

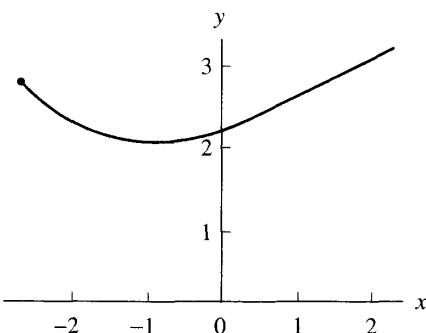


Figura 9.14 La solución del sistema $x' = 2x + 3y$ e $y' = 2x + y$ en $[0.0, 1.0]$.

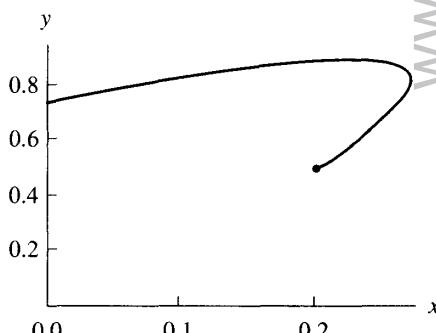


Figura 9.15 La solución del sistema $x' = 3x - y$ e $y' = 4x - y$ en $[0.0, 2.0]$.

4. Resuelva el sistema $x' = y - 4x$, $y' = x + y$ con la condición inicial $x(0) = 1$ e $y(0) = 1$ en el intervalo $0 \leq t \leq 1.2$ usando como tamaño de paso $h = 0.05$. La curva poligonal formada por las coordenadas de la solución numérica obtenida se muestra en la Figura 9.17 y puede compararse con la solución exacta

$$x(t) = \frac{3e^{-\sqrt{29}t/2} - 3e^{\sqrt{29}t/2}}{2\sqrt{29}e^{3t/2}} + \frac{e^{-\sqrt{29}t/2} + e^{\sqrt{29}t/2}}{2e^{3t/2}},$$

$$y(t) = \frac{-7e^{-\sqrt{29}t/2} + 7e^{\sqrt{29}t/2}}{2\sqrt{29}e^{3t/2}} + \frac{e^{-\sqrt{29}t/2} + e^{\sqrt{29}t/2}}{2e^{3t/2}}.$$

En los Ejercicios 5 a 8:

- (a) Compruebe que la función $x(t)$ es la solución.
 - (b) Reformule la ecuación diferencial de segundo orden como un sistema de dos ecuaciones de primer orden.
 - (c) Use el método de Euler con tamaño de paso $h = 0.1$ para calcular a mano x_1 y x_2 .
 - (d) Use el método de Runge-Kutta con tamaño de paso $h = 0.05$ para calcular a mano x_1 .
5. $2x''(t) - 5x'(t) - 3x(t) = 45e^{2t}$ con $x(0) = 2$ y $x'(0) = 1$
 $x(t) = 4e^{-t/2} + 7e^{3t} - 9e^{2t}$
6. $x''(t) + 6x'(t) + 9x(t) = 0$ con $x(0) = 4$ y $x'(0) = -4$
 $x(t) = 4e^{-3t} + 8te^{-3t}$
7. $x''(t) + x(t) = 6 \cos(t)$ con $x(0) = 2$ e $x'(0) = 3$
 $x(t) = 2 \cos(t) + 3 \sin(t) + 3t \sin(t)$

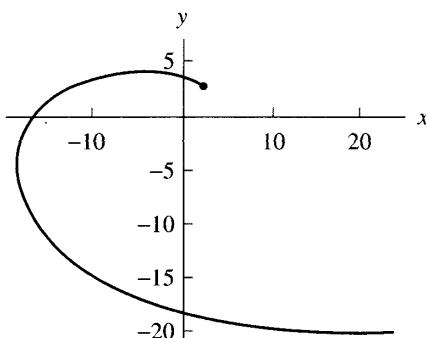


Figura 9.16 La solución del sistema $x' = x - 4y$ e $y' = x + y$ en $[0.0, 2.0]$.

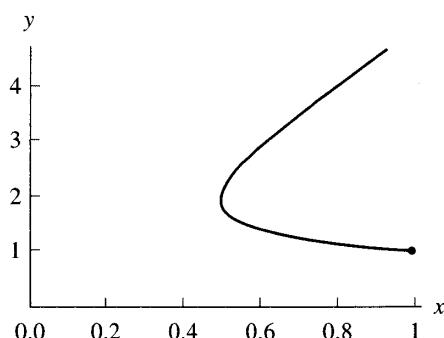


Figura 9.17 La solución del sistema $x' = y - 4x$ e $y' = x + y$ en $[0.0, 1.2]$.

8. $x''(t) + 3x'(t) = 12$ con $x(0) = 5$ y $x'(0) = 1$
 $x(t) = 4 + 4t + e^{-3t}$

Algoritmos y programas

1. Escriba un programa que permita resolver un sistema de ecuaciones diferenciales mediante el método de Runge-Kutta de orden $N = 4$ dado en (8).

En los Problemas 2 a 5, utilice su programa del Problema 1 para resolver el sistema correspondiente usando el método de Runge-Kutta con tamaño de paso $h = 0.05$. Dibuje la aproximación obtenida y la solución exacta en una misma gráfica.

2. $x' = 2x + 3y$, $y' = 2x + y$ con $x(0) = -2.7$, $y(0) = 2.8$ en $0 \leq t \leq 1.0$
 $x(t) = -\frac{69}{25}e^{-t} + \frac{3}{50}e^{4t}$ e $y(t) = \frac{69}{25}e^{-t} + \frac{1}{25}e^{4t}$

3. $x' = 3x - y$, $y' = 4x - y$ con $x(0) = 0.2$, $y(0) = 0.5$ en $0 \leq t \leq 2$
 $x(t) = \frac{1}{5}e^t - \frac{1}{10}te^t$ e $y(t) = \frac{1}{2}e^t - \frac{1}{5}te^t$

4. $x' = x - 4y$, $y' = x + y$ con $x(0) = 2$, $y(0) = 3$ en $0 \leq t \leq 2$
 $x(t) = -2e^t + 4e^t \cos^2(t) - 12e^t \cos(t) \sin(t)$
 $y(t) = -3e^t + 6e^t \cos^2(t) + 2e^t \cos(t) \sin(t)$

5. $x' = y - 4x$, $y' = x + y$ con $x(0) = 1$, $y(0) = 1$ en $0 \leq t \leq 1.2$
 $x(t) = \frac{3e^{-\sqrt{29}t/2} - 3e^{\sqrt{29}t/2}}{2\sqrt{29}e^{3t/2}} + \frac{e^{-\sqrt{29}t/2} + e^{\sqrt{29}t/2}}{2e^{3t/2}}$
 $y(t) = \frac{-7e^{-\sqrt{29}t/2} + 7e^{\sqrt{29}t/2}}{2\sqrt{29}e^{3t/2}} + \frac{e^{-\sqrt{29}t/2} + e^{\sqrt{29}t/2}}{2e^{3t/2}}$

En los Problemas 6 a 9:

- (a) Reformule la ecuación diferencial de segundo orden como un sistema de dos ecuaciones de primer orden.
(b) Use su programa del método de Runge-Kutta para resolver cada sistema en el intervalo $[0, 2]$ tomando como tamaño de paso $h = 0.05$.

(c) Dibuje la aproximación obtenida y la solución exacta en una misma gráfica.

6. $2x''(t) - 5x'(t) - 3x(t) = 45e^{2t}$ con $x(0) = 2$ y $x'(0) = 1$
 $x(t) = 4e^{-t/2} + 7e^{3t} - 9e^{2t}$

7. $x''(t) + 6x'(t) + 9x(t) = 0$ con $x(0) = 4$ y $x'(0) = -4$
 $x(t) = 4e^{-3t} + 8te^{-3t}$

8. $x''(t) + x(t) = 6 \cos(t)$ con $x(0) = 2$ y $x'(0) = 3$
 $x(t) = 2 \cos(t) + 3 \sin(t) + 3t \sin(t)$

9. $x''(t) + 3x'(t) = 12$ con $x(0) = 5$ y $x'(0) = 1$
 $x(t) = 4 + 4t + e^{-3t}$

En los Problemas 10 a 19, use su programa del método de Runge-Kutta de orden $N = 4$ para resolver la ecuación diferencial o el sistema de ecuaciones diferenciales que se da y dibuje la solución obtenida.

- 10.** Un cierto sistema resonante de muelles sobre el que se ejerce una fuerza externa periódica se modela mediante la ecuación

$$x''(t) + 25x(t) = 8 \operatorname{sen}(5t) \quad \text{con } x(0) = 0 \text{ y } x'(0) = 0.$$

Use el método de Runge-Kutta para resolver la ecuación diferencial en el intervalo $[0, 2]$ usando $M = 40$ pasos con $h = 0.05$.

- 11.** El modelo matemático de un cierto circuito eléctrico RCL (resistencia, condensador e inductancia) es

$$Q''(t) + 20Q'(t) + 125Q(t) = 9 \operatorname{sen}(5t)$$

con $Q(0) = 0$ y $Q'(0) = 0$. Use el método de Runge-Kutta para resolver la ecuación diferencial en el intervalo $[0, 2]$ usando $M = 40$ pasos con $h = 0.05$. *Observación.* $I(t) = Q'(t)$ es la intensidad de corriente en el instante t .

- 12.** En un instante t , un péndulo forma un ángulo $x(t)$ con el eje vertical. Suponiendo despreciable la fricción, la ecuación del movimiento del péndulo es

$$mlx''(t) = -mg \operatorname{sen}(x(t)),$$

donde m es la masa y l es la longitud de la cuerda. Use el método de Runge-Kutta para resolver la ecuación diferencial en el intervalo $[0, 2]$ usando $M = 40$ pasos con $h = 0.05$, sabiendo que $g = 9.8 \text{ m/s}^2$, en los casos

- (a) $l = 1 \text{ m}$, $x(0) = 0.3$ y $x'(0) = 0$.
 (b) $l = 0.25 \text{ m}$, $x(0) = 0.3$ y $x'(0) = 0$.

- 13.** *El modelo predador-presa.* Un ejemplo de un sistema de ecuaciones diferenciales no lineales es el problema “predador-presa.” En un cierto hábitat viven conejos y lince, cuyas poblaciones en un instante t denotamos por $x(t)$ e $y(t)$, respectivamente. El modelo depredador-presa establece que $x(t)$ e $y(t)$ verifican el sistema

$$\begin{aligned} x'(t) &= Ax(t) - Bx(t)y(t), \\ y'(t) &= Cx(t)y(t) - Dy(t). \end{aligned}$$

Una simulación típica con un computador usaría como coeficientes, por ejemplo,

$$A = 2, \quad B = 0.02, \quad C = 0.0002, \quad D = 0.8.$$

Use el método de Runge-Kutta para resolver el sistema en el intervalo $[0, 5]$ usando $M = 40$ pasos con $h = 0.2$ en los siguientes casos

- (a) $x(0) = 3000$ conejos e $y(0) = 120$ lince.
 (b) $x(0) = 5000$ conejos e $y(0) = 100$ lince.

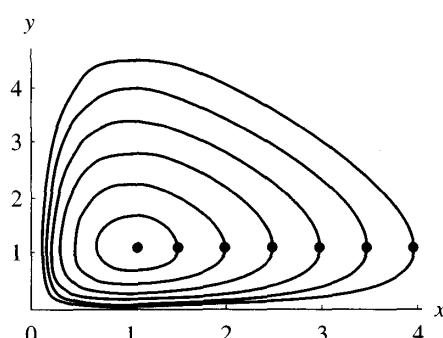


Figura 9.18 Soluciones del sistema $x' = x - xy$ e $y' = -y + xy$.

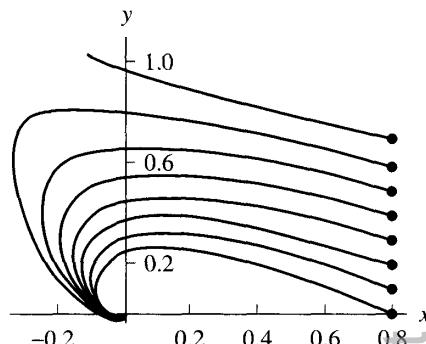


Figura 9.19 Soluciones del sistema $x' = -3x - 2y - 2xy^2$ e $y' = 2x - y + 2y^3$.

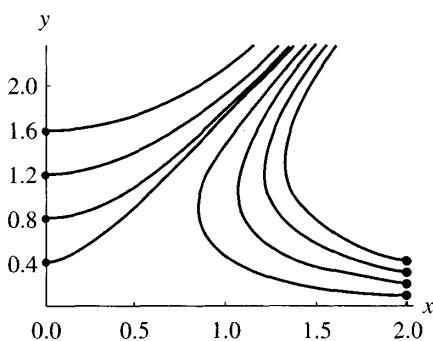


Figura 9.20 Soluciones del sistema $x' = y^2 - x^2$ e $y' = 2xy$.

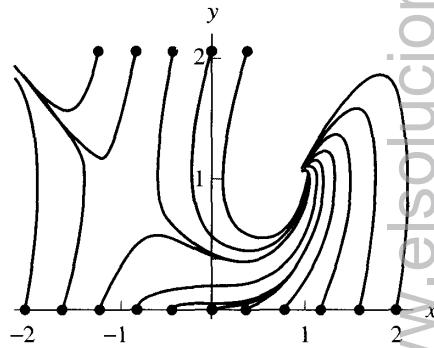


Figura 9.21 Soluciones del sistema $x' = 1 - y$ e $y' = x^2 - y^2$.

14. Resuelva $x' = x - xy$, $y' = -y + xy$ con $x(0) = 4$ e $y(0) = 1$ en $[0, 8]$ tomando $h = 0.1$. Las trayectorias de este sistema son curvas cerradas y la trayectoria poligonal obtenida con la solución numérica es una de las curvas de la Figura 9.18.
15. Resuelva $x' = -3x - 2y - 2xy^2$, $y' = 2x - y + 2y^3$ con $x(0) = 0.8$ e $y(0) = 0.6$ en $[0, 4]$ tomando $h = 0.1$. De acuerdo con la teoría cualitativa, el origen se clasifica, para este sistema, como un foco asintóticamente estable. La trayectoria poligonal obtenida con la solución numérica es una de las curvas de la Figura 9.19.

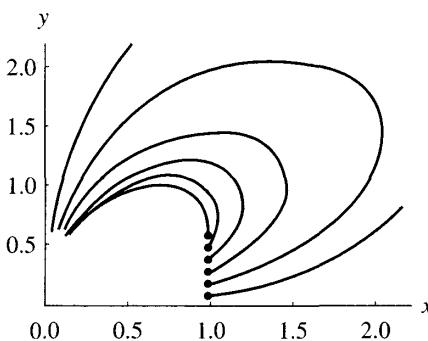


Figura 9.22 Soluciones del sistema $x' = x^3 - 2xy^2$, $y' = 2x^2y - y^3$.

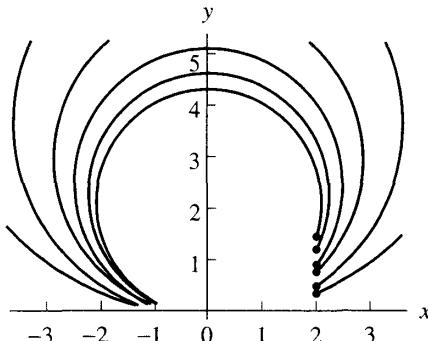


Figura 9.23 Soluciones del sistema $x' = x^2 - y^2$, $y' = 2xy$.

16. Resuelva $x' = y^2 - x^2$, $y' = 2xy$ con $x(0) = 2.0$ e $y(0) = 0.1$ en $[0.0, 1.5]$ tomando $h = 0.05$. De acuerdo con la teoría cualitativa, el origen se clasifica, para este sistema, como un punto de silla inestable. La trayectoria poligonal obtenida con la solución numérica es una de las curvas de la Figura 9.20.
17. Resuelva $x' = 1 - y$, $y' = x^2 - y^2$ con $x(0) = -1.2$ e $y(0) = 0.0$ en $[0, 5]$ tomando $h = 0.1$. De acuerdo con la teoría cualitativa, el punto $(1, 1)$ se clasifica, para este sistema, como un foco asintóticamente estable. La trayectoria poligonal obtenida con la solución numérica es una de las curvas de la Figura 9.21.
18. Resuelva $x' = x^3 - 2xy^2$, $y' = 2x^2y - y^3$ con $x(0) = 1.0$ e $y(0) = 0.2$ en $[0, 2]$ tomando $h = 0.025$. Este sistema tiene un punto crítico inestable en el origen. La trayectoria poligonal obtenida con la solución numérica es una de las curvas de la Figura 9.22.
19. Resuelva $x' = x^2 - y^2$, $y' = 2xy$ con $x(0) = 2.0$ e $y(0) = 0.6$ en $[0.0, 1.6]$ tomando $h = 0.02$. Este sistema tiene un punto crítico inestable en el origen. La trayectoria poligonal obtenida con la solución numérica es una de las curvas de la Figura 9.23.

8 Problemas de contorno

Otro tipo de ecuaciones diferenciales son de la forma

$$(1) \quad x'' = f(t, x, x') \quad \text{para} \quad a \leq t \leq b,$$

con la condición de contorno (o frontera)

$$(2) \quad x(a) = \alpha \quad \text{y} \quad x(b) = \beta.$$

Esto es lo que se conoce como **problema de contorno** o **problema de valores en la frontera**.

Hay que comprobar que se cumplen las condiciones recogidas en el siguiente teorema, que garantizan que un problema de contorno tiene solución antes de emplear un método numérico; si no se hace, puede que obtengamos resultados absurdos.

Teorema 9.8 (Problema de contorno). Supongamos que $f(t, x, y)$ es continua en la región $R = \{(t, x, y) : a \leq t \leq b, -\infty < x < \infty, -\infty < y < \infty\}$ y que las derivadas parciales $\partial f / \partial x = f_x(t, x, y)$ y $\partial f / \partial y = f_y(t, x, y)$ son continuas en R . Si

$$(3) \quad f_x(t, x, y) > 0 \quad \text{para todo } (t, x, y) \in R$$

y existe una constante $M > 0$ tal que

$$(4) \quad |f_y(t, x, y)| \leq M \quad \text{para todo } (t, x, y) \in R,$$

entonces el problema de contorno

$$(5) \quad x'' = f(t, x, x') \quad \text{con } x(a) = \alpha \text{ y } x(b) = \beta$$

tiene solución única $x = x(t)$ en $a \leq t \leq b$.

Hemos usado la notación $y = x'(t)$ para distinguir la tercera variable de la función $f(t, x, x')$. Merece la pena destacar el caso especial de las ecuaciones diferenciales lineales.

Corolario 9.1 (Problemas de contorno lineales). Supongamos que la función f del Teorema 9.8 es de la forma

$$f(t, x, y) = p(t)y + q(t)x + r(t)$$

y que f y sus derivadas parciales $\partial f / \partial x = q(t)$ y $\partial f / \partial y = p(t)$ son continuas en R (lo que garantiza, en particular, que $|p(t)| \leq M = \max_{a \leq t \leq b} \{|p(t)|\}$). Si

$$(6) \quad q(t) > 0 \quad \text{para todo } t \in [a, b]$$

entonces el **problema de contorno lineal**

$$(7) \quad x'' = p(t)x'(t) + q(t)x(t) + r(t) \quad \text{con } x(a) = \alpha \text{ y } x(b) = \beta$$

tiene solución única $x = x(t)$ en $a \leq t \leq b$.

El método de disparo lineal

Una de las formas de calcular la solución de un problema de contorno lineal es utilizar su estructura lineal para descomponerlo en dos problemas de valor inicial especiales. Supongamos que $u(t)$ es la solución única del problema de valor inicial

$$(8) \quad u'' = p(t)u'(t) + q(t)u(t) + r(t) \quad \text{con } u(a) = \alpha \text{ y } u'(a) = 0.$$

Supongamos, además, que $v(t)$ es la solución única del problema de valor inicial

$$(9) \quad v'' = p(t)v'(t) + q(t)v(t) \quad \text{con } v(a) = 0 \text{ y } v'(a) = 1.$$

Entonces la combinación lineal

$$(10) \quad x(t) = u(t) + Cv(t)$$

es una solución de $x'' = p(t)x'(t) + q(t)x(t) + r(t)$, como podemos comprobar haciendo los cálculos

$$\begin{aligned} x'' &= u'' + Cv'' = p(t)u'(t) + q(t)u(t) + r(t) + p(t)Cv'(t) + q(t)Cv(t) \\ &= p(t)(u'(t) + Cv'(t)) + q(t)(u(t) + Cv(t)) + r(t) \\ &= p(t)x'(t) + q(t)x(t) + r(t). \end{aligned}$$

La solución $x(t)$ de la ecuación (8) toma los siguientes valores en la frontera del intervalo

$$(11) \quad \begin{aligned} x(a) &= u(a) + Cv(a) = \alpha + 0 = \alpha, \\ x(b) &= u(b) + Cv(b). \end{aligned}$$

Si imponemos la condición de contorno $x(b) = \beta$ en (11), obtenemos $C = (\beta - u(b))/v(b)$. En consecuencia, si $v(b) \neq 0$, entonces la solución única del problema de contorno (7) es

$$(12) \quad x(t) = u(t) + \frac{\beta - u(b)}{v(b)}v(t).$$

Observación. Si q verifica las hipótesis del Corolario 9.1, entonces no se da el caso problemático de que $v(t) \equiv 0$, de manera que la solución buscada es la dada por la expresión (12); los detalles de comprobación se dejan como ejercicio (véase el Ejercicio 3).

Ejemplo 9.17. Vamos a resolver el problema de contorno

$$x''(t) = \frac{2t}{1+t^2}x'(t) - \frac{2}{1+t^2}x(t) + 1$$

con $x(0) = 1.25$ y $x(4) = -0.95$ en el intervalo $[0, 4]$.

Las funciones p , q y r son $p(t) = 2t/(1+t^2)$, $q(t) = -2/(1+t^2)$ y $r(t) = 1$, respectivamente. Usando el método de Runge-Kutta de orden 4 con tamaño de paso $h = 0.2$ calculamos soluciones numéricas $\{u_j\}$ y $\{v_j\}$ de los problemas (8) y (9), respectivamente. Las aproximaciones $\{u_j\}$ a $u(t)$ se muestran en la primera columna de la Tabla 9.15; por otro lado, tomando $u(4) \approx u_{20} = -2.893535$ y $v(4) \approx v_{20} = 4$ en (12) construimos

$$w_j = \frac{b - u(4)}{v(4)} v_j = 0.485884 v_j.$$

Entonces la solución numérica del problema de contorno viene dada por $\{x_j\} = \{u_j + w_j\}$. En la Tabla 9.15 se muestran algunas de estas aproximaciones y en la Figura 9.24 se muestran sus gráficas. Puede comprobarse que $v(t) = t$ es la solución exacta del problema (9); es decir,

$$v''(t) = \frac{2t}{1+t^2} v'(t) - \frac{2}{1+t^2} v(t)$$

con la condición inicial $v(0) = 0$ y $v'(0) = 1$.

En la Tabla 9.16 se comparan las aproximaciones obtenidas con el método de disparo lineal tomando tamaños de paso $h = 0.2$ y $h = 0.1$ y la solución exacta

$$x(t) = 1.25 + 0.4860896526t - 2.25t^2 + 2t \arctan(t) + \frac{1}{2}(t^2 - 1) \ln(1 + t^2).$$

En la tabla también se incluyen las columnas de los errores; puesto que el error en el método de Runge-Kutta es de orden $O(h^4)$, el error de las aproximaciones con el tamaño de paso menor $h = 0.1$ es, aproximadamente, $\frac{1}{16}$ del error de las aproximaciones con el tamaño de paso mayor $h = 0.2$. En la Figura 9.25 se muestra la gráfica de la solución aproximada cuando $h = 0.2$.

MATLAB

El Programa 9.10 incluye una llamada al Programa 9.9, con el que se resuelven los problemas de valor inicial (8) y (9). El Programa 9.9 proporciona soluciones aproximadas de sistemas de ecuaciones diferenciales usando la modificación pertinente del método de Runge-Kutta de orden $N = 4$; por tanto, es necesario almacenar las ecuaciones (8) y (9) como un sistema de ecuaciones igual que el (11) de la Sección 9.7. A modo de ilustración, consideremos el problema de contorno del Ejemplo 9.17: El siguiente archivo, llamado F1.m, almacena el problema de valor inicial (8) como un sistema.

```
function Z=F1(t,Z)
x=Z(1);y=Z(2);
Z=[y,2*t*y/(1+t^2)-2*x/(1+t^2)+1];
```

Un archivo similar, llamado F2.m, almacenará el problema de valor inicial dado en (9) (basta poner $r(t) = 0$ en F1) en la forma adecuada.

Finalmente, utilizando la instrucción `plot(L(:,1), L(:,2))`, podremos dibujar la gráfica de la aproximación obtenida con el Programa 9.10.

Tabla 9.15 Solución aproximada $\{x_j\} = \{u_j + w_j\}$ de la ecuación $x''(t) = \frac{2t}{1+t^2}x'(t) - \frac{2}{1+t^2} + 1$.

t_j	u_j	w_j	$x_j = u_j + w_j$
0.0	1.250000	0.000000	1.250000
0.2	1.220131	0.097177	1.317308
0.4	1.132073	0.194353	1.326426
0.6	0.990122	0.291530	1.281652
0.8	0.800569	0.388707	1.189276
1.0	0.570844	0.485884	1.056728
1.2	0.308850	0.583061	0.891911
1.4	0.022522	0.680237	0.702759
1.6	-0.280424	0.777413	0.496989
1.8	-0.592609	0.874591	0.281982
2.0	-0.907039	0.971767	0.064728
2.2	-1.217121	1.068944	-0.148177
2.4	-1.516639	1.166121	-0.350518
2.6	-1.799740	1.263297	-0.536443
2.8	-2.060904	1.360474	-0.700430
3.0	-2.294916	1.457651	-0.837265
3.2	-2.496842	1.554828	-0.942014
3.4	-2.662004	1.652004	-1.010000
3.6	-2.785960	1.749181	-1.036779
3.8	-2.864481	1.846358	-1.018123
4.0	-2.893535	1.943535	-0.950000

Programa 9.9 (Método de Runge-Kutta de orden $N = 4$ para sistemas). Construcción de aproximaciones a la solución del sistema de ecuaciones diferenciales

$$x'_1(t) = f_1(t, x_1(t), \dots, x_n(t))$$

$$\vdots \qquad \vdots$$

$$x'_n(t) = f_n(t, x_1(t), \dots, x_n(t))$$

con $x_1(a) = \alpha_1, \dots, x_n(a) = \alpha_n$ en el intervalo $[a, b]$.

```
function [T,Z]=rks4(F,a,b,Za,M)
% Datos
% - F es la función, almacenada como una
%   cadena de caracteres 'F'
% - a y b son los extremos derecho e izquierdo
%   del intervalo
```

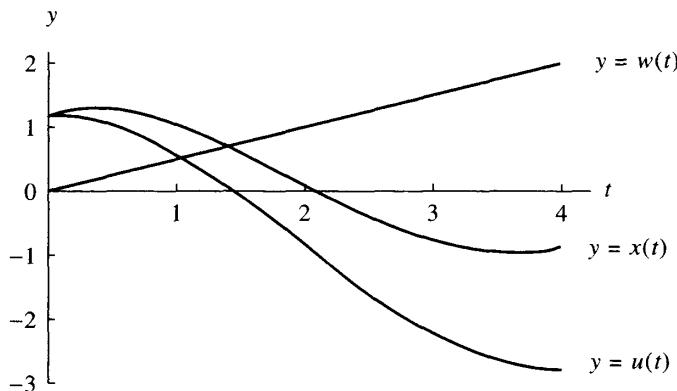


Figura 9.24 Aproximaciones numéricas $u(t)$ y $w(t)$ usadas para formar $x(t) = u(t) + w(t)$, solución de

$$x''(t) = \frac{2t}{1+t^2}x'(t) - \frac{2}{1+t^2}x(t) + 1.$$

Tabla 9.16 Soluciones numéricas de $x''(t) = \frac{2t}{1+t^2}x'(t) - \frac{2}{1+t^2}x(t) + 1$.

t_j	x_j	$x(t_j)$	$x(t_j) - x_j$	t_j	x_j	$x(t_j)$	$x(t_j) - x_j$
	$h = 0.2$	exacto	error		$h = 0.1$	exacto	error
0.0	1.250000	1.250000	0.000000	0.0	1.250000	1.250000	0.000000
0.2	1.317308	1.317350	0.000042	0.1	1.291116	1.291117	0.000001
0.4	1.326426	1.326505	0.000079	0.2	1.317348	1.317350	0.000002
0.6	1.281652	1.281762	0.000110	0.3	1.328986	1.328990	0.000004
0.8	1.189276	1.189412	0.000136	0.4	1.326500	1.326505	0.000005
1.0	1.056728	1.056886	0.000158	0.5	1.310508	1.310514	0.000006
1.2	0.891911	0.892086	0.000175	0.6	1.281756	1.281762	0.000006
1.6	0.496989	0.497187	0.000198	0.8	1.189404	1.189412	0.000008
2.0	0.064728	0.064931	0.000203	1.0	1.056876	1.056886	0.000010
2.4	-0.350518	-0.350325	0.000193	1.2	0.892076	0.892086	0.000010
2.8	-0.700430	-0.700262	0.000168	1.6	0.497175	0.497187	0.000012
3.2	-0.942014	-0.941888	0.000126	2.0	0.064919	0.064931	0.000012
3.6	-1.036779	-1.036708	0.000071	2.4	-0.350337	-0.350325	0.000012
4.0	-0.950000	-0.950000	0.000000	2.8	-0.700273	-0.700262	0.000011
				3.2	-0.941895	-0.941888	0.000007
				3.6	-1.036713	-1.036708	0.000005
				4.0	-0.950000	-0.950000	0.000000

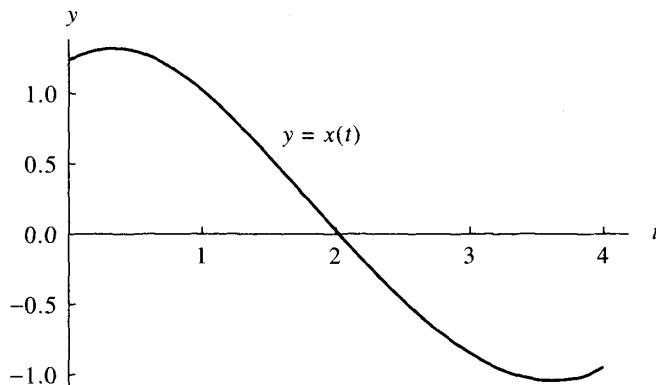


Figura 9.25 La gráfica de la solución numérica de

$$x''(t) = \frac{2t}{1+t^2}x'(t) - \frac{2}{1+t^2}x(t) + 1$$

(tomando $h = 0.2$).

```
%      - Z=[x1(a)...xn(a)] es la condición inicial
%      - M es el número de pasos
% Resultados
%      - T el vector de los nodos
%      - Z=[x1(t)...xn(t)]; donde xk(t) es la aproximación
%          a la k-ésima variable dependiente
h=(b-a)/M;
T=zeros(1,M+1);
Z=zeros(M+1,length(Za));
T=a:h:b;
Z(1,:)=Za;
for j=1:M
    k1=h*feval(F,T(j),Z(j,:));
    k2=h*feval(F,T(j)+h/2,Z(j,:)+k1/2);
    k3=h*feval(F,T(j)+h/2,Z(j,:)+k2/2);
    k4=h*feval(F,T(j)+h,Z(j,:)+k3);
    Z(j+1,:)=Z(j,:)+(k1+2*k2+2*k3+k4)/6;
end
```

Programa 9.10 (Método de disparo lineal). Construcción de aproximaciones a la solución del problema de contorno $x'' = p(t)x'(t) + q(t)x(t) + r(t)$, con $x(a) = \alpha$ y $x(b) = \beta$ en el intervalo $[a, b]$, mediante el método de disparo lineal, usando el método de Runge-Kutta de orden $N = 4$ para resolver los problemas de valor inicial adecuados.

```

function L=linsht(F1,F2,a,b,alpha,beta,M)
% Datos
% - F1 y F2 son los sistemas de ecuaciones de primer
%   orden que representan los problemas de valor
%   inicial (9) y (10) almacenados como cadenas de
%   caracteres 'F1' y 'F2'
% - a y b son los extremos derecho e izquierdo
%   del intervalo
% - alpha = x(a) y beta = x(b) son las condiciones
%   de contorno
% - M es el número de pasos
% Resultado
% - L =[T' X]; siendo T' el vector de dimensión (M+1)x1
%   de las abscisas y X el vector de dimensión (M+1)x1
%   de las ordenadas
% Resolución del sistema F1
Za=[alpha,0];
[T,Z]=rks4(F1,a,b,Za,M);
U=Z(:,1);
% Resolución del sistema F2
Za=[0,1];
[T,Z]=rks4(F2,a,b,Za,M);
V=Z(:,1);
% Cálculo de la solución del problema de contorno
X=U+(beta-U(M+1))*V/V(M+1);
L=[T' X];

```

Ejercicios

1. Compruebe que la función $x(t)$ dada es la solución del correspondiente problema de contorno.

- (a) $x'' = (-2/t)x' + (2/t^2)x + (10 \cos(\ln(t)))/t^2$ en $[1, 3]$ con $x(1) = 1$ y $x(3) = -1$.

$$x(t) = \frac{4.335950689 - 0.3359506908t^3 - 3t^2 \cos(\ln(t)) + t^2 \sin(\ln(t))}{t^2}$$

- (b) $x'' = -2x' - 2x + e^{-t} + \sin(2t)$ en $[0, 4]$ con $x(0) = 0.6$ y $x(4) = -0.1$.

$$\begin{aligned}x(t) = & \frac{1}{5} + e^{-t} - \frac{1}{5}e^{-t} \cos(t) - \frac{2}{5} \cos^2(t) \\& + 3.670227413e^{-t} \sin(t) - \frac{1}{5} \cos(t) \sin(t)\end{aligned}$$

- (c) $x'' = -4x' - 4x + 5 \cos(4t) + \sin(2t)$ en $[0, 2]$ con $x(0) = 0.75$ y $x(2) = 0.25$.

$$\begin{aligned}x(t) = & -\frac{1}{40} + 1.025e^{-2t} - 1.915729975te^{-2t} + \frac{19}{20} \cos^2(t) \\& - \frac{6}{5} \cos^4(t) - \frac{4}{5} \cos(t) \sin(t) + \frac{8}{5} \cos^3(t) \sin(t)\end{aligned}$$

- (d) $x'' + (1/t)x' + (1 - 1/(4t^2))x = 0$ en $[1, 6]$ con $x(1) = 1$ y $x(6) = 0$.

$$x(t) = \frac{0.2913843206 \cos(t) + 1.001299385 \sin(t)}{\sqrt{t}}$$

- (e) $x'' - (1/t)x' + (1/t^2)x = 1$ en $[0.5, 4.5]$ con $x(0.5) = 1$ y $x(4.5) = 2$.

$$x(t) = t^2 - 0.2525826491t - 2.528442297t \ln(t)$$

2. ¿Verifica el problema de contorno del Ejercicio 1(e) las hipótesis del Corolario 9.1? Razona la respuesta.
3. Pruebe que si q verifica las hipótesis del Corolario 9.1, entonces $v(t) \equiv 0$ es la solución única del problema de contorno

$$v'' = p(t)v'(t) + q(t)v(t) \quad \text{con } v(a) = 0 \text{ y } v(b) = 0.$$

Algoritmos y programas

- (a) Use los Programas 9.9 y 9.10 para resolver cada uno de los problemas de contorno del Ejercicio 1, tomando como tamaño de paso $h = 0.05$.
 (b) Dibuje su solución y la solución exacta en una misma gráfica.
- Diseñe programas análogos al Programa 9.9 basados en
 - el método de Heun,
 - el método de Adams-Bashforth-Moulton y
 - el método de Hamming.
- (a) Modifique el Programa 9.10 de manera que utilice cada uno de sus programas del Problema 2.
 (b) Use sus programas para resolver cada uno de los cinco problemas de contorno del Ejercicio 1 tomando como tamaño de paso $h = 0.05$.
 (c) Dibuje sus soluciones y la solución exacta en una misma gráfica.

9.9 El método de las diferencias finitas

Para resolver algunos problemas de contorno de segundo orden pueden utilizarse las fórmulas de diferencias finitas que proporcionan aproximaciones a las derivadas. Consideremos la ecuación lineal

$$(1) \quad x'' = p(t)x'(t) + q(t)x(t) + r(t)$$

en $[a, b]$ con $x(a) = \alpha$ y $x(b) = \beta$. Hagamos una partición de $[a, b]$ usando los nodos $a = t_0 < t_1 < \dots < t_N = b$, siendo $h = (b - a)/N$ y $t_j = a + jh$ para $j = 0, 1, \dots, N$. Usando las fórmulas de diferencias centradas dadas en el Capítulo 6 para aproximar las derivadas

$$(2) \quad x'(t_j) = \frac{x(t_{j+1}) - x(t_{j-1})}{2h} + O(h^2)$$

y

$$(3) \quad x''(t_j) = \frac{x(t_{j+1}) - 2x(t_j) + x(t_{j-1})}{h^2} + O(h^2),$$

lo que hacemos ahora es reemplazar cada término $x(t_j)$ del miembro derecho de las fórmulas (2) y (3) por x_j y sustituir el resultado en la ecuación (1), lo que nos da la relación

$$(4) \quad \frac{x_{j+1} - 2x_j + x_{j-1}}{h^2} + O(h^2) = p(t_j) \left(\frac{x_{j+1} - x_{j-1}}{2h} + O(h^2) \right) + q(t_j)x_j + r(t_j).$$

Eliminando los términos de orden $O(h^2)$ en (4) e introduciendo la notación $p_j = p(t_j)$, $q_j = q(t_j)$ y $r_j = r(t_j)$, obtenemos la ecuación en diferencias

$$(5) \quad \frac{x_{j+1} - 2x_j + x_{j-1}}{h^2} = p_j \frac{x_{j+1} - x_{j-1}}{2h} + q_j x_j + r_j,$$

que se usa para calcular aproximaciones numéricas a la solución de la ecuación diferencial (1). Para ello, multiplicamos cada miembro de (5) por h^2 , agrupamos los términos que contienen las incógnitas x_{j-1} , x_j y x_{j+1} y los disponemos como un sistema de ecuaciones lineales:

$$(6) \quad \left(\frac{-h}{2}p_j - 1 \right) x_{j-1} + (2 + h^2 q_j)x_j + \left(\frac{h}{2}p_j - 1 \right) x_{j+1} = -h^2 r_j,$$

para $j = 1, 2, \dots, N - 1$, siendo $x_0 = \alpha$ y $x_N = \beta$. El sistema (6) es un sistema tridiagonal de $N - 1$ ecuaciones con otras tantas incógnitas, lo que se ve más

claramente si escribimos el sistema usando la notación matricial:

$$\left[\begin{array}{ccc|c} 2 + h^2 q_1 & \frac{h}{2} p_1 - 1 & & x_1 \\ \frac{-h}{2} p_2 - 1 & 2 + h^2 q_2 & \frac{h}{2} p_2 - 1 & 0 \\ & & \ddots & x_2 \\ & & & \dots \\ & \frac{-h}{2} p_j - 1 & 2 + h^2 q_j & \frac{h}{2} p_j - 1 & x_j \\ & & & \ddots & \dots \\ \mathbf{0} & \frac{-h}{2} p_{N-2} - 1 & 2 + h^2 q_{N-2} & \frac{h}{2} p_{N-2} - 1 & x_{N-2} \\ & & \frac{-h}{2} p_{N-1} - 1 & 2 + h^2 q_{N-1} & x_{N-1} \end{array} \right] = \begin{bmatrix} -h^2 r_1 + e_0 \\ -h^2 r_2 \\ \vdots \\ -h^2 r_j \\ \vdots \\ -h^2 r_{N-2} \\ -h^2 r_{N-1} + e_N \end{bmatrix},$$

siendo

$$e_0 = \left(\frac{h}{2} p_1 + 1 \right) \alpha \quad \text{y} \quad e_N = \left(\frac{-h}{2} p_{N-1} + 1 \right) \beta.$$

Cuando se realizan los cálculos con un tamaño de paso h , la aproximación numérica que se obtiene es un conjunto finito de puntos $\{(t_j, x_j)\}$ y si se conoce la solución exacta $x(t_j)$, entonces podemos comparar x_j con $x(t_j)$.

Ejemplo 9.18. Vamos a resolver el problema de contorno

$$x''(t) = \frac{2t}{1+t^2} x'(t) - \frac{2}{1+t^2} x(t) + 1$$

con $x(0) = 1.25$ y $x(4) = -0.95$ en el intervalo $[0, 4]$.

Las funciones p , q y r son $p(t) = 2t/(1+t^2)$, $q(t) = -2/(1+t^2)$ y $r(t) = 1$, respectivamente. Resolviendo el correspondiente sistema de ecuaciones lineales (6), el método de diferencias finitas proporciona las soluciones numéricas $\{x_j\}$. En la Tabla 9.17 se recoge una muestra de las aproximaciones $\{x_{j,1}\}$, $\{x_{j,2}\}$, $\{x_{j,3}\}$ y $\{x_{j,4}\}$ correspondientes a los tamaños de paso $h_1 = 0.2$, $h_2 = 0.1$, $h_3 = 0.05$ y $h_4 = 0.025$. La sucesión $\{x_{j,2}\}$ generada tomando $h_2 = 0.1$ contiene 41 términos de los que sólo

Tabla 9.17 Aproximaciones numéricas para $x''(t) = \frac{2t}{1+t^2}x'(t) - \frac{2}{1+t^2}x(t) + 1$.

t_j	$x_{j,1}$ $h = 0.2$	$x_{j,2}$ $h = 0.1$	$x_{j,3}$ $h = 0.05$	$x_{j,4}$ $h = 0.025$	$x(t_j)$ exacto
0.0	1.250000	1.250000	1.250000	1.250000	1.250000
0.2	1.314503	1.316646	1.317174	1.317306	1.317350
0.4	1.320607	1.325045	1.326141	1.326414	1.326505
0.6	1.272755	1.279533	1.281206	1.281623	1.281762
0.8	1.177399	1.186438	1.188670	1.189227	1.189412
1.0	1.042106	1.053226	1.055973	1.056658	1.056886
1.2	0.874878	0.887823	0.891023	0.891821	0.892086
1.4	0.683712	0.698181	0.701758	0.702650	0.702947
1.6	0.476372	0.492027	0.495900	0.496865	0.497187
1.8	0.260264	0.276749	0.280828	0.281846	0.282184
2.0	0.042399	0.059343	0.063537	0.064583	0.064931
2.2	-0.170616	-0.153592	-0.149378	-0.148327	-0.147977
2.4	-0.372557	-0.355841	-0.351702	-0.350669	-0.350325
2.6	-0.557565	-0.541546	-0.537580	-0.536590	-0.536261
2.8	-0.720114	-0.705188	-0.701492	-0.700570	-0.700262
3.0	-0.854988	-0.841551	-0.838223	-0.837393	-0.837116
3.2	-0.957250	-0.945700	-0.942839	-0.942125	-0.941888
3.4	-1.022221	-1.012958	-1.010662	-1.010090	-1.009899
3.6	-1.045457	-1.038880	-1.037250	-1.036844	-1.036709
3.8	-1.022727	-1.019238	-1.018373	-1.018158	-1.018086
4.0	-0.950000	-0.950000	-0.950000	-0.950000	-0.950000

se muestran uno de cada dos, los que corresponden a los 21 valores de $\{t_j\}$ dados en la Tabla 9.17 y que son los generados tomando $h_1 = 0.2$. Análogamente, lo que se muestra de las sucesiones $\{x_{j,3}\}$ y $\{x_{j,4}\}$ es sólo una porción de todos los valores generados tomando los tamaños de paso $h_3 = 0.05$ y $h_4 = 0.025$, respectivamente, y corresponden a los mismos 21 nodos $\{t_j\}$. En la Figura 9.26 se muestran las gráficas de la poligonal formada con los puntos $\{(t_j, x_{j,1})\}$ para el caso $h_1 = 0.2$.

Ahora comparamos las soluciones numéricas de la Tabla 9.17 con la exacta

$$x(t) = 1.25 + 0.486089652t - 2.25t^2 + 2t \arctan(t) + \frac{1}{2}(t^2 - 1) \ln(1 + t^2).$$

Puede probarse que las soluciones numéricas tienen un error de $O(h^2)$; por tanto, la reducción del tamaño de paso a su mitad produce una disminución del error a, más o menos, su cuarta parte. Un escrutinio detallado de la Tabla 9.18 revela que eso es lo que ocurre. Por ejemplo, en el punto $t_j = 1.0$ los errores de las aproximaciones correspondientes a los tamaños de paso h_1 , h_2 , h_3 y h_4 son $e_{j,1} = 0.014780$, $e_{j,2} = 0.003660$, $e_{j,3} = 0.000913$ y $e_{j,4} = 0.000228$, respectivamente; los cocientes sucesivos de estos errores son $e_{j,2}/e_{j,1} = 0.003660/0.014780 = 0.2476$, $e_{j,3}/e_{j,2} = 0.000913/0.003660 = 0.2495$ y $e_{j,4}/e_{j,3} = 0.000228/0.000913 = 0.2497$, que se acercan a $\frac{1}{4}$.

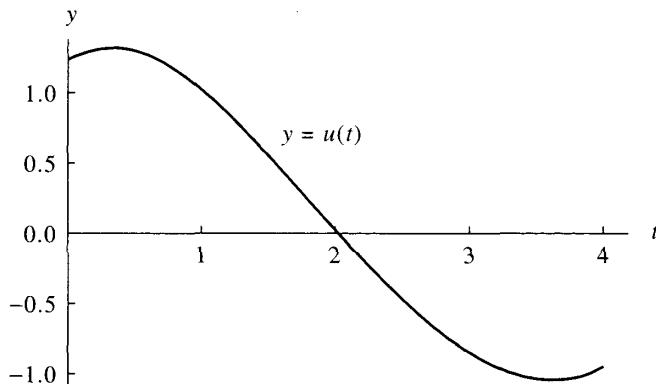


Figura 9.26 Gráfica de la aproximación numérica, tomando $h = 0.2$, a la solución de

$$x''(t) = \frac{2t}{1+t^2} x'(t) - \frac{2}{1+t^2} x(t) + 1.$$

Finalmente, vamos a mostrar cómo puede usarse el esquema de mejora de Richardson para extrapolar los valores poco aproximados $\{x_{j,1}\}$, $\{x_{j,2}\}$, $\{x_{j,3}\}$ y $\{x_{j,4}\}$ y conseguir seis cifras decimales de precisión. Eliminando los términos de orden $O(h^2)$ y $O((h/2)^2)$ en las aproximaciones $\{x_{j,1}\}$ y $\{x_{j,2}\}$ generamos los valores extrapolados correspondientes $\{z_{j,1}\} = \{(4x_{j,2} - x_{j,1})/3\}$. De manera parecida, eliminando los términos del error de orden $O((h/2)^2)$ y $O((h/4)^2)$ en $\{x_{j,2}\}$ y $\{x_{j,3}\}$ se generan los valores extrapolados $\{z_{j,2}\} = \{(4x_{j,3} - x_{j,2})/3\}$. Puede probarse que se puede aplicar el segundo nivel del esquema de mejora de Richardson a las sucesiones $\{z_{j,1}\}$ y $\{z_{j,2}\}$ y generar una tercera mejora $\{(16z_{j,2} - z_{j,1})/15\}$ (véase la Referencia [41]). Vamos a ilustrar la situación hallando los valores extrapolados que corresponden a $t_j = 1.0$. El primer valor extrapolado es

$$\frac{4x_{j,2} - x_{j,1}}{3} = \frac{4(1.053226) - 1.042106}{3} = 1.056932 = z_{j,1}.$$

El segundo valor extrapolado es

$$\frac{4x_{j,3} - x_{j,2}}{3} = \frac{4(1.055973) - 1.053226}{3} = 1.056889 = z_{j,2}.$$

Finalmente, la tercera extrapolación involucra los términos $z_{j,1}$ y $z_{j,2}$:

$$\frac{16z_{j,2} - z_{j,1}}{15} = \frac{16(1.056889) - 1.056932}{15} = 1.056886.$$

Este último valor tiene ya seis cifras decimales de precisión. Los valores mejorados en los demás puntos se recogen en la Tabla 9.19.

Tabla 9.18 Errores de las aproximaciones numéricas obtenidas con el método de diferencias finitas.

t_j	$x(t_j) - x_{j,1}$ $= e_{j,1}$	$x(t_j) - x_{j,2}$ $= e_{j,2}$	$x(t_j) - x_{j,3}$ $= e_{j,3}$	$x(t_j) - x_{j,4}$ $= e_{j,4}$
	$h_1 = 0.2$	$h_2 = 0.1$	$h_3 = 0.05$	$h_4 = 0.025$
0.0	0.000000	0.000000	0.000000	0.000000
0.2	0.002847	0.000704	0.000176	0.000044
0.4	0.005898	0.001460	0.000364	0.000091
0.6	0.009007	0.002229	0.000556	0.000139
0.8	0.012013	0.002974	0.000742	0.000185
1.0	0.014780	0.003660	0.000913	0.000228
1.2	0.017208	0.004263	0.001063	0.000265
1.4	0.019235	0.004766	0.001189	0.000297
1.6	0.020815	0.005160	0.001287	0.000322
1.8	0.021920	0.005435	0.001356	0.000338
2.0	0.022533	0.005588	0.001394	0.000348
2.2	0.022639	0.005615	0.001401	0.000350
2.4	0.022232	0.005516	0.001377	0.000344
2.6	0.021304	0.005285	0.001319	0.000329
2.8	0.019852	0.004926	0.001230	0.000308
3.0	0.017872	0.004435	0.001107	0.000277
3.2	0.015362	0.003812	0.000951	0.000237
3.4	0.012322	0.003059	0.000763	0.000191
3.6	0.008749	0.002171	0.000541	0.000135
3.8	0.004641	0.001152	0.000287	0.000072
4.0	0.000000	0.000000	0.000000	0.000000

MATLAB

El Programa 9.12 incluye una llamada al Programa 9.11 para que se resuelva el sistema tridiagonal (6), también es necesario que las funciones coeficientes $p(t)$, $q(t)$ y $r(t)$ (del problema de contorno (1)) se almacenen en archivos **p.m**, **q.m** y **r.m**, respectivamente.

Programa 9.11 (Sistemas tridiagonales). Resolución del sistema tri-diagonal $CX = B$, donde C es una matriz tridiagonal.

```
function X=trisys(A,D,C,B)
% Datos
%
% - A es la subdiagonal de la matriz de los coeficientes
% - D es la diagonal principal de la matriz de
%   los coeficientes
% - C es la superdiagonal de la matriz de
```

Tabla 9.19 Extrapolación de Richardson de las aproximaciones numéricas $\{x_{j,1}\}$, $\{x_{j,2}\}$, $\{x_{j,3}\}$ obtenidas con el método de las diferencias finitas.

t_j	$\frac{4x_{j,2} - x_{j,1}}{3} = z_{j,1}$	$\frac{4x_{j,3} - x_{j,2}}{3} = z_{j,2}$	$\frac{16z_{j,2} - z_{j,1}}{3}$	$x(t_j)$ Solución exacta
0.0	1.250000	1.250000	1.250000	1.250000
0.2	1.317360	1.317351	1.317350	1.317350
0.4	1.326524	1.326506	1.326504	1.326505
0.6	1.281792	1.281764	1.281762	1.281762
0.8	1.189451	1.189414	1.189412	1.189412
1.0	1.056932	1.056889	1.056886	1.056886
1.2	0.892138	0.892090	0.892086	0.892086
1.4	0.703003	0.702951	0.702947	0.702948
1.6	0.497246	0.497191	0.497187	0.497187
1.8	0.282244	0.282188	0.282184	0.282184
2.0	0.064991	0.064935	0.064931	0.064931
2.2	-0.147918	-0.147973	-0.147977	-0.147977
2.4	-0.350268	-0.350322	-0.350325	-0.350325
2.6	-0.536207	-0.536258	-0.536261	-0.536261
2.8	-0.700213	-0.700259	-0.700263	-0.700262
3.0	-0.837072	-0.837113	-0.837116	-0.837116
3.2	-0.941850	-0.941885	-0.941888	-0.941888
3.4	-1.009870	-1.009898	-1.009899	-1.009899
3.6	-1.036688	-1.036707	-1.036708	-1.036708
3.8	-1.018075	-1.018085	-1.018086	-1.018086
4.0	-0.950000	-0.950000	-0.950000	-0.950000

```

% los coeficientes
% - B es el vector de los términos independientes
%   del sistema lineal
% Resultado
% - X es el vector solución

N=length(B);
for k=2:N
    mult=A(k-1)/D(k-1);
    D(k)=D(k)-mult*C(k-1);
    B(k)=B(k)-mult*B(k-1);
end
X(N)=B(N)/D(N);
for k= N-1:-1:1
    X(k)=(B(k)-C(k)*X(k+1))/D(k);
end

```

Programa 9.12 (Método de diferencias finitas). Construcción de una aproximación a la solución del problema de contorno $x'' = p(t)x'(t) + q(t)x(t) + r(t)$, con $x(a) = \alpha$ y $x(b) = \beta$ en el intervalo $[a, b]$, usando el método de diferencias finitas de orden $O(h^2)$.

Observación. Los nodos son $a = t_1 < \dots < t_{N+1} = b$ y los puntos de la solución numérica son $\{(t_j, x_j)\}_{j=1}^{N+1}$.

```

function F=findiff(p,q,r,a,b,alpha,beta,N)
% Datos
% - p, q y r son las funciones coeficientes de la ecuación
%   almacenadas como cadenas de caracteres 'p', 'q' y 'r'
% - a y b son los extremos izquierdo y derecho
%   del intervalo
% - alpha=x(a) y beta=x(b) son los valores en la frontera
% - N es el número de pasos
% Resultado
% - F=[T' X'] siendo T' el vector de orden 1 x N de
%   los nodos y X' el vector 1 x N de los valores
% Inicialización de los vectores y de h
T=zeros(1,N+1);
X=zeros(1,N-1);
Va=zeros(1,N-2);
Vb=zeros(1,N-1);
Vc=zeros(1,N-2);
Vd=zeros(1,N-1);
h=(b-a)/N;
% Cálculo del vector de los términos independientes B en AX=B
Vt=a+h:h:a+h*(N-1);
Vb=-h^2*feval(r,Vt);
Vb(1)=Vb(1)+(1+h/2*feval(p,Vt(1)))*alpha;
Vb(N-1)=Vb(N-1)+(1-h/2*feval(p,Vt(N-1)))*beta;
% Cálculo de la diagonal principal de A en AX=B
Vd=2+h^2*feval(q,Vt);
% Cálculo de la superdiagonal de A en AX=B
Vta=Vt(1,2:N-1);
Va=-1-h/2*feval(p,Vta);
% Cálculo de la subdiagonal de A en AX=B
Vtc=Vt(1,1:N-2);
Vc=-1+h/2*feval(p,Vtc);
% Resolución de AX=B usando trisys
X=trisys(Va,Vd,Vc,Vb);
T=[a,Vt,b];

```

```
X=[alpha,X,beta];
F=[T' X'];
```

Ejercicios

En los Ejercicios 1 a 3, use el método de diferencias finitas para aproximar $x(a+0.5)$.

- (a) Tome $h_1 = 0.5$ y realice un paso haciendo las operaciones a mano. Luego tome $h_2 = 0.25$ y realice dos pasos haciendo las operaciones a mano
 - (b) Use la técnica de extrapolación de Richardson con los valores obtenidos en el apartado (a) para generar una aproximación mejor (o sea, la dada por $z_{j,1} = (4x_{j,2} - x_{j,1})/3$).
 - (c) Compare sus resultados de los apartados (a) y (b) con el valor exacto $x(a+0.5)$.
1. $x'' = 2x' - x + t^2 - 1$ en $[0, 1]$ con $x(0) = 5$ y $x(1) = 10$
 $x(t) = t^2 + 4t + 5$
 2. $x'' + (1/t)x' + (1 - 1/(4t^2))x = 0$ en $[1, 6]$ con $x(1) = 1$ y $x(6) = 0$
 $x(t) = \frac{0.2913843206\cos(t) + 1.001299385\sin(t)}{\sqrt{t}}$
 3. $x'' - (1/t)x' + (1/t^2)x = 1$ en $[0.5, 4.5]$ con $x(0.5) = 1$ y $x(4.5) = 2$
 $x(t) = t^2 - 0.2525826491t - 2.528442297t \ln(t)$
 4. Supongamos que p , q y r son continuas en el intervalo $[a, b]$ y que $q(t) \geq 0$ para $a \leq t \leq b$. Pruebe que si h verifica $0 < h < 2/M$, donde $M = \max_{a \leq t \leq b} \{|p(t)|\}$, entonces la matriz de los coeficientes del sistema (6) es de diagonal estrictamente dominante y el sistema tiene solución única.
 5. Supongamos que $p(t) \equiv C_1 > 0$ y $q(t) \equiv C_2 > 0$. (a) Escriba el sistema tridiagonal lineal correspondiente a esta situación. (b) Pruebe que el sistema es de diagonal estrictamente dominante y, por tanto, tiene solución única supuesto que $C_1/C_2 \leq h$.

Algoritmos y programas

1. Use los Programas 9.11 y 9.12 para resolver el correspondiente problema de contorno tomando los tamaños de paso $h = 0.1$ y $h = 0.01$. Dibuje sus soluciones aproximadas y la solución exacta en una misma gráfica.
 - (a) $x'' = 2x' - x + t^2 - 1$ en $[0, 1]$ con $x(0) = 5$ y $x(1) = 10$
 $x(t) = t^2 + 4t + 5$
 - (b) $x'' + (1/t)x' + (1 - 1/(4t^2))x = 0$ en $[1, 6]$ con $x(1) = 1$ y $x(6) = 0$
 $x(t) = \frac{0.2913843206\cos(t) + 1.001299385\sin(t)}{\sqrt{t}}$

556 CAP. 9 ECUACIONES DIFERENCIALES ORDINARIAS

- (c) $x'' - (1/t)x' + (1/t^2)x = 1$ en $[0.5, 4.5]$ con $x(0.5) = 1$ y $x(4.5) = 2$
 $x(t) = t^2 - 0.2525826491t - 2.528442297t \ln(t)$

En los Problemas 2 a 7, use los Programas 9.11 y 9.12 para resolver el correspondiente problema de contorno tomando tamaños de paso $h = 0.2$, $h = 0.1$ y $h = 0.05$. En cada problema, dibuje las tres soluciones en una misma gráfica.

2. $x'' = (-2/t)x' + (2/t^2)x + (10 \cos(\ln(t)))/t^2$ en $[1, 3]$ con $x(1) = 1$ y $x(3) = -1$
3. $x'' = -5x' - 6x + te^{-2t} + 3.9 \cos(3t)$ en $[0, 3]$ con $x(0) = 0.95$ y $x(3) = 0.15$
4. $x'' = -4x' - 4x + 5 \cos(4t) + \operatorname{sen}(2t)$ en $[0, 2]$ con $x(0) = 0.75$ y $x(2) = 0.25$
5. $x'' = -2x' - 2x + e^{-t} + \operatorname{sen}(2t)$ en $[0, 4]$ con $x(0) = 0.6$ y $x(4) = -0.1$
6. $x'' + (2/t)x' - (2/t^2)x = \operatorname{sen}(t)/t^2$ en $[1, 6]$ con $x(1) = -0.02$ y $x(6) = 0.02$
7. $x'' + (1/t)x' + (1 - 1/(4t^2))x = \sqrt{t} \cos(t)$ en $[1, 6]$ con $x(1) = 1.0$ y $x(6) = -0.5$
8. Diseñe un programa que utilice los Programas 9.11 y 9.12 como subprogramas y permita llevar a cabo el proceso de extrapolación ilustrado en el Ejemplo 9.18 y en la Tabla 9.19.
9. Con cada uno de los problemas de contorno que se plantean, utilice su programa del Problema 8 tomando los tamaños de paso $h = 0.1$, $h = 0.05$ y $h = 0.025$ para construir una tabla análoga a la Tabla 9.19. Dibuje su solución extrapolada y la solución exacta en una misma gráfica.
 - (a) $x'' = 2x' - x + t^2 - 1$ en $[0, 1]$ con $x(0) = 5$ y $x(1) = 10$
 $x(t) = t^2 + 4t + 5$
 - (b) $x'' + (1/t)x' + (1 - 1/(4t^2))x = 0$ en $[1, 6]$ con $x(1) = 1$ y $x(6) = 0$
 $x(t) = \frac{0.2913843206 \cos(t) + 1.001299385 \operatorname{sen}(t)}{\sqrt{t}}$
 - (c) $x'' - (1/t)x' + (1/t^2)x = 1$ en $[0.5, 4.5]$ con $x(0.5) = 1$ y $x(4.5) = 2$
 $x(t) = t^2 - 0.2525826491t - 2.528442297t \ln(t)$

Ecuaciones en derivadas parciales

Muchos problemas en ciencia aplicada, física e ingeniería se modelan matemáticamente mediante ecuaciones en derivadas parciales. Una ecuación diferencial en la que aparecen dos o más variables independientes se llama **ecuación en derivadas parciales**. No es necesario haber seguido un curso especializado de ecuaciones en derivadas parciales para entender los principios rudimentarios involucrados en la obtención de soluciones numéricas. En este capítulo estudiaremos los métodos de las diferencias finitas que se basan en las fórmulas para aproximar las derivadas primera y segunda de una función. Vamos a empezar clasificando los tres tipos de ecuaciones que investigaremos e introduciendo un problema físico típico de cada clase. Una ecuación en derivadas parciales de la forma

$$(1) \quad A\Phi_{xx} + B\Phi_{xy} + C\Phi_{yy} = f(x, y, \Phi, \Phi_x, \Phi_y),$$

donde A , B y C son constantes, se llama **casi-lineal** y hay tres tipos de ecuaciones casi-lineales:

- (2) Si $B^2 - 4AC < 0$, la ecuación se llama **elíptica**.
- (3) Si $B^2 - 4AC = 0$, la ecuación se llama **parabólica**.
- (4) Si $B^2 - 4AC > 0$, la ecuación se llama **hiperbólica**.

Como ejemplo de una ecuación hiperbólica consideraremos el modelo unidimensional de la cuerda vibrante. El desplazamiento $u(x, t)$ de la cuerda viene

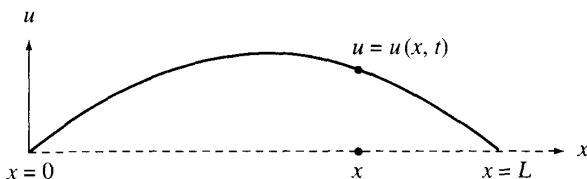


Figura 10.1 La ecuación de ondas modela las vibraciones de una cuerda.

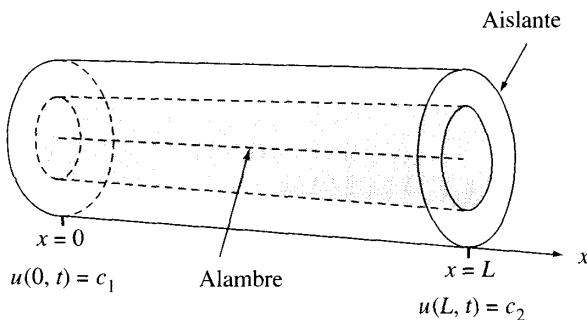


Figura 10.2 La ecuación del calor modela la temperatura de un alambre aislado.

gobernado por la ecuación de ondas

$$(5) \quad \rho u_{tt}(x, y) = T u_{xx}(x, t) \quad \text{para } 0 < x < L \text{ y } 0 < t < \infty,$$

con posición y velocidad iniciales dadas por

$$(6) \quad \begin{aligned} u(x, 0) &= f(x) && \text{para } t = 0 \text{ y } 0 \leq x \leq L, \\ u_t(x, 0) &= g(x) && \text{para } t = 0 \text{ y } 0 < x < L, \end{aligned}$$

y siendo los valores en los extremos de la cuerda

$$(7) \quad \begin{aligned} u(0, t) &= 0 && \text{para } x = 0 \text{ y } 0 \leq t < \infty, \\ u(L, t) &= 0 && \text{para } x = L \text{ y } 0 \leq t < \infty. \end{aligned}$$

La constante ρ es la masa de la cuerda por unidad de longitud y T es la tensión de la cuerda. En la Figura 10.1 se muestra el diagrama de una cuerda con extremos fijos situados en los puntos $(0, 0)$ y $(L, 0)$.

Como ejemplo de ecuación parabólica consideramos el modelo uni-dimensional del flujo de calor en un alambre aislado de longitud L (véase la Figura 10.2). La ecuación del calor, que nos da la temperatura $u(x, t)$ en la posición x del alambre y en el instante t , es

$$(8) \quad \kappa u_{xx}(x, t) = \sigma \rho u_t(x, t) \quad \text{para } 0 < x < L \text{ y } 0 < t < \infty,$$

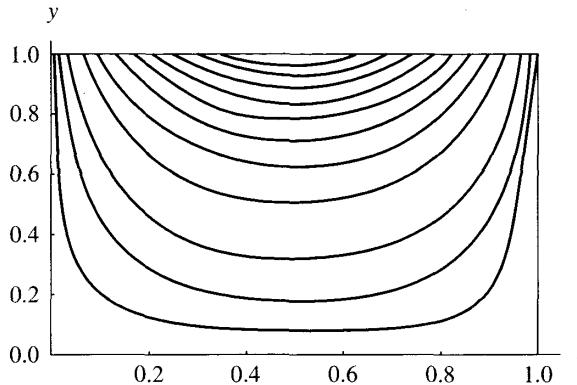


Figura 10.3 Curvas de nivel $u(x, y) = C$ de la solución de la ecuación de Laplace.

la distribución inicial de temperaturas es

$$(9) \quad u(x, 0) = f(x) \quad \text{para } t = 0 \text{ y } 0 \leq x \leq L,$$

y la condición de contorno en los extremos del alambre es

$$(10) \quad \begin{aligned} u(0, t) &= c_1 && \text{para } x = 0 \text{ y } 0 \leq t < \infty, \\ u(L, t) &= c_2 && \text{para } x = L \text{ y } 0 \leq t < \infty. \end{aligned}$$

La constante κ es el coeficiente de conductividad térmica, σ es el calor específico y ρ es la densidad del material.

Como ejemplo de una ecuación elíptica consideramos la función potencial $u(x, y)$, que podría representar el régimen permanente de un potencial electrostático o el régimen permanente de la distribución de la temperatura en una región rectangular del plano. Estas situaciones se modelan mediante la ecuación de Laplace en un rectángulo:

$$(11) \quad u_{xx}(x, y) + u_{yy}(x, y) = 0 \quad \text{para } 0 < x < 1 \text{ y } 0 < y < 1,$$

con condiciones de contorno especificadas:

$$\begin{aligned} u(x, 0) &= f_1(x) && \text{para } y = 0 \text{ y } 0 \leq x \leq 1 \text{ (abajo),} \\ u(x, 1) &= f_2(x) && \text{para } y = 1 \text{ y } 0 \leq x \leq 1 \text{ (arriba),} \\ u(0, y) &= f_3(y) && \text{para } x = 0 \text{ y } 0 \leq y \leq 1 \text{ (a la izquierda),} \\ u(1, y) &= f_4(y) && \text{para } x = 1 \text{ y } 0 \leq y \leq 1 \text{ (a la derecha).} \end{aligned}$$

En la Figura 10.3 se muestran algunas curvas de nivel de la función $u(x, y)$ con condiciones de contorno $f_1(x) = 0$, $f_2(x) = \sin(\pi x)$, $f_3(y) = 0$ y $f_4(y) = 0$ en el cuadrado $R = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$.

10.1 Ecuaciones hiperbólicas

La ecuación de ondas

Como ejemplo de una ecuación en derivadas parciales hiperbólica, consideraremos la ecuación de ondas

$$(1) \quad u_{tt}(x, t) = c^2 u_{xx}(x, t) \quad \text{para } 0 < x < a \text{ y } 0 < t < b,$$

con las condiciones de contorno

$$(2) \quad \begin{aligned} u(0, t) &= 0 & \text{y} & \quad u(a, t) = 0 & \quad \text{para } 0 \leq t \leq b, \\ u(x, 0) &= f(x) & & & \quad \text{para } 0 \leq x \leq a, \\ u_t(x, 0) &= g(x) & & & \quad \text{para } 0 < x < a. \end{aligned}$$

La ecuación de ondas modela el desplazamiento u desde su posición de equilibrio de una cuerda elástica vibrante cuyos extremos, de coordenadas $x = 0$ y $x = a$, están fijos. Aunque es posible determinar la solución exacta de la ecuación de ondas por medio de las series de Fourier, vamos a usar este problema como prototipo de la situación que se da en las ecuaciones hiperbólicas.

Construcción de la ecuación en diferencias

Hacemos una partición del rectángulo $R = \{(x, t) : 0 \leq x \leq a, 0 \leq t \leq b\}$ en una malla que consta de $n - 1$ por $m - 1$ rectángulos de lados $\Delta x = h$ y $\Delta t = k$, como la que se muestra en la Figura 10.4. Empezamos por la fila de abajo, donde $t = t_1 = 0$ y sabemos que la solución es $u(x_i, t_1) = f(x_i)$. Ahora usaremos una ecuación en diferencias para calcular, en las filas sucesivas, las aproximaciones a la solución exacta, que en los puntos de la malla es $u(x_i, t_j)$. O sea, para cada $j = 2, 3, \dots, m$, calcularemos

$$\{u_{i,j} \approx u(x_i, t_j) : i = 1, 2, \dots, n\}.$$

Las fórmulas de diferencias centradas para aproximar $u_{tt}(x, t)$ y $u_{xx}(x, t)$ son

$$(3) \quad u_{tt}(x, t) = \frac{u(x, t + k) - 2u(x, t) + u(x, t - k)}{k^2} + O(k^2)$$

y

$$(4) \quad u_{xx}(x, t) = \frac{u(x + h, t) - 2u(x, t) + u(x - h, t)}{h^2} + O(h^2).$$

El espaciado entre los puntos de la malla es uniforme en todas las filas: $x_{i+1} = x_i + h$ (y $x_{i-1} = x_i - h$) y también es uniforme en todas las columnas:

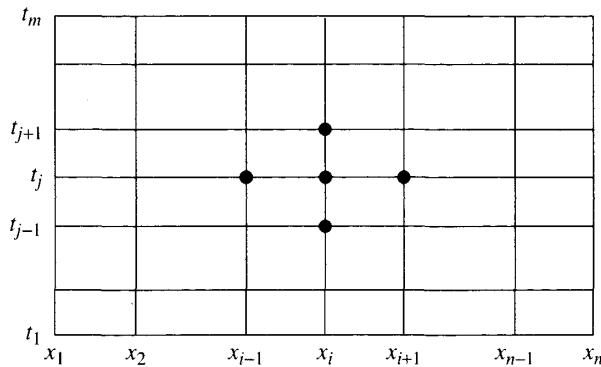


Figura 10.4 La malla para resolver $u_{tt}(x, t) = c^2 u_{xx}(x, t)$ en la región R .

$t_{j+1} = t_j + k$ (y $t_{j-1} = t_j - k$). Teniendo esto en cuenta, obtenemos la ecuación en diferencias eliminando los términos de orden $\mathcal{O}(k^2)$ y $\mathcal{O}(h^2)$ de las relaciones (3) y (4), usando la aproximación $u_{i,j}$ en vez de $u(x_i, t_j)$ en dichas relaciones (3) y (4) y sustituyendo las relaciones resultantes en (1); todo lo cual produce

$$(5) \quad \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{k^2} = c^2 \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2},$$

que es la ecuación en diferencias que usaremos como aproximación a la ecuación diferencial (1). Por comodidad, introducimos en la ecuación (5) la constante $r = ck/h$, obteniendo

$$(6) \quad u_{i,j+1} - 2u_{i,j} + u_{i,j-1} = r^2(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}).$$

Reordenando un poco los términos, observamos que podemos emplear la ecuación (6) para determinar las aproximaciones a la solución en los puntos de la fila $(j+1)$ -ésima de la malla, supuesto que conocemos las aproximaciones a la solución en los puntos de las dos filas anteriores, la j -ésima y la $(j-1)$ -ésima:

$$(7) \quad u_{i,j+1} = (2 - 2r^2)u_{i,j} + r^2(u_{i+1,j} + u_{i-1,j}) - u_{i,j-1},$$

para $i = 2, 3, \dots, n-1$. En la Figura 10.5 se muestra la posición en la malla de los cuatro valores conocidos que aparecen en el miembro derecho de la relación (7), los que se usan para determinar la aproximación $u_{i,j+1}$.

Hay que tener cuidado al usar la fórmula (7). Si el error cometido en una etapa de los cálculos no se amplifica en las etapas posteriores, entonces se dice que el método es estable. Para garantizar la estabilidad de la fórmula (7) es necesario que $r = ck/h \leq 1$. Existen otros esquemas, llamados métodos implícitos, que son de desarrollo más complejo pero no imponen restricciones sobre r para que se tenga estabilidad (véase la Referencia [90]).

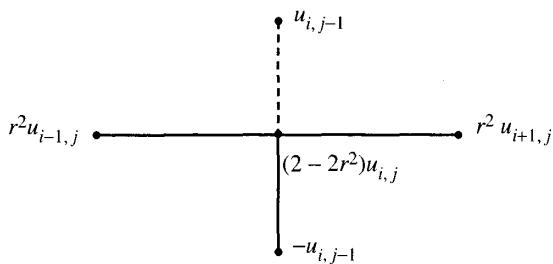


Figura 10.5 El esquema de la ecuación en diferencias para la ecuación de ondas.

Los valores iniciales

Si queremos usar la fórmula (7) para calcular las aproximaciones en los puntos de la tercera fila de la malla, es necesario disponer de las aproximaciones en los puntos de las filas primera y segunda. Los valores de la primera fila ya los tenemos, vienen dados por la función f . Sin embargo, los valores de la segunda fila no se suelen proporcionar, así que se usa la función $g(x)$, dada en el contorno, para conseguir las aproximaciones en los puntos de esta segunda fila. Fijemos $x = x_i$ en la frontera inferior de R y apliquemos la fórmula de Taylor de orden 1 para desarrollar $u(x, t)$ alrededor de $(x_i, 0)$; para el valor $u(x_i, k)$ se verifica

$$(8) \quad u(x_i, k) = u(x_i, 0) + u_t(x_i, 0)k + O(k^2).$$

Ahora usamos $u(x_i, 0) = f(x_i) = f_i$ y $u_t(x_i, 0) = g(x_i) = g_i$ en (8) para obtener la fórmula con la que conseguimos las aproximaciones numéricas en los puntos de la segunda fila (recordemos que $t_2 = k$):

$$(9) \quad u_{i,2} = f_i + kg_i \quad \text{para } i = 2, 3, \dots, n-1.$$

Normalmente, $u(x_i, t_2) \neq u_{i,2}$, así que el error introducido al usar la fórmula (9) se propagará a toda la malla sin atenuarse cuando usemos el esquema dado por la fórmula (7). Por consiguiente, para evitar que los valores $u_{i,2}$ calculados con la fórmula (9) introduzcan en el proceso un error de truncamiento apreciable, es aconsejable que se tome un tamaño de paso k muy pequeño.

A menudo se da el caso de que la función $f(x)$ dada en el contorno es dos veces derivable en el intervalo, con lo cual tenemos que $u_{xx}(x, 0) = f''(x)$, igualdad que nos permite usar la fórmula de Taylor de orden $n = 2$ para obtener una aproximación mejorada a los valores de la segunda fila de la malla. Para hacer esto, volvemos a la ecuación de ondas y, usando la relación entre las derivadas parciales segundas, obtenemos

$$(10) \quad u_{tt}(x_i, 0) = c^2 u_{xx}(x_i, 0) = c^2 f''(x_i) = c^2 \frac{f_{i+1} - 2f_i + f_{i-1}}{h^2} + O(h^2).$$

Recordando que la fórmula de Taylor de orden 2 es

$$(11) \quad u(x, k) = u(x, 0) + u_t(x, 0)k + \frac{u_{tt}(x, 0)k^2}{2} + \mathcal{O}(k^3)$$

y aplicando esta expresión (11) en el punto $x = x_i$ junto con las expresiones (9) y (10), obtenemos

$$(12) \quad u(x_i, k) = f_i + kg_i + \frac{c^2 k^2}{2h^2} (f_{i+1} - 2f_i + f_{i-1}) + \mathcal{O}(h^2)\mathcal{O}(k^2) + \mathcal{O}(k^3).$$

Puesto que $r = ck/h$, podemos simplificar la expresión (12) y obtener la siguiente fórmula de diferencias que nos proporciona aproximaciones numéricicas mejoradas a los elementos de la segunda fila:

$$(13) \quad u_{i,2} = (1 - r^2)f_i + kg_i + \frac{r^2}{2}(f_{i+1} + f_{i-1})$$

para $i = 2, 3, \dots, n - 1$.

La solución de D'Alembert

El matemático francés Jean Le Rond d'Alembert (1717–1783) descubrió que

$$(14) \quad u(x, t) = F(x + ct) + G(x - ct)$$

es una solución de la ecuación de ondas (1) en el intervalo $0 \leq x \leq a$, si ambas funciones F y G son dos veces derivables. Podemos comprobar esto sustituyendo directamente en la ecuación de ondas (1) las derivadas parciales de segundo orden de la función definida en la expresión (14), que son

$$(15) \quad u_{tt}(x, t) = c^2 F''(x + ct) + c^2 G''(x - ct),$$

$$(16) \quad u_{xx}(x, t) = F''(x + ct) + G''(x - ct).$$

En efecto, al sustituir estas expresiones en (1) obtenemos

$$\begin{aligned} u_{tt}(x, t) &= c^2 F''(x + ct) + c^2 G''(x - ct) \\ &= c^2 (F''(x + ct) + G''(x - ct)) \\ &= c^2 u_{xx}(x, t). \end{aligned}$$

La solución particular que verifica las condiciones iniciales y las condiciones de contorno $u(x, 0) = f(x)$ y $u_t(x, 0) = 0$ viene dada por las extensiones impares y periódicas de período $2a$ definidas para $0 \leq x \leq a$ por $F(x) = G(x) = f(x)/2$; comprobación que dejamos como ejercicio.

Cuando se conocen dos filas exactamente

La precisión de las aproximaciones numéricas obtenidas mediante la fórmula (7) depende del error de truncamiento de las fórmulas que se utilizan para convertir la ecuación diferencial en una ecuación en diferencias. Aunque es raro que se conozcan los valores de la solución exacta en la segunda fila, si esto fuera posible entonces, tomando como incremento $k = ch$ en el eje de la variable t , el proceso generaría la solución exacta en todos los demás puntos de la malla.

Teorema 10.1. Supongamos que los valores en las dos primeras filas de la malla $u_{i,1} = u(x_i, 0)$ y $u_{i,2} = u(x_i, k)$, para $i = 1, 2, \dots, n$, son los que toma la solución exacta de la ecuación de ondas (1). Si el tamaño de paso en el eje de la variable t es $k = h/c$, entonces $r = 1$ y la fórmula (7) se transforma en

$$(17) \quad u_{i,j+1} = u_{i+1,j} + u_{i-1,j} - u_{i,j-1}.$$

Además, en este caso, las soluciones obtenidas mediante el método de las diferencias finitas (17) son exactas (si no tenemos en cuenta los errores de redondeo introducidos por el computador) en todos los nodos de la malla.

Demostración. Usando la solución de d'Alembert y la relación $ck = h$, tenemos $x_i - ct_j = (i-1)h - c(j-1)k = (i-1)h - (j-1)h = (i-j)h$ y, de manera similar, $x_i + ct_j = (i+j-2)h$. Ahora usamos estas igualdades en la ecuación (14) y obtenemos la siguiente fórmula particular para el valor $u(t_i, x_j)$:

$$(18) \quad u(t_i, x_j) = F((i-j)h) + G((i+j-2)h),$$

para $i = 1, 2, \dots, n$ y $j = 1, 2, \dots, m$. Supongamos, por inducción en j , que $u_{i,k} = u(t_i, x_k)$ para $i = 1, 2, \dots, n$ y $k = 1, 2, \dots, j$ con $j \geq 2$. Entonces, usando la hipótesis de inducción con los términos $u_{i+1,j}$, $u_{i-1,j}$ y $u_{i,j-1}$ del miembro derecho de la expresión (17), obtenemos

$$\begin{aligned} u_{i,j+1} &= u_{i+1,j} + u_{i-1,j} - u_{i,j-1} \\ &= F((i+1-j)h) + F((i-1-j)h) \\ &\quad - F((i-(j-1))h) + G((i+1+j-2)h) \\ &\quad + G((i-1+j-2)h) - G((i+j-1-2)h) \\ &= F((i-(j+1))h) + G((i+j+1-2)h) = u(t_i, x_{j+1}), \end{aligned}$$

para $i = 1, 2, \dots, n$.

Atención. El Teorema 10.1 no garantiza que las soluciones numéricas sean exactas cuando los cálculos se realizan usando las fórmulas (9) y (13) como aproximaciones de los valores $u_{i,2}$ de la segunda fila. De hecho, se introduce un error de truncamiento si $u_{i,2} \neq u(x_i, k)$ para algún i , con $1 \leq i \leq n$. Por esta razón recomendamos que se calculen las mejores aproximaciones posibles a los valores de la segunda fila usando las aproximaciones de Taylor de segundo orden dadas por la expresión (13).

Ejemplo 10.1. Vamos a usar el método de las diferencias finitas para resolver la ecuación de ondas de una cuerda vibrante:

$$(19) \quad u_{tt}(x, t) = 4u_{xx}(x, t) \quad \text{para } 0 < x < 1 \text{ y } 0 < t < 0.5,$$

con las condiciones de contorno

$$(20) \quad \begin{aligned} u(0, t) &= 0 & \text{y} & u(1, t) = 0 & \text{para } 0 \leq t \leq 0.5, \\ u(x, 0) &= f(x) = \operatorname{sen}(\pi x) + \operatorname{sen}(2\pi x) & \text{para } 0 \leq x \leq 1, \\ u_t(x, 0) &= g(x) = 0 & \text{para } 0 \leq x \leq 1. \end{aligned}$$

Por conveniencia tomamos $h = 0.1$ y $k = 0.05$. Puesto que $c = 2$, entonces $r = ck/h = 2(0.05)/0.1 = 1$. Como $g(x) = 0$ y $r = 1$, la fórmula (13) para calcular los valores de la segunda fila queda

$$(21) \quad u_{i,2} = \frac{f_{i-1} + f_{i+1}}{2} \quad \text{para } i = 2, 3, \dots, 9.$$

Sustituyendo $r = 1$ en la ecuación (7) obtenemos la ecuación en diferencias, ya simplificada,

$$(22) \quad u_{i,j+1} = u_{i+1,j} + u_{i-1,j} - u_{i,j-1}.$$

Usando las fórmulas dadas en (21) y, sucesivamente, en (22) generamos las aproximaciones a los valores $u(x, t)$ que se recogen en la Tabla 10.1 para $0 < x_i < 1$ y $0 \leq t_j \leq 0.50$.

Los valores numéricos dados en la Tabla 10.1 coinciden en más de seis cifras decimales con los correspondientes a la solución exacta

$$u(x, t) = \operatorname{sen}(\pi x) \cos(2\pi t) + \operatorname{sen}(2\pi x) \cos(4\pi t).$$

En la Figura 10.6 se muestra una representación tridimensional de los valores recogidos en la Tabla 10.1. ■

Ejemplo 10.2. Vamos a usar el método de las diferencias finitas para resolver la ecuación de ondas de una cuerda vibrante:

$$(23) \quad u_{tt}(x, t) = 4u_{xx}(x, t) \quad \text{para } 0 < x < 1 \text{ y } 0 < t < 0.5,$$

con las condiciones de contorno

$$(24) \quad \begin{aligned} u(0, t) &= 0 & \text{y} & u(1, t) = 0 & \text{para } 0 \leq t \leq 1, \\ u(x, 0) &= f(x) = \begin{cases} x & \text{para } 0 \leq x \leq \frac{3}{5} \\ 1.5 - 1.5x & \text{para } \frac{3}{5} \leq x \leq 1, \end{cases} \\ u_t(x, 0) &= g(x) = 0 & \text{para } 0 < x < 1. \end{aligned}$$

Por conveniencia tomamos $h = 0.1$ y $k = 0.05$. Puesto que $c = 2$, entonces tenemos otra vez que $r = 1$. Usando las fórmulas (21) y, sucesivamente, (22) generamos las aproximaciones a los valores $u(x, t)$ que se recogen en la Tabla 10.2 para $0 \leq x_i \leq 1$ y $0 \leq t_j \leq 0.50$. En la Figura 10.7 se muestra una representación tridimensional de los valores recogidos en la Tabla 10.2. ■

Tabla 10.1 Solución de la ecuación de ondas (19) con las condiciones de contorno dadas en (20).

t_j	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
0.00	0.896802	1.538842	1.760074	1.538842	1.000000	0.363271	-0.142040	-0.363271	-0.278768
0.05	0.769421	1.328438	1.538842	1.380037	0.951056	0.428980	0.000000	-0.210404	-0.181636
0.10	0.431636	0.769421	0.948401	0.951056	0.809017	0.587785	0.360616	0.181636	0.068364
0.15	0.000000	0.051599	0.181636	0.377381	0.587785	0.740653	0.769421	0.639384	0.363271
0.20	-0.380037	-0.587785	-0.519421	-0.181636	0.309017	0.769421	1.019421	0.951056	0.571020
0.25	-0.587785	-0.951056	-0.951056	-0.587785	0.000000	0.587785	0.951056	0.951056	0.587785
0.30	-0.571020	-0.951056	-1.019421	-0.769421	-0.309017	0.181636	0.519421	0.587785	0.380037
0.35	-0.363271	-0.639384	-0.769421	-0.740653	-0.587785	-0.377381	-0.181636	-0.051599	0.000000
0.40	-0.068364	-0.181636	-0.360616	-0.587785	-0.809017	-0.951056	-0.948401	-0.769421	-0.431636
0.45	0.181636	0.210404	0.000000	-0.428980	-0.951056	-1.380037	-1.538842	-1.328438	-0.769421
0.50	0.278768	0.363271	0.142040	-0.363271	-1.000000	-1.538842	-1.760074	-1.538842	-0.896802

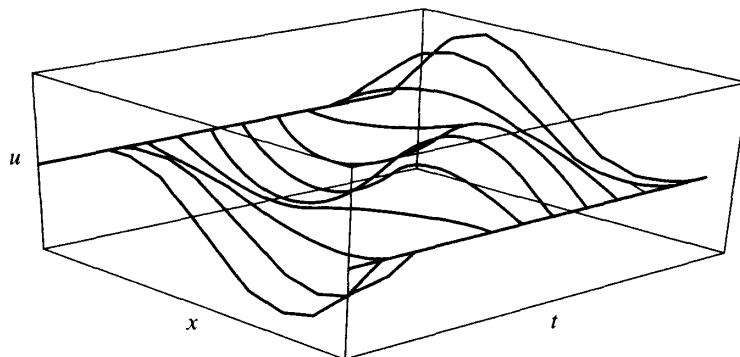


Figura 10.6 La cuerda vibrante de la ecuación (19) con las condiciones de contorno dadas en (20).

MATLAB

El Programa 10.1 permite aproximar la solución de la ecuación de ondas (1) con las condiciones de contorno (2). Usando las instrucciones `mesh(U)` o `surf(U)` podemos conseguir una representación tridimensional de la matriz U en la que se almacena la solución numérica; además, la instrucción `contour(U)` permite obtener una gráfica análoga a la de la Figura 10.3.

Programa 10.1 (Resolución de la ecuación de ondas por el método de las diferencias finitas). Construcción de la solución numérica de $u_{tt}(x, t) = c^2 u_{xx}(x, t)$ en $R = \{(x, t) : 0 \leq x \leq a, 0 \leq t \leq b\}$ con $u(0, t) = 0$, $u(a, t) = 0$ para $0 \leq t \leq b$ y $u(x, 0) = f(x)$, $u_t(x, 0) = g(x)$ para $0 \leq x \leq a$.

```
function U = finedif(f,g,a,b,c,n,m)
% Datos
%      - f=u(x,0) dada como una cadena de caracteres 'f'
```

Tabla 10.2 Solución de la ecuación de ondas (23) con las condiciones de contorno dadas en (24).

t_j	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
0.00	0.100	0.200	0.300	0.400	0.500	0.600	0.450	0.300	0.150
0.05	0.100	0.200	0.300	0.400	0.500	0.475	0.450	0.300	0.150
0.10	0.100	0.200	0.300	0.400	0.375	0.350	0.325	0.300	0.150
0.15	0.100	0.200	0.300	0.275	0.250	0.225	0.200	0.175	0.150
0.20	0.100	0.200	0.175	0.150	0.125	0.100	0.075	0.050	0.025
0.25	0.100	0.075	0.050	0.025	0.000	-0.025	-0.050	-0.075	-0.100
0.30	-0.025	-0.050	-0.075	-0.100	-0.125	-0.150	-0.175	-0.200	-0.100
0.35	-0.150	-0.175	-0.200	-0.225	-0.250	-0.275	-0.300	-0.200	-0.100
0.40	-0.150	-0.300	-0.325	-0.350	-0.375	-0.400	-0.300	-0.200	-0.100
0.45	-0.150	-0.300	-0.450	-0.475	-0.500	-0.400	-0.300	-0.200	-0.100
0.50	-0.150	-0.300	-0.450	-0.600	-0.500	-0.400	-0.300	-0.200	-0.100

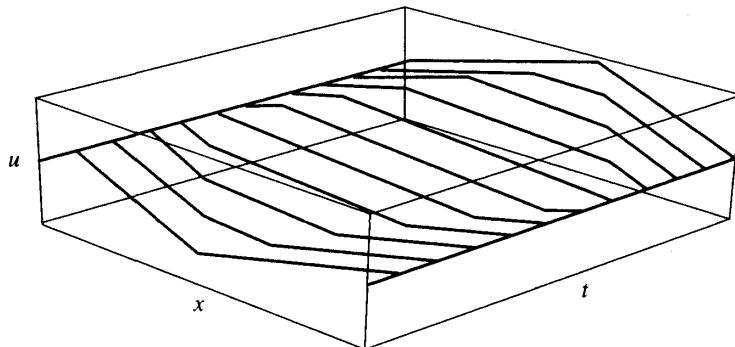


Figura 10.7 La cuerda vibrante de la ecuación de ondas (23) con las condiciones de contorno dadas en las ecuaciones (24).

```
% - g=ut(x,0) dada como una cadena de caracteres 'g'
% - a y b son los extremos superiores de
%   los intervalos [0,a] y [0,b]
% - c es la constante de la ecuación de ondas
% - n y m es el número de nodos en [0,a] y [0,b]
% Resultado
% - U es la matriz, análoga a la de la Tabla 10.1,
%   en la que se almacena la solución numérica
%
% Inicialización de los parámetros y de U
h=a/(n-1);
k=b/(m-1);
```

```

r=c*k/h;
r2=r^2;
r22=r^2/2;
s1=1-r^2;
s2=2-2*r^2;
U=zeros(n,m);

% Cálculo de las dos primeras filas
for i=2:n-1
    U(i,1)=feval(f,h*(i-1));
    U(i,2)=s1*feval(f,h*(i-1))+k*feval(g,h*(i-1)) ...
        +r22*(feval(f,h*i)+feval(f,h*(i-2)));
end

% Cálculo de las demás filas
for j=3:m,
    for i=2:(n-1),
        U(i,j) = s2*U(i,j-1)+r2*(U(i-1,j-1)+U(i+1,j-1))-U(i,j-2);
    end
end
U=U';

```

Ejercicios

- (a) Verifique por sustitución directa que $u(x, t) = \operatorname{sen}(n\pi x) \cos(2n\pi t)$ es una solución de la ecuación de ondas $u_{tt}(x, t) = 4u_{xx}(x, t)$ para cada número natural $n = 1, 2, \dots$
(b) Verifique por sustitución directa que $u(x, t) = \operatorname{sen}(n\pi x) \cos(cn\pi t)$ es una solución de la ecuación de ondas $u_{tt}(x, t) = c^2 u_{xx}(x, t)$ para cada número natural $n = 1, 2, \dots$
- Supongamos que la posición y velocidad iniciales de la cuerda son $u(x, 0) = f(x)$ y $u_t(x, 0) \equiv 0$, respectivamente. Demuestre que la solución de d'Alembert para este caso es

$$u(x, t) = \frac{f(x + ct) + f(x - ct)}{2}.$$

- Obtenga una forma más simple de la ecuación en diferencias (7) para $h = 2ck$.

En los Ejercicios 4 y 5, use el método de las diferencias finitas para calcular las tres primeras filas de la solución aproximada de la ecuación de ondas dada. Realice las operaciones a mano (o con calculadora).

4. $u_{tt}(x, t) = 4u_{xx}(x, t)$, para $0 \leq x \leq 1$ y $0 \leq t \leq 0.5$, con las condiciones de contorno

$$\begin{aligned} u(0, t) &= 0 & u(1, t) &= 0 & \text{para } 0 \leq t \leq 0.5, \\ u(x, 0) &= f(x) = \operatorname{sen}(\pi x) & & \text{para } 0 \leq x \leq 1, \\ u_t(x, 0) &= g(x) = 0 & & \text{para } 0 \leq x \leq 1. \end{aligned}$$

Tome $h = 0.2$, $k = 0.1$ y $r = 1$.

5. $u_{tt}(x, t) = 4u_{xx}(x, t)$, para $0 \leq x \leq 1$ y $0 \leq t \leq 0.5$, con las condiciones de contorno

$$\begin{aligned} u(0, t) &= 0 & u(1, t) &= 0 & \text{para } 0 \leq t \leq 0.5, \\ u(x, 0) &= f(x) = \begin{cases} \frac{5x}{2} & \text{para } 0 \leq x \leq \frac{3}{5}, \\ \frac{15 - 15x}{4} & \text{para } \frac{3}{5} \leq x \leq 1, \end{cases} \\ u_t(x, 0) &= g(x) = 0 & \text{para } 0 < x < 1. \end{aligned}$$

Tome $h = 0.2$, $k = 0.1$ y $r = 1$.

6. Supongamos que la posición y velocidad iniciales de la cuerda son $u(x, 0) = f(x)$ y $u_t(x, 0) = g(x)$, respectivamente. Demuestre que la solución de d'Alembert para este caso es

$$u(x, t) = \frac{f(x + ct) + f(x - ct)}{2} + \frac{1}{2c} \int_{x-ct}^{x+ct} g(s) ds.$$

7. En la ecuación $u_{tt}(x, t) = 9u_{xx}(x, t)$, ¿qué relación debe existir entre h y k para que la ecuación en diferencias que se obtenga venga dada por $u_{i,j+1} = u_{i+1,j} + u_{i-1,j} - u_{i,j-1}$?
8. ¿Qué dificultad puede aparecer cuando se intenta usar el método de las diferencias finitas para resolver $u_{tt}(x, t) = 4u_{xx}(x, t)$ tomando $k = 0.02$ y $h = 0.03$?

Algoritmos y programas

En los Problemas 1 a 8, use el Programa 10.1 para resolver la ecuación de ondas $u_{tt}(x, t) = c^2u_{xx}(x, t)$, para $0 \leq x \leq a$ y $0 \leq t \leq b$, con las condiciones de contorno

$$\begin{aligned} u(0, t) &= 0 & u(a, t) &= 0 & \text{para } 0 \leq t \leq b, \\ u(x, 0) &= f(x) & & \text{para } 0 \leq x \leq a, \\ u_t(x, 0) &= g(x) & & \text{para } 0 \leq x \leq a, \end{aligned}$$

empleando los valores dados en cada caso. Use las instrucciones `surf` y `contour` para dibujar sus soluciones aproximadas.

1. Use $a = 1$, $b = 1$, $c = 1$, $f(x) = \operatorname{sen}(\pi x)$ y $g(x) = 0$. Por conveniencia, tome $h = 0.1$ y $k = 0.1$.

2. Use $a = 1$, $b = 1$, $c = 1$, $f(x) = x - x^2$ y $g(x) = 0$. Por conveniencia, tome $h = 0.1$ y $k = 0.1$.
3. Use $a = 1$, $b = 1$, $c = 1$, $f(x) = \begin{cases} 2x & \text{para } 0 \leq x \leq \frac{1}{2}, \\ 2 - 2x & \text{para } \frac{1}{2} \leq x \leq 1, \end{cases}$, $g(x) = 0$, $h = 0.1$ y $k = 0.1$.
4. Use $a = 1$, $b = 1$, $c = 2$, $f(x) = \sin(\pi x)$, $g(x) = 0$, $h = 0.1$ y $k = 0.05$.
5. Use $a = 1$, $b = 1$, $c = 2$, $f(x) = x - x^2$, $g(x) = 0$, $h = 0.1$ y $k = 0.05$.
6. Repita el Problema 3, pero tomando $c = 2$ y $k = 0.05$.
7. Repita el Problema 1, pero tomando $f(x) = \sin(2\pi x) + \sin(4\pi x)$.
8. Repita el Problema 2, pero tomando $c = 2$, $f(x) = \sin(2\pi x) + \sin(4\pi x)$ y $k = 0.05$.

10.2 Ecuaciones parabólicas

La ecuación del calor

Como ejemplo de ecuación en derivadas parciales parabólica, consideramos la ecuación del calor unidimensional

$$(1) \quad u_t(x, t) = c^2 u_{xx}(x, t) \quad \text{para } 0 \leq x < a \text{ y } 0 < t < b,$$

con la condición inicial

$$(2) \quad u(x, 0) = f(x) \quad \text{para } t = 0 \text{ y } 0 \leq x \leq a$$

y las condiciones de contorno

$$(3) \quad \begin{aligned} u(0, t) &= g_1(t) \equiv c_1 && \text{para } x = 0 \text{ y } 0 \leq t \leq b, \\ u(a, t) &= g_2(t) \equiv c_2 && \text{para } x = a \text{ y } 0 \leq t \leq b. \end{aligned}$$

La ecuación del calor modela la distribución de temperaturas en un alambre aislado, cuyos extremos se mantienen a temperatura constante c_1 y c_2 , a partir de una distribución inicial de temperaturas a lo largo del alambre $f(x)$. Aunque se pueden calcular soluciones exactas de la ecuación del calor usando series de Fourier, vamos a usar este problema como prototipo de la resolución numérica de ecuaciones parabólicas.

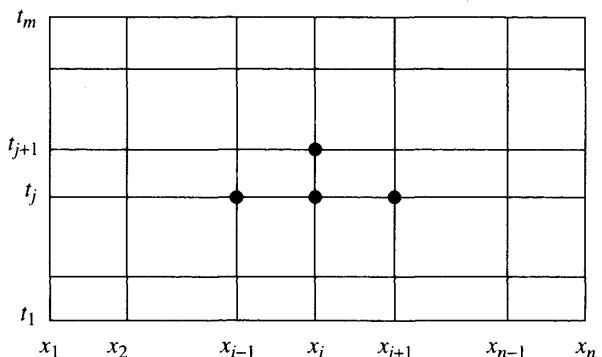


Figura 10.8 La malla para resolver $u_t(x, t) = c^2 u_{xx}(x, t)$ en la región R .

Construcción de la ecuación en diferencias

Dividimos el rectángulo $R = \{(x, t) : 0 \leq x \leq a, 0 \leq t \leq b\}$ en $n - 1$ por $m - 1$ rectángulos de lados $\Delta x = h$ y $\Delta t = k$, como se muestra en la Figura 10.8. Empezando en la fila de más abajo, donde $t = t_1 = 0$ y la solución es $u(x_i, t_1) = f(x_i)$, desarrollaremos un método para calcular las aproximaciones a los valores exactos $u(x, t)$ en los puntos de la malla: $\{u_{i,j} \approx u(x_i, t_j) : i = 1, 2, \dots, n\}$, para $j = 2, 3, \dots, m$.

Las fórmulas de diferencias que usamos para $u_t(x, t)$ y $u_{xx}(x, t)$ son, respectivamente,

$$(4) \quad u_t(x, t) = \frac{u(x, t+k) - u(x, t)}{k} + O(k)$$

y

$$(5) \quad u_{xx}(x, t) = \frac{u(x-h, t) - 2u(x, t) + u(x+h, t)}{h^2} + O(h^2).$$

Teniendo en cuenta que el tamaño de los rectángulos de la malla es uniforme en cada fila: $x_{i+1} = x_i + h$ (y $x_{i-1} = x_i - h$) y en cada columna: $t_{j+1} = t_j + k$, despreciando los términos $O(k)$ y $O(h^2)$, usando la aproximación $u_{i,j}$ en vez de $u(x_i, t_j)$ en las ecuaciones (4) y (5) y sustituyendo lo que se obtiene en la ecuación del calor (1), nos queda:

$$(6) \quad \frac{u_{i,j+1} - u_{i,j}}{k} = c^2 \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2},$$

que es una aproximación a la relación (1). Por comodidad, tomamos $r = c^2 k / h^2$ en (6) y reordenamos un poco los términos para obtener la ecuación en diferencias progresivas explícita

$$(7) \quad u_{i,j+1} = (1 - 2r)u_{i,j} + r(u_{i-1,j} + u_{i+1,j}).$$

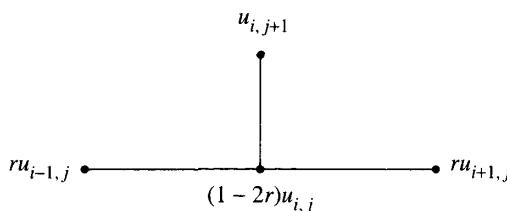


Figura 10.9 El esquema de diferencias progresivas.

La ecuación en diferencias (7) se emplea para calcular las aproximaciones en la fila $(j + 1)$ -ésima de la malla a partir de las aproximaciones de la fila anterior; hagamos notar que esta fórmula proporciona explícitamente el valor $u_{i,j+1}$ en función de $u_{i-1,j}$, $u_{i,j}$ y $u_{i+1,j}$. En la Figura 10.9 se representa el esquema computacional correspondiente a la fórmula (7).

La simplicidad de la fórmula (7) nos invita a usarla inmediatamente. Sin embargo, es importante usar técnicas numéricas que sean estables y la fórmula (7) no siempre lo es. Recordemos que un método es estable si cuando se introduce un error en una etapa del proceso, este error se va amortiguando hasta prácticamente desaparecer. La fórmula de diferencias progresivas (7) es estable si, y sólo si, $0 \leq r \leq \frac{1}{2}$. Esto significa que el tamaño de paso k debe cumplir $k \leq h^2/(2c^2)$; si esto no se cumple, entonces puede ocurrir que los errores introducidos en la fila $\{u_{i,j}\}$ se amplifiquen en alguna fila posterior $\{u_{i,p}\}$ para algún $p > j$. El siguiente ejemplo ilustra esta situación.

Ejemplo 10.3. Vamos a usar el método de las diferencias progresivas para resolver la ecuación del calor

$$(8) \quad u_t(x, t) = u_{xx}(x, t) \quad \text{para } 0 < x < 1 \text{ y } 0 < t < 0.20,$$

con las condiciones iniciales

$$(9) \quad u(x, 0) = f(x) = 4x - 4x^2 \quad \text{para } t = 0 \text{ y } 0 \leq x \leq 1$$

y las condiciones de contorno

$$(10) \quad \begin{aligned} u(0, t) &= g_1(t) \equiv 0 && \text{para } x = 0 \text{ y } 0 \leq t \leq 0.20, \\ u(1, t) &= g_2(t) \equiv 0 && \text{para } x = 1 \text{ y } 0 \leq t \leq 0.20. \end{aligned}$$

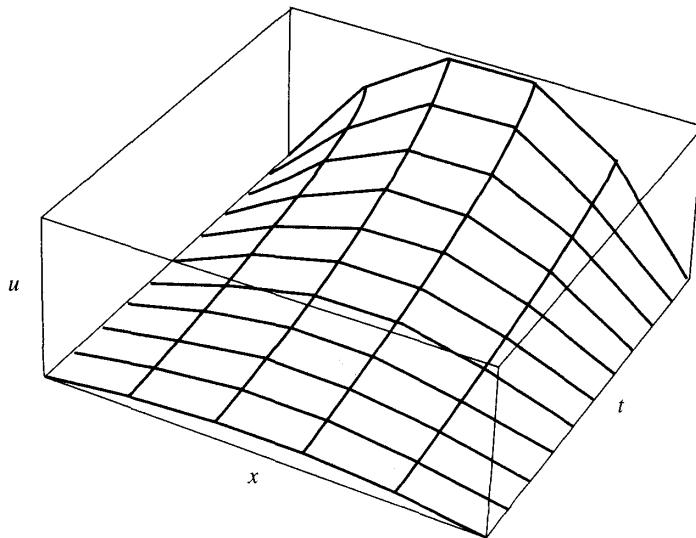
La primera vez usamos tamaños de paso $\Delta x = h = 0.2$ y $\Delta t = k = 0.02$ y $c = 1$, de manera que $r = 0.5$. La malla tendrá $n = 6$ columnas de ancho y $m = 11$ filas de alto. En este caso, la fórmula (7) queda:

$$(11) \quad u_{i,j+1} = \frac{u_{i-1,j} + u_{i+1,j}}{2}.$$

La fórmula (11) es estable para $r = 0.5$ y puede ser usada con garantías de éxito para generar aproximaciones razonablemente precisas a $u(x, t)$. En la Tabla 10.3 se

Tabla 10.3 Resultados obtenidos con el método de las diferencias progresivas para $r = 0.5$.

	$x_1 = 0.00$	$x_2 = 0.20$	$x_3 = 0.40$	$x_4 = 0.60$	$x_5 = 0.80$	$x_6 = 1.00$
$t_1 = 0.00$	0.000000	0.640000	0.960000	0.960000	0.640000	0.000000
$t_2 = 0.02$	0.000000	0.480000	0.800000	0.800000	0.480000	0.000000
$t_3 = 0.04$	0.000000	0.400000	0.640000	0.640000	0.400000	0.000000
$t_4 = 0.06$	0.000000	0.320000	0.520000	0.520000	0.320000	0.000000
$t_5 = 0.08$	0.000000	0.260000	0.420000	0.420000	0.260000	0.000000
$t_6 = 0.10$	0.000000	0.210000	0.340000	0.340000	0.210000	0.000000
$t_7 = 0.12$	0.000000	0.170000	0.275000	0.275000	0.170000	0.000000
$t_8 = 0.14$	0.000000	0.137500	0.222500	0.222500	0.137500	0.000000
$t_9 = 0.16$	0.000000	0.111250	0.180000	0.180000	0.111250	0.000000
$t_{10} = 0.18$	0.000000	0.090000	0.145625	0.145625	0.090000	0.000000
$t_{11} = 0.20$	0.000000	0.072812	0.117813	0.117813	0.072812	0.000000

**Figura 10.10** Resultados obtenidos con el método de las diferencias progresivas para $r = 0.5$.

recogen las aproximaciones en las filas sucesivas de la malla y en la Figura 10.10 se da una representación tridimensional de estos resultados.

La segunda vez, tomamos como tamaños de paso $\Delta x = h = 0.2$ y $\Delta t = k = \frac{1}{30} \approx 0.033333$, de manera que $r = 0.833333$. En este caso, la fórmula (7) queda:

$$(12) \quad u_{i,j+1} = -0.666665u_{i,j} + 0.833333(u_{i-1,j} + u_{i+1,j}).$$

Tabla 10.4 Resultados obtenidos con el método de las diferencias progresivas para $r = 0.833333$.

	$x_1 = 0.00$	$x_2 = 0.20$	$x_3 = 0.40$	$x_4 = 0.60$	$x_5 = 0.80$	$x_6 = 1.00$
$t_1 = 0.000000$	0.000000	0.640000	0.960000	0.960000	0.640000	0.000000
$t_2 = 0.033333$	0.000000	0.373333	0.693333	0.693333	0.373333	0.000000
$t_3 = 0.066667$	0.000000	0.328889	0.426667	0.426667	0.328889	0.000000
$t_4 = 0.100000$	0.000000	0.136296	0.345185	0.345185	0.136296	0.000000
$t_5 = 0.133333$	0.000000	0.196790	0.171111	0.171111	0.196790	0.000000
$t_6 = 0.166667$	0.000000	0.011399	0.192510	0.192510	0.011399	0.000000
$t_7 = 0.200000$	0.000000	0.152826	0.041584	0.041584	0.152826	0.000000
$t_8 = 0.233333$	0.000000	-0.067230	0.134286	0.134286	-0.067230	0.000000
$t_9 = 0.266667$	0.000000	0.156725	-0.033644	-0.033644	0.156725	0.000000
$t_{10} = 0.300000$	0.000000	-0.132520	0.124997	0.124997	-0.132520	0.000000
$t_{11} = 0.333333$	0.000000	0.192511	-0.089601	-0.089601	0.192511	0.000000

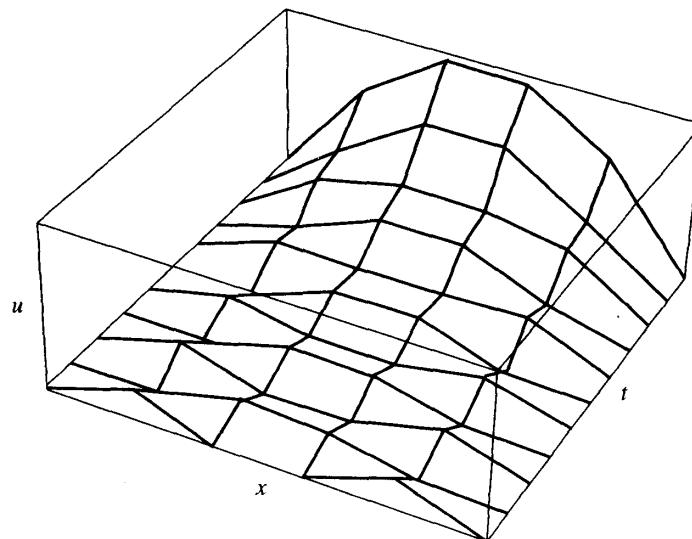


Figura 10.11 Resultados obtenidos con el método de las diferencias progresivas para $r = 0.833333$.

La fórmula (12) es inestable porque $r > \frac{1}{2}$ y los errores introducidos en una fila se amplificarán en las filas posteriores. Los valores numéricos que se obtienen, que son aproximaciones a $u(x, t)$ bastante poco precisas para $0 \leq t \leq 0.33333$, se recogen en la Tabla 10.4. En la Figura 10.11 se da una representación tridimensional de estos resultados.

La precisión de la ecuación en diferencias (7) es de orden $\mathcal{O}(k) + \mathcal{O}(h^2)$ y, como el término $\mathcal{O}(k)$ tiende a cero linealmente, no es sorprendente que k deba tomarse muy pequeño para obtener buenas aproximaciones. Aún así, la necesidad de que el método sea estable nos plantea consideraciones adicionales. Supongamos que las aproximaciones obtenidas en la malla no son suficientemente precisas y que debemos reducir los tamaños de paso $\Delta x = h_0$ y $\Delta t = k_0$. Si tomamos como nuevo tamaño de paso para la coordenada x simplemente $\Delta x = h_1 = h_0/2$ y queremos mantener el mismo valor del cociente r , entonces k_1 debe cumplir

$$k_1 = \frac{r(h_1)^2}{c^2} = \frac{r(h_0)^2}{4c^2} = \frac{k_0}{4}.$$

En consecuencia, hay que doblar el número de nodos de la malla en el eje de la variable x y cuadruplicarlo en el eje de la variable t , con lo cual el esfuerzo computacional será ocho veces mayor. Este esfuerzo extra es, normalmente, prohibitivo y nos invita a buscar métodos más eficaces que no estén sujetos a restricciones de estabilidad tan exigentes. El método que vamos a proponer es un método implícito, no explícito, pero este incremento en el nivel de complejidad tendrá como contrapartida la garantía de la estabilidad sin condiciones adicionales.

El método de Crank-Nicholson

Este esquema implícito, inventado por John Crank y Phyllis Nicolson (véase la Referencia [29]), se basa en la construcción de una aproximación numérica al valor de la solución de la ecuación del calor (1) en $(x, t + k/2)$ que es un punto situado entre dos filas de la malla. Concretamente, para $u_t(x, t + k/2)$ usamos la aproximación que se obtiene a partir de la fórmula de diferencias centradas

$$(13) \quad u_t\left(x, t + \frac{k}{2}\right) = \frac{u(x, t + k) - u(x, t)}{k} + \mathcal{O}(k^2)$$

y para $u_{xx}(x, t + k/2)$ usamos como aproximación el valor medio de las aproximaciones a $u_{xx}(x, t)$ y $u_{xx}(x, t + k)$; este valor medio tiene una precisión del orden de $\mathcal{O}(h^2)$:

$$(14) \quad u_{xx}\left(x, t + \frac{k}{2}\right) = \frac{1}{2h^2}(u(x - h, t + k) - 2u(x, t + k) + u(x + h, t + k) \\ + u(x - h, t) - 2u(x, t) + u(x + h, t)) + \mathcal{O}(h^2).$$

Trabajando de manera similar a como lo hicimos para obtener el esquema de diferencias progresivas, sustituimos las expresiones (13) y (14) en la ecuación

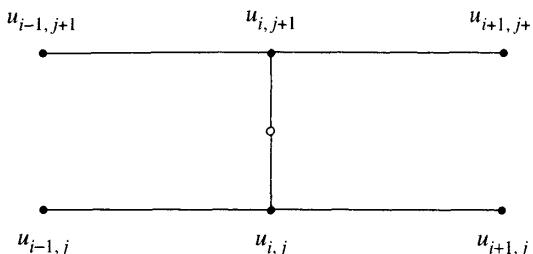


Figura 10.12 El esquema de Crank-Nicholson.

del calor (1) y despreciamos los términos del error $O(h^2)$ y $O(k^2)$. Entonces, manteniendo la notación $u_{i,j} \approx u(x_i, t_j)$, obtenemos la ecuación en diferencias

$$(15) \quad \frac{u_{i,j+1} - u_{i,j}}{k} = c^2 \frac{u_{i-1,j+1} - 2u_{i,j+1} + u_{i+1,j+1} + u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{2h^2}.$$

Volvemos a tomar $r = c^2 k / h^2$ en (15) y despejamos los tres valores “aún por calcular” $u_{i-1,j+1}$, $u_{i,j+1}$ y $u_{i+1,j+1}$ escribiéndolos en el miembro de la izquierda de la ecuación. Esta reordenación de los términos de la ecuación (15) produce la siguiente ecuación en diferencias implícita

$$(16) \quad -ru_{i-1,j+1} + (2 + 2r)u_{i,j+1} - ru_{i+1,j+1} \\ = (2 - 2r)u_{i,j} + r(u_{i-1,j} + u_{i+1,j}).$$

para $i = 2, 3, \dots, n - 1$. Los términos del miembro derecho de la ecuación (16) son todos conocidos, así que estas ecuaciones forman un sistema lineal tridiagonal $\mathbf{AX} = \mathbf{B}$. En la Figura 10.12 se muestran los seis puntos que se usan en la fórmula (16) de Crank-Nicholson así como el punto intermedio en el que se basan las aproximaciones numéricas.

Cuando se trabaja con la fórmula (16) se suele tomar como cociente $r = 1$; en este caso, el tamaño de paso en el eje de la variable t es $\Delta t = k = h^2/c^2$ y las ecuaciones de (16) se pueden escribir de manera más simple como

$$(17) \quad -u_{i-1,j+1} + 4u_{i,j+1} - u_{i+1,j+1} = u_{i-1,j} + u_{i+1,j},$$

para $i = 2, 3, \dots, n - 1$. En la primera y última de estas ecuaciones hay que usar las condiciones de contorno, es decir, $u_{1,j} = u_{1,j+1} = c_1$ y $u_{n,j} = u_{n,j+1} = c_2$, respectivamente. Las ecuaciones de (17) se escriben de forma especialmente

atractiva en su forma matricial tridiagonal $\mathbf{AX} = \mathbf{B}$:

$$\begin{bmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & & & \\ & -1 & 4 & -1 & \\ & & \ddots & & \\ \mathbf{O} & & -1 & 4 & -1 \\ & & & -1 & 4 \end{bmatrix} \begin{bmatrix} u_{2,j+1} \\ u_{3,j+1} \\ \vdots \\ u_{i,j+1} \\ \vdots \\ u_{n-2,j+1} \\ u_{n-1,j+1} \end{bmatrix} = \begin{bmatrix} 2c_1 + u_{3,j} \\ u_{2,j} + u_{4,j} \\ \vdots \\ u_{i-1,j} + u_{i+1,j} \\ \vdots \\ u_{n-3,j} + u_{n-1,j} \\ u_{n-2,j} + 2c_2 \end{bmatrix}.$$

Cuando se utiliza un computador para llevar a cabo el método de Crank-Nicholson, el sistema lineal tridiagonal $\mathbf{AX} = \mathbf{B}$ puede resolverse bien por métodos directos, bien de forma iterativa.

Ejemplo 10.4. Vamos a usar el método de Crank-Nicholson para resolver la ecuación

$$(18) \quad u_t(x, t) = u_{xx}(x, t) \quad \text{para } 0 < x < 1 \text{ y } 0 < t < 0.1,$$

con las condiciones iniciales

$$(19) \quad u(x, 0) = f(x) = \operatorname{sen}(\pi x) + \operatorname{sen}(3\pi x) \quad \text{para } t = 0 \text{ y } 0 \leq x \leq 1,$$

y las condiciones de contorno

$$\begin{aligned} u(0, t) &= g_1(t) \equiv 0 && \text{para } x = 0 \text{ y } 0 \leq t \leq 0.1, \\ u(1, t) &= g_2(t) \equiv 0 && \text{para } x = 1 \text{ y } 0 \leq t \leq 0.1. \end{aligned}$$

Por simplicidad, tomamos como tamaños de paso $\Delta x = h = 0.1$ y $\Delta t = k = 0.01$ de manera que el cociente es $r = 1$. La malla tendrá $n = 11$ columnas de ancho y $m = 11$ filas de alto. En la Tabla 10.5 se muestran los resultados obtenidos con el algoritmo para $0 < x_i < 1$ y $0 \leq t_j \leq 0.1$.

Las aproximaciones obtenidas con el método de Crank-Nicholson son buenas aproximaciones de los valores exactos

$$u(x, t) = \operatorname{sen}(\pi x)e^{-\pi^2 t} + \operatorname{sen}(3\pi x)e^{-9\pi^2 t}$$

que, en la última fila, son

t_{11}	0.115285	0.219204	0.301570	0.354385	0.372569	0.354385	0.301570	0.219204	0.115285
----------	----------	----------	----------	----------	----------	----------	----------	----------	----------

En la Figura 10.13 se da una representación tridimensional de los valores recogidos en la Tabla 10.5.

Tabla 10.5 Los valores $u(x_i, t_j)$ obtenidos con el método de Crank-Nicholson para $t_j = (j - 1)/100$.

	$x_2 = 0.1$	$x_3 = 0.2$	$x_4 = 0.3$	$x_5 = 0.4$	$x_6 = 0.5$	$x_7 = 0.6$	$x_8 = 0.7$	$x_9 = 0.8$	$x_{10} = 0.9$
t_1	1.118034	1.538842	1.118034	0.363271	0.000000	0.363271	1.118034	1.538842	1.118034
t_2	0.616905	0.928778	0.862137	0.617659	0.490465	0.617659	0.862137	0.928778	0.616905
t_3	0.394184	0.647957	0.718601	0.680009	0.648834	0.680009	0.718601	0.647957	0.394184
t_4	0.288660	0.506682	0.625285	0.666493	0.673251	0.666493	0.625285	0.506682	0.288660
t_5	0.233112	0.425766	0.556006	0.625082	0.645788	0.625082	0.556006	0.425766	0.233112
t_6	0.199450	0.372035	0.499571	0.575402	0.600242	0.575402	0.499571	0.372035	0.199450
t_7	0.175881	0.331490	0.451058	0.525306	0.550354	0.525306	0.451058	0.331490	0.175881
t_8	0.157405	0.298131	0.408178	0.477784	0.501545	0.477784	0.408178	0.298131	0.157405
t_9	0.141858	0.269300	0.369759	0.433821	0.455802	0.433821	0.369759	0.269300	0.141858
t_{10}	0.128262	0.243749	0.335117	0.393597	0.413709	0.393597	0.335117	0.243749	0.128262
t_{11}	0.116144	0.220827	0.303787	0.356974	0.375286	0.356974	0.303787	0.220827	0.116144

MATLAB

Programa 10.2 (Método de diferencias progresivas para la ecuación del calor). Construcción de aproximaciones a la solución de $u_t(x, t) = c^2 u_{xx}(x, t)$ en $R = \{(x, t) : 0 \leq x \leq a, 0 \leq t \leq b\}$, con $u(x, 0) = f(x)$ para $0 \leq x \leq a$ y $u(0, t) = c_1$, $u(a, t) = c_2$ para $0 \leq t \leq b$.

```

function U=forwdif(f,c1,c2,a,b,c,n,m)

% Datos
%   - f=u(x,0) almacenada como una cadena de caracteres 'f'
%   - c1=u(0,t) y c2=u(a,t)
%   - a y b son los extremos derechos de [0,a] y [0,b]
%   - c es la constante de la ecuación del calor
%   - n y m son el número de nodos en [0,a] y [0,b]
% Resultado
%   - U es la matriz de aproximaciones;
%   análoga a la de la Tabla 10.4

% Inicialización de los parámetros y de U
h=a/(n-1);
k=b/(m-1);
r=c^2*k/h^2;
s=1-2*r;
U=zeros(n,m);

% Condiciones de contorno
U(1,1:m)=c1;
U(n,1:m)=c2;

% Construcción de la primera fila de U

```

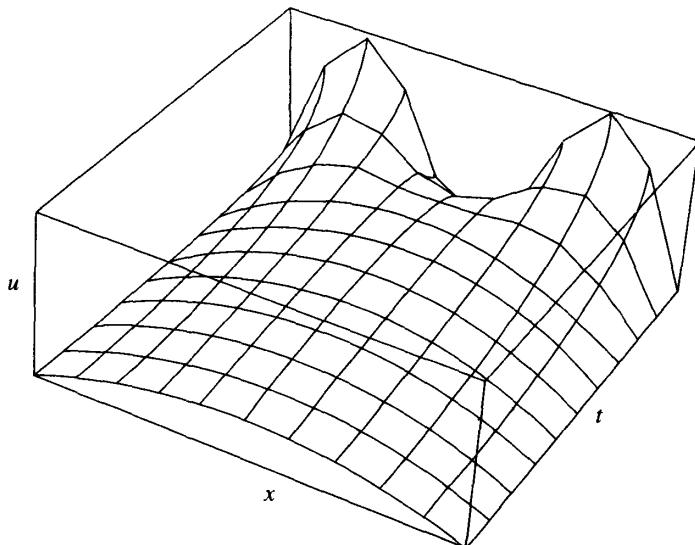


Figura 10.13 $u = u(x_i, t_j)$ obtenidos con el método de Crank-Nicholson.

```

U(2:n-1,1)=feval(f,h:h:(n-2)*h)';
% Construcción de las demás filas de U
for j=2:m
    for i=2:n-1
        U(i,j)=s*U(i,j-1)+r*(U(i-1,j-1)+U(i+1,j-1));
    end
end
U=U';

```

Programa 10.3 (Método de Crank-Nicholson para la ecuación del calor). Construcción de las aproximaciones a la solución de $u_t(x, t) = c^2 u_{xx}(x, t)$ en $R = \{(x, t) : 0 \leq x \leq a, 0 \leq t \leq b\}$, con $u(x, 0) = f(x)$ para $0 \leq x \leq a$ y $u(0, t) = c_1$, $u(a, t) = c_2$ para $0 \leq t \leq b$.

```

function U=crnich(f,c1,c2,a,b,c,n,m)
% Datos
% - f=u(x,0) almacenada como una cadena de caracteres 'f'
% - c1=u(0,t) y c2=u(a,t)
% - a y b son los extremos derechos de [0,a] y [0,b]
% - c es la constante de la ecuación del calor
% - n y m son el número de nodos en [0,a] y [0,b]

```

```
% Resultado
% - U es la matriz de las aproximaciones;
% análoga a la de la Tabla 10.5

% Inicialización de los parámetros y de U
h=a/(n-1);
k=b/(m-1);
r=c^2*k/h^2;
s1=2+2/r;
s2=2/r-2;
U=zeros(n,m);

% Condiciones de contorno
U(1,1:m)=c1;
U(n,1:m)=c2;

% Generación de la primera fila de U
U(2:n-1,1)=feval(f,h:h:(n-2)*h)';

% Construcción de los elementos diagonales y no diagonales de
% A y del vector de términos independientes B y resolución del
% sistema AX=B
Vd(1,1:n)=s1*ones(1,n);
Vd(1)=1;
Vd(n)=1;
Va=-ones(1,n-1);
Va(n-1)=0;
Vc=-ones(1,n-1);
Vc(1)=0;
Vb(1)=c1;
Vb(n)=c2;
for j=2:m
    for i=2:n-1
        Vb(i)=U(i-1,j-1)+U(i+1,j-1)+s2*U(i,j-1);
    end
    X=trisys(Va,Vd,Vc,Vb);
    U(1:n,j)=X';
end
U=U'
```

Ejercicios

1. (a) Verifique, sustituyendo directamente en la ecuación, que, para cada número natural $n = 1, 2, \dots$, la función $u(x, t) = \sin(n\pi x)e^{-4n^2\pi^2t}$ es una solución de la ecuación del calor $u_t(x, t) = 4u_{xx}(x, t)$.

- (b) Verifique, sustituyendo directamente en la ecuación, que, para cada número natural $n = 1, 2, \dots$, la función $u(x, t) = \sin(n\pi x)e^{-(cn\pi)^2 t}$ es una solución de la ecuación del calor $u_t(x, t) = c^2 u_{xx}(x, t)$.

2. ¿Qué dificultades podrían aparecer si se usa $\Delta t = k = h^2/c^2$ en la fórmula (7)?

En los Ejercicios 3 y 4, use el método de las diferencias progresivas para calcular las tres primeras filas de la malla que se construye para la ecuación del calor que se da. Realice las operaciones a mano (o con una calculadora).

3. $u_t(x, t) = u_{xx}(x, t)$ para $0 < x < 1$ y $0 \leq t \leq 0.1$, con la condición inicial $u(x, 0) = f(x) = \sin(\pi x)$ para $t = 0$ y $0 \leq x \leq 1$ y las condiciones de contorno

$$\begin{aligned} u(0, t) &= c_1 = 0 && \text{para } x = 0 \text{ y } 0 \leq t \leq 0.1, \\ u(1, t) &= c_2 = 0 && \text{para } x = 1 \text{ y } 0 \leq t \leq 0.1. \end{aligned}$$

Tome $h = 0.2$, $k = 0.02$ y $r = 0.5$.

4. $u_t(x, t) = u_{xx}(x, t)$ para $0 < x < 1$ y $0 \leq t \leq 0.1$, con la condición inicial $u(x, 0) = f(x) = 1 - |2x - 1|$ para $t = 0$ y $0 \leq x \leq 1$ y las de contorno

$$\begin{aligned} u(0, t) &= c_1 = 0 && \text{para } x = 0 \text{ y } 0 \leq t \leq 0.1, \\ u(1, t) &= c_2 = 0 && \text{para } x = 1 \text{ y } 0 \leq t \leq 0.1. \end{aligned}$$

5. Supongamos que $\Delta t = k = h^2/(2c^2)$.

- (a) Use esta igualdad en la fórmula (16) y simplifique la ecuación resultante.
 (b) Exprese las ecuaciones del apartado (a) matricialmente $\mathbf{AX} = \mathbf{B}$.
 (c) ¿Es la matriz del apartado (b) de diagonal estrictamente dominante?

6. Pruebe que $u(x, t) = \sum_{j=1}^N a_j e^{-(j\pi)^2 t} \sin(j\pi x)$ es una solución de la ecuación $u_t(x, t) = u_{xx}(x, t)$ para $0 \leq x \leq 1$ y $0 < t$, que cumple las condiciones de contorno $u(0, t) = 0$, $u(1, t) = 0$ con valores iniciales $u(x, 0) = \sum_{j=1}^N a_j \sin(j\pi x)$.

7. Considere la solución exacta $u(x, t) = \sin(\pi x)e^{-\pi^2 t} + \sin(3\pi x)e^{-(3\pi)^2 t}$ tratada en el Ejemplo 10.4.

- (a) Fijando x , calcule el valor de $\lim_{t \rightarrow \infty} u(x, t)$.
 (b) ¿Qué significa físicamente este límite?

8. Deseamos resolver la ecuación parabólica $u_t(x, t) - u_{xx}(x, t) = h(x)$.

- (a) Desarrolle la ecuación en diferencias progresivas explícita para esta situación.
 (b) Desarrolle la ecuación en diferencias implícita para esta situación.

9. Supongamos que usamos la ecuación en diferencias (11) y que $f(x) \geq 0$, $g_1(t) = 0$ y $g_2(t) = 0$.

- (a) Pruebe que el valor máximo de $u(x_i, t_{j+1})$ en la fila $(j+1)$ -ésima es menor o igual que el valor máximo de $u(x_i, t_j)$ en la fila j -ésima.
 (b) Formule una conjectura sobre el valor máximo de $u(x_i, t_n)$ en la fila n -ésima cuando n tiende a infinito.

Algoritmos y programas

En los Problemas 1 y 2, use el Programa 10.3 para resolver la ecuación del calor $u_t(x, t) = c^2 u_{xx}(x, t)$ para $0 < x < 1$ y $0 < t < 0.1$, con la condición inicial $u(x, 0) = f(x)$ para $t = 0$ y $0 \leq x \leq 1$, y las condiciones de contorno

$$\begin{aligned} u(0, t) &= c_1 = 0 && \text{para } x = 0 \text{ y } 0 \leq t \leq 0.1, \\ u(1, t) &= c_2 = 0 && \text{para } x = 1 \text{ y } 0 \leq t \leq 0.1, \end{aligned}$$

para los valores que se dan. Use las instrucciones `surf` y `contour` del paquete de programas MATLAB para dibujar sus soluciones aproximadas.

1. Use $f(x) = \sin(\pi x) + \sin(2\pi x)$, $h = 0.1$, $k = 0.01$ y $r = 1$.
2. Use $f(x) = 3 - |3x - 1| - |3x - 2|$, $h = 0.1$, $k = 0.01$ y $r = 1$.
3. (a) Modifique los Programas 10.2 y 10.3 de manera que permitan resolver problemas con condiciones de contorno generales $u(0, t) = g_1(t) \neq 0$ y $u(a, t) = g_2(t) \neq 0$.
(b) Use su modificación del Programa 10.3 para resolver las ecuaciones del calor de los Problemas 1 y 2 con las condiciones de contorno

$$\begin{aligned} u(0, t) &= g_1(t) = t^2 && \text{para } x = 0 \text{ y } 0 \leq t < 0.1, \\ u(1, t) &= g_2(t) = e^t && \text{para } x = 1 \text{ y } 0 \leq t \leq 0.1. \end{aligned}$$

- (c) Use las instrucciones `surf` y `contour` del paquete de programas MATLAB para dibujar sus soluciones aproximadas.
4. Construya programas en MATLAB que permitan llevar a cabo, respectivamente, el método explícito de diferencias progresivas y el método implícito de diferencias de los apartados (a) y (b) del Ejercicio 8.
5. Use sus programas del Problema 4 para resolver la ecuación del calor $u_t(x, t) - u_{xx}(x, t) = \sin(x)$ para $0 < x < 1$ y $0 < t < 0.20$, con la condición inicial $u(x, 0) = f(x) = \sin(\pi x) + \sin(3\pi x)$ y las condiciones de contorno

$$\begin{aligned} u(0, t) &= c_2 = 0 && \text{para } x = 0 \text{ y } 0 \leq t \leq 0.20, \\ u(1, t) &= c_2 = 0 && \text{para } x = 1 \text{ y } 0 \leq t \leq 0.20. \end{aligned}$$

Tome $h = 0.2$, $k = 0.02$ y $r = 0.5$.

10.3 Ecuaciones elípticas

Como ejemplos de ecuaciones en derivadas parciales elípticas, consideraremos las ecuaciones de Laplace, Poisson y Helmholtz. Recordemos que la laplaciana de una función $u(x, y)$ es

$$(1) \quad \nabla^2 u = u_{xx} + u_{yy}.$$

Con esta notación, las ecuaciones de Laplace, Poisson y Helmholtz pueden expresarse de la siguiente manera:

- (2) $\nabla^2 u = 0$ ecuación de Laplace,
 (3) $\nabla^2 u = g(x, y)$ ecuación de Poisson,
 (4) $\nabla^2 u + f(x, y)u = g(x, y)$ ecuación de Helmholtz.

Si se conocen los valores que debe tomar la función u (problema de Dirichlet) o su derivada normal $\partial u(x, y)/\partial N = 0$ (problema de Neumann) en la frontera de una región rectangular R del plano, entonces cada uno de estos problemas puede resolverse mediante la técnica numérica conocida como el método de las diferencias finitas.

La ecuación en diferencias para la laplaciana

El primer paso consiste en obtener una versión discretizada del operador de Laplace que nos permita usarlo numéricamente. La fórmula para $f''(x)$ es

$$(5) \quad f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} + O(h^2),$$

así que, al aplicar esta fórmula a la función $u(x, y)$ para aproximar $u_{xx}(x, y)$ y $u_{yy}(x, y)$ y sumar los resultados, obtenemos

$$(6) \quad \nabla^2 u = \frac{u(x+h, y) + u(x-h, y) + u(x, y+h) + u(x, y-h) - 4u(x, y)}{h^2} + O(h^2).$$

Ahora dividimos el rectángulo $R = \{(x, y) : 0 \leq x \leq a, 0 \leq y \leq b\}$ en $n-1 \times m-1$ cuadrados de lado h (o sea, $a = nh$ y $b = mh$), como se muestra en la Figura 10.14.

Para resolver la ecuación de Laplace, imponemos la aproximación

$$(7) \quad \frac{u(x+h, y) + u(x-h, y) + u(x, y+h) + u(x, y-h) - 4u(x, y)}{h^2} = 0,$$

que tiene una precisión de orden $O(h^2)$ en los puntos interiores de la malla $(x, y) = (x_i, y_j)$ para $i = 2, \dots, n-1$ y $j = 2, \dots, m-1$. Como los puntos de la malla están espaciados uniformemente: $x_{i+1} = x_i + h$, $x_{i-1} = x_i - h$, $y_{j+1} = y_j + h$ e $y_{j-1} = y_j - h$; denotando por $u_{i,j}$ la aproximación al valor $u(x_i, y_j)$, la ecuación (7) queda

$$(8) \quad \nabla^2 u_{i,j} \approx \frac{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}}{h^2} = 0;$$

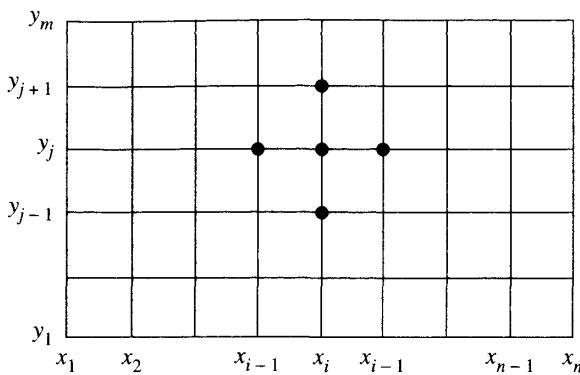


Figura 10.14 La malla usada en la ecuación en diferencias de Laplace.

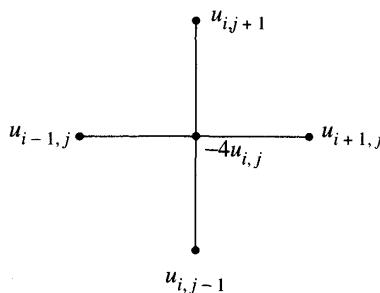


Figura 10.15 Esquema para la ecuación de Laplace.

expresión que se conoce como la **fórmula de diferencias con cinco puntos** para la laplaciana. Esta fórmula relaciona el valor de la función $u_{i,j}$ con sus cuatro valores adyacentes $u_{i+1,j}$, $u_{i-1,j}$, $u_{i,j+1}$ y $u_{i,j-1}$, como se muestra en la Figura 10.15. Eliminando de la expresión (8) el denominador h^2 obtenemos la fórmula de aproximación para la ecuación de Laplace

$$(9) \quad u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0.$$

Construcción del sistema lineal

Supongamos que tenemos un problema de Dirichlet, es decir, que conocemos los valores de la función $u(x, y)$ en la frontera de la región R :

$$\begin{aligned} u(x_1, y_j) &= u_{1,j} && \text{para } 2 \leq j \leq m-1 && (\text{a la izquierda}), \\ u(x_i, y_1) &= u_{i,1} && \text{para } 2 \leq i \leq n-1 && (\text{abajo}), \\ u(x_n, y_j) &= u_{n,j} && \text{para } 2 \leq j \leq m-1 && (\text{a la derecha}), \\ u(x_i, y_m) &= u_{i,m} && \text{para } 2 \leq i \leq n-1 && (\text{arriba}). \end{aligned}$$

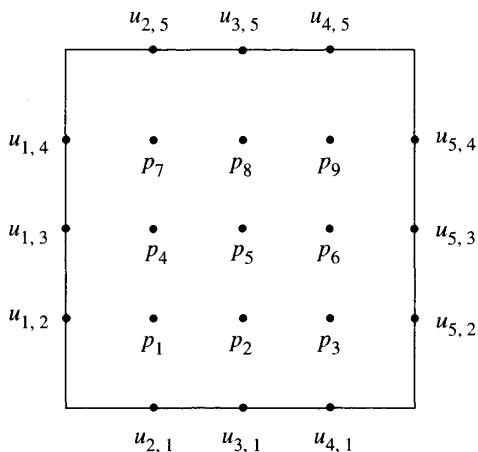


Figura 10.16 Una malla de orden 5×5 para un problema de contorno.

Al aplicar la fórmula (9) en cada uno de los puntos de la malla que son interiores a R , obtenemos un sistema de $(n - 2)$ ecuaciones lineales con $(n - 2)$ incógnitas, cuya solución nos proporciona las aproximaciones a $u(x, y)$ en los puntos interiores de R . Por ejemplo, supongamos que la región es un cuadrado, que $n = m = 5$ y que los valores desconocidos $u(x_i, y_j)$ en los nueve puntos interiores de la malla se etiquetan p_1, p_2, \dots, p_9 como se indica en la Figura 10.16.

Aplicando la fórmula (9) de aproximación a la ecuación de Laplace en cada uno de los puntos interiores de la malla, obtenemos el sistema de nueve ecuaciones lineales $\mathbf{AP} = \mathbf{B}$:

$$\begin{aligned}
 -4p_1 + p_2 &+ p_4 &= -u_{2,1} - u_{1,2} \\
 p_1 - 4p_2 + p_3 &+ p_5 &= -u_{3,1} \\
 p_2 - 4p_3 &+ p_6 &= -u_{4,1} - u_{5,2} \\
 p_1 &- 4p_4 + p_5 &+ p_7 &= -u_{1,3} \\
 p_2 &+ p_4 - 4p_5 + p_6 &+ p_8 &= 0 \\
 p_3 &+ p_5 - 4p_6 &+ p_9 &= -u_{5,3} \\
 p_4 &&- 4p_7 + p_8 &= -u_{2,5} - u_{1,4} \\
 p_5 &&+ p_7 - 4p_8 + p_9 &= -u_{3,5} \\
 p_6 &&+ p_8 - 4p_9 &= -u_{4,5} - u_{5,4}.
 \end{aligned}$$

Ejemplo 10.5. Vamos a determinar la solución aproximada de la ecuación de Laplace $\nabla^2 u = 0$ en el rectángulo $R = \{(x, y) : 0 \leq x \leq 4, 0 \leq y \leq 4\}$, donde $u(x, y)$ denota la temperatura en un punto (x, y) , los valores en la frontera son

$$u(x, 0) = 20 \quad \text{y} \quad u(x, 4) = 180 \quad \text{para} \quad 0 < x < 4,$$

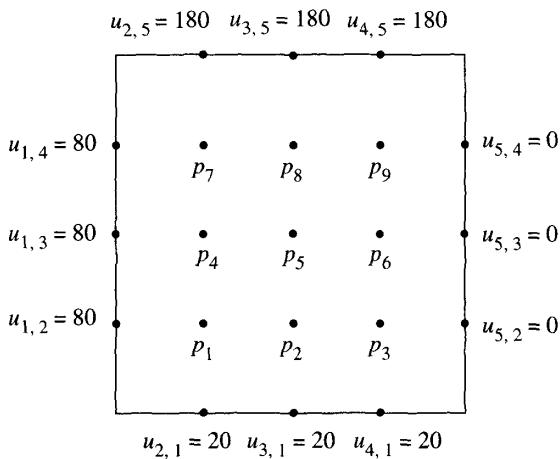


Figura 10.17 La malla de orden \$5 \times 5\$ en el Ejemplo 10.5.

y

$$u(0, y) = 80 \quad \text{y} \quad u(4, y) = 0 \quad \text{para} \quad 0 < y < 4$$

y la malla que se usa es la que se muestra en la Figura 10.17.

Al aplicar la fórmula (9) en este caso, el sistema \$\mathbf{AP} = \mathbf{B}\$ que se obtiene es

$$\begin{aligned} -4p_1 + p_2 + p_4 &= -100 \\ p_1 - 4p_2 + p_3 + p_5 &= -20 \\ p_2 - 4p_3 + p_4 + p_6 &= -20 \\ p_1 - 4p_4 + p_5 + p_7 &= -80 \\ p_2 + p_4 - 4p_5 + p_6 + p_8 &= 0 \\ p_3 + p_5 - 4p_6 + p_9 &= 0 \\ p_4 - 4p_7 + p_8 &= -260 \\ p_5 + p_7 - 4p_8 + p_9 &= -180 \\ p_6 + p_8 - 4p_9 &= -180. \end{aligned}$$

El vector solución \$\mathbf{P}\$ puede obtenerse mediante el método de eliminación de Gauss (también pueden diseñarse esquemas más eficientes, como la extensión del algoritmo tridiagonal a sistemas pentadiagonales). Las temperaturas en los puntos interiores de la malla, expresadas en forma vectorial, son

$$\begin{aligned} \mathbf{P} &= [p_1 \ p_2 \ p_3 \ p_4 \ p_5 \ p_6 \ p_7 \ p_8 \ p_9]' \\ &= [55.7143 \ 43.2143 \ 27.1429 \ 79.6429 \ 70.0000 \\ &\quad 45.3571 \ 112.857 \ 111.786 \ 84.2857]'. \end{aligned}$$

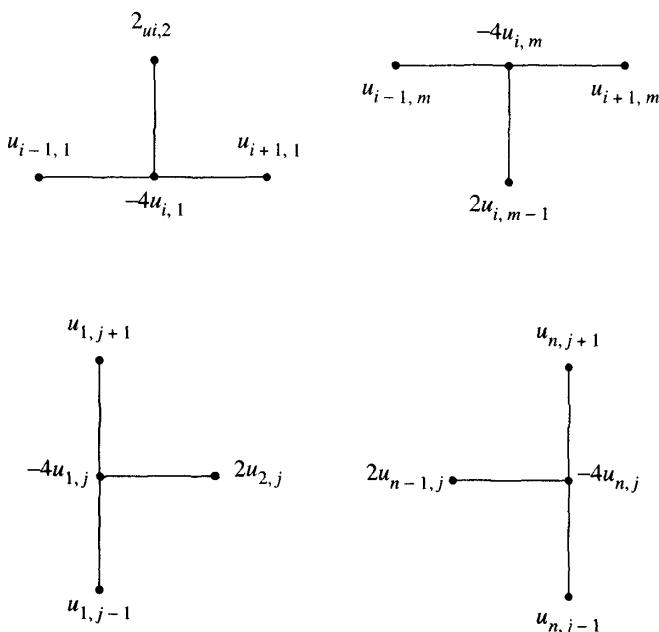


Figura 10.18 Los esquemas para la condición de contorno de Neumann.

Condiciones de contorno sobre la derivada

Cuando se especifican los valores de la derivada direccional de $u(x, y)$ en la dirección perpendicular al contorno de R , se dice que tenemos un problema con condiciones de contorno de Neumann. Como ejemplo vamos a resolver un caso en el que la derivada normal es nula

$$(10) \quad \frac{\partial}{\partial N} u(x, y) = 0.$$

En el contexto de los problemas de distribución de temperaturas, esto significa que el contorno está aislado y que no hay flujo de calor a través de él.

Supongamos que fijamos $x = x_n$, de manera que consideraremos el lado derecho $x = a$ del rectángulo $R = \{(x, y) : 0 \leq x \leq a, 0 \leq y \leq b\}$. La condición de contorno sobre la derivada normal en este lado es, entonces,

$$(11) \quad \frac{\partial}{\partial x} u(x_n, y_j) = u_x(x_n, y_j) = 0.$$

La ecuación de diferencias de Laplace en el punto (x_n, y_j) es

$$(12) \quad u_{n+1,j} + u_{n-1,j} + u_{n,j+1} + u_{n,j-1} - 4u_{n,j} = 0,$$

en la que el valor $u_{n+1,j}$ es desconocido porque el punto correspondiente está fuera de la región R . Sin embargo, podemos usar la fórmula de derivación numérica

$$(13) \quad \frac{u_{n+1,j} - u_{n-1,j}}{2h} \approx u_x(x_n, y_j) = 0$$

y obtener la aproximación $u_{n+1,j} \approx u_{n-1,j}$, cuyo orden de precisión es $O(h^2)$. Al usar esta aproximación en la expresión (12), el resultado es

$$2u_{n-1,j} + u_{n,j+1} + u_{n,j-1} - 4u_{n,j} = 0.$$

En esta fórmula se relaciona el valor de la función $u_{n,j}$ con sus tres valores adyacentes $u_{n-1,j}$, $u_{n,j+1}$ y $u_{n,j-1}$.

Los esquemas computacionales de Neumann para los puntos de los demás lados se deducen de forma parecida (véase la Figura 10.18), de manera que los cuatro casos son:

$$(14) \quad 2u_{i,2} + u_{i-1,1} + u_{i+1,1} - 4u_{i,1} = 0 \quad (\text{lado inferior}),$$

$$(15) \quad 2u_{i,m-1} + u_{i-1,m} + u_{i+1,m} - 4u_{i,m} = 0 \quad (\text{lado superior}),$$

$$(16) \quad 2u_{2,j} + u_{1,j-1} + u_{1,j+1} - 4u_{1,j} = 0 \quad (\text{lado izquierdo}),$$

$$(17) \quad 2u_{n-1,j} + u_{n,j-1} + u_{n,j+1} - 4u_{n,j} = 0 \quad (\text{lado derecho}).$$

Podemos abordar también problemas mixtos, en los que se usa la condición sobre la derivada normal $\partial u(x, y)/\partial N = 0$ en una parte de la frontera de R y valores de contorno $u(x, y)$ especificados en el resto de la frontera. Las ecuaciones para determinar las aproximaciones a $u(x_i, y_j)$ en la parte del contorno en la que se aplica la condición sobre la derivada normal son las dadas por el esquema de Neumann (14)–(17) apropiado. Para las aproximaciones a $u(x_i, y_j)$ en los puntos interiores a R seguimos usando la fórmula (9) de aproximación a la ecuación de Laplace.

Ejemplo 10.6. Vamos a calcular una solución aproximada de la ecuación de Laplace $\nabla^2 u = 0$ en el rectángulo $R = \{(x, y) : 0 \leq x \leq 4, 0 \leq y \leq 4\}$, donde $u(x, y)$ denota la temperatura en el punto (x, y) y las condiciones de contorno son las que se muestran en la Figura 10.19:

$$\begin{aligned} u(x, 4) &= 180 && \text{para } 0 < x < 4, \\ u_y(x, 0) &= 0 && \text{para } 0 < x < 4, \\ u(0, y) &= 80 && \text{para } 0 \leq y < 4, \\ u(4, y) &= 0 && \text{para } 0 \leq y < 4. \end{aligned}$$

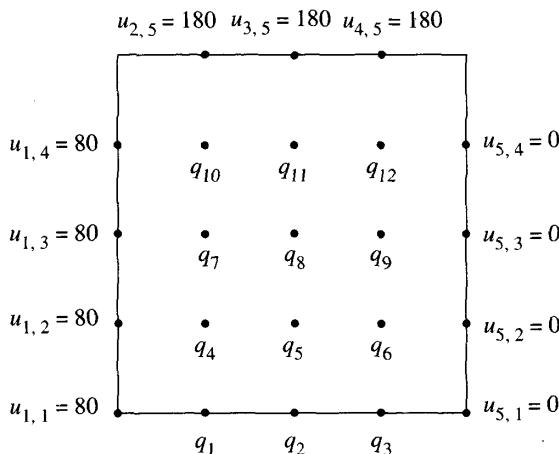


Figura 10.19 La malla de orden 5×5 en el Ejemplo 10.6.

En los puntos q_1 , q_2 y q_3 del contorno aplicamos la fórmula de Neumann (14) y en los puntos q_4, q_5, \dots, q_{12} aplicamos el esquema de Laplace (9). Lo que obtenemos es un sistema lineal $\mathbf{AQ} = \mathbf{B}$ de 12 ecuaciones con 12 incógnitas:

$$\begin{aligned}
 -4q_1 + q_2 &+ 2q_4 &= -80 \\
 q_1 - 4q_2 + q_3 &+ 2q_5 &= 0 \\
 q_2 - 4q_3 &+ 2q_6 &= 0 \\
 q_1 &- 4q_4 + q_5 + q_7 &= -80 \\
 q_2 &+ q_4 - 4q_5 + q_6 + q_8 &= 0 \\
 q_3 &+ q_5 - 4q_6 + q_9 &= 0 \\
 q_4 &- 4q_7 + q_8 + q_{10} &= -80 \\
 q_5 &+ q_7 - 4q_8 + q_9 + q_{11} &= 0 \\
 q_6 &+ q_8 - 4q_9 + q_{12} &= 0 \\
 q_7 &- 4q_{10} + q_{11} &= -260 \\
 q_8 &+ q_{10} - 4q_{11} + q_{12} &= -180 \\
 q_9 &+ q_{11} - 4q_{12} &= -180.
 \end{aligned}$$

El vector solución \mathbf{Q} puede obtenerse mediante el método de eliminación de Gauss (también pueden diseñarse esquemas más eficientes, como la extensión del algoritmo tridiagonal a sistemas pentadiagonales). Las temperaturas en los puntos interiores de la malla y en los puntos del borde inferior, expresadas en forma vectorial, son

$$\begin{aligned}
 \mathbf{Q} &= [q_1 \ q_2 \ q_3 \ q_4 \ q_5 \ q_6 \ q_7 \ q_8 \ q_9 \ q_{10} \ q_{11} \ q_{12}]' \\
 &= [71.8218 \ 56.8543 \ 32.2342 \ 75.2165 \ 61.6806 \ 36.0412 \\
 &\quad 87.3636 \ 78.6103 \ 50.2502 \ 115.628 \ 115.147 \ 86.3492]'.
 \end{aligned}$$

Métodos iterativos

Acabamos de ver cómo podemos resolver la ecuación en diferencias de Laplace construyendo un cierto sistema de ecuaciones lineales y resolviéndolo. El inconveniente que presenta este método es el almacenamiento: Puesto que para obtener resultados mejores hay que trabajar con una malla más fina, es posible que el número de ecuaciones sea muy elevado. Por ejemplo, el cálculo numérico de la solución de un problema de Dirichlet requiere la resolución de un sistema de $(n - 2)(m - 2)$ ecuaciones; si dividimos R en un número modesto de cuadrados, digamos 10 por 10, entonces tenemos un sistema de 91 ecuaciones con 91 incógnitas. En consecuencia, parece sensato trabajar con técnicas que reduzcan la cantidad de datos que se deben almacenar; así, un método iterativo sólo requeriría que se almacenaran las 100 aproximaciones numéricas $\{u_{i,j}\}$ correspondientes a los puntos de la malla.

Empecemos con la ecuación en diferencias de Laplace

$$(18) \quad u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0$$

y supongamos que conocemos los valores de $u(x, y)$ en el contorno:

$$(19) \quad \begin{aligned} u(x_1, y_j) &= u_{1,j} && \text{para } 2 \leq j \leq m - 1 && \text{(a la izquierda),} \\ u(x_i, y_1) &= u_{i,1} && \text{para } 2 \leq i \leq n - 1 && \text{(abajo),} \\ u(x_n, y_j) &= u_{n,j} && \text{para } 2 \leq j \leq m - 1 && \text{(a la derecha),} \\ u(x_i, y_m) &= u_{i,m} && \text{para } 2 \leq i \leq n - 1 && \text{(arriba).} \end{aligned}$$

Ahora escribimos la ecuación (18) de forma adecuada para iterar:

$$(20) \quad u_{i,j} = u_{i,j} + r_{i,j},$$

siendo

$$(21) \quad r_{i,j} = \frac{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}}{4},$$

para $2 \leq i \leq n - 1$ y $2 \leq j \leq m - 1$.

Es necesario disponer de valores iniciales en los puntos interiores de la malla; para ello puede valer la constante K , definida como la media de los $2n + 2m - 4$ valores en el contorno dados por (19). Cada paso de la iteración consiste en hacer un barrido de todos los puntos interiores de la malla con la fórmula recursiva (20) hasta que el término residual $r_{i,j}$ que aparece en el miembro derecho de (20) se “reduzca a cero” (o sea, hasta que se tenga $|r_{i,j}| < \varepsilon$ para cada $2 \leq i \leq n - 1$ y $2 \leq j \leq m - 1$, siendo ε una tolerancia prefijada). Podemos aumentar la velocidad de convergencia a cero de los términos residuales $\{r_{i,j}\}$ usando el método conocido como método de sobrerelajación sucesiva. Este

método es el que resulta de aplicar la fórmula recursiva

$$(22) \quad u_{i,j} = u_{i,j} + \omega \left(\frac{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j}}{4} \right)$$

$$= u_{i,j} + \omega r_{i,j},$$

en la que el parámetro ω verifica $1 \leq \omega < 2$. En el método de sobrerrelajación sucesiva, cada paso de la iteración consiste en hacer un barrido de la malla con la fórmula recursiva (22) hasta que se tenga $|r_{i,j}| < \varepsilon$. Para elegir el valor óptimo del parámetro ω hay que estudiar los autovalores de la matriz que caracteriza el método iterativo que estamos usando para resolver un sistema lineal; en nuestro caso, dicho valor óptimo viene dado por la fórmula

$$(23) \quad \omega = \frac{4}{2 + \sqrt{4 - \left(\cos\left(\frac{\pi}{n-1}\right) + \cos\left(\frac{\pi}{m-1}\right) \right)^2}}.$$

Si lo que se especifica en algún trozo de la frontera es una condición de Neumann, entonces tenemos que escribir las expresiones (14) a (17) de forma adecuada para iterar; los cuatro casos, incluyendo ya el parámetro de sobrerrelajación ω son:

$$(24) \quad u_{i,1} = u_{i,1} + \omega \left(\frac{2u_{i,2} + u_{i-1,1} + u_{i+1,1} - 4u_{i,1}}{4} \right) \quad (\text{lado inferior}),$$

$$(25) \quad u_{i,m} = u_{i,m} + \omega \left(\frac{2u_{i,m-1} + u_{i-1,m} + u_{i+1,m} - 4u_{i,m}}{4} \right) \quad (\text{lado superior}),$$

$$(26) \quad u_{i,j} = u_{i,j} + \omega \left(\frac{2u_{2,j} + u_{1,j-1} + u_{1,j+1} - 4u_{1,j}}{4} \right) \quad (\text{lado izquierdo}),$$

$$(27) \quad u_{n,j} = u_{n,j} + \omega \left(\frac{2u_{n-1,j} + u_{n,j-1} + u_{n,j+1} - 4u_{n,j}}{4} \right) \quad (\text{lado derecho}).$$

Ejemplo 10.7. Vamos a usar un método iterativo para hallar una solución aproximada de la ecuación de Laplace $\nabla^2 u = 0$ en el cuadrado R definido por $R = \{(x, y) : 0 \leq x \leq 4, 0 \leq y \leq 4\}$, con las condiciones de contorno

$$u(x, 0) = 20 \quad \text{y} \quad u(x, 4) = 180 \quad \text{para } 0 < x < 4,$$

y

$$u(0, y) = 80 \quad \text{y} \quad u(4, y) = 0 \quad \text{para } 0 < y < 4.$$

Tabla 10.6 Solución aproximada de la ecuación de Laplace con condiciones de Dirichlet.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
y_9	130.000	180.000	180.000	180.000	180.000	180.000	180.000	180.000	90.0000
y_8	80.000	124.821	141.172	145.414	144.005	137.478	122.642	88.6070	0.0000
y_7	80.000	102.112	113.453	116.479	113.126	103.266	84.4844	51.7856	0.0000
y_6	80.000	89.1736	94.0499	93.9210	88.7553	77.9737	60.2439	34.0510	0.0000
y_5	80.000	80.5319	79.6515	76.3999	70.0003	59.6301	44.4667	24.1744	0.0000
y_4	80.000	73.3023	67.6241	62.0267	55.2159	46.0796	33.8184	18.1798	0.0000
y_3	80.000	65.0528	55.5159	48.8671	42.7568	35.6543	26.5473	14.7266	0.0000
y_2	80.000	51.3931	40.5195	35.1691	31.2899	27.2335	21.9900	14.1791	0.0000
y_1	50.000	20.0000	20.0000	20.0000	20.0000	20.0000	20.0000	20.0000	10.0000

Dividimos el cuadrado en 64 cuadrados de lado $\Delta x = \Delta y = h = 0.5$ y tomamos como valor inicial en los puntos interiores de la malla $u_{i,j} = 70$ para cada $i = 2, \dots, 8$ y $j = 2, \dots, 8$. Usamos el método de sobrerelajación sucesiva con el parámetro $\omega = 1.44646$ (que se obtiene al sustituir $n = 9$ y $m = 9$ en la fórmula (23)); después de 19 iteraciones, los valores residuales son todos menores que una milésima (de hecho, $|r_{i,j}| \leq 0.000606 < 0.001$). Las aproximaciones que se obtienen se muestran en la Tabla 10.6 y su representación tridimensional se muestra en la Figura 10.20. Puesto que las funciones de contorno son discontinuas en las esquinas y los valores correspondientes no se utilizan en los cálculos, hemos tomado $u_{1,1} = 50$, $u_{9,1} = 10$, $u_{1,9} = 130$ y $u_{9,9} = 90$ para completar la Tabla 10.6 y la Figura 10.20. ■

Ejemplo 10.8. Vamos a usar un método iterativo para hallar una solución aproximada de la ecuación de Laplace $\nabla^2 u = 0$ en el cuadrado R dado por $R = \{(x, y) : 0 \leq x \leq 4, 0 \leq y \leq 4\}$, con las condiciones de contorno

$$\begin{aligned} u(x, 4) &= 180 && \text{para } y = 4 && y && 0 < x < 4, \\ u_y(x, 0) &= 0 && \text{para } y = 0 && y && 0 < x < 4, \\ u(0, y) &= 80 && \text{para } x = 0 && y && 0 \leq y < 4, \\ u(4, y) &= 0 && \text{para } x = 4 && y && 0 \leq y < 4. \end{aligned}$$

Dividimos el cuadrado en 64 cuadrados de lado $\Delta x = \Delta y = h = 0.5$, tomamos como valores iniciales en el lado inferior, donde $y = y_1 = 0$, los que resultan de aplicar el método de interpolación lineal y tomamos como valor inicial en los puntos interiores de la malla $u_{i,j} = 70$ para cada $i = 2, \dots, 8$ y $j = 2, \dots, 8$. Usamos el método de sobrerelajación sucesiva con el parámetro $\omega = 1.44646$ (como en el Ejemplo 10.7); después de 29 iteraciones, los valores residuales son todos menores que una milésima (de hecho, $|r_{i,j}| \leq 0.000998 < 0.001$). Las aproximaciones que se obtienen se muestran en la Tabla 10.7 y su representación tridimensional se muestra en la Figura 10.21. Puesto que las funciones de contorno son discontinuas en las esquinas y los valores correspondientes no se utilizan en los cálculos, hemos tomado $u_{1,9} = 130$ y $u_{9,9} = 90$ para completar la Tabla 10.7 y la Figura 10.21. ■

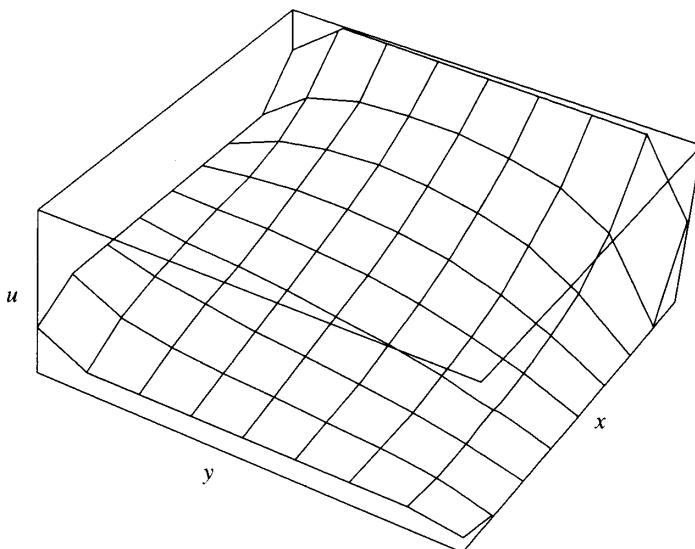


Figura 10.20 Solución $u = u(x, y)$ de un problema de Dirichlet.

Tabla 10.7 Solución aproximada de la ecuación de Laplace con condiciones de contorno mixtas.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
y_9	130.000	180.000	180.000	180.000	180.000	180.000	180.000	180.000	90.0000
y_8	80.000	126.457	142.311	146.837	145.468	138.762	123.583	89.1008	0.0000
y_7	80.000	103.518	115.951	119.568	116.270	105.999	86.4683	52.8201	0.0000
y_6	80.000	91.6621	98.4053	99.2137	94.0461	82.4936	63.4715	35.7113	0.0000
y_5	80.000	84.7247	86.7936	84.8347	78.2063	66.4578	49.2124	26.5538	0.0000
y_4	80.000	80.4424	79.2089	75.1245	67.4860	55.9185	40.3665	21.2915	0.0000
y_3	80.000	77.8354	74.4742	68.9677	60.6944	49.3635	35.0435	18.2459	0.0000
y_2	80.000	76.4244	71.8842	65.5772	56.9600	45.7972	32.1981	16.6485	0.0000
y_1	80.000	75.9774	71.0605	64.4964	55.7707	44.6670	31.3032	16.1500	0.0000

Las ecuaciones de Poisson y Helmholtz

Consideremos la ecuación de Poisson

$$(28) \quad \nabla^2 u = g(x, y).$$

Usando la notación $g_{i,j} = g(x_i, y_j)$, la extensión de la fórmula (20) para resolver la ecuación (28) sobre una malla rectangular es

$$(29) \quad u_{i,j} = u_{i,j} + \frac{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} - h^2 g_{i,j}}{4}.$$

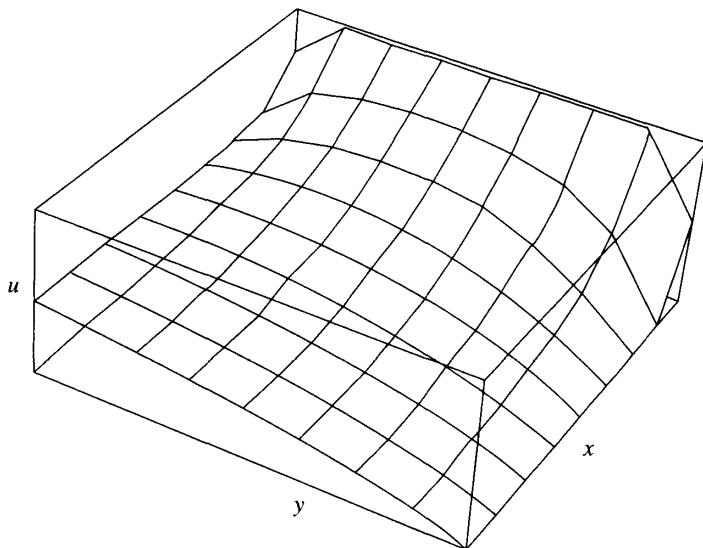


Figura 10.21 Solución $u = u(x, y)$ de un problema mixto.

Consideremos la ecuación de Helmholtz

$$(30) \quad \nabla^2 u + f(x, y)u = g(x, y).$$

Usando la notación $f_{i,j} = f(x_i, y_j)$, la extensión de la fórmula (20) para resolver la ecuación (30) sobre una malla rectangular es

$$(31) \quad u_{i,j} = u_{i,j} + \frac{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - (4 - h^2 f_{i,j})u_{i,j} - h^2 g_{i,j}}{4 - h^2 f_{i,j}}.$$

Estas fórmulas se analizarán con más detalle en los ejercicios.

Mejoras

Una modificación de la aproximación (8) que podemos emplear es la **fórmula de diferencias con nueve puntos** para aproximar la laplaciana:

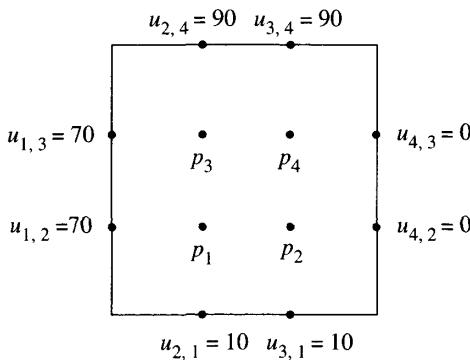
$$\begin{aligned} \nabla^2 u_{i,j} &\approx \frac{1}{6h^2}(u_{i+1,j-1} + u_{i-1,j-1} + u_{i+1,j+1} + u_{i-1,j+1} \\ &\quad + 4u_{i+1,j} + 4u_{i-1,j} + 4u_{i,j+1} + 4u_{i,j-1} - 20u_{i,j}) = 0. \end{aligned}$$

El error de truncamiento para esta fórmula es de orden $O(h^4)$ cuando se usa para resolver las ecuaciones de Poisson o Helmholtz, así que no se produce una mejora frente a la fórmula de cinco puntos. Sin embargo, cuando se utiliza la fórmula de nueve puntos para resolver la ecuación de Laplace $\nabla^2 u = 0$, el error de truncamiento es de orden $O(h^6)$ y sí se obtiene una mejora apreciable.

MATLAB

Programa 10.4 (Resolución de un problema de Dirichlet con la ecuación de Laplace). Construcción de una aproximación a la solución de $u_{xx}(x, y) + u_{yy}(x, y) = 0$ en $R = \{(x, y) : 0 \leq x \leq a, 0 \leq y \leq b\}$ con las condiciones de contorno $u(x, 0) = f_1(x)$, $u(x, b) = f_2(x)$ para $0 \leq x \leq a$ y $u(0, y) = f_3(y)$, $u(a, y) = f_4(y)$ para $0 \leq y \leq b$. Se supone que $\Delta x = \Delta y = h$ y que existen dos números naturales n y m tales que $a = nh$ y $b = mh$.

```
function U=dirich(f1,f2,f3,f4,a,b,h,tol,max1)
% Datos
% - f1,f2,f3,f4 son las funciones en el contorno
% almacenadas como cadenas de caracteres
% - a y b son los extremos superiores de los
% intervalos [0,a] y [0,b]
% - h es el incremento
% - tol es la tolerancia
% Resultado
% - U es la matriz, análoga a la de la Tabla 10.6,
% en la que se almacena la solución numérica
% Inicialización de los parámetros y de U
n=fix(a/h)+1;
m=fix(b/h)+1;
ave=(a*(feval(f1,0)+feval(f2,0)) ...
+ b*(feval(f3,0)+feval(f4,0)))/(2*a+2*b);
U=ave*ones(n,m);
% Condiciones de contorno
U(1,1:m)=feval(f3,0:h:(m-1)*h)';
U(n,1:m)=feval(f4,0:h:(m-1)*h)';
U(1:n,1)=feval(f1,0:h:(n-1)*h);
U(1:n,m)=feval(f2,0:h:(n-1)*h);
U(1,1)=(U(1,2)+U(2,1))/2;
U(1,m)=(U(1,m-1)+U(2,m))/2;
U(n,1)=(U(n-1,1)+U(n,2))/2;
U(n,m)=(U(n-1,m)+U(n,m-1))/2;
% Parámetro de sobrerelajación
w=4/(2+sqrt(4-(cos(pi/(n-1))+cos(pi/(m-1)))^2));
% Mejora de las aproximaciones
err=1;
cnt=0;
while((err>tol)&(cnt<=max1))
    err=0;
```

**Figura 10.22** La malla para el Ejercicio 1.

```

for j=2:m-1
    for i=2:n-1
        relx=w*(U(i,j+1)+U(i,j-1)+U(i+1,j)+U(i-1,j) ...
            -4*U(i,j))/4;
        U(i,j)=U(i,j)+relx;
        if (err<=abs(relx))
            err=abs(relx);
        end
    end
end
cnt=cnt+1;
U=flipud(U');

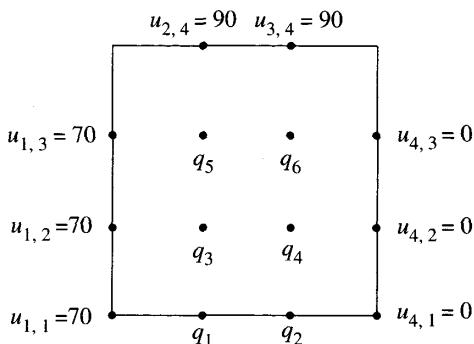
```

Ejercicios

- 1. (a)** Determine el sistema de cuatro ecuaciones con cuatro incógnitas \$p_1\$, \$p_2\$, \$p_3\$ y \$p_4\$ que se usa para calcular las aproximaciones a la función armónica \$u(x, y)\$ en el cuadrado \$R = \{(x, y) : 0 \leq x \leq 3, 0 \leq y \leq 3\}\$ (véase la Figura 10.22). Los valores en la frontera son

$$\begin{aligned} u(x, 0) &= 10 & u(x, 3) &= 90 & \text{para } 0 < x < 3, \\ u(0, y) &= 70 & u(3, y) &= 0 & \text{para } 0 < y < 3. \end{aligned}$$

- (b)** Resuelva las ecuaciones del apartado (a) para hallar \$p_1\$, \$p_2\$, \$p_3\$ y \$p_4\$.
- 2. (a)** Determine el sistema de seis ecuaciones con seis incógnitas \$q_1\$, \$q_2\$, \$\dots\$, \$q_6\$ que se usa para calcular las aproximaciones a la función armónica \$u(x, y)\$

**Figura 10.23** La malla para el Ejercicio 2.

en el cuadrado $R = \{(x, y) : 0 \leq x \leq 3, 0 \leq y \leq 3\}$ (véase la Figura 10.23). Los valores en la frontera son

$$\begin{aligned} u(x, 3) &= 90 & \text{y} & \quad u_y(x, 0) = 90 & \quad \text{para} & \quad 0 < x < 3, \\ u(0, y) &= 70 & \text{y} & \quad u(3, y) = 0 & \quad \text{para} & \quad 0 \leq y < 3. \end{aligned}$$

- (b) Resuelva las ecuaciones del apartado (a) para hallar q_1, q_2, \dots, q_6 .
- 3. (a) Pruebe que $u(x, y) = a_1 \operatorname{sen}(x) \operatorname{senh}(y) + b_1 \operatorname{senh}(x) \operatorname{sen}(y)$ es una solución de la ecuación de Laplace.
- (b) Pruebe que $u(x, y) = a_n \operatorname{sen}(nx) \operatorname{senh}(ny) + b_n \operatorname{senh}(nx) \operatorname{sen}(ny)$ es una solución de la ecuación de Laplace para cada número natural $n = 1, 2, \dots$
- 4. Sea $u(x, y) = x^2 - y^2$. Calcule los valores $u(x+h, y)$, $u(x-h, y)$, $u(x, y+h)$ y $u(x, y-h)$, sustitúyalos en la ecuación (7) y simplifique el resultado.
- 5. (a) Supongamos que u es una función de la forma $u(x, y) = ax^2 + bxy + cy^2 + dx + ey + f$. Determine qué relación entre los coeficientes garantiza que $u_{xx} + u_{yy} = 0$.
- (b) Supongamos que u es una función de la forma dada en el apartado (a). Determine qué relación entre los coeficientes garantiza que $u_{xx} + u_{yy} = -1$.
- (c) Determine los coeficientes de la función polinomial $u(x, y)$ dada en el apartado (a) que verifica la ecuación de Laplace con las condiciones de contorno $u(x, 0) = 0$ y $u(x, \beta) = 0$.
- (d) Determine los coeficientes de la función polinomial $u(x, y)$ dada en el apartado (a) que verifica la ecuación en derivadas parciales del apartado (b) con las condiciones de contorno $u(x, 0) = 0$ y $u(x, \beta) = 0$.
- 6. Resuelva $u_{xx} + u_{yy} = -4u$ en $R = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$ con las condiciones de contorno

$$\begin{aligned} u(x, 0) &= \cos(2x) & \text{y} & \quad u(x, 1) = \cos(2x) + \operatorname{sen}(2) & \quad \text{para} & \quad 0 \leq x \leq 1, \\ u(0, y) &= \operatorname{sen}(2y) & \text{y} & \quad u(1, y) = \operatorname{sen}(2x) + \cos(2) & \quad \text{para} & \quad 0 \leq y \leq 1. \end{aligned}$$

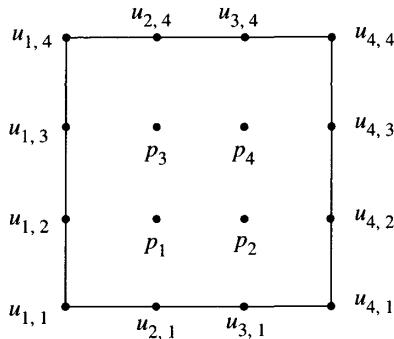


Figura 10.24 La malla para el Ejercicio 7.

7. Determine el sistema de cuatro ecuaciones con cuatro incógnitas p_1 , p_2 , p_3 y p_4 que aparece al utilizar la fórmula de diferencias con nueve puntos sobre la malla de orden 4×4 que se muestra en la Figura 10.24.

Algoritmos y programas

1. (a) Use el Programa 10.4 para calcular aproximaciones a la función armónica $u(x, y)$ en el cuadrado $R = \{(x, y) : 0 \leq x \leq 1.5, 0 \leq y \leq 1.5\}$. Tome $h = 0.5$ y los valores en la frontera

$$\begin{aligned} u(x, 0) &= x^4, & u(x, 1.5) &= x^4 - 13.5x^2 + 5.0625 \text{ para } 0 \leq x \leq 1.5, \\ u(0, y) &= y^4, & u(1.5, y) &= y^4 - 13.5y^2 + 5.0625 \text{ para } 0 \leq y \leq 1.5. \end{aligned}$$

- (b) Use la instrucción `surf` para dibujar la aproximación obtenida en el apartado (a) y compárela con la solución exacta, que viene dada por $u(x, y) = x^4 - 6x^2y^2 + y^4$.

2. Modifique el Programa 9.11 (sistemas tridiagonales) de manera que pueda resolver sistemas pentadiagonales

3. (a) Use una malla de orden 5×5 parecida a la del Ejemplo 10.5 y determine el sistema de nueve ecuaciones con nueve incógnitas $p_1, p_2, p_3, \dots, p_9$ para calcular aproximaciones a la función $u(x, y)$ que es armónica en el cuadrado $R = \{(x, y) : 0 \leq x \leq 4, 0 \leq y \leq 4\}$ y verifica las condiciones de contorno

$$\begin{aligned} u(x, 0) &= 10 & u(x, 4) &= 120 & \text{para } 0 < x < 4, \\ u(0, y) &= 90 & u(4, y) &= 40 & \text{para } 0 < y < 4. \end{aligned}$$

- (b) Use su modificación del Programa 9.11 para calcular p_1, p_2, \dots, p_9 .
(c) Use el Programa 10.4 para calcular las aproximaciones.

- (d) Use una malla de orden 9×9 como la del Ejemplo 10.7 y el Programa 10.4 para calcular las aproximaciones.
4. (a) Use una malla de orden 5×5 como la del Ejemplo 10.6 y determine el sistema de doce ecuaciones con doce incógnitas q_1, q_2, \dots, q_{12} para calcular aproximaciones a la función $u(x, y)$ que es armónica en el cuadrado $R = \{(x, y) : 0 \leq x \leq 4, 0 \leq y \leq 4\}$ y verifica las condiciones de contorno

$$\begin{aligned} u(x, 4) &= 120 & y & \quad u_y(x, 0) = 0 & \quad \text{para} & \quad 0 < x < 4, \\ u(0, y) &= 90 & y & \quad u(4, y) = 40 & \quad \text{para} & \quad 0 \leq y < 4. \end{aligned}$$

- (b) Use su modificación del Programa 9.11 para calcular q_1, q_2, \dots, q_{12} .
 (c) Modifique el Programa 10.4 para calcular las aproximaciones.
 (d) Use una malla de orden 9×9 como la del Ejemplo 10.8 y una modificación adecuada del Programa 10.4 para calcular las aproximaciones.
5. (a) Usando una malla de orden 5×5 , deduzca el sistema de nueve ecuaciones con nueve incógnitas $p_1, p_2, p_3, \dots, p_9$ para calcular aproximaciones a la solución $u(x, y)$ de la ecuación de Poisson con $g(x, y) = 2$ en el cuadrado $R = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$ con las condiciones de contorno

$$\begin{aligned} u(x, 0) &= x^2 & y & \quad u(x, 1) = (x - 1)^2 & \quad \text{para} & \quad 0 \leq x \leq 1, \\ u(0, y) &= y^2 & y & \quad u(1, y) = (y - 1)^2 & \quad \text{para} & \quad 0 \leq y \leq 1. \end{aligned}$$

- (b) Use su modificación del Programa 9.11 para calcular p_1, p_2, \dots, p_9 .
 (c) Modifique el Programa 10.4 para calcular las aproximaciones.
 (d) Use una malla de orden 9×9 y su modificación del Programa 10.4 para calcular las aproximaciones.
6. (a) Usando una malla de orden 5×5 , deduzca el sistema de nueve ecuaciones con nueve incógnitas $p_1, p_2, p_3, \dots, p_9$ para calcular aproximaciones a la solución $u(x, y)$ de la ecuación de Poisson con $g(x, y) = y$ en el cuadrado $R = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$ con las condiciones de contorno

$$\begin{aligned} u(x, 0) &= x^3 & y & \quad u(x, 1) = x^3 & \quad \text{para} & \quad 0 \leq x \leq 1, \\ u(0, y) &= 0 & y & \quad u(1, y) = 1 & \quad \text{para} & \quad 0 \leq y \leq 1. \end{aligned}$$

- (b) Use su modificación del Programa 9.11 para calcular p_1, p_2, \dots, p_9 .
 (c) Modifique el Programa 10.4 para calcular las aproximaciones.
 (d) Use una malla de orden 9×9 y su modificación del Programa 10.4 para calcular las aproximaciones.

Autovalores y autovectores

El diseño de algunos sistemas en ingeniería tiene en cuenta lo que se conoce como el “**criterio de fallo de la tensión principal máxima**”. Esta teoría se basa en la hipótesis de que la tensión máxima que actúa sobre un cuerpo es la que determina su rotura. El resultado matemático relacionado es el teorema de los ejes principales, o teorema espectral, para una aplicación lineal $\mathbf{Y} = \mathbf{AX}$. Cuando la dimensión es dos y la matriz \mathbf{A} es simétrica, el teorema dice que existe una base formada por dos vectores ortogonales \mathbf{U}_1 y \mathbf{U}_2 de manera que el efecto de la aplicación consiste en estirar el espacio, multiplicando por unos factores λ_1 y λ_2 , llamados autovalores, en las direcciones señaladas por \mathbf{U}_1 y

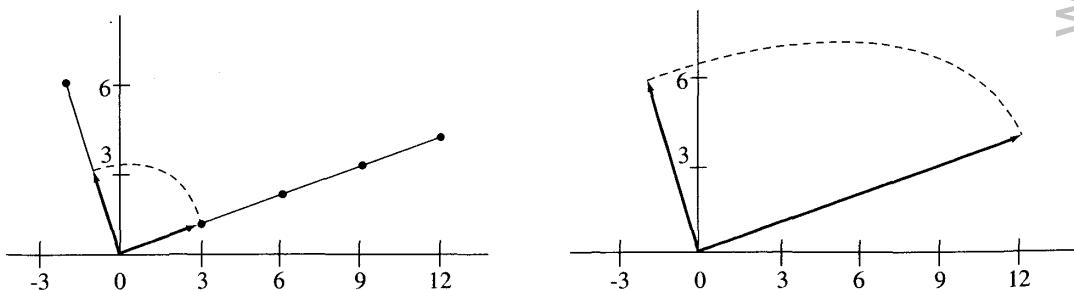


Figura 11.1 (a) Preimágenes $\mathbf{U}_1 = [3 \ 1]'$ y $\mathbf{U}_2 = [-1 \ 3]'$ de la aplicación $\mathbf{Y} = \mathbf{AX}$.
(b) Los vectores imágenes $\mathbf{V}_1 = \mathbf{AU}_1 = [12 \ 4]'$ y $\mathbf{V}_2 = \mathbf{AU}_2 = [-2 \ 6]'$.

\mathbf{U}_2 , respectivamente. Consideremos las matriz simétrica

$$\begin{bmatrix} 3.8 & 0.6 \\ 0.6 & 2.2 \end{bmatrix};$$

sus direcciones principales son $\mathbf{U}_1 = [3 \ 1]'$ y $\mathbf{U}_2 = [-1 \ 3]'$, con autovalores correspondientes $\lambda_1 = 4$ y $\lambda_2 = 2$, respectivamente. Las imágenes de estos vectores son $\mathbf{V}_1 = \mathbf{A}\mathbf{U}_1 = [12 \ 4]' = 4[3 \ 1]'$ y $\mathbf{V}_2 = \mathbf{A}\mathbf{U}_2 = [-2 \ 6]' = 2[-1 \ 3]'$. Esta aplicación transforma el cuarto de círculo mostrado en la Figura 11.1(a) en el cuarto de elipse que se muestra en la Figura 11.1(b).

11.1 El problema de los autovalores

Repaso

Daremos un breve repaso de algunas ideas y conceptos del álgebra lineal. Las demostraciones de los teoremas quedan propuestas como ejercicios o bien pueden consultarse en cualquier texto estándar de álgebra lineal (véase la Referencia [132]).

En el Capítulo 3 vimos métodos para resolver un sistema de n ecuaciones lineales con n incógnitas. Allí se asumía como hipótesis que el determinante de la matriz de los coeficientes era distinto de cero y, en consecuencia, que la solución era única. En el caso de un sistema homogéneo $\mathbf{AX} = \mathbf{0}$, si $\det(\mathbf{A}) \neq 0$, entonces la única solución es la solución trivial $\mathbf{X} = \mathbf{0}$. Por el contrario, si $\det(\mathbf{A}) = 0$, entonces existen soluciones no triviales de $\mathbf{AX} = \mathbf{0}$. Supongamos pues, que $\det(\mathbf{A}) = 0$, y vamos a ver cómo podemos resolver el sistema lineal homogéneo

$$(1) \quad \begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= 0 \\ \vdots &\quad \vdots \quad \vdots \quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= 0. \end{aligned}$$

El sistema de ecuaciones (1) siempre tiene la solución trivial $x_1 = 0, x_2 = 0, \dots, x_n = 0$. Usando el método de eliminación de Gauss podemos obtener otras soluciones formando un conjunto de relaciones entre las variables.

Ejemplo 11.1. Vamos a determinar las soluciones no triviales del sistema homogéneo

$$x_1 + 2x_2 - x_3 = 0$$

$$2x_1 + x_2 + x_3 = 0$$

$$5x_1 + 4x_2 + x_3 = 0.$$

Eliminamos x_1 usando el método de Gauss, obteniendo

$$\begin{aligned}x_1 + 2x_2 - x_3 &= 0 \\-3x_2 + 3x_3 &= 0 \\-6x_2 + 6x_3 &= 0.\end{aligned}$$

Dado que la tercera ecuación es múltiplo de la segunda, el sistema se reduce a dos ecuaciones con tres incógnitas:

$$\begin{aligned}x_1 + x_2 &= 0 \\-x_2 + x_3 &= 0.\end{aligned}$$

Podemos elegir una de las incógnitas y usarla como parámetro; por ejemplo, tomando $x_3 = t$, la segunda ecuación implica que $x_2 = t$ y la primera ecuación nos dice que $x_1 = -t$. Por tanto, la solución puede expresarse como el conjunto de relaciones:

$$\begin{aligned}x_1 &= -t \\x_2 &= t \quad \text{o bien} \\x_3 &= t\end{aligned} \quad \mathbf{X} = \begin{bmatrix} -t \\ t \\ t \end{bmatrix} = t \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix},$$

siendo t cualquier número real.

Definición 11.1 (Independencia lineal). Se dice que los vectores $\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_n$ son *linealmente independientes* si la igualdad

$$(2) \quad c_1 \mathbf{U}_1 + c_2 \mathbf{U}_2 + \cdots + c_n \mathbf{U}_n = \mathbf{0}$$

implica que $c_1 = 0, c_2 = 0, \dots, c_n = 0$. Si los vectores no son linealmente independientes, entonces se dice que son linealmente dependientes. En otras palabras, los vectores son *linealmente dependientes* si existe un conjunto de números $\{c_1, c_2, \dots, c_n\}$ no todos cero, tales que la igualdad (2) se verifica. ▲

Dos vectores en el espacio euclídeo bidimensional \mathbb{R}^2 son linealmente independientes si, y sólo si, no son paralelos. Tres vectores en el espacio euclídeo tridimensional \mathbb{R}^3 son linealmente independientes si, y sólo si, no están sobre un mismo plano.

Teorema 11.1. Los vectores $\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_n$ son linealmente dependientes si, y sólo si, alguno de ellos puede escribirse como combinación lineal de los demás.

Un aspecto muy importante es la posibilidad de expresar cada vector de un espacio vectorial como una combinación lineal de vectores extraídos de un conjunto pequeño; esto da lugar a la siguiente definición.

Definición 11.2 (Base). Supongamos que $S = \{\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_m\}$ es un conjunto de m vectores en el espacio euclídeo n -dimensional \mathbb{R}^n . Se dice que el conjunto S es una base de \mathbb{R}^n si para cada vector \mathbf{X} de \mathbb{R}^n existe un único conjunto de escalares $\{c_1, c_2, \dots, c_m\}$ tales que \mathbf{X} puede expresarse como la combinación lineal

$$(3) \quad \mathbf{X} = c_1 \mathbf{U}_1 + c_2 \mathbf{U}_2 + \cdots + c_m \mathbf{U}_m. \quad \blacktriangle$$

Teorema 11.2. En \mathbb{R}^n , cualquier conjunto de n vectores linealmente independientes forma una base de \mathbb{R}^n . Cada vector \mathbf{X} de \mathbb{R}^n se expresa de manera única como combinación lineal, dada en (3), de los vectores de la base.

Teorema 11.3. Sean $\mathbf{K}_1, \mathbf{K}_2, \dots, \mathbf{K}_m$ vectores en \mathbb{R}^n .

- (4) Si $m > n$, entonces los vectores son linealmente dependientes.
- (5) Si $m = n$, entonces los vectores son linealmente dependientes si, y sólo si, $\det(\mathbf{K}) = 0$, siendo $\mathbf{K} = [\mathbf{K}_1 \ \mathbf{K}_2 \ \dots \ \mathbf{K}_m]$.

Autovalores

En el análisis de algunos modelos matemáticos surgen preguntas como las siguientes: ¿Cuándo es singular la matriz $\mathbf{A} - \lambda \mathbf{I}$, siendo λ un parámetro? ¿Cuál es el comportamiento de la sucesión de vectores $\{\mathbf{A}^j \mathbf{X}_0\}_{j=0}^{\infty}$? ¿Cuál es la interpretación geométrica de una transformación? Las soluciones de muchos problemas que se plantean en disciplinas diferentes, como la economía, la ingeniería o la física, involucran este tipo de preguntas. La teoría de autovalores y autovectores es suficientemente potente como para resolver estos problemas que, de otra forma, serían intratables.

Sea \mathbf{A} una matriz cuadrada de dimensión $n \times n$ y sea \mathbf{X} un vector de dimensión n . El producto $\mathbf{Y} = \mathbf{AX}$ puede verse como una transformación del espacio n -dimensional en sí mismo y nos planteamos la búsqueda de escalares λ para los que exista un vector no nulo \mathbf{X} tal que

$$(6) \quad \mathbf{AX} = \lambda \mathbf{X};$$

es decir, tales que la aplicación lineal $T(\mathbf{X}) = \mathbf{AX}$ transforme \mathbf{X} en su múltiplo $\lambda \mathbf{X}$. Cuando esto ocurre, se dice que \mathbf{X} es un autovector correspondiente al autovalor λ y juntos forman un “autopar” (λ, \mathbf{X}) de \mathbf{A} . En general, el escalar λ y las componentes del vector \mathbf{X} son números complejos pero, por simplicidad, la mayoría de nuestros ejemplos sólo tratarán casos con números reales. No obstante, las técnicas pueden extenderse fácilmente al caso complejo. La matriz identidad \mathbf{I} puede usarse para expresar la ecuación (6) como $\mathbf{AX} = \lambda \mathbf{IX}$ que, a su vez, puede escribirse en la forma habitual de un sistema lineal

$$(7) \quad (\mathbf{A} - \lambda \mathbf{I}) \mathbf{X} = \mathbf{0}.$$

Lo importante de la ecuación (7) es que el producto de la matriz $(\mathbf{A} - \lambda \mathbf{I})$ por el vector no nulo \mathbf{X} es el vector cero! De acuerdo con el Teorema 3.5, este sistema lineal tiene soluciones no triviales si, y sólo si, la matriz $\mathbf{A} - \lambda \mathbf{I}$ es singular, es decir,

$$(8) \quad \det(\mathbf{A} - \lambda \mathbf{I}) = 0.$$

Este determinante puede escribirse como

$$(9) \quad \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0$$

y, al desarrollarlo, se forma un polinomio de grado n que se llama polinomio característico de la matriz:

$$(10) \quad \begin{aligned} p(\lambda) &= \det(\mathbf{A} - \lambda \mathbf{I}) \\ &= (-1)^n (\lambda^n + c_1 \lambda^{n-1} + c_2 \lambda^{n-2} + \cdots + c_{n-1} \lambda + c_n). \end{aligned}$$

Todo polinomio de grado n tiene exactamente n raíces (no necesariamente distintas). Sustituyendo cada raíz λ en la ecuación (7), obtenemos un sistema de ecuaciones subdeterminado que tiene una solución no trivial \mathbf{X} y, si λ es real, entonces podemos encontrar un autovector real \mathbf{X} . Para poner énfasis en estos conceptos, damos las siguientes definiciones.

Definición 11.3 (Autovalor). Si \mathbf{A} es una matriz real de orden $n \times n$, entonces sus n autovalores $\lambda_1, \lambda_2, \dots, \lambda_n$ son las raíces, reales o complejas, de su polinomio característico

$$(11) \quad p(\lambda) = \det(\mathbf{A} - \lambda \mathbf{I}).$$

Definición 11.4 (Autowector). Si λ es un autovalor de \mathbf{A} y el vector no nulo \mathbf{V} verifica

$$(12) \quad \mathbf{AV} = \lambda \mathbf{V},$$

entonces se dice que \mathbf{V} es un autowector de \mathbf{A} correspondiente al autovalor λ , lo que también expresaremos diciendo que (λ, \mathbf{V}) es una pareja autovalor-autowector de \mathbf{A} .

El polinomio característico (11) puede factorizarse como

$$(13) \quad p(\lambda) = (-1)^n (\lambda - \lambda_1)^{m_1} (\lambda - \lambda_2)^{m_2} \cdots (\lambda - \lambda_k)^{m_k},$$

donde m_j se llama multiplicidad del autovalor λ_j ; la suma de las multiplicidades de todos los autovalores es n , o sea,

$$n = m_1 + m_2 + \cdots + m_k.$$

Recordamos ahora tres resultados sobre la existencia de autowectores.

Teorema 11.4. (a) Cada autovalor λ tiene, al menos, un autovector \mathbf{V} que le corresponde.

(b) Si λ tiene multiplicidad r , entonces le corresponden como mucho r autovectores linealmente independientes $\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_r$.

Teorema 11.5. Supongamos que \mathbf{A} es una matriz cuadrada y que $\lambda_1, \lambda_2, \dots, \lambda_k$ son autovalores distintos de \mathbf{A} con autovectores asociados $\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_k$, respectivamente. Entonces $\{\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_k\}$ es un conjunto de vectores linealmente independientes.

Teorema 11.6. Si los autovalores de una matriz \mathbf{A} de orden $n \times n$ son todos distintos, entonces existen n autovectores $\{\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_n\}$ linealmente independientes.

El Teorema 11.4 se suele aplicar cuando se hacen los cálculos a mano de la siguiente manera: Sustituimos el autovalor λ de multiplicidad $r \geq 1$ en la ecuación

$$(14) \quad (\mathbf{A} - \lambda \mathbf{I})\mathbf{V} = \mathbf{0}$$

y llevamos a cabo el método de eliminación de Gauss para obtener la forma reducida de este sistema, en la que aparecerán $n - k$ ecuaciones con n incógnitas, siendo $1 \leq k \leq r$. Por tanto, tenemos k variables libres que escogemos de forma adecuada para obtener k soluciones linealmente independientes $\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_k$, que serán autovectores correspondientes a λ .

Ejemplo 11.2. Vamos a calcular las parejas autovalor-autovector $(\lambda_j, \mathbf{V}_j)$ de la matriz

$$\mathbf{A} = \begin{bmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{bmatrix}.$$

Mostraremos también que los autovectores son independientes.

La ecuación característica $\det(\mathbf{A} - \lambda \mathbf{I}) = 0$ es:

$$(15) \quad \begin{vmatrix} 3 - \lambda & -1 & 0 \\ -1 & 2 - \lambda & -1 \\ 0 & -1 & 3 - \lambda \end{vmatrix} = -\lambda^3 + 8\lambda^2 - 19\lambda + 12 = 0,$$

que puede escribirse como $-(\lambda - 1)(\lambda - 3)(\lambda - 4) = 0$. Por tanto, los tres autovalores son $\lambda_1 = 1$, $\lambda_2 = 3$ y $\lambda_3 = 4$.

Caso (i): Sustituyendo $\lambda_1 = 1$ en la ecuación (14) obtenemos

$$\begin{aligned} 2x_1 - x_2 &= 0 \\ -x_1 + x_2 - x_3 &= 0 \\ -x_2 + 2x_3 &= 0. \end{aligned}$$

Puesto que la suma de la primera ecuación más dos veces la segunda más la tercera es idénticamente cero, el sistema se reduce a dos ecuaciones con tres incógnitas:

$$\begin{aligned} 2x_1 - x_2 &= 0 \\ -x_2 + 2x_3 &= 0. \end{aligned}$$

Tomando $x_2 = 2a$, siendo a una constante no nula arbitraria, podemos usar estas ecuaciones para obtener $x_1 = a$ y $x_3 = a$, respectivamente. Por tanto, el primer par es $(\lambda_1 = 1, \mathbf{V}_1 = [a \ 2a \ a]' = a[1 \ 2 \ 1]')$.

Caso (ii): Sustituyendo $\lambda_2 = 3$ en la ecuación (14) obtenemos

$$\begin{aligned} -x_2 &= 0 \\ -x_1 - x_2 - x_3 &= 0 \\ -x_2 &= 0, \end{aligned}$$

que es equivalente al sistema de dos ecuaciones

$$\begin{aligned} x_1 &+ x_3 = 0 \\ x_2 &= 0. \end{aligned}$$

Tomando $x_1 = b$, con b una constante no nula arbitraria, tenemos $x_3 = -b$, así que la segunda pareja es $(\lambda_2 = 3, \mathbf{V}_2 = [b \ 0 \ -b]' = b[1 \ 0 \ -1]')$.

Caso (iii): Sustituyendo $\lambda_3 = 4$ en (14) nos queda

$$\begin{aligned} -x_1 - x_2 &= 0 \\ -x_1 - 2x_2 - x_3 &= 0 \\ -x_2 - x_3 &= 0, \end{aligned}$$

que es equivalente al sistema de dos ecuaciones

$$\begin{aligned} x_1 + x_2 &= 0 \\ x_2 + x_3 &= 0. \end{aligned}$$

Tomando $x_3 = c$, siendo c una constante no nula, podemos usar la segunda ecuación para obtener $x_2 = -c$ y, luego, la primera para obtener $x_1 = c$. Por tanto, la tercera pareja es $(\lambda_3 = 4, \mathbf{V}_3 = [c \ -c \ c]' = c[1 \ -1 \ 1]')$.

Para probar que los vectores son linealmente independientes, basta aplicar el Teorema 11.5. Sin embargo, nunca viene mal revisar las técnicas del álgebra lineal y utilizar el Teorema 11.3. Formamos el determinante

$$\det([\mathbf{V}_1 \ \mathbf{V}_2 \ \mathbf{V}_3]) = \begin{vmatrix} a & b & c \\ 2a & 0 & -c \\ a & -b & c \end{vmatrix} = -6abc$$

y, puesto que $\det([\mathbf{V}_1 \ \mathbf{V}_2 \ \mathbf{V}_3]) \neq 0$, el Teorema 11.3 implica que los vectores \mathbf{V}_1 , \mathbf{V}_2 y \mathbf{V}_3 son linealmente independientes ■

En el Ejemplo 11.2 se muestra cómo podemos realizar las operaciones a mano para calcular autovalores y autovectores cuando la dimensión n es baja: (1) se determinan los coeficientes del polinomio característico; (2) se calculan sus raíces; (3) se hallan las soluciones no triviales de cada sistema de ecuaciones lineales $(\mathbf{A} - \lambda \mathbf{I})\mathbf{V} = \mathbf{0}$. Cuando la dimensión es alta existen varios métodos para resolver el problema del cálculo de autovalores; siguiendo la tendencia marcada en los últimos años, presentaremos los métodos de las potencias, el método de Jacobi y el algoritmo QR . El algoritmo QR y sus mejoras son los que se utilizan en los paquetes de programas para profesionales como el EISPACK o el MATLAB (véase la Referencia [178]).

Puesto que en la igualdad (12) el vector \mathbf{V} se multiplica por la derecha de \mathbf{A} , se dice a veces que es un **autovector derecho** correspondiente a λ . Existen también los autovectores izquierdos \mathbf{Y} , que son los que verifican

$$(16) \quad \mathbf{Y}' \mathbf{A} = \lambda \mathbf{Y}'.$$

En general, un autovector izquierdo no tiene por qué ser un autovector derecho. Sin embargo, si \mathbf{A} es real y simétrica ($\mathbf{A}' = \mathbf{A}$), entonces

$$(17) \quad (\mathbf{AV})' = \mathbf{V}' \mathbf{A}' = \mathbf{V}' \mathbf{A}, \\ (\lambda \mathbf{V})' = \lambda \mathbf{V}'.$$

Por tanto, todo autovector derecho \mathbf{V} es también un autovector izquierdo cuando la matriz \mathbf{A} es simétrica. En el resto del capítulo trabajaremos sólo con autovectores derechos.

Si tenemos un autovector \mathbf{V} y c es un escalar, entonces $c\mathbf{V}$ también es un autovector como prueba el siguiente cálculo

$$(18) \quad \mathbf{A}(c\mathbf{V}) = c(\mathbf{AV}) = c(\lambda \mathbf{V}) = \lambda(c\mathbf{V}),$$

así que, para fijar ideas y tener algún tipo de unicidad, normalizaremos el autovector usando alguna de las normas

$$(19) \quad \|\mathbf{X}\|_\infty = \max\{|x_k| : 1 \leq k \leq n\}$$

o bien

$$(20) \quad \|\mathbf{X}\|_2 = \left(\sum_{k=1}^n |x_k|^2 \right)^{1/2}$$

y exigiendo que o bien $\|\mathbf{X}\|_\infty = 1$, o bien $\|\mathbf{X}\|_2 = 1$.

Matrices diagonalizables

El problema de los autovalores se entiende muy fácilmente cuando la matriz es una matriz diagonal \mathbf{D} dada como

$$(21) \quad \mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$

Sea $\mathbf{E}_j = [0 \ 0 \ \cdots \ 0 \ 1 \ 0 \ \cdots \ 0]'$, donde la componente j -ésima es 1 y todas las demás son 0, el vector j -ésimo de la base canónica. Entonces

$$(22) \quad \mathbf{D}\mathbf{E}_j = [0 \ 0 \ \cdots \ 0 \ \lambda_j \ 0 \ \cdots \ 0]' = \lambda_j \mathbf{E}_j,$$

así que las parejas autovalor-autovector de \mathbf{D} son $(\lambda_j, \mathbf{E}_j)$ para $j = 1, 2, \dots, n$. Sería entonces deseable disponer de alguna forma simple de transformar una matriz cualquiera \mathbf{A} en una matriz diagonal que tenga los mismos autovalores.

Definición 11.5. Se dice que dos matrices \mathbf{A} y \mathbf{B} de orden $n \times n$ son *semejantes* si existe una matriz invertible \mathbf{K} tal que

$$(23) \quad \mathbf{B} = \mathbf{K}^{-1} \mathbf{A} \mathbf{K}. \quad \blacktriangle$$

Teorema 11.7. Supongamos que \mathbf{A} y \mathbf{B} son matrices semejantes y que λ es un autovalor de \mathbf{A} con autovector correspondiente \mathbf{V} . Entonces λ es también un autovalor de \mathbf{B} . Si, además, $\mathbf{K}^{-1} \mathbf{A} \mathbf{K} = \mathbf{B}$, entonces $\mathbf{Y} = \mathbf{K}^{-1} \mathbf{V}$ es un autovector de \mathbf{B} asociado al autovalor λ .

Se dice que una matriz \mathbf{A} de orden $n \times n$ es *diagonalizable* si es semejante a una matriz diagonal. El siguiente teorema pone de manifiesto el papel crucial que juegan los autovectores en este proceso.

Teorema 11.8 (Diagonalización). Una matriz \mathbf{A} de orden $n \times n$ es semejante a una matriz diagonal \mathbf{D} si, y sólo si, tiene n autovectores linealmente independientes. En ese caso, se tiene

$$(24) \quad \begin{aligned} \mathbf{V}^{-1} \mathbf{A} \mathbf{V} &= \mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) \\ \mathbf{V} &= [\mathbf{V}_1 \ \mathbf{V}_2 \ \dots \ \mathbf{V}_n], \end{aligned}$$

donde los n pares autovalor-autovector son $(\lambda_j, \mathbf{V}_j)$, para $j = 1, 2, \dots, n$.

El Teorema 11.8 implica que toda matriz \mathbf{A} de orden $n \times n$ que tenga n autovalores distintos es diagonalizable.

Ejemplo 11.3. Veamos que la siguiente matriz es diagonalizable.

$$\mathbf{A} = \begin{bmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{bmatrix}.$$

En el Ejemplo 11.2 calculamos los autovalores de \mathbf{A} , que son $\lambda_1 = 1$, $\lambda_2 = 3$ y $\lambda_3 = 4$, y la matriz de los autovectores

$$\mathbf{V} = [\mathbf{V}_1 \ \mathbf{V}_2 \ \mathbf{V}_3] = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 0 & -1 \\ 1 & -1 & 1 \end{bmatrix},$$

cuya matriz inversa \mathbf{V}^{-1} es

$$\mathbf{V}^{-1} = \begin{bmatrix} \frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{2} & 0 & -\frac{1}{2} \\ \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} \end{bmatrix}.$$

Dejamos como ejercicio completar los detalles del cálculo del producto que aparece en la relación (24):

$$\begin{bmatrix} \frac{1}{6} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{2} & 0 & -\frac{1}{2} \\ \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 0 & -1 \\ 1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix},$$

lo que prueba que \mathbf{A} es diagonalizable: $\mathbf{V}^{-1}\mathbf{AV} = \mathbf{D} = \text{diag}(1, 3, 4)$.

Un resultado más general que relaciona la estructura de una matriz con sus autovalores es el siguiente teorema.

Teorema 11.9 (Teorema de Schur). Supongamos que \mathbf{A} es una matriz cualquiera de orden $n \times n$. Entonces existen una matriz invertible \mathbf{P} y una matriz triangular superior \mathbf{T} cuyos elementos diagonales son los autovalores de \mathbf{A} , tales que $\mathbf{T} = \mathbf{P}^{-1}\mathbf{AP}$.

En algunos problemas de análisis de estructuras en ingeniería se requiere disponer de una base de \mathbb{R}^n que esté formada por autovectores de una matriz \mathbf{A} . Esto hace que resulte más fácil entender cómo se transforma el espacio mediante la aplicación $\mathbf{Y} = \mathbf{T}(\mathbf{X}) = \mathbf{AX}$: Recordemos que un par autovalor-autovector $(\lambda_j, \mathbf{V}_j)$ tiene la propiedad de que T transforma \mathbf{V}_j en su múltiplo $\lambda_j \mathbf{V}_j$. Esta propiedad se explota en el siguiente teorema.

Teorema 11.10. Sea \mathbf{A} una matriz de orden $n \times n$ que tiene n parejas autovalor-autovector linealmente independientes $(\lambda_j, \mathbf{V}_j)$ para $j = 1, 2, \dots, n$. Entonces cualquier vector \mathbf{X} de \mathbb{R}^n puede escribirse de forma única como combinación lineal de los autovectores:

$$(25) \quad \mathbf{X} = c_1 \mathbf{V}_1 + c_2 \mathbf{V}_2 + \cdots + c_n \mathbf{V}_n.$$

La aplicación lineal $T(\mathbf{X}) = \mathbf{AX}$ transforma \mathbf{X} en el vector

$$(26) \quad \mathbf{Y} = T(\mathbf{X}) = c_1 \lambda_1 \mathbf{V}_1 + c_2 \lambda_2 \mathbf{V}_2 + \cdots + c_n \lambda_n \mathbf{V}_n.$$

Ejemplo 11.4. Supongamos que una matriz \mathbf{A} de orden 3×3 tiene como autovalores $\lambda_1 = 2$, $\lambda_2 = -1$ y $\lambda_3 = 4$, con autovectores correspondientes $\mathbf{V}_1 = [1 \ 2 \ -2]', \mathbf{V}_2 = [-2 \ 1 \ 1]', \text{ y } \mathbf{V}_3 = [1 \ 3 \ -4]',$ respectivamente. Vamos a determinar la imagen de $\mathbf{X} = [-1 \ 2 \ 1]'$ mediante la aplicación lineal $T(\mathbf{X}) = \mathbf{AX}$.

Tenemos que expresar \mathbf{X} como combinación lineal de los autovectores, es decir, hay que resolver la ecuación

$$[-1 \ 2 \ 1]' = c_1 [1 \ 2 \ -2]' + c_2 [-2 \ 1 \ 1]' + c_3 [1 \ 3 \ -4]'$$

cuyas incógnitas son c_1 , c_2 y c_3 . Observemos que esto es equivalente a resolver el sistema lineal

$$\begin{aligned} c_1 - 2c_2 + c_3 &= -1 \\ 2c_1 + c_2 + 3c_3 &= 2 \\ -2c_1 + c_2 - 4c_3 &= 1, \end{aligned}$$

cuya solución es $c_1 = 2$, $c_2 = 1$ y $c_3 = -1$. Usando la Definición 11.4 de autovector, podemos calcular $T(\mathbf{X})$ como sigue

$$\begin{aligned} T(\mathbf{X}) &= \mathbf{A}(2\mathbf{V}_1 + \mathbf{V}_2 - \mathbf{V}_3) \\ &= 2\mathbf{AV}_1 + \mathbf{AV}_2 - \mathbf{AV}_3 \\ &= 2(2\mathbf{V}_1) - \mathbf{V}_2 - 4\mathbf{V}_3 \\ &= [2 \ -5 \ 7]'. \end{aligned}$$

Propiedades de las matrices simétricas

No es fácil determinar cuántos autovectores independientes tiene una matriz cualquiera, lo mejor es utilizar los potentes algoritmos que se encuentran en los paquetes de programas para profesionales como el EISPACK o el MATLAB. Sin embargo, si se sabe que toda matriz real y simétrica de orden $n \times n$ tiene n autovectores reales, de manera que a un autovalor de multiplicidad m_j le corresponden m_j autovectores linealmente independientes. En consecuencia, toda matriz real y simétrica es diagonalizable.

Definición 11.6 (Ortogonal). Se dice que los vectores $\{\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_n\}$ son ortogonales si se verifica que

$$(27) \quad \mathbf{V}'_j \mathbf{V}_k = 0 \quad \text{siempre que} \quad j \neq k.$$

Definición 11.7 (Ortonormal). Supongamos que $\{\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_n\}$ es un conjunto de vectores ortogonales. Entonces se dice que son ortonormales si todos son de norma unidad, es decir,

$$(28) \quad \mathbf{V}'_j \mathbf{V}_k = 0 \quad \text{siempre que} \quad j \neq k.$$

$$\mathbf{V}'_j \mathbf{V}_j = 1 \quad \text{para todo } j = 1, 2, \dots, n.$$

Teorema 11.11. Vectores ortonormales son linealmente independientes.

Observación. El vector cero no puede estar en ningún conjunto de vectores ortonormales.

Definición 11.8 (Matriz ortogonal). Se dice que una matriz \mathbf{A} de orden $n \times n$ es ortogonal si su traspuesta \mathbf{A}' es la inversa de \mathbf{A} ; o sea,

$$(29) \quad \mathbf{A}' \mathbf{A} = I,$$

que es equivalente a

$$(30) \quad \mathbf{A}^{-1} = \mathbf{A}'.$$

En otros términos, \mathbf{A} es ortogonal si, y sólo si, sus columnas (o filas) son ortonormales.

Teorema 11.12 (Teorema espectral o de los ejes principales). Si \mathbf{A} es una matriz real y simétrica, entonces existe una matriz ortogonal \mathbf{K} tal que

$$(31) \quad \mathbf{K}' \mathbf{A} \mathbf{K} = \mathbf{K}^{-1} \mathbf{A} \mathbf{K} = \mathbf{D},$$

donde \mathbf{D} es una matriz diagonal en la que se recogen los autovalores de \mathbf{A} .

Corolario 11.1. Si \mathbf{A} es una matriz real y simétrica de orden $n \times n$ entonces existen n autovectores linealmente independientes de \mathbf{A} , que son ortogonales.

Corolario 11.2. Los autovalores de una matriz real y simétrica son números reales.

Teorema 11.13. Si \mathbf{A} es una matriz real y simétrica, entonces autovectores correspondientes a autovalores distintos son ortogonales.

Teorema 11.14. Una matriz simétrica \mathbf{A} es definida positiva si, y sólo si, todos sus autovalores son positivos.

Estimaciones del tamaño de los autovalores

Es útil disponer de cotas del tamaño de los autovalores de una matriz \mathbf{A} . Los siguientes resultados proporcionan algunas ideas sobre este tema.

Definición 11.9 (Norma matricial). A cada norma vectorial $\|\mathbf{X}\|$ se le asocia, de manera natural, la norma matricial dada por

$$(32) \quad \|\mathbf{A}\| = \max \{\|\mathbf{AX}\| : \|\mathbf{X}\| = 1\}$$

que, a veces, se llama norma matricial subordinada. Por ejemplo, para la norma subordinada a la norma $\|\mathbf{X}\|_\infty$ se tiene la siguiente fórmula:

$$(33) \quad \|\mathbf{A}\|_\infty = \max \left\{ \sum_{j=1}^n |a_{ij}| : 1 \leq i \leq n \right\}.$$

Teorema 11.15. Si λ es un autovalor cualquiera de \mathbf{A} , entonces

$$(34) \quad |\lambda| \leq \|\mathbf{A}\|,$$

para cualquier norma matricial $\|\mathbf{A}\|$ subordinada a una norma vectorial.

Teorema 11.16 (Teorema de los círculos de Gershgorin). Sea \mathbf{A} una matriz de orden $n \times n$ y sea C_j el círculo del plano complejo de centro a_{jj} y radio

$$(35) \quad r_j = \sum_{k=1, k \neq j}^n |a_{jk}| \quad \text{para cada } j = 1, 2, \dots, n;$$

o sea, C_j está formado por todos los números complejos $z = x + iy$ tales que $|z - a_{jj}| \leq r_j$,

$$(36) \quad C_j = \{z : |z - a_{jj}| \leq r_j\}.$$

Si $S = \bigcup_{i=1}^n C_i$, entonces todos los autovalores de \mathbf{A} están en el conjunto S . Es más, si la unión de k de estos círculos no se intersecta con los $n - k$ restantes, entonces debe contener precisamente k autovalores (contando sus multiplicidades).

Teorema 11.17 (Teorema del radio espectral). Sea \mathbf{A} una matriz real y simétrica. Entonces el radio espectral de \mathbf{A} coincide con su norma $\|\mathbf{A}\|_2$; es decir, se tiene la relación

$$(37) \quad \max\{|\lambda_1|, |\lambda_2|, \dots, |\lambda_n|\} = \|\mathbf{A}\|_2.$$

Una visión de conjunto de los métodos

Para problemas con matrices simétricas de tamaño moderado, se puede usar el método de Jacobi. Para problemas con matrices simétricas grandes (con n del orden de las centenas), es mejor usar el método de Householder para obtener una matriz tridiagonal semejante y, después, usar el método QR . A diferencia de lo que pasa con las matrices simétricas, las matrices reales que no son simétricas pueden tener autovalores y autovectores complejos. Si la matriz tiene un autovalor dominante, entonces el método de las potencias permite calcular dicho autovalor dominante; después podemos usar las técnicas de deflación para ir determinando los siguientes autovalores dominantes. Para matrices reales que no son simétricas, el método de Householder permite calcular una matriz semejante que es de Hessenberg, a la que podemos aplicar el algoritmo LR o el QR .

Ejercicios

- Para cada una de las siguientes matrices determine: (i) su polinomio característico $p(\lambda)$, (ii) sus autovalores y (iii) un autovector para cada autovalor.
- (a) $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix}$
- (b) $\mathbf{A} = \begin{bmatrix} 1 & 6 \\ 9 & 2 \end{bmatrix}$
- (c) $\mathbf{A} = \begin{bmatrix} -2 & 3 \\ 3 & -2 \end{bmatrix}$
- (d) $\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 2 \\ -1 & 3 & 2 \end{bmatrix}$
- (e) $\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 2 & 2 & 3 \\ 0 & 0 & 3 & 2 \\ 0 & 0 & 0 & 4 \end{bmatrix}$
- Determine el radio espectral de cada una de las matrices del Ejercicio 1.
- Determine las normas $\|\mathbf{A}\|_2$ y $\|\mathbf{A}\|_\infty$ de cada una de las matrices del Ejercicio 1.
- Determine cuáles de las matrices del Ejercicio 1 son diagonalizables, si es que hay alguna. Para las que sean diagonalizables, halle las matrices \mathbf{V} y \mathbf{D} del Teorema 11.8 y lleve a cabo el producto que se muestra en la expresión (24).
- (a) Pruebe que para cada número θ fijo se tiene que la matriz

$$\mathbf{R} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$$

- es ortogonal (\mathbf{R} se llama matriz de la rotación de ángulo θ).
- (b) Determine los valores de θ para los que los autovalores de \mathbf{R} son reales.
 - En la Sección 3.2 se introdujeron las rotaciones planas $\mathbf{R}_x(\alpha)$, $\mathbf{R}_y(\beta)$ y $\mathbf{R}_z(\gamma)$.
 - Pruebe que para cualesquiera α , β y γ fijos, las matrices respectivas $\mathbf{R}_x(\alpha)$, $\mathbf{R}_y(\beta)$ y $\mathbf{R}_z(\gamma)$ son ortogonales.
 - Determine los valores de α , β y γ para los que, respectivamente, los autovalores de $\mathbf{R}_x(\alpha)$, $\mathbf{R}_y(\beta)$ y $\mathbf{R}_z(\gamma)$ son reales.

7. Sea $\mathbf{A} = \begin{bmatrix} a+3 & 2 \\ 2 & a \end{bmatrix}$.

(a) Pruebe que el polinomio característico de la matriz \mathbf{A} es

$$p(\lambda) = \lambda^2 - (3 + 2a)\lambda + a^2 - 3a - 4.$$

(b) Pruebe que los autovalores de \mathbf{A} son $\lambda_1 = a + 4$ y $\lambda_2 = a - 1$.

(c) Pruebe que los autovectores de \mathbf{A} son $\mathbf{V}_1 = [2 \ 1]'$ y $\mathbf{V}_2 = [-1 \ 2]'$.

8. Sea (λ, \mathbf{V}) una pareja autovalor-autovector de una matriz \mathbf{A} . Pruebe que si k es un número natural, entonces (λ^k, \mathbf{V}) forman una pareja autovalor-autovector de la matriz \mathbf{A}^k .

9. Supongamos que \mathbf{V} es un autovector de \mathbf{A} correspondiente al autovalor $\lambda = 3$. Pruebe que $\lambda = 9$ es el autovalor de \mathbf{A}^2 correspondiente a \mathbf{V} .

10. Supongamos que \mathbf{V} es un autovector de \mathbf{A} correspondiente al autovalor $\lambda = 2$. Pruebe que $\lambda = \frac{1}{2}$ es el autovalor de \mathbf{A}^{-1} correspondiente a \mathbf{V} .

11. Supongamos que \mathbf{V} es un autovector de \mathbf{A} correspondiente al autovalor $\lambda = 5$. Pruebe que $\lambda = 4$ es el autovalor de $\mathbf{A} - \mathbf{I}$ correspondiente a \mathbf{V} .

12. Sea \mathbf{A} una matriz cuadrada de orden $n \times n$ cuyo polinomio característico $p(\lambda)$ viene dado por

$$\begin{aligned} p(\lambda) &= \det(\mathbf{A} - \lambda\mathbf{I}) \\ &= (-1)^n(\lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \cdots + c_{n-1}\lambda + c_n). \end{aligned}$$

(a) Pruebe que el término constante de $p(\lambda)$ es $c_n = (-1)^n \det(\mathbf{A})$.

(b) Pruebe que el coeficiente de λ^{n-1} es $c_1 = -(a_{11} + a_{22} + \cdots + a_{nn})$.

13. Supongamos que \mathbf{A} es semejante a una matriz diagonal, o sea,

$$\mathbf{K}^{-1}\mathbf{A}\mathbf{K} = \mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n).$$

Pruebe que si k es un número natural, entonces

$$\mathbf{A}^k = \mathbf{K} \text{diag}(\lambda_1^k, \lambda_2^k, \dots, \lambda_n^k) \mathbf{K}^{-1}.$$

2 Los métodos de las potencias

Vamos a describir ahora el método de las potencias para calcular la pareja autovalor-autovector dominante. Veremos luego su extensión, el método de las potencias inversas, que será de utilidad para calcular un autovalor para el que se dispone de una buena aproximación inicial. Esto ocurre, por ejemplo, cuando aplicamos algún otro método que converge rápidamente pero es de precisión limitada; se puede aplicar el método de las potencias inversas para refinar los valores numéricos obtenidos y aumentar la precisión. Para discutir estos procesos, necesitamos dar algunas definiciones.

Definición 11.10. Si λ_1 es un autovalor de \mathbf{A} que, en valor absoluto, es mayor que cualquier otro autovalor, entonces se dice que es un **autovalor dominante** y sus autovectores correspondientes se llaman **autovectores dominantes**. ▲

Definición 11.11. Diremos que un autovector \mathbf{V} está normalizado cuando su coordenada de mayor tamaño es igual a 1. ▲

Es fácil normalizar un autovector $[v_1 \ v_2 \ \dots \ v_n]'$, basta construir el nuevo vector $\mathbf{V} = (1/c)[v_1 \ v_2 \ \dots \ v_n]'$ donde $c = v_j$ y $|v_j| = \max\{|v_i| : 1 \leq i \leq n\}$.

Supongamos que la matriz \mathbf{A} tiene un autovalor dominante λ y que hay un único (salvo el signo) autovector real normalizado \mathbf{V} correspondiente a λ . Este par autovalor-autovector (λ, \mathbf{V}) puede encontrarse mediante el siguiente procedimiento iterativo conocido como el **método de las potencias**. Empezamos con el vector

$$(1) \quad \mathbf{X}_0 = [1 \ 1 \ \dots \ 1]'$$

y generamos recursivamente una sucesión $\{\mathbf{X}_k\}$ de acuerdo con el siguiente esquema:

$$(2) \quad \begin{aligned} \mathbf{Y}_k &= \mathbf{AX}_k, \\ \mathbf{X}_{k+1} &= \frac{1}{c_{k+1}} \mathbf{Y}_k, \end{aligned}$$

donde c_{k+1} es la coordenada de \mathbf{Y}_k de mayor tamaño (en caso de empate, se toma la que aparezca en primer lugar). Las sucesiones $\{\mathbf{X}_k\}$ y $\{c_k\}$ convergerán, respectivamente, a \mathbf{V} y λ :

$$(3) \quad \lim_{k \rightarrow \infty} \mathbf{X}_k = \mathbf{V} \quad \text{y} \quad \lim_{k \rightarrow \infty} c_k = \lambda.$$

Ejemplo 11.5. Vamos a usar el método de las potencias para determinar el autovalor y el autovector dominantes de la matriz

$$\mathbf{A} = \begin{bmatrix} 0 & 11 & -5 \\ -2 & 17 & -7 \\ -4 & 26 & -10 \end{bmatrix}.$$

Empezamos con $\mathbf{X}_0 = [1 \ 1 \ 1]'$ y usamos las fórmulas dadas en (2) para generar la sucesión de vectores $\{\mathbf{X}_k\}$ y la de constantes $\{c_k\}$. El resultado de la primera iteración es

$$\begin{bmatrix} 0 & 11 & -5 \\ -2 & 17 & -7 \\ -4 & 26 & -10 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 8 \\ 12 \end{bmatrix} = 12 \begin{bmatrix} \frac{1}{2} \\ \frac{2}{3} \\ 1 \end{bmatrix} = c_1 \mathbf{X}_1.$$

Tabla 11.1 Método de las potencias usado en el Ejemplo 11.5 para hallar el autovector dominante normalizado $\mathbf{V} = \left[\begin{smallmatrix} \frac{2}{5} & \frac{3}{5} & 1 \end{smallmatrix} \right]'$ y su autovalor $\lambda = 4$.

$\mathbf{AX}_k =$	\mathbf{Y}_k	=	$c_{k+1}\mathbf{X}_{k+1}$
$\mathbf{AX}_0 = [6.000000 \quad 8.000000 \quad 12.000000]'$	$= 12.00000[0.500000 \quad 0.666667 \quad 1]'$	$= c_1\mathbf{X}_1$	
$\mathbf{AX}_1 = [2.333333 \quad 3.333333 \quad 5.333333]'$	$= 5.333333[0.437500 \quad 0.625000 \quad 1]'$	$= c_2\mathbf{X}_2$	
$\mathbf{AX}_2 = [1.875000 \quad 2.750000 \quad 4.500000]'$	$= 4.500000[0.416667 \quad 0.611111 \quad 1]'$	$= c_3\mathbf{X}_3$	
$\mathbf{AX}_3 = [1.722222 \quad 2.555556 \quad 4.222222]'$	$= 4.222222[0.407895 \quad 0.605263 \quad 1]'$	$= c_4\mathbf{X}_4$	
$\mathbf{AX}_4 = [1.657895 \quad 2.473684 \quad 4.105263]'$	$= 4.105263[0.403846 \quad 0.602564 \quad 1]'$	$= c_5\mathbf{X}_5$	
$\mathbf{AX}_5 = [1.628205 \quad 2.435897 \quad 4.051282]'$	$= 4.051282[0.401899 \quad 0.601266 \quad 1]'$	$= c_6\mathbf{X}_6$	
$\mathbf{AX}_6 = [1.613924 \quad 2.417722 \quad 4.025316]'$	$= 4.025316[0.400943 \quad 0.600629 \quad 1]'$	$= c_7\mathbf{X}_7$	
$\mathbf{AX}_7 = [1.606918 \quad 2.408805 \quad 4.012579]'$	$= 4.012579[0.400470 \quad 0.600313 \quad 1]'$	$= c_8\mathbf{X}_8$	
$\mathbf{AX}_8 = [1.603448 \quad 2.404389 \quad 4.006270]'$	$= 4.006270[0.400235 \quad 0.600156 \quad 1]'$	$= c_9\mathbf{X}_9$	
$\mathbf{AX}_9 = [1.601721 \quad 2.402191 \quad 4.003130]'$	$= 4.003130[0.400117 \quad 0.600078 \quad 1]'$	$= c_{10}\mathbf{X}_{10}$	
$\mathbf{AX}_{10} = [1.600860 \quad 2.401095 \quad 4.001564]'$	$= 4.001564[0.400059 \quad 0.600039 \quad 1]'$	$= c_{11}\mathbf{X}_{11}$	

El resultado de la segunda iteración es

$$\begin{bmatrix} 0 & 11 & -5 \\ -2 & 17 & -7 \\ -4 & 26 & -10 \end{bmatrix} \begin{bmatrix} \frac{1}{2} \\ \frac{2}{3} \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{7}{3} \\ \frac{10}{3} \\ \frac{16}{3} \end{bmatrix} = \frac{16}{3} \begin{bmatrix} \frac{7}{16} \\ \frac{5}{8} \\ 1 \end{bmatrix} = c_2\mathbf{X}_2.$$

El proceso iterativo produce la sucesión $\{\mathbf{X}_k\}$ (donde \mathbf{X}_k es un vector normalizado):

$$12 \begin{bmatrix} \frac{1}{2} \\ \frac{2}{3} \\ \frac{3}{1} \end{bmatrix}, \frac{16}{3} \begin{bmatrix} \frac{7}{16} \\ \frac{5}{8} \\ 1 \end{bmatrix}, \frac{9}{2} \begin{bmatrix} \frac{5}{12} \\ \frac{11}{18} \\ 1 \end{bmatrix}, \frac{38}{9} \begin{bmatrix} \frac{31}{76} \\ \frac{23}{38} \\ 1 \end{bmatrix}, \frac{78}{19} \begin{bmatrix} \frac{21}{52} \\ \frac{47}{78} \\ 1 \end{bmatrix}, \frac{158}{39} \begin{bmatrix} \frac{127}{316} \\ \frac{95}{158} \\ 1 \end{bmatrix}, \dots$$

La sucesión de vectores converge a $\mathbf{V} = \left[\begin{smallmatrix} \frac{2}{5} & \frac{3}{5} & 1 \end{smallmatrix} \right]'$, y la sucesión de escalares converge a $\lambda = 4$ (véase la Tabla 11.1). Puede probarse que la velocidad de convergencia es lineal. ■

Teorema 11.18 (Método de las potencias). Supongamos que una matriz \mathbf{A} de orden $n \times n$ tiene n autovalores distintos $\lambda_1, \lambda_2, \dots, \lambda_n$ ordenados en tamaño decreciente

$$(4) \quad |\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_n|.$$

Si \mathbf{X}_0 se escoge adecuadamente (veremos luego qué significa esto), entonces las sucesiones $\{\mathbf{X}_k = [x_1^{(k)} \ x_2^{(k)} \ \dots \ x_n^{(k)}]'\}$ y $\{c_k\}$ generadas recursivamente con el esquema iterativo

$$(5) \quad \mathbf{Y}_k = \mathbf{AX}_k$$

y

$$(6) \quad \mathbf{X}_{k+1} = \frac{1}{c_{k+1}} \mathbf{Y}_k,$$

donde

$$(7) \quad c_{k+1} = x_j^{(k)} \quad \text{y} \quad x_j^{(k)} = \max\{|x_i^{(k)}| : 1 \leq i \leq n\},$$

convergen, respectivamente, al autovector dominante \mathbf{V}_1 y al autovalor dominante λ_1 . Es decir,

$$(8) \quad \lim_{k \rightarrow \infty} \mathbf{X}_k = \mathbf{V}_1 \quad \text{y} \quad \lim_{k \rightarrow \infty} c_k = \lambda_1.$$

Demostración. Puesto que \mathbf{A} tiene n autovalores distintos, existen n autovectores normalizados correspondientes \mathbf{V}_j , para $j = 1, 2, \dots, n$, que son linealmente independientes y forman una base de \mathbb{R}^n . Por tanto, el vector inicial \mathbf{X}_0 puede expresarse como una combinación lineal

$$(9) \quad \mathbf{X}_0 = b_1 \mathbf{V}_1 + b_2 \mathbf{V}_2 + \cdots + b_n \mathbf{V}_n.$$

Supongamos que elegimos $\mathbf{X}_0 = [x_1 \ x_2 \ \dots \ x_n]'$ de manera que $b_1 \neq 0$ (esto es lo que significa “adecuadamente” en el enunciado). Supongamos también que \mathbf{X}_0 está normalizado; o sea, $\max\{|x_j| : 1 \leq j \leq n\} = 1$. Como $\{\mathbf{V}_j\}_{j=1}^n$ son autovectores de \mathbf{A} , el producto $\mathbf{A}\mathbf{X}_0$ y la normalización que se realiza a continuación producen

$$(10) \quad \begin{aligned} \mathbf{Y}_0 &= \mathbf{A}\mathbf{X}_0 = \mathbf{A}(b_1 \mathbf{V}_1 + b_2 \mathbf{V}_2 + \cdots + b_n \mathbf{V}_n) \\ &= b_1 \mathbf{A}\mathbf{V}_1 + b_2 \mathbf{A}\mathbf{V}_2 + \cdots + b_n \mathbf{A}\mathbf{V}_n \\ &= b_1 \lambda_1 \mathbf{V}_1 + b_2 \lambda_2 \mathbf{V}_2 + \cdots + b_n \lambda_n \mathbf{V}_n \\ &= \lambda_1 \left(b_1 \mathbf{V}_1 + b_2 \left(\frac{\lambda_2}{\lambda_1} \right) \mathbf{V}_2 + \cdots + b_n \left(\frac{\lambda_n}{\lambda_1} \right) \mathbf{V}_n \right) \end{aligned}$$

y

$$\mathbf{X}_1 = \frac{\lambda_1}{c_1} \left(b_1 \mathbf{V}_1 + b_2 \left(\frac{\lambda_2}{\lambda_1} \right) \mathbf{V}_2 + \cdots + b_n \left(\frac{\lambda_n}{\lambda_1} \right) \mathbf{V}_n \right).$$

Después de k iteraciones obtenemos

$$\begin{aligned}
 (11) \quad & \mathbf{Y}_{k-1} = \mathbf{A}\mathbf{X}_{k-1} \\
 &= \mathbf{A} \frac{\lambda_1^{k-1}}{c_1 c_2 \cdots c_{k-1}} \left(b_1 \mathbf{V}_1 + b_2 \left(\frac{\lambda_2}{\lambda_1} \right)^{k-1} \mathbf{V}_2 + \cdots + b_n \left(\frac{\lambda_n}{\lambda_1} \right)^{k-1} \mathbf{V}_n \right) \\
 &= \frac{\lambda_1^{k-1}}{c_1 c_2 \cdots c_{k-1}} \left(b_1 \mathbf{A}\mathbf{V}_1 + b_2 \left(\frac{\lambda_2}{\lambda_1} \right)^{k-1} \mathbf{A}\mathbf{V}_2 + \cdots + b_n \left(\frac{\lambda_n}{\lambda_1} \right)^{k-1} \mathbf{A}\mathbf{V}_n \right) \\
 &= \frac{\lambda_1^{k-1}}{c_1 c_2 \cdots c_{k-1}} \left(b_1 \lambda_1 \mathbf{V}_1 + b_2 \left(\frac{\lambda_2}{\lambda_1} \right)^{k-1} \lambda_2 \mathbf{V}_2 + \cdots + b_n \left(\frac{\lambda_n}{\lambda_1} \right)^{k-1} \lambda_n \mathbf{V}_n \right) \\
 &= \frac{\lambda_1^k}{c_1 c_2 \cdots c_{k-1}} \left(b_1 \mathbf{V}_1 + b_2 \left(\frac{\lambda_2}{\lambda_1} \right)^k \mathbf{V}_2 + \cdots + b_n \left(\frac{\lambda_n}{\lambda_1} \right)^k \mathbf{V}_n \right)
 \end{aligned}$$

y

$$\mathbf{X}_k = \frac{\lambda_1^k}{c_1 c_2 \cdots c_k} \left(b_1 \mathbf{V}_1 + b_2 \left(\frac{\lambda_2}{\lambda_1} \right)^{k-1} \mathbf{V}_2 + \cdots + b_n \left(\frac{\lambda_n}{\lambda_1} \right)^{k-1} \mathbf{V}_n \right).$$

Como hemos supuesto que $|\lambda_j|/|\lambda_1| < 1$ para cada $j = 2, 3, \dots, n$, tenemos

$$(12) \quad \lim_{k \rightarrow \infty} b_j \left(\frac{\lambda_j}{\lambda_1} \right)^k \mathbf{V}_j = \mathbf{0} \quad \text{para cada } j = 2, 3, \dots, n.$$

Por tanto, deducimos que

$$(13) \quad \lim_{k \rightarrow \infty} \mathbf{X}_k = \lim_{k \rightarrow \infty} \frac{b_1 \lambda_1^k}{c_1 c_2 \cdots c_k} \mathbf{V}_1.$$

Como tenemos que tanto \mathbf{X}_k como \mathbf{V}_1 son vectores normalizados de manera que sus componentes de mayor tamaño deben ser ambas iguales a 1, esto implica que la sucesión de escalares que multiplica a \mathbf{V}_1 en el miembro derecho de la relación (13) debe converger a 1; es decir,

$$(14) \quad \lim_{k \rightarrow \infty} \frac{b_1 \lambda_1^k}{c_1 c_2 \cdots c_k} = 1.$$

En consecuencia, la sucesión de vectores $\{\mathbf{X}_k\}$ converge al autovector dominante:

$$(15) \quad \lim_{k \rightarrow \infty} \mathbf{X}_k = \mathbf{V}_1.$$

Reemplazando k por $k-1$ en los términos de la sucesión dada en (14) obtenemos

$$\lim_{k \rightarrow \infty} \frac{b_1 \lambda_1^{k-1}}{c_1 c_2 \cdots c_{k-1}} = 1,$$

así que, dividiendo la relación (14) entre la que acabamos de obtener, nos queda

$$\lim_{k \rightarrow \infty} \frac{\lambda_1}{c_k} = \lim_{k \rightarrow \infty} \frac{b_1 \lambda_1^k / (c_1 c_2 \cdots c_k)}{b_1 \lambda_1^{k-1} / (c_1 c_2 \cdots c_{k-1})} = \frac{1}{1} = 1.$$

Por consiguiente, la sucesión de constantes $\{c_k\}$ converge al autovalor dominante:

$$(16) \quad \lim_{k \rightarrow \infty} c_k = \lambda_1,$$

lo que completa la demostración del teorema. •

Velocidad de convergencia

A la luz de la relación (12), observamos que el coeficiente de \mathbf{V}_j en \mathbf{X}_k tiende a cero de manera proporcional a $(\lambda_j/\lambda_1)^k$, así que la velocidad de convergencia de $\{\mathbf{X}_k\}$ a \mathbf{V}_1 depende de los términos $(\lambda_2/\lambda_1)^k$ y, por tanto, el orden de convergencia es lineal. De manera similar, puede observarse que la convergencia de la sucesión de constantes $\{c_k\}$ a λ_1 también es lineal. Sabemos que el método Δ^2 de Aitken puede usarse con cualquier sucesión linealmente convergente $\{p_k\}$ para formar una nueva sucesión

$$\left\{ \hat{p}_k = \frac{(p_{k+1} - p_k)^2}{p_{k+2} - 2p_{k+1} + p_k} \right\},$$

que converge más rápidamente. Aplicando este método en el Ejemplo 11.5 logramos acelerar la convergencia de las constantes $\{c_k\}$ y también la convergencia de las dos primeras componentes de $\{\mathbf{X}_k\}$. En la Tabla 11.2 se muestran los resultados de este proceso de aceleración y se comparan con los obtenidos previamente.

Método de las potencias inversas con traslación

Vamos a describir ahora el método de las potencias inversas con traslación. Para aplicar este método es necesario disponer de una buena aproximación inicial a un autovalor y lo que se consigue mediante las iteraciones del método de las potencias inversas con traslación es obtener una solución más precisa. La aproximación inicial puede obtenerse mediante otros procedimientos que no veremos aquí, como el método QM o el método de Givens. Hay que tener en cuenta, sin embargo, que cuando el autovalor es complejo, cuando es múltiple o cuando hay una pareja de autovalores del mismo tamaño, aparecen dificultades computacionales que deben abordarse usando métodos más avanzados; los ejemplos que presentaremos ilustrarán el caso en que los autovalores son todos distintos. El método de las potencias inversas con traslación se basa en los siguientes tres resultados (cuyas demostraciones dejamos como ejercicios).

Tabla 11.2 Comparación de las velocidades de convergencia del método de las potencias y de la técnica de aceleración Δ^2 de Aitken.

	$c_k \mathbf{Y}_k$	$\hat{c}_k \widehat{\mathbf{X}}_k$
$c_1 \mathbf{X}_1$	$= 12.000000[0.5000000, 0.6666667, 1]'$; $4.3809524[0.4062500, 0.6041667, 1]'$ = $\hat{c}_1 \widehat{\mathbf{X}}_1$	
$c_2 \mathbf{X}_2$	$= 5.3333333[0.4375000, 0.6250000, 1]'$; $4.0833333[0.4015152, 0.6010101, 1]'$ = $\hat{c}_2 \widehat{\mathbf{X}}_2$	
$c_3 \mathbf{X}_3$	$= 4.5000000[0.4166667, 0.6111111, 1]'$; $4.0202020[0.4003759, 0.6002506, 1]'$ = $\hat{c}_3 \widehat{\mathbf{X}}_3$	
$c_4 \mathbf{X}_4$	$= 4.2222222[0.4078947, 0.6052632, 1]'$; $4.0050125[0.4000938, 0.6000625, 1]'$ = $\hat{c}_4 \widehat{\mathbf{X}}_4$	
$c_5 \mathbf{X}_5$	$= 4.1052632[0.4038462, 0.6025641, 1]'$; $4.0012508[0.4000234, 0.6000156, 1]'$ = $\hat{c}_5 \widehat{\mathbf{X}}_5$	
$c_6 \mathbf{X}_6$	$= 4.0512821[0.4018987, 0.6012658, 1]'$; $4.0003125[0.4000059, 0.6000039, 1]'$ = $\hat{c}_6 \widehat{\mathbf{X}}_6$	
$c_7 \mathbf{X}_7$	$= 4.0253165[0.4009434, 0.6006289, 1]'$; $4.0000781[0.4000015, 0.6000010, 1]'$ = $\hat{c}_7 \widehat{\mathbf{X}}_7$	
$c_8 \mathbf{X}_8$	$= 4.0125786[0.4004702, 0.6003135, 1]'$; $4.0000195[0.400004, 0.600002, 1]'$ = $\hat{c}_8 \widehat{\mathbf{X}}_8$	
$c_9 \mathbf{X}_9$	$= 4.0062696[0.4002347, 0.6001565, 1]'$; $4.0000049[0.400001, 0.600001, 1]'$ = $\hat{c}_9 \widehat{\mathbf{X}}_9$	
$c_{10} \mathbf{X}_{10}$	$= 4.0031299[0.4001173, 0.6000782, 1]'$; $4.0000012[0.400000, 0.600000, 1]'$ = $\hat{c}_{10} \widehat{\mathbf{X}}_{10}$	

Teorema 11.19 (Autovalores trasladados). Supongamos que (λ, \mathbf{X}) es una pareja autovalor-autovector de una matriz \mathbf{A} . Si α es una constante cualquiera, entonces $(\lambda - \alpha, \mathbf{V})$ es una pareja autovalor-autovector de la matriz $\mathbf{A} - \alpha\mathbf{I}$.

Teorema 11.20 (Autovalores inversos). Supongamos que (λ, \mathbf{X}) es una pareja autovalor-autovector de una matriz invertible \mathbf{A} . Entonces $(1/\lambda, \mathbf{V})$ es una pareja autovalor-autovector de la matriz \mathbf{A}^{-1} .

Teorema 11.21. Supongamos que (λ, \mathbf{X}) es una pareja autovalor-autovector de una matriz \mathbf{A} . Si α no es un autovalor de \mathbf{A} , entonces $(1/(\lambda - \alpha), \mathbf{V})$ es una pareja autovalor-autovector de la matriz $(\mathbf{A} - \alpha\mathbf{I})^{-1}$.

Teorema 11.22 (Método de las potencias inversas con traslación). Supongamos que una matriz \mathbf{A} de orden $n \times n$ posee n autovalores distintos $\lambda_1, \lambda_2, \dots, \lambda_n$ y fijemos uno de estos autovalores λ_j . Entonces existe una constante α tal que $\mu_1 = 1/(\lambda_j - \alpha)$ es el autovalor dominante de la matriz $(\mathbf{A} - \alpha\mathbf{I})^{-1}$. Es más, si \mathbf{X}_0 se escoge adecuadamente, entonces las sucesiones $\{\mathbf{X}_k = [x_1^{(k)} \ x_2^{(k)} \ \dots \ x_n^{(k)}]'\}$ y $\{c_k\}$ generadas recursivamente por las fórmulas

$$(17) \quad \mathbf{Y}_k = (\mathbf{A} - \alpha\mathbf{I})^{-1} \mathbf{X}_k$$

y

$$(18) \quad \mathbf{X}_{k+1} = \frac{1}{c_{k+1}} \mathbf{Y}_k,$$

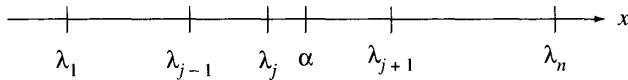


Figura 11.2 Localización de α en el método de las potencias inversas con traslación.

donde

$$(19) \quad c_{k+1} = x_j^{(k)} \quad \text{y} \quad x_j^{(k)} = \max\{|x_i^{(k)}| : 1 \leq i \leq n\}$$

convergen a la pareja autovalor-autovector dominante (μ_1, \mathbf{V}_j) de la matriz $(\mathbf{A} - \alpha \mathbf{I})^{-1}$. Finalmente, el autovalor de \mathbf{A} que hemos fijado viene dado por

$$(20) \quad \lambda_j = \frac{1}{\mu_1} + \alpha.$$

Observación. Cuando se lleva el Teorema 11.22 a la práctica, lo que se hace es emplear un método de resolución de sistemas lineales para calcular el vector \mathbf{Y}_k de cada paso como la solución del sistema $(\mathbf{A} - \alpha \mathbf{I})\mathbf{Y}_k = \mathbf{X}_k$.

Demostración. Supongamos, sin perder generalidad, que $\lambda_1 < \lambda_2 < \dots < \lambda_n$. Tomemos una constante α ($\alpha \neq \lambda_j$) que esté más cerca de λ_j que de cualquiera de los otros autovalores (véase la Figura 11.2); o sea,

$$(21) \quad |\lambda_j - \alpha| < |\lambda_i - \alpha| \quad \text{para cada } i = 1, 2, \dots, j-1, j+1, \dots, n.$$

De acuerdo con el Teorema 11.21, $(1/(\lambda_j - \alpha), \mathbf{V})$ es un par autovalor-autovector de la matriz $(\mathbf{A} - \alpha \mathbf{I})^{-1}$. Las relaciones dadas en (21) implican que $1/|\lambda_i - \alpha| < 1/|\lambda_j - \alpha|$ para cada $i \neq j$ de manera que $\mu_1 = 1/(\lambda_j - \alpha)$ es el autovalor dominante de la matriz $(\mathbf{A} - \alpha \mathbf{I})^{-1}$. El método de las potencias inversas con traslación consiste, entonces, en la aplicación del método de las potencias para determinar la pareja (μ_1, \mathbf{V}_j) . Finalmente, la relación $\lambda_j = 1/\mu_1 + \alpha$ proporciona el autovalor de \mathbf{A} que queríamos calcular. •

Ejemplo 11.6. Vamos a ver cómo funciona el método de las potencias inversas con traslación para calcular las parejas autovalor-autovector de la matriz

$$\mathbf{A} = \begin{bmatrix} 0 & 11 & -5 \\ -2 & 17 & -7 \\ -4 & 26 & -10 \end{bmatrix}.$$

Para ello usaremos que los autovalores de esta matriz \mathbf{A} son $\lambda_1 = 4$, $\lambda_2 = 2$ y $\lambda_3 = 1$, y elegiremos un valor de α y un vector de partida adecuados en cada caso.

Tabla 11.3 Método de las potencias inversas con traslación para la matriz $(\mathbf{A} - 4.2\mathbf{I})^{-1}$ del Ejemplo 11.6: Convergencia hacia el autovector $\mathbf{V} = \left[\frac{2}{5} \quad \frac{3}{5} \quad 1\right]'$ y hacia $\mu_1 = -5$.

$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_k =$	$c_{k+1}\mathbf{X}_{k+1}$
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_0 = -23.18181818$	$[0.4117647059 \quad 0.6078431373 \quad 1]'$ = $c_1\mathbf{X}_1$
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_1 = -5.356506239$	$[0.4009983361 \quad 0.6006655574 \quad 1]'$ = $c_2\mathbf{X}_2$
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_2 = -5.030252609$	$[0.4000902120 \quad 0.6000601413 \quad 1]'$ = $c_3\mathbf{X}_3$
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_3 = -5.002733697$	$[0.4000081966 \quad 0.6000054644 \quad 1]'$ = $c_4\mathbf{X}_4$
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_4 = -5.000248382$	$[0.4000007451 \quad 0.6000004967 \quad 1]'$ = $c_5\mathbf{X}_5$
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_5 = -5.000022579$	$[0.4000000677 \quad 0.6000000452 \quad 1]'$ = $c_6\mathbf{X}_6$
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_6 = -5.000002053$	$[0.4000000062 \quad 0.6000000041 \quad 1]'$ = $c_7\mathbf{X}_7$
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_7 = -5.000000187$	$[0.4000000006 \quad 0.6000000004 \quad 1]'$ = $c_8\mathbf{X}_8$
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_8 = -5.000000017$	$[0.4000000001 \quad 0.6000000000 \quad 1]'$ = $c_9\mathbf{X}_9$

Caso (i): Para el autovalor $\lambda_1 = 4$, elegimos $\alpha = 4.2$ y el vector inicial $\mathbf{X}_0 = [1 \quad 1 \quad 1]'$. En primer lugar calculamos la matriz $\mathbf{A} - 4.2\mathbf{I}$ y la solución del sistema

$$\begin{bmatrix} -4.2 & 11 & -5 \\ -2 & 12.8 & -7 \\ -4 & 26 & -14.2 \end{bmatrix} \mathbf{Y}_0 = \mathbf{X}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

obteniendo el vector $\mathbf{Y}_0 = [-9.545454545 \quad -14.09090909 \quad -23.18181818]'$. Entonces, $c_1 = -23.18181818$ y $\mathbf{X}_1 = [0.4117647059 \quad 0.6078431373 \quad 1]'$. El método iterativo genera los valores y vectores que se muestran en la Tabla 11.3: La sucesión $\{c_k\}$ converge a $\mu_1 = -5$, que es el autovalor dominante de $(\mathbf{A} - 4.2\mathbf{I})^{-1}$, y la sucesión $\{\mathbf{X}_k\}$ converge a $\mathbf{V}_1 = \left[\frac{2}{5} \quad \frac{3}{5} \quad 1\right]'$. El autovalor λ_1 de \mathbf{A} viene dado por $\lambda_1 = 1/\mu_1 + \alpha = 1/(-5) + 4.2 = -0.2 + 4.2 = 4$.

Caso (ii): Para el autovalor $\lambda_2 = 2$, elegimos $\alpha = 2.1$ y el vector inicial $\mathbf{X}_0 = [1 \quad 1 \quad 1]'$. En primer lugar calculamos la matriz $\mathbf{A} - 2.1\mathbf{I}$ y la solución del sistema

$$\begin{bmatrix} -2.1 & 11 & -5 \\ -2 & 14.9 & -7 \\ -4 & 26 & -12.1 \end{bmatrix} \mathbf{Y}_0 = \mathbf{X}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

obteniendo el vector $\mathbf{Y}_0 = [11.05263158 \quad 21.57894737 \quad 42.63157895]'$. Entonces, $c_1 = 42.63157895$ y $\mathbf{X}_1 = [0.2592592593 \quad 0.5061728395 \quad 1]'$. El método iterativo genera los valores y vectores que se muestran en la Tabla 11.4: El autovalor dominante de $(\mathbf{A} - 2.1\mathbf{I})^{-1}$ es $\mu_1 = -10$ y la segunda pareja autovalor-autovector de la matriz \mathbf{A} viene dada por $\lambda_2 = 1/(-10) + 2.1 = -0.1 + 2.1 = 2$ y $\mathbf{V}_2 = [\frac{1}{4} \quad \frac{1}{2} \quad 1]'$.

Caso (iii): Para el autovalor $\lambda_3 = 1$, tomamos $\alpha = 0.875$ y el vector inicial $\mathbf{X}_0 = [0 \quad 1 \quad 1]'$. Los valores que se obtienen en la iteración se muestran en la

Tabla 11.4 Método de las potencias inversas con traslación para la matriz $(\mathbf{A} - 2.1\mathbf{I})^{-1}$ del Ejemplo 11.6: Convergencia hacia el autovector $\mathbf{V} = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & 1 \end{bmatrix}'$ y hacia $\mu_1 = -10$.

$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_k =$	$c_{k+1}\mathbf{X}_{k+1}$
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_0 = 42.63157895$	$[0.2592592593 \quad 0.5061728395 \quad 1]'$
$= c_1\mathbf{X}_1$	
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_1 = -9.350227420$	$[0.2494788047 \quad 0.4996525365 \quad 1]'$
$= c_2\mathbf{X}_2$	
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_2 = -10.03657511$	$[0.2500273314 \quad 0.5000182209 \quad 1]'$
$= c_3\mathbf{X}_3$	
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_3 = -9.998082009$	$[0.2499985612 \quad 0.4999990408 \quad 1]'$
$= c_4\mathbf{X}_4$	
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_4 = -10.00010097$	$[0.2500000757 \quad 0.5000000505 \quad 1]'$
$= c_5\mathbf{X}_5$	
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_5 = -9.999994686$	$[0.2499999960 \quad 0.4999999973 \quad 1]'$
$= c_6\mathbf{X}_6$	
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_6 = -10.00000028$	$[0.2500000002 \quad 0.5000000001 \quad 1]'$
$= c_7\mathbf{X}_7$	

Tabla 11.5 Método de las potencias inversas con traslación para la matriz $(\mathbf{A} - 0.875\mathbf{I})^{-1}$ del Ejemplo 11.6: Convergencia hacia el autovector $\mathbf{V} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}'$ y hacia $\mu_1 = 8$.

$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_k =$	$c_{k+1}\mathbf{X}_{k+1}$
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_0 = -30.40000000$	$[0.5052631579 \quad 0.4947368421 \quad 1]'$
$= c_1\mathbf{X}_1$	
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_1 = 8.404210526$	$[0.5002004008 \quad 0.4997995992 \quad 1]'$
$= c_2\mathbf{X}_2$	
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_2 = 8.015390782$	$[0.5000080006 \quad 0.4999919994 \quad 1]'$
$= c_3\mathbf{X}_3$	
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_3 = 8.000614449$	$[0.5000003200 \quad 0.4999996800 \quad 1]'$
$= c_4\mathbf{X}_4$	
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_4 = 8.000024576$	$[0.5000000128 \quad 0.4999999872 \quad 1]'$
$= c_5\mathbf{X}_5$	
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_5 = 8.000000983$	$[0.5000000005 \quad 0.4999999995 \quad 1]'$
$= c_6\mathbf{X}_6$	
$(\mathbf{A} - \alpha\mathbf{I})^{-1}\mathbf{X}_6 = 8.000000039$	$[0.5000000000 \quad 0.5000000000 \quad 1]'$
$= c_7\mathbf{X}_7$	

Tabla 11.5. El autovalor dominante de $(\mathbf{A} - 0.875\mathbf{I})^{-1}$ es $\mu_1 = 8$ y la tercera pareja autovalor-autovector de \mathbf{A} es $\lambda_3 = 1/8 + 0.875 = 0.125 + 0.875 = 1$ y $\mathbf{V}_3 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}'$. La sucesión de vectores $\{\mathbf{X}_k\}$ que se obtiene a partir del vector inicial $[0 \ 1 \ 1]'$ resulta ser convergente en siete pasos. (Sin embargo, partiendo del vector $\mathbf{X}_0 = [1 \ 1 \ 1]'$, como hemos hecho en los casos anteriores, aparecen dificultades computacionales y la convergencia es mucho más lenta. La razón es que la componente de $[1 \ 1 \ 1]'$ en la dirección de \mathbf{V}_3 es cero, por lo que este vector inicial no es “adecuado” en el sentido del Teorema 11.18. Sin embargo, a largo plazo, los errores de redondeo acaban produciendo en la iteración un vector que sí tiene una componente no nula, pero muy pequeña, en la dirección de \mathbf{V}_3 así que, finalmente, hay convergencia a \mathbf{V}_3 pero ésta resulta ser muy lenta debido a la pequeñez de dicha componente.) ■

MATLAB

Programa 11.1 (Método de las potencias). Cálculo del autovalor dominante λ_1 y del autovector asociado V_1 de una matriz A de orden $n \times n$. Se supone que los n autovalores verifican la condición de dominación: $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_n| > 0$.

```
function [lambda,V]=power1(A,X,epsilon,max1)

% Datos
% - A es una matriz de orden n x n
% - X es el vector inicial de orden n x 1
% - epsilon es la tolerancia
% - max1 es el número máximo de iteraciones
% Resultados
% - lambda es el autovalor dominante
% - V es el autovector dominante

% Inicialización de los parámetros
lambda=0;
cnt=0;
err=1;
state=1;

while ((cnt<=max1)&(state==1))
    Y=A*X;
    % Normalización de Y
    [m j]=max(abs(Y));
    c1=m;
    dc=abs(lambda-c1);
    Y=(1/c1)*Y;
    % Actualización de X y de lambda y criterio de convergencia
    dv=norm(X-Y);
    err=max(dc,dv);
    X=Y;
    lambda=c1;
    state=0;
    if(err>epsilon)
        state=1;
    end
    cnt=cnt+1;
end
V=X;
```

Programa 11.2 (Método de las potencias inversas con traslación). Cálculo del autovector λ_j de una matriz A de orden $n \times n$ que está más cerca de una constante α y de su autovector asociado V_j . Se supone que los n autovalores verifican $\lambda_1 < \lambda_2 < \dots < \lambda_n$ y que α es un número real tal que $|\lambda_j - \alpha| < |\lambda_i - \alpha|$, para cada $i = 1, 2, \dots, j-1, j+1, \dots, n$.

```

function [lambda,V]=invpow(A,X,alpha,epsilon,max1)

% Datos
%      - A es una matriz de orden n x n
%      - X es el vector inicial de orden n x 1
%      - alfa es la traslación
%      - epsilon es la tolerancia
%      - max1 es el número máximo de iteraciones
% Resultados
%      - lambda es el autovalor que está más cerca de alfa
%      - V es el autovector asociado a lambda

% Valores iniciales de la matriz A-alfa I y de los parámetros
[n n]=size(A);
A=A-alfa*eye(n);
lambda=0;
cnt=0;
err=1;
state=1;
[L,U,P]=lu(A);
while ((cnt<=max1)&(state==1))
    % Resolución del sistema AY=X
    Y=L\ (P*X);
    Y=U\ Y;
    % Normalización de Y
    [m j]=max(abs(Y));
    c1=m;
    dc=abs(lambda-c1);
    Y=(1/c1)*Y;
    % Actualización de X y de lambda y criterio de convergencia
    dv=norm(X-Y);
    err=max(dc,dv);
    X=Y;
    lambda=c1;
    state=0;
    if (err>epsilon)
        state=1;
    end
    cnt=cnt+1;

```

```

end
lambda=alfa+1/c1;
V=X;

```

Ejercicios

- Sea (λ, \mathbf{V}) una pareja autovalor-autovector de una matriz \mathbf{A} . Pruebe que si α es una constante cualquiera, entonces $(\lambda - \alpha, \mathbf{V})$ es una pareja autovalor-autovector de la matriz $\mathbf{A} - \alpha\mathbf{I}$.
- Sea (λ, \mathbf{V}) una pareja autovalor-autovector de una matriz \mathbf{A} . Pruebe que si \mathbf{A} es invertible, entonces $(1/\lambda, \mathbf{V})$ es una pareja autovalor-autovector de la matriz \mathbf{A}^{-1} .
- Sea (λ, \mathbf{V}) una pareja autovalor-autovector de una matriz \mathbf{A} . Pruebe que si α no es un autovalor de \mathbf{A} , entonces $(1/(\lambda - \alpha), \mathbf{V})$ es una pareja autovalor-autovector de la matriz $(\mathbf{A} - \alpha\mathbf{I})^{-1}$.
- Técnicas de deflación. Supongamos que $\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_n$ son los autovalores de una matriz \mathbf{A} con autovectores asociados respectivos $\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3, \dots, \mathbf{V}_n$ y que λ_1 tiene multiplicidad 1. Pruebe que si \mathbf{X} es cualquier vector tal que $\mathbf{X}'\mathbf{V}_1 = 1$, entonces la matriz

$$\mathbf{B} = \mathbf{A} - \lambda_1 \mathbf{V}_1 \mathbf{X}'$$

tiene como autovalores $0, \lambda_2, \lambda_3, \dots, \lambda_n$ con autovectores asociados respectivos $\mathbf{V}_1, \mathbf{W}_2, \mathbf{W}_3, \dots, \mathbf{W}_n$, donde \mathbf{V}_j y \mathbf{W}_j están relacionados por la igualdad

$$\mathbf{V}_j = (\lambda - \lambda_1)\mathbf{W}_j + \lambda_1(\mathbf{X}'\mathbf{W}_j)\mathbf{V}_1 \quad \text{para cada } j = 2, 3, \dots, n.$$

- Cadenas de Markov y autovalores. Una cadena de Markov puede describirse mediante una matriz cuadrada \mathbf{A} cuyos términos son todos positivos y tales que la suma de los términos de cada columna es igual a 1. Por ejemplo, sea $\mathbf{P}_0 = [x^{(0)} \ y^{(0)}]'$ un vector que representa el número de personas de una ciudad que consumen, respectivamente, dos marcas X e Y de un cierto producto. Cada mes la gente decide si sigue comprando la misma marca o si cambia, de manera que la probabilidad de que alguien que compra X cambie a Y es 0.3 y, por otro lado, la probabilidad de que alguien que compra Y cambie a X es 0.2. La ecuación de transición para esta cadena mensual es

$$\mathbf{P}_{k+1} = \mathbf{AP}_k = \begin{bmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{bmatrix} \begin{bmatrix} x^{(k)} \\ y^{(k)} \end{bmatrix}.$$

Si se verifica que $\mathbf{AV} = \mathbf{V}$, entonces se dice que \mathbf{V} es una distribución o un estado estable de la cadena de Markov. Si existe un estado estable \mathbf{V} ,

entonces $\lambda = 1$ debe ser un autovalor de \mathbf{A} y \mathbf{V} es un autovector asociado, o sea, $(\mathbf{A} - \mathbf{I})\mathbf{V} = \mathbf{0}$.

- Compruebe que para el ejemplo anterior se tiene que $\lambda = 1$ es un autovalor de la matriz de transición \mathbf{A} .
- Compruebe que el conjunto de autovectores asociados al autovalor $\lambda = 1$ es $\{t[3/2 \ 1]': t \in \mathbb{R}, t \neq 0\}$.
- Suponiendo que la población de la ciudad es de 50 000 habitantes, use el resultado del apartado (b) para comprobar que la distribución estable correspondiente es $[30\ 000 \ 20\ 000]'$.

Algoritmos y programas

En los Problemas 1–4 utilice:

- El Programa 11.1 para hallar el par autovalor-autovector dominante de la matriz dada.
- El Programa 11.2 para hallar los demás pares autovalor-autovector.

$$1. \mathbf{A} = \begin{bmatrix} 7 & 6 & -3 \\ -12 & -20 & 24 \\ -6 & -12 & 16 \end{bmatrix}$$

$$2. \mathbf{A} = \begin{bmatrix} -14 & -30 & 42 \\ 24 & 49 & -66 \\ 12 & 24 & -32 \end{bmatrix}$$

$$3. \mathbf{A} = \begin{bmatrix} 2.5 & -2.5 & 3.0 & 0.5 \\ 0.0 & 5.0 & -2.0 & 2.0 \\ -0.5 & -0.5 & 4.0 & 2.5 \\ -2.5 & -2.5 & 5.0 & 3.5 \end{bmatrix}$$

$$4. \mathbf{A} = \begin{bmatrix} 2.5 & -2.0 & 2.5 & 0.5 \\ 0.5 & 5.0 & -2.5 & -0.5 \\ -1.5 & 1.0 & 3.5 & -2.5 \\ 2.0 & 3.0 & -5.0 & 3.0 \end{bmatrix}$$

- Supongamos que las probabilidades de que una persona que utiliza una marca X se cambie a la marca Y o a la Z son, respectivamente, 0.4 y 0.2; que las probabilidades de que una persona que utiliza la marca Y se cambie a la marca X o a la Z son, respectivamente, 0.2 y 0.2 y que las probabilidades de que una persona que utiliza la marca Z se cambie a la marca X o a la Y son, respectivamente, 0.1 y 0.1. La ecuación de transición para esta cadena es

$$\mathbf{P}_{k+1} = \mathbf{AP}_k = \begin{bmatrix} 0.4 & 0.2 & 0.1 \\ 0.4 & 0.6 & 0.1 \\ 0.2 & 0.2 & 0.8 \end{bmatrix} \begin{bmatrix} x^{(k)} \\ y^{(k)} \\ z^{(k)} \end{bmatrix}.$$

- Compruebe que $\lambda = 1$ es un autovalor de \mathbf{A} .
 - Determine la distribución estable para una ciudad de 80 000 habitantes.
- Supongamos que la industria del café en una cierta ciudad se concentra en cinco marcas B_1, B_2, B_3, B_4 y B_5 . Supongamos que se venden 60 toneladas de café al mes de manera que el consumo es de 3 kilos de café al mes por persona y que, independientemente de la marca, cada kilo de café vendido produce un beneficio de un euro. Se ha determinado empíricamente que la

matriz de transición mensual entre las marcas es la matriz \mathbf{A} siguiente en la que a_{ij} representa la probabilidad de que una persona que compra la marca B_j pase a comprar la marca B_i

$$\mathbf{A} = \begin{bmatrix} 0.1 & 0.2 & 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.3 & 0.4 & 0.1 & 0.2 \\ 0.3 & 0.3 & 0.1 & 0.1 & 0.2 \\ 0.4 & 0.1 & 0.2 & 0.1 & 0.2 \end{bmatrix}.$$

Una agencia de publicidad garantiza a la empresa que produce la marca B_1 que, por 40 millones de euros al año, puede cambiar la primera columna de la matriz de manera que pase a ser $[0.3 \ 0.1 \ 0.1 \ 0.2 \ 0.3]'$. ¿Debe la empresa que produce la marca B_1 aceptar la oferta de la agencia de publicidad?

7. Escriba un programa, basado en la técnica de deflación dada en el Ejercicio 4, para calcular todos los autovalores de una matriz. Para determinar la pareja autovalor-autovector dominante de cada paso, su programa debería llamar al Programa 11.1.
8. Use su programa del Problema 7 para calcular todos los autovalores de las siguientes matrices

$$(a) \quad \mathbf{A} = \begin{bmatrix} 1 & 2 & -1 \\ 1 & 0 & 1 \\ 4 & -4 & 5 \end{bmatrix}$$

$$(b) \quad \mathbf{A} = [a_{ij}], \text{ siendo } a_{ij} = \begin{cases} i + j & i = j \\ ij & i \neq j \end{cases} \quad \text{con } i, j = 1, 2, \dots, 15.$$

El método de Jacobi

El método de Jacobi es un algoritmo para calcular todas las parejas autovalor-autovector de una matriz simétrica que es muy fácil de entender. Es un método fiable que proporciona respuestas uniformemente precisas y que para matrices de orden menor o igual que 10 es competitivo frente a otros métodos más sofisticados; también es aceptable para matrices de orden menor o igual que 20, si la velocidad de convergencia no es una cuestión muy relevante.

El método de Jacobi funciona para todas las matrices simétricas reales; esta limitación no es muy severa ya que, en la práctica, hay un gran número de problemas en la ingeniería y en la matemática aplicada que involucran el cálculo de los autovalores de una matriz simétrica. Desde un punto de vista teórico, el método de Jacobi incorpora algunas técnicas que se utilizan en algoritmos más sofisticados por lo que merece la pena estudiarlo con detalle.

Rotaciones planas

Empezamos con una revisión geométrica sobre los cambios de coordenadas. Sea \mathbf{X} un vector en el espacio n -dimensional y consideremos la aplicación lineal $\mathbf{Y} = \mathbf{R}\mathbf{X}$, donde \mathbf{R} es la matriz de orden $n \times n$ dada por:

$$\mathbf{R} = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \cos \phi & \cdots & \sin \phi & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & -\sin \phi & \cdots & \cos \phi & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \begin{array}{l} \leftarrow \text{fila } p \\ \leftarrow \text{fila } q \\ \uparrow \text{col } p \qquad \uparrow \text{col } q \end{array}$$

En esta expresión todos los términos de \mathbf{R} que están fuera de la diagonal son cero salvo los dos que valen $\pm \sin \phi$ y todos los términos de la diagonal son 1 excepto los dos que valen $\cos \phi$. El efecto de esta transformación $\mathbf{Y} = \mathbf{R}\mathbf{X}$ puede verse fácilmente:

$$\begin{aligned} y_j &= x_j && \text{cuando } j \neq p \text{ y } j \neq q, \\ y_p &= x_p \cos \phi + x_q \sin \phi, \\ y_q &= -x_p \sin \phi + x_q \cos \phi. \end{aligned}$$

Tenemos entonces que esta aplicación lineal es una rotación de ángulo ϕ en el plano coordenado X_pOX_q . Eligiendo adecuadamente el ángulo ϕ , podemos conseguir que $y_p = 0$ o bien que $y_q = 0$ en el vector imagen. La transformación inversa $\mathbf{X} = \mathbf{R}^{-1}\mathbf{Y}$ es una rotación de ángulo $-\phi$ en el mismo plano coordenado X_pOX_q . Observemos también que \mathbf{R} es una matriz ortogonal, o sea,

$$\mathbf{R}^{-1} = \mathbf{R}' \quad \text{o bien} \quad \mathbf{R}'\mathbf{R} = \mathbf{I}.$$

Semejanza de matrices y transformaciones ortogonales

Consideremos el problema de autovalores

$$(1) \quad \mathbf{A}\mathbf{X} = \lambda\mathbf{X}.$$

Supongamos que \mathbf{K} es una matriz invertible y definamos \mathbf{B} mediante

$$(2) \quad \mathbf{B} = \mathbf{K}^{-1}\mathbf{A}\mathbf{K}.$$

Multiplicando por $\mathbf{K}^{-1}\mathbf{X}$ a la derecha de la relación (2) obtenemos

$$(3) \quad \begin{aligned} \mathbf{B}\mathbf{K}^{-1}\mathbf{X} &= \mathbf{K}^{-1}\mathbf{A}\mathbf{K}\mathbf{K}^{-1}\mathbf{X} = \mathbf{K}^{-1}\mathbf{AX} \\ &= \mathbf{K}^{-1}\lambda\mathbf{X} = \lambda\mathbf{K}^{-1}\mathbf{X}. \end{aligned}$$

Hacemos ahora el cambio de variable

$$(4) \quad \mathbf{Y} = \mathbf{K}^{-1}\mathbf{X} \quad \text{o bien} \quad \mathbf{X} = \mathbf{KY}$$

que, al sustituirlo en (3), produce el problema de autovalores

$$(5) \quad \mathbf{BY} = \lambda\mathbf{Y}.$$

Comparando los problemas (1) y (5), vemos que la transformación de semejanza (2) conserva el autovalor λ y que, aunque los autovectores son diferentes, están relacionados por el cambio de variable (4).

Supongamos que la matriz \mathbf{R} es ortogonal (o sea, $\mathbf{R}^{-1} = \mathbf{R}'$) y que \mathbf{C} se define mediante

$$(6) \quad \mathbf{C} = \mathbf{R}'\mathbf{AR}.$$

Multiplicando por $\mathbf{R}'\mathbf{X}$ a la derecha de la relación (6) obtenemos

$$(7) \quad \mathbf{CR}'\mathbf{X} = \mathbf{R}'\mathbf{ARR}'\mathbf{X} = \mathbf{R}'\mathbf{AX} = \mathbf{R}'\lambda\mathbf{X} = \lambda\mathbf{R}'\mathbf{X}.$$

Hacemos ahora el cambio de variable

$$(8) \quad \mathbf{Y} = \mathbf{R}'\mathbf{X} \quad \text{o} \quad \mathbf{X} = \mathbf{RY}$$

que, al sustituirlo en (7), produce el problema de autovalores

$$(9) \quad \mathbf{CY} = \lambda\mathbf{Y}.$$

Como antes, los autovalores de (1) y (9) son los mismos. Sin embargo, en el problema (9) el cambio de variables (8) es más fácil de deshacer ya que $\mathbf{R}^{-1} = \mathbf{R}'$.

Si, además, suponemos que la matriz \mathbf{A} es simétrica (es decir, $\mathbf{A} = \mathbf{A}'$), entonces tenemos

$$(10) \quad \mathbf{C}' = (\mathbf{R}'\mathbf{AR})' = \mathbf{R}'\mathbf{A}(\mathbf{R}')' = \mathbf{R}'\mathbf{AR} = \mathbf{C}.$$

Así que \mathbf{C} es también una matriz simétrica. En consecuencia, si \mathbf{A} es una matriz simétrica y \mathbf{R} es una matriz ortogonal, entonces la transformación de \mathbf{A} en \mathbf{C} dada por (6) conserva la simetría y los autovalores y la relación entre los autovectores viene dada por el cambio de variable (8). (Estos comentarios valen también para el método QR que veremos en la siguiente sección.)

Sucesión de transformaciones de Jacobi

Empezando con una matriz simétrica \mathbf{A} , construimos una sucesión de matrices ortogonales: $\mathbf{R}_1, \mathbf{R}_2, \dots$, de la siguiente manera:

$$(11) \quad \begin{aligned} \mathbf{D}_0 &= \mathbf{A}, \\ \mathbf{D}_j &= \mathbf{R}'_j \mathbf{D}_{j-1} \mathbf{R}_j \quad \text{para } j = 1, 2, \dots \end{aligned}$$

Vamos a mostrar cómo hay que construir la sucesión $\{\mathbf{R}_j\}$ de manera que

$$(12) \quad \lim_{j \rightarrow \infty} \mathbf{D}_j = \mathbf{D} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n).$$

En la práctica el proceso se detiene cuando los elementos que están fuera de la diagonal son suficientemente pequeños, de manera que

$$(13) \quad \mathbf{D}_k \approx \mathbf{D},$$

siendo

$$(14) \quad \mathbf{D}_k = \mathbf{R}'_k \mathbf{R}'_{k-1} \cdots \mathbf{R}'_1 \mathbf{A} \mathbf{R}_1 \mathbf{R}_2 \cdots \mathbf{R}_{k-1} \mathbf{R}_k.$$

Si definimos

$$(15) \quad \mathbf{R} = \mathbf{R}_1 \mathbf{R}_2 \cdots \mathbf{R}_{k-1} \mathbf{R}_k,$$

entonces $\mathbf{R}^{-1} \mathbf{A} \mathbf{R} = \mathbf{D}_k$, así que

$$(16) \quad \mathbf{A} \mathbf{R} = \mathbf{R} \mathbf{D}_k \approx \mathbf{R} \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n).$$

Si expresamos las columnas de \mathbf{R} como vectores: $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$, entonces podemos expresar \mathbf{R} como un vector fila de vectores columna:

$$(17) \quad \mathbf{R} = [\mathbf{X}_1 \ \mathbf{X}_2 \ \dots \ \mathbf{X}_n].$$

Esto nos permite escribir la relación (16) columna a columna:

$$(18) \quad [\mathbf{A} \mathbf{X}_1 \ \mathbf{A} \mathbf{X}_2 \ \dots \ \mathbf{A} \mathbf{X}_n] \approx [\lambda_1 \mathbf{X}_1 \ \lambda_2 \mathbf{X}_2 \ \dots \ \lambda_n \mathbf{X}_n].$$

De las relaciones (17) y (18) se deduce que \mathbf{X}_j , que es la j -ésima columna de \mathbf{R} , es una aproximación del autovector correspondiente al autovalor λ_j .

El paso general

El objetivo en cada paso del método de Jacobi es reducir a cero dos de los términos simétricos que están fuera de la diagonal, digamos los que ocupan las posiciones (p, q) y (q, p) . Denotemos por \mathbf{R}_1 la primera matriz ortogonal que vamos a usar, de manera que en la matriz

$$(19) \quad \mathbf{D}_1 = \mathbf{R}'_1 \mathbf{A} \mathbf{R}_1$$

los elementos d_{pq} y d_{qp} sean cero, y donde \mathbf{R}_1 es de la forma

$$(20) \quad \mathbf{R}_1 = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & s & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & -s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \begin{array}{l} \leftarrow \text{fila } p \\ \leftarrow \text{fila } q \end{array}$$

$\overset{\uparrow}{\text{col } p} \quad \overset{\uparrow}{\text{col } q}$

Aquí todos los elementos de \mathbf{R}_1 que están fuera de la diagonal son cero excepto: el que vale s , colocado en la fila p , columna q , y el que vale $-s$, colocado en la fila q , columna p . Por otro lado, todos los elementos diagonales valen 1 excepto los dos que valen c , en las posiciones p -ésima y q -ésima de la diagonal. Esta matriz es la matriz de una rotación plana, en la que hemos usado la notación $c = \cos \phi$ y $s = \sin \phi$.

Debemos verificar que la transformación (19) sólo altera los elementos de las filas y las columnas p -ésimas y q -ésimas. Consideraremos la multiplicación de \mathbf{A} por \mathbf{R}_1 , o sea, el producto $\mathbf{B} = \mathbf{A} \mathbf{R}_1$:

$$(21) \quad \mathbf{B} = \begin{bmatrix} a_{11} & \cdots & a_{1p} & \cdots & a_{1q} & \cdots & a_{1n} \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{p1} & \cdots & a_{pp} & \cdots & a_{pq} & \cdots & a_{pn} \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{q1} & \cdots & a_{qp} & \cdots & a_{qq} & \cdots & a_{qn} \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{n1} & \cdots & a_{np} & \cdots & a_{nq} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & s & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & -s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}.$$

Aplicando la regla de multiplicación de matrices, filas de la primera por columnas de la segunda, observamos que no se producen cambios en las columnas

primera a $(p - 1)$ -ésima, $(p + 1)$ -ésima a $(q - 1)$ -ésima y $(q + 1)$ -ésima a n -ésima; sólo se alteran las columnas p -ésima y q -ésima:

$$(22) \quad \begin{aligned} b_{jm} &= a_{jm} && \text{cuando } m \neq p \text{ y } m \neq q, \\ b_{jp} &= ca_{jp} - sa_{jq} && \text{para } j = 1, 2, \dots, n, \\ b_{jq} &= sa_{jp} + ca_{jq} && \text{para } j = 1, 2, \dots, n. \end{aligned}$$

Un argumento parecido prueba que al multiplicar \mathbf{R}'_1 por \mathbf{A} sólo se alteran las filas p -ésima y q -ésima. En consecuencia, la transformación

$$(23) \quad \mathbf{D}_1 = \mathbf{R}'_1 \mathbf{A} \mathbf{R}_1$$

sólo altera los elementos de las filas y las columnas p -ésimas y q -ésimas. Los elementos d_{jk} de la matriz \mathbf{D}_1 que no son iguales que los correspondientes de \mathbf{A} se calculan mediante las fórmulas

$$(24) \quad \begin{aligned} d_{jp} &= ca_{jp} - sa_{jq} && \text{para } j \neq p \text{ y } j \neq q, \\ d_{jq} &= sa_{jp} + ca_{jq} && \text{para } j \neq p \text{ y } j \neq q, \\ d_{pp} &= c^2 a_{pp} + s^2 a_{qq} - 2cs a_{pq}, \\ d_{qq} &= s^2 a_{pp} + c^2 a_{qq} + 2cs a_{pq}, \\ d_{pq} &= (c^2 - s^2)a_{pq} + cs(a_{pp} - a_{qq}), \end{aligned}$$

y los demás se hallan por simetría.

Cómo hacer d_{pq} y d_{qp} iguales a cero

El objetivo en cada paso del método de Jacobi es reducir a cero los elementos d_{pq} y d_{qp} que están fuera de la diagonal. La estrategia obvia es tomar

$$(25) \quad c = \cos \phi \quad \text{y} \quad s = \sin \phi,$$

siendo ϕ el ángulo de rotación que produce el efecto deseado. Sin embargo, para calcular este ángulo hay que realizar algunas manipulaciones ingeniosas con las identidades trigonométricas. Usando (25) para calcular la cotangente del ángulo doble obtenemos

$$(26) \quad \theta = \cot 2\phi = \frac{c^2 - s^2}{2cs}.$$

Supongamos que $a_{pq} \neq 0$ y que deseamos obtener $d_{pq} = 0$. Entonces, usando la última ecuación de (24), obtenemos

$$(27) \quad 0 = (c^2 - s^2)a_{pq} + cs(a_{pp} - a_{qq}).$$

Ordenando un poco los términos nos queda $(c^2 - s^2)/(cs) = (a_{qq} - a_{pp})/a_{pq}$ que, junto con (26), nos permite calcular θ :

$$(28) \quad \theta = \frac{a_{qq} - a_{pp}}{2a_{pq}}.$$

Aunque podemos usar la fórmula (28) junto con las fórmulas (25) y (26) para calcular c y s , puede probarse que se propaga menos error de redondeo si calculamos $\tan \phi$ y usamos este valor en cálculos posteriores. Así pues tomamos

$$(29) \quad t = \tan \phi = \frac{s}{c}.$$

Dividiendo el numerador y el denominador de (26) entre c^2 obtenemos

$$\theta = \frac{1 - s^2/c^2}{2s/c} = \frac{1 - t^2}{2t},$$

que nos da la ecuación

$$(30) \quad t^2 + 2t\theta - 1 = 0.$$

Puesto que $t = \tan \phi$, la menor de las raíces de la ecuación (30) corresponde al menor ángulo de rotación que podemos tomar y que cumple $|\phi| \leq \pi/4$. La forma especial que tiene la ecuación de segundo grado nos permite trabajar con la siguiente fórmula para hallar la menor de sus raíces

$$(31) \quad t = -\theta \pm (\theta^2 + 1)^{1/2} = \frac{\text{sign}(\theta)}{|\theta| + (\theta^2 + 1)^{1/2}},$$

siendo $\text{sign}(\theta) = 1$ cuando $\theta \geq 0$ y $\text{sign}(\theta) = -1$ cuando $\theta < 0$. Una vez hallado t , los valores c y s se calculan usando las fórmulas

$$(32) \quad c = \frac{1}{(t^2 + 1)^{1/2}} \quad \text{y} \quad s = ct.$$

Resumen del paso general

Resumimos ahora los cálculos que hay que realizar para hacer cero el elemento a_{pq} . En primer lugar hay que elegir la fila p y la columna q de manera que $a_{pq} \neq 0$. En segundo lugar se determinan los valores preliminares

$$(33) \quad \begin{aligned} \theta &= \frac{a_{qq} - a_{pp}}{2a_{pq}}, \\ t &= \frac{\text{sign}(\theta)}{|\theta| + (\theta^2 + 1)^{1/2}}, \\ c &= \frac{1}{(t^2 + 1)^{1/2}}, \\ s &= ct. \end{aligned}$$

En tercer lugar, para construir $\mathbf{D} = \mathbf{D}_1$, se toman (escribiéndolo como si fuera un programa hecho con el paquete MATLAB):

$$(34) \quad \begin{aligned} d_{pq} &= 0; \\ d_{qp} &= 0; \\ d_{pp} &= c^2 a_{pp} + s^2 a_{qq} - 2csa_{pq}; \\ d_{qq} &= s^2 a_{pp} + c^2 a_{qq} + 2csa_{pq}; \\ \text{for } j &= 1 : n \\ \text{if } (j \neq p) \text{ y } (j \neq q) \\ d_{jp} &= ca_{jp} - sa_{jq}; \\ d_{pj} &= d_{jp}; \\ d_{jq} &= ca_{jq} + sa_{jp}; \\ d_{qj} &= d_{jq}; \\ \text{end} \\ \text{end} \end{aligned}$$

Actualización de la matriz de los autovectores

Es necesario también ir calculando el producto $\mathbf{R}_1 \mathbf{R}_2 \cdots \mathbf{R}_j$. Cuando se detenga el algoritmo en la k -ésima iteración, entonces tendremos la matriz ortogonal

$$(35) \quad \mathbf{V}_k = \mathbf{R}_1 \mathbf{R}_2 \cdots \mathbf{R}_k.$$

Para ir calculando estas matrices \mathbf{V}_j , con $j = 1, 2, \dots, k$, empezamos tomando como matriz inicial $\mathbf{V} = \mathbf{I}$. Usando unas variables vectoriales \mathbf{P} y \mathbf{Q} para almacenar las columnas p -ésima y q -ésima de \mathbf{V} , respectivamente, entonces en cada paso se realizan las siguientes operaciones:

$$(36) \quad \begin{aligned} \text{for } j &= 1 : n \\ \mathbf{P}_j &= v_{jp}; \\ \mathbf{Q}_j &= v_{jq}; \\ \text{end} \\ \text{for } j &= 1 : n \\ v_{jp} &= c\mathbf{P}_j - s\mathbf{Q}_j; \\ v_{jq} &= s\mathbf{P}_j + c\mathbf{Q}_j; \\ \text{end} \end{aligned}$$

Estrategia para decidir qué a_{pq} se elimina

La velocidad de convergencia del método de Jacobi se estima mediante la suma de los cuadrados de los elementos que están fuera de la diagonal:

$$(37) \quad S_1 = \sum_{\substack{j,k=1 \\ k \neq j}}^n |a_{jk}|^2$$

$$(38) \quad S_2 = \sum_{\substack{j,k=1 \\ k \neq j}}^n |d_{jk}|^2, \quad \text{siendo} \quad \mathbf{D}_1 = \mathbf{R}' \mathbf{A} \mathbf{R}.$$

Dejamos como ejercicio la demostración, a partir de las relaciones dadas en (34), de que

$$(39) \quad S_2 = S_1 - 2|a_{pq}|^2.$$

Si en cada paso denotamos por S_j la suma de los cuadrados de los elementos que están fuera de la diagonal de \mathbf{D}_j , entonces la sucesión $\{S_j\}$ decrece monótonamente y está acotada inferiormente por cero. En el algoritmo original de Jacobi, dado en 1846, el elemento que hay que convertir en cero es, en cada paso, el elemento de fuera de la diagonal que tiene mayor valor absoluto, lo que involucra la necesidad de calcular el valor

$$(40) \quad \max\{\mathbf{A}\} = \max\{|a_{pq}| : p < q\}.$$

Esta elección garantiza que $\{S_j\}$ converge a cero y, en consecuencia, que $\{\mathbf{D}_j\}$ converge a \mathbf{D} y que $\{\mathbf{V}_j\}$ converge a la matriz \mathbf{V} de los autovectores (véase la Referencia [68]).

El procedimiento de selección propuesto por Jacobi puede requerir mucho tiempo de cálculo ya que hay que realizar del orden de $(n^2 - n)/2$ comparaciones en cada paso; lo que resulta prohibitivo para valores grandes de n . Una estrategia algo mejor es el método de Jacobi cíclico que consiste en ir eliminando los elementos por orden estricto de filas. Se elige un valor ε para la tolerancia, se hace un barrido de toda la matriz y si se encuentra un elemento a_{pq} de mayor tamaño que ε , entonces se elimina. En cada barrido se empieza por los elementos de la primera fila: $a_{12}, a_{13}, \dots, a_{1n}$; luego los de la segunda: $a_{23}, a_{24}, \dots, a_{2n}$; y así sucesivamente. Puede probarse que la convergencia de ambos métodos de Jacobi, el original y el cíclico, es cuadrática. Para utilizar el método cíclico hay que tener en cuenta que la suma de los cuadrados de los elementos de la diagonal se va incrementando en cada iteración; es decir, si

$$(41) \quad T_0 = \sum_{j=1}^n |a_{jj}|^2$$

y

$$T_1 = \sum_{j=1}^n |d_{jj}|^2$$

entonces

$$T_1 = T_0 + 2|a_{pq}|^2.$$

En consecuencia, la sucesión $\{\mathbf{D}_j\}$ converge a la matriz diagonal \mathbf{D} . Hágamos notar también que el tamaño medio de un elemento de la diagonal puede calcularse mediante la fórmula $(T_0/n)^{1/2}$ y que los tamaños de los elementos que están fuera de la diagonal deben compararse con $\varepsilon(T_0/n)^{1/2}$, siendo ε la tolerancia fijada de antemano. Por tanto, eliminamos a_{pq} si

$$(42) \quad |a_{pq}| > \varepsilon \left(\frac{T_0}{n} \right)^{1/2}.$$

Otra variación del método, llamado el método de Jacobi con umbral, queda propuesto como tarea de investigación (véase la Referencia [178]).

Ejemplo 11.7. Vamos a usar el método de Jacobi para transformar la siguiente matriz simétrica en una matriz diagonal:

$$\begin{bmatrix} 8 & -1 & 3 & -1 \\ -1 & 6 & 2 & 0 \\ 3 & 2 & 9 & 1 \\ -1 & 0 & 1 & 7 \end{bmatrix}$$

Dejamos los detalles computacionales como ejercicio. La primera rotación, que servirá para hacer cero el elemento $a_{13} = 3$, es

$$\mathbf{R}_1 = \begin{bmatrix} 0.763020 & 0.000000 & 0.646375 & 0.000000 \\ 0.000000 & 1.000000 & 0.000000 & 0.000000 \\ -0.646375 & 0.000000 & 0.763020 & 0.000000 \\ 0.000000 & 0.000000 & 0.000000 & 1.000000 \end{bmatrix}.$$

Haciendo el cálculo $\mathbf{A}_2 = \mathbf{R}_1 \mathbf{A}_1 \mathbf{R}_1$, llegamos a

$$\mathbf{A}_2 = \begin{bmatrix} 5.458619 & -2.055770 & 0.000000 & -1.409395 \\ -2.055770 & 6.000000 & 0.879665 & 0.000000 \\ 0.000000 & 0.879665 & 11.541381 & 0.116645 \\ -1.409395 & 0.000000 & 0.116645 & 7.000000 \end{bmatrix}.$$

A continuación hacemos cero el elemento $a_{12} = -2.055770$, obteniendo

$$\mathbf{A}_3 = \begin{bmatrix} 3.655795 & 0.000000 & 0.579997 & -1.059649 \\ 0.000000 & 7.802824 & 0.661373 & 0.929268 \\ 0.579997 & 0.661373 & 11.541381 & 0.116645 \\ -1.059649 & 0.929268 & 0.116645 & 7.000000 \end{bmatrix}.$$

Después de 10 iteraciones llegamos a

$$A_{10} = \begin{bmatrix} 3.295870 & 0.002521 & 0.037859 & 0.000000 \\ 0.002521 & 8.405210 & -0.004957 & 0.066758 \\ 0.037859 & -0.004957 & 11.704123 & -0.001430 \\ 0.000000 & 0.066758 & -0.001430 & 6.594797 \end{bmatrix}.$$

Hacen falta seis iteraciones más para que los elementos de la diagonal se acerquen a los autovalores

$$D = \text{diag}(3.295699, 8.407662, 11.704301, 6.592338).$$

Sin embargo, en este paso los elementos que están fuera de la diagonal no son aún lo bastante pequeños y harán falta tres iteraciones más para que sean de tamaño menor que 10^{-6} . En ese paso, las aproximaciones a los autovectores son las columnas de la matriz $V = R_1 R_2 \cdots R_{18}$, dada por

$$V = \begin{bmatrix} 0.528779 & -0.573042 & 0.582298 & 0.230097 \\ 0.591967 & 0.472301 & 0.175776 & -0.628975 \\ -0.536039 & 0.282050 & 0.792487 & -0.071235 \\ 0.287454 & 0.607455 & 0.044680 & 0.739169 \end{bmatrix}.$$

MATLAB

Programa 11.3 (Método de Jacobi para aproximar autovalores y autovectores). Cálculo de un conjunto completo de pares autovalor-autovector $\{(\lambda_j, V_j)\}_{j=1}^n$ de una matriz A de orden $n \times n$, real y simétrica mediante el método de Jacobi.

```
function [V,D]=jacobi1(A,epsilon)
% Datos
%     - A es una matriz de orden n x n
%     - epsilon es la tolerancia
% Resultados
%     - D es la matriz diagonal de orden n x n
%         de los autovalores
%     - V es la matriz de orden n x n de los autovectores
% Inicialización de V, D y los parámetros
D=A;
[n,n]=size(A);
V=eye(n);
state=1;
% Cálculo del elemento (p,q) de A que tiene mayor
% magnitud entre los que están fuera de la diagonal
```

```
[m1 p]=max(abs(D-diag(diag(D))));  

[m2 q]=max(m1);  

p=p(q);  

while(state==1)  

    % Eliminación de Dpq y Dqp  

    theta=(D(q,q)-D(p,p))/2*D(p,q);  

    t=sign(theta)/(abs(theta)+sqrt(theta^2+1));  

    c=1/sqrt(t^2+1);  

    s=c*t;  

    R=[c s;-s c];  

    D([p q],:)=R'*D([p q],:);  

    D(:,[p q])=D(:,[p q])*R;  

    V(:,[p q])=V(:,[p q])*R;  

    [m1 p]=max(abs(D-diag(diag(D))));  

    [m2 q]=max(m1);  

    p=p(q);  

    if (abs(D(p,q))<epsilon*sqrt(sum(diag(D).^2)/n))  

        state=0;  

    end  

end  

D=diag(diag(D));
```

Ejercicios

1. *Sistemas de masas y muelles.* Consideremos el sistema de masas enlazadas por muelles que se muestra en la Figura 11.3. El modelo matemático que describe los desplazamientos de las masas cuando se pierde la posición de equilibrio estático es

$$\begin{bmatrix} k_1 + k_2 & -k_2 & 0 \\ -k_2 & k_2 + k_3 & -k_3 \\ 0 & -k_3 & k_3 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} + \begin{bmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{bmatrix} \begin{bmatrix} x_1''(t) \\ x_2''(t) \\ x_3''(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

- (a) Haciendo el cambio de variables $x_j(t) = v_j \operatorname{sen}(\omega t + \theta)$ para $j = 1, 2, 3$, donde θ es una constante, pruebe que el modelo puede reformularse de la

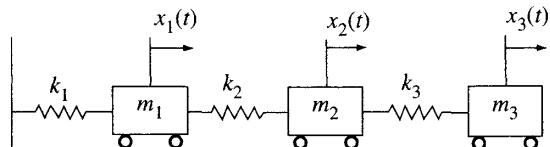


Figura 11.3 Un sistema de masas enlazadas por muelles.

siguiente manera:

$$\begin{bmatrix} \frac{k_1 + k_2}{m_1} & \frac{-k_2}{m_1} & 0 \\ \frac{-k_2}{m_2} & \frac{k_2 + k_3}{m_2} & \frac{-k_3}{m_2} \\ 0 & \frac{-k_3}{m_3} & \frac{k_3}{m_3} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \omega^2 \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}.$$

- (b) Tomando $\lambda = \omega^2$, las tres soluciones del apartado (a) son las parejas autovalor-autovector $(\lambda_j, \mathbf{V}_j = [v_1^{(j)} \ v_2^{(j)} \ v_3^{(j)}]')$ para $j = 1, 2, 3$. Pruebe que estas parejas permiten formar un sistema fundamental de soluciones:

$$\mathbf{X}_j(t) = \begin{bmatrix} v_1^{(j)} \sin(\omega_j t + \theta) \\ v_2^{(j)} \sin(\omega_j t + \theta) \\ v_3^{(j)} \sin(\omega_j t + \theta) \end{bmatrix} = \sin(\omega_j t + \theta) \begin{bmatrix} v_1^{(j)} \\ v_2^{(j)} \\ v_3^{(j)} \end{bmatrix},$$

donde $\omega_j = \sqrt{\lambda_j}$ para $j = 1, 2, 3$.

Observación. Estas tres soluciones se llaman **modos principales de vibración**.

2. El sistema lineal homogéneo de ecuaciones diferenciales

$$\begin{aligned} x'_1(t) &= x_1(t) + x_2(t) \\ x'_2(t) &= -2x_1(t) + 4x_2(t) \end{aligned}$$

puede escribirse de forma matricial como

$$\mathbf{X}'(t) = \begin{bmatrix} x'_1(t) \\ x'_2(t) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \mathbf{A}\mathbf{X}(t).$$

- (a) Compruebe que $(2, [1 \ 1]')$ y $(3, [1 \ 2]')$ son parejas autovalor-autovector de la matriz \mathbf{A} .
- (b) Compruebe, sustituyendo directamente en la forma matricial del sistema, que tanto $\mathbf{X}(t) = e^{2t}[1 \ 1]'$ como $\mathbf{X}(t) = e^{3t}[1 \ 2]'$ son soluciones del sistema de ecuaciones diferenciales.
- (c) Compruebe, sustituyendo directamente en la forma matricial del sistema, que $\mathbf{X}(t) = c_1 e^{2t}[1 \ 1]' + c_2 e^{3t}[1 \ 2]'$ es la solución general del sistema de ecuaciones diferenciales.

Observación. Si la matriz \mathbf{A} tiene n autovalores distintos, entonces tiene n autovectores independientes. En ese caso, la solución general del correspondiente sistema lineal homogéneo de ecuaciones diferenciales puede escribirse como una combinación lineal:

$$\mathbf{X}(t) = c_1 e^{\lambda_1 t} \mathbf{V}_1 + c_2 e^{\lambda_2 t} \mathbf{V}_2 + \cdots + c_n e^{\lambda_n t} \mathbf{V}_n.$$

3. Use la técnica descrita en el Ejercicio 2 para resolver, a mano, cada uno de los problemas de valor inicial siguientes:

(a) $\begin{aligned}x'_1 &= 4x_1 + 2x_2 \\x'_2 &= 3x_1 - x_2\end{aligned}$ con $\begin{cases}x_1(0) = 1 \\x_2(0) = 2\end{cases}$

(b) $\begin{aligned}x'_1 &= 2x_1 - 12x_2 \\x'_2 &= x_1 - 5x_2\end{aligned}$ con $\begin{cases}x_1(0) = 2 \\x_2(0) = 2\end{cases}$

(c) $\begin{aligned}x'_1 &= x_2 \\x'_2 &= x_3 \\x'_3 &= 8x_1 - 14x_2 + 7x_3\end{aligned}$ con $\begin{cases}x_1(0) = 1 \\x_2(0) = 2 \\x_3(0) = 3\end{cases}$

Algoritmos y programas

1. Use el Programa 11.3 para hallar las parejas autovalor-autovector de la matriz dada con una tolerancia $\varepsilon = 10^{-7}$. Compare sus resultados con los que se obtienen al aplicar la instrucción `eig` del paquete MATLAB escribiendo `[V,D]=eig(A)` en la ventana de trabajo del programa

(a) $A = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}$

(b) $A = \begin{bmatrix} 2.25 & -0.25 & -1.25 & 2.75 \\ -0.25 & 2.25 & 2.75 & 1.25 \\ -1.25 & 2.75 & 2.25 & -0.25 \\ 2.75 & 1.25 & -0.25 & 2.25 \end{bmatrix}$

(c) $A = [a_{ij}]$, con $a_{ij} = \begin{cases} i+j & i=j \\ ij & i \neq j \end{cases} \quad i, j = 1, 2, \dots, 30$

(d) $A = [a_{ij}]$, con $a_{ij} = \begin{cases} \cos(\operatorname{sen}(i+j)) & i=j \\ i+ij+j & i \neq j \end{cases} \quad i, j = 1, 2, \dots, 40$

2. Use la técnica descrita en el Ejercicio 1 y el Programa 11.3 para hallar los modos principales de vibración de los sistemas de masas enlazadas por muelles cuyos coeficientes son

(a) $k_1 = 3, k_2 = 2, k_3 = 1, m_1 = 1, m_2 = 1, m_3 = 1$.

(b) $k_1 = \frac{1}{2}, k_2 = \frac{1}{4}, k_3 = \frac{1}{4}, m_1 = 4, m_2 = 4, m_3 = 4$.

(c) $k_1 = 0.2, k_2 = 0.4, k_3 = 0.3, m_1 = 2.5, m_2 = 2.5, m_3 = 2.5$.

3. Use la técnica descrita en el Ejercicio 2 y el Programa 11.3 para hallar la solución general de los sistemas lineales homogéneos de ecuaciones diferenciales siguientes:
- (a) $x'_1 = 4x_1 + 3x_2 + 2x_3 + x_4,$
 $x'_2 = 3x_1 + 4x_2 + 3x_3 + 2x_4,$
 $x'_3 = 2x_1 + 3x_2 + 4x_3 + 3x_4,$
 $x'_4 = x_1 + 2x_2 + 3x_3 + 4x_4.$
- (b) $x'_1 = 5x_1 + 4x_2 + 3x_3 + 2x_4 + x_5,$
 $x'_2 = 4x_1 + 5x_2 + 4x_3 + 3x_4 + 2x_5,$
 $x'_3 = 3x_1 + 4x_2 + 5x_3 + 4x_4 + 3x_5,$
 $x'_4 = 2x_1 + 3x_2 + 4x_3 + 5x_4 + 4x_5,$
 $x'_5 = x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5.$
4. Modifique el Programa 11.3 de manera que lleve a cabo el método cíclico de Jacobi.
5. Use su programa del Problema 4 con las matrices simétricas del Problema 1. En particular, compare el número de iteraciones que se realizan en su programa y en el Programa 11.3 hasta satisfacer la tolerancia que se da.

11.4 Autovalores de matrices simétricas

El método de Householder

Cada paso del método de Jacobi produce dos ceros fuera de la diagonal, pero en las siguientes iteraciones estos ceros podrían desaparecer, así que es necesario realizar muchas iteraciones para conseguir que todos los elementos de fuera de la diagonal sean suficientemente pequeños. Desarrollaremos ahora un método que produce varios ceros fuera de la diagonal en cada iteración, ceros que permanecen a lo largo de las iteraciones sucesivas. Empezamos describiendo el paso fundamental del proceso.

Teorema 11.23 (Reflexión de Householder). Si los vectores \mathbf{X} e \mathbf{Y} tienen la misma norma, entonces existe una matriz ortogonal y simétrica \mathbf{P} tal que

$$(1) \quad \mathbf{Y} = \mathbf{P}\mathbf{X},$$

donde

$$(2) \quad \mathbf{P} = \mathbf{I} - 2\mathbf{W}\mathbf{W}'$$

siendo

$$(3) \quad \mathbf{W} = \frac{\mathbf{X} - \mathbf{Y}}{\|\mathbf{X} - \mathbf{Y}\|_2}.$$

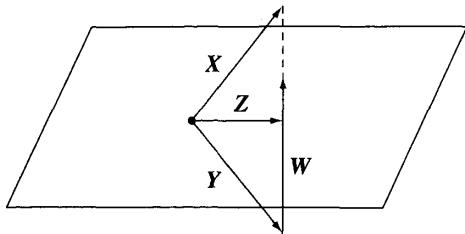


Figura 11.4 Los vectores \mathbf{W} , \mathbf{X} , \mathbf{Y} y \mathbf{Z} que aparecen en la reflexión de Householder.

Puesto que \mathbf{P} es ortogonal y simétrica, se tiene que

$$(4) \quad \mathbf{P}^{-1} = \mathbf{P}.$$

*Demuestra*ción. El vector \mathbf{W} definido por la relación (3) es el vector unitario en la dirección de $\mathbf{X} - \mathbf{Y}$; por tanto,

$$(5) \quad \mathbf{W}'\mathbf{W} = 1$$

y

$$(6) \quad \mathbf{Y} = \mathbf{X} + c\mathbf{W},$$

siendo $c = -\|\mathbf{X} - \mathbf{Y}\|_2$. Como \mathbf{X} e \mathbf{Y} tienen la misma norma, podemos usar la regla del paralelogramo para la suma de vectores y deducir que el vector $\mathbf{Z} = (\mathbf{X} + \mathbf{Y})/2 = \mathbf{X} + (c/2)\mathbf{W}$ es ortogonal al vector \mathbf{W} (véase la Figura 11.4). Esto implica que

$$\mathbf{W}'(\mathbf{X} + \frac{c}{2}\mathbf{W}) = 0.$$

Usando ahora la relación (5) podemos desarrollar la igualdad anterior y obtener

$$(7) \quad \mathbf{W}'\mathbf{X} + \frac{c}{2}\mathbf{W}'\mathbf{W} = \mathbf{W}'\mathbf{X} + \frac{c}{2} = 0.$$

El paso crucial es, ahora, usar la igualdad (7) para expresar c como

$$(8) \quad c = -2(\mathbf{W}'\mathbf{X})$$

y sustituir esta expresión en (6), lo que nos da

$$\mathbf{Y} = \mathbf{X} + c\mathbf{W} = \mathbf{X} - 2\mathbf{W}'\mathbf{X}\mathbf{W}.$$

Como el producto $\mathbf{W}'\mathbf{X}$ es un escalar, esta última igualdad puede escribirse de la siguiente manera

$$(9) \quad \mathbf{Y} = \mathbf{X} - 2\mathbf{W}\mathbf{W}'\mathbf{X} = (\mathbf{I} - 2\mathbf{W}\mathbf{W}')\mathbf{X},$$

con lo cual, mirando en (9), vemos que $\mathbf{P} = \mathbf{I} - 2\mathbf{W}\mathbf{W}'$ es la matriz buscada. La matriz \mathbf{P} es simétrica porque

$$\begin{aligned}\mathbf{P}' &= (\mathbf{I} - 2\mathbf{W}\mathbf{W}')' = \mathbf{I} - 2(\mathbf{W}\mathbf{W}')' \\ &= \mathbf{I} - 2\mathbf{W}\mathbf{W}' = \mathbf{P}\end{aligned}$$

y es ortogonal porque

$$\begin{aligned}\mathbf{P}'\mathbf{P} &= (\mathbf{I} - 2\mathbf{W}\mathbf{W}')(\mathbf{I} - 2\mathbf{W}\mathbf{W}') \\ &= \mathbf{I} - 4\mathbf{W}\mathbf{W}' + 4\mathbf{W}\mathbf{W}'\mathbf{W}\mathbf{W}' \\ &= \mathbf{I} - 4\mathbf{W}\mathbf{W}' + 4\mathbf{W}\mathbf{W}' = \mathbf{I},\end{aligned}$$

lo que completa la prueba del teorema. •

Hagamos notar que el efecto de la aplicación $\mathbf{Y} = \mathbf{P}\mathbf{X}$ es reflejar \mathbf{X} tomando la dirección \mathbf{Z} como eje de reflexión, de ahí el nombre de **reflexión de Householder**.

Corolario 11.3 (La k -ésima matriz de Householder). Sea \mathbf{X} el vector $[x_1 \dots x_n]'$ de orden $n \times 1$. Si k es un número natural con $1 \leq k \leq n - 2$, entonces existen un vector \mathbf{W}_k y una matriz $\mathbf{P}_k = \mathbf{I} - 2\mathbf{W}_k\mathbf{W}_k'$ tales que

$$(10) \quad \mathbf{P}_k \mathbf{X} = \mathbf{P}_k \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} \\ x_{k+2} \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ -S \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{Y}.$$

Demostración. La clave es definir el valor S de manera que $\|\mathbf{X}\|_2 = \|\mathbf{Y}\|_2$ y, entonces, aplicar el Teorema 11.23. El valor adecuado de S es el que verifica

$$(11) \quad S^2 = x_{k+1}^2 + x_{k+2}^2 + \dots + x_n^2,$$

lo que se comprueba calculando las normas de \mathbf{X} e \mathbf{Y} :

$$\begin{aligned}(12) \quad \|\mathbf{X}\|_2 &= x_1^2 + x_2^2 + \dots + x_n^2 \\ &= x_1^2 + x_2^2 + \dots + x_k^2 + S^2 \\ &= \|\mathbf{Y}\|_2.\end{aligned}$$

Usando la relación (3) del Teorema 11.23 calculamos el vector \mathbf{W} :

$$\begin{aligned}(13) \quad \mathbf{W} &= \frac{1}{R}(\mathbf{X} - \mathbf{Y}) \\ &= \frac{1}{R}[0 \ \dots \ 0 \ (x_{k+1} + S) \ x_{k+2} \ \dots \ x_n]'.\end{aligned}$$

Puede probarse que se propaga un error de redondeo menor si se toma el signo de S igual que el signo de x_{k+1} ; así que

$$(14) \quad S = \text{sign}(x_{k+1})(x_{k+1}^2 + x_{k+2}^2 + \cdots + x_n^2)^{1/2}.$$

Puesto que el escalar R de (13) hay que elegirlo de manera que $\|W\|_2 = 1$, debe cumplirse que

$$(15) \quad \begin{aligned} R^2 &= (x_{k+1} + S)^2 + x_{k+2}^2 + \cdots + x_n^2 \\ &= 2x_{k+1}S + S^2 + x_{k+1}^2 + x_{k+2}^2 + \cdots + x_n^2 \\ &= 2x_{k+1}S + 2S^2. \end{aligned}$$

En definitiva, la matriz P_k viene dada por

$$(16) \quad P_k = I - 2WW'$$

lo que completa la prueba.

Transformaciones de Householder

Supongamos que A es una matriz simétrica de orden $n \times n$. Entonces una sucesión de $n - 2$ transformaciones de semejanza de la forma PAP , donde las matrices P son de Householder, permite reducir A a una matriz simétrica y tridiagonal que tiene los mismos autovalores. Vamos a visualizar el proceso cuando $n = 5$. La primera transformación es de la forma P_1AP_1 , donde P_1 se construye aplicando el Corolario 11.3 con la primera columna de la matriz A en el papel de vector X . Es fácil comprobar que la forma general de P_1 es

$$(17) \quad P_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & p & p & p & p \\ 0 & p & p & p & p \\ 0 & p & p & p & p \\ 0 & p & p & p & p \end{bmatrix},$$

donde la letra p representa los elementos, no necesariamente iguales, de P_1 ; por tanto, la transformación P_1AP_1 no altera el valor del elemento a_{11} de A :

$$(18) \quad P_1AP_1 = \begin{bmatrix} a_{11} & v_1 & 0 & 0 & 0 \\ u_1 & w_1 & w & w & w \\ 0 & w & w & w & w \\ 0 & w & w & w & w \\ 0 & w & w & w & w \end{bmatrix} = A_1.$$

El elemento denotado por u_1 ha cambiado como resultado de multiplicar por P_1 a la izquierda y el denotado por v_1 ha cambiado como resultado de multiplicar

por \mathbf{P}_1 a la derecha; puesto que \mathbf{A}_1 es simétrica, tenemos $u_1 = v_1$. Los elementos denotados genéricamente por w se ven alterados como resultado de las dos multiplicaciones. Además, como \mathbf{X} es la primera columna de \mathbf{A} , la relación (10) nos dice que $u_1 = -S$.

La segunda transformación de Householder se aplica a la matriz \mathbf{A}_1 definida por la igualdad (18) y se denota por $\mathbf{P}_2 \mathbf{A}_1 \mathbf{P}_2$, donde \mathbf{P}_2 se construye aplicando el Corolario 11.3 con la segunda columna de la matriz \mathbf{A}_1 en el papel de vector \mathbf{X} . De nuevo, es fácil comprobar que la forma general de \mathbf{P}_2 es

$$(19) \quad \mathbf{P}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & p & p & p \\ 0 & 0 & p & p & p \\ 0 & 0 & p & p & p \end{bmatrix},$$

donde p representa un elemento genérico de \mathbf{P}_2 . La matriz identidad de orden 2×2 que aparece en la esquina superior izquierda asegura que los ceros ya conseguidos en el primer paso no se verán alterados por la segunda transformación $\mathbf{P}_2 \mathbf{A}_1 \mathbf{P}_2$ cuyo resultado es

$$(20) \quad \mathbf{P}_2 \mathbf{A}_1 \mathbf{P}_2 = \begin{bmatrix} a_{11} & v_1 & 0 & 0 & 0 \\ u_1 & w_1 & v_2 & 0 & 0 \\ 0 & u_2 & w_2 & w & w \\ 0 & 0 & w & w & w \\ 0 & 0 & w & w & w \end{bmatrix} = \mathbf{A}_2.$$

Los elementos u_2 y v_2 se vieron alterados como resultado de las multiplicaciones por \mathbf{P}_2 a la izquierda y la derecha, respectivamente. El resto de los elementos w se vieron afectados por las dos multiplicaciones.

La tercera transformación de Householder, $\mathbf{P}_3 \mathbf{A}_2 \mathbf{P}_3$, se aplica a la matriz \mathbf{A}_2 dada en (20), aplicando el corolario cuando \mathbf{X} es la tercera columna de \mathbf{A}_2 . La forma general de \mathbf{P}_3 es

$$(21) \quad \mathbf{P}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & p & p \\ 0 & 0 & 0 & p & p \end{bmatrix}.$$

De nuevo, la matriz identidad de orden 3×3 asegura que la transformación $\mathbf{P}_3 \mathbf{A}_2 \mathbf{P}_3$ no altera los elementos que ya hemos hecho igual a cero en \mathbf{A}_2 , de

manera que obtenemos

$$(22) \quad \mathbf{P}_3 \mathbf{A}_2 \mathbf{P}_3 = \begin{bmatrix} a_{11} & v_1 & 0 & 0 & 0 \\ u_1 & w_1 & v_2 & 0 & 0 \\ 0 & u_2 & w_2 & v_3 & 0 \\ 0 & 0 & u_3 & w & w \\ 0 & 0 & 0 & w & w \end{bmatrix} = \mathbf{A}_3.$$

Así que con tres transformaciones podemos reducir \mathbf{A} a una forma tridiagonal.

La transformación \mathbf{PAP} no se realiza multiplicando las matrices; es mucho más eficiente llevarla a cabo haciendo algunas manipulaciones con los vectores involucrados, como se muestra en el siguiente teorema.

Teorema 11.24 (Cálculo de una transformación de Householder). Si $\mathbf{P} = \mathbf{I} - 2\mathbf{WW}'$ es una matriz de Householder, entonces la transformación \mathbf{PAP} puede obtenerse como sigue: Tomando

$$(23) \quad \mathbf{V} = \mathbf{AW},$$

calculamos

$$(24) \quad c = \mathbf{W}'\mathbf{V}$$

y

$$(25) \quad \mathbf{Q} = \mathbf{V} - c\mathbf{W}.$$

Entonces

$$(26) \quad \mathbf{PAP} = \mathbf{A} - 2\mathbf{W}\mathbf{Q}' - 2\mathbf{Q}\mathbf{W}'.$$

Demostración. Primero formamos el producto

$$\mathbf{AP} = \mathbf{A}(\mathbf{I} - 2\mathbf{WW}') = \mathbf{A} - 2\mathbf{AWW}'.$$

Usando la relación (23), podemos escribir esto como

$$(27) \quad \mathbf{AP} = \mathbf{A} - 2\mathbf{VW}'.$$

Ahora usamos la igualdad (27) para escribir

$$(28) \quad \mathbf{PAP} = (\mathbf{I} - 2\mathbf{WW}')(\mathbf{A} - 2\mathbf{VW}')$$

que, desarrollando el producto y escribiendo el término $2(2\mathbf{WW}'\mathbf{VW}')$ en dos porciones,

$$(29) \quad \mathbf{PAP} = \mathbf{A} - 2\mathbf{W}(\mathbf{W}'\mathbf{A}) + 2\mathbf{W}(\mathbf{W}'\mathbf{VW}') - 2\mathbf{VW}' + 2\mathbf{W}(\mathbf{W}'\mathbf{V})\mathbf{W}'.$$

Como \mathbf{A} es simétrica, podemos usar la igualdad $(\mathbf{W}' \mathbf{A}) = (\mathbf{W}' \mathbf{A}') = \mathbf{V}'$. La parte delicada es observar ahora que $(\mathbf{W}' \mathbf{V})$ es un escalar, así que conmuta libremente con cualquiera de los demás términos de los productos. La otra igualdad escalar $\mathbf{W}' \mathbf{V} = (\mathbf{W}' \mathbf{V})'$ nos permite obtener la relación

$$(\mathbf{W}' \mathbf{V}) \mathbf{W}' = \mathbf{W}' (\mathbf{W}' \mathbf{V}) = \mathbf{W}' (\mathbf{W}' \mathbf{V})' = ((\mathbf{W}' \mathbf{V}) \mathbf{W})' = (\mathbf{W}' \mathbf{V} \mathbf{W})'.$$

Usando estos resultados en los términos de (29) que están entre paréntesis, deducimos

$$(30) \quad \mathbf{P} \mathbf{A} \mathbf{P} = \mathbf{A} - 2\mathbf{W} \mathbf{V}' + 2\mathbf{W} (\mathbf{W}' \mathbf{V} \mathbf{W})' - 2\mathbf{V} \mathbf{W}' + 2\mathbf{W}' \mathbf{V} \mathbf{W} \mathbf{W}'$$

que, aplicando la propiedad distributiva, queda

$$(31) \quad \mathbf{P} \mathbf{A} \mathbf{P} = \mathbf{A} - 2\mathbf{W} (\mathbf{V}' - (\mathbf{W}' \mathbf{V} \mathbf{W})') - 2(\mathbf{V} - \mathbf{W}' \mathbf{V} \mathbf{W}) \mathbf{W}'.$$

Finalmente, la definición de \mathbf{Q} dada en (25) junto con (31) prueban la igualdad (26). •

Reducción a forma tridiagonal

Supongamos que \mathbf{A} es una matriz simétrica de orden $n \times n$. Empezando con

$$(32) \quad \mathbf{A}_0 = \mathbf{A},$$

se puede construir una sucesión $\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_{n-2}$ de matrices de Householder tales que los elementos de la matriz

$$(33) \quad \mathbf{A}_k = \mathbf{P}_k \mathbf{A}_{k-1} \mathbf{P}_k \quad \text{para } k = 1, 2, \dots, n-2,$$

que están debajo de la primera subdiagonal en las columnas $1, 2, \dots, k$ son todos iguales a cero. En particular, la matriz \mathbf{A}_{n-2} es una matriz simétrica y tridiagonal que es semejante a la matriz \mathbf{A} y que se llama, a veces, **forma de Hessenberg** de \mathbf{A} . Este proceso se conoce como **Método de Householder** y también puede usarse con una matriz cualquiera para obtener una matriz semejante cuyos elementos por debajo de la primera subdiagonal son todos iguales a cero.

Ejemplo 11.8. Vamos a usar el método de Householder para reducir a forma tridiagonal la matriz

$$\mathbf{A}_0 = \begin{bmatrix} 4 & 2 & 2 & 1 \\ 2 & -3 & 1 & 1 \\ 2 & 1 & 3 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix}$$

Haremos un resumen de los cálculos dejando los detalles como ejercicio. En el primer paso usamos los valores $S = 3$ y $R = 30^{1/2} = 5.477226$ para construir el vector

$$\mathbf{W}' = \frac{1}{\sqrt{30}} [0 \ 5 \ 2 \ 1] = [0.000000 \ 0.912871 \ 0.365148 \ 0.182574].$$

El producto $\mathbf{V} = \mathbf{A}\mathbf{W}$ nos da

$$\begin{aligned}\mathbf{V}' &= \frac{1}{\sqrt{30}} [0 \ -12 \ 12 \ 9] \\ &= [0.000000 \ -2.190890 \ 2.190890 \ 1.643168].,\end{aligned}$$

de manera que la constante $c = \mathbf{W}'\mathbf{V}$ es

$$c = -0.9.$$

Formamos ahora el vector $\mathbf{Q} = \mathbf{V} - c\mathbf{W} = \mathbf{V} + 0.9\mathbf{W}$:

$$\begin{aligned}\mathbf{Q}' &= \frac{1}{\sqrt{30}} [0.000000 \ -7.500000 \ 13.800000 \ 9.900000] \\ &= [0.000000 \ -1.369306 \ 2.519524 \ 1.807484]\end{aligned}$$

y calculamos $\mathbf{A}_1 = \mathbf{A}_0 - 2\mathbf{W}\mathbf{Q}' - 2\mathbf{Q}\mathbf{W}'$

$$\mathbf{A}_1 = \begin{bmatrix} 4.0 & -3.0 & 0.0 & 0.0 \\ -3.0 & 2.0 & -2.6 & -1.8 \\ 0.0 & -2.6 & -0.68 & -1.24 \\ 0.0 & -1.8 & -1.24 & 0.68 \end{bmatrix}.$$

En el segundo, y último, paso las constantes son $S = -3.1622777$, $R = 6.0368737$ y $c = -1.2649111$ y los vectores son

$$\begin{aligned}\mathbf{W}' &= [0.000000 \ 0.000000 \ -0.954514 \ -0.298168], \\ \mathbf{V}' &= [0.000000 \ 0.000000 \ 1.018797 \ 0.980843], \\ \mathbf{Q}' &= [0.000000 \ 0.000000 \ -0.188578 \ 0.603687].\end{aligned}$$

La matriz tridiagonal $\mathbf{A}_2 = \mathbf{A}_1 - 2\mathbf{W}\mathbf{Q}' - 2\mathbf{Q}\mathbf{W}'$ que se obtiene es

$$\mathbf{A}_2 = \begin{bmatrix} 4.0 & -3.0 & 0.0 & 0.0 \\ -3.0 & 2.0 & 3.162278 & 0.0 \\ 0.0 & 3.162278 & -1.4 & -0.2 \\ 0.0 & 0.0 & -0.2 & 1.4 \end{bmatrix}.$$

MATLAB

Programa 11.4 (Reducción a la forma tridiagonal). Reducción de una matriz A simétrica y de orden $n \times n$ a forma tridiagonal usando $n - 2$ transformaciones de Householder.

```
function T=house(A)
% Dato
%      - A es una matriz simétrica de orden n x n
% Resultado
%      - T es una matriz tridiagonal
[n,n]=size(A);
for k=1:n-2
    % Construcción de W
    s=norm(A(k+1:n,k));
    if (A(k+1,k)<0)
        s=-s;
    end
    r=sqrt(2*s*(A(k+1,k)+s));
    W(1:k)=zeros(1,k);
    W(k+1)=(A(k+1,k)+s)/r;
    W(k+2:n)=A(k+2:n,k)'/r;
    % Construcción de V
    V(1:k)=zeros(1,k);
    V(k+1:n)=A(k+1:n,k+1:n)*W(k+1:n)';
    % Construcción de Q
    c=W(k+1:n)*V(k+1:n)';
    Q(1:k)=zeros(1,k);
    Q(k+1:n)=V(k+1:n)-c*W(k+1:n);
    % Cálculo de la matriz T
    A(k+2:n,k)=zeros(n-k-1,1);
    A(k,k+2:n)=zeros(1,n-k-1);
    A(k+1,k)=-s;
    A(k,k+1)=-s;
    A(k+1:n,k+1:n)=A(k+1:n,k+1:n) ...
    -2*W(k+1:n)'*Q(k+1:n)-2*Q(k+1:n)'*W(k+1:n);
end
T=A;
```

El método QR

Supongamos que A es una matriz real y simétrica de orden n . En el apartado anterior hemos visto el método de Householder para construir una matriz tri-

diagonal simétrica semejante a la matriz \mathbf{A} . Para hallar todos los autovalores de una matriz tridiagonal se emplea el método QR (que también puede aplicarse a matrices cualesquiera). Este método se basa en que \mathbf{A} puede factorizarse como el producto $\mathbf{A} = \mathbf{QR}$ de una matriz ortogonal \mathbf{Q} por una matriz triangular superior \mathbf{R} , lo que se conoce como **factorización QR** de la matriz \mathbf{A} . Veremos más adelante cómo se puede obtener esta factorización mediante una sucesión de $n - 1$ transformaciones ortogonales (que pueden ser rotaciones como las que se utilizan en el método de Jacobi o bien reflexiones de Householder).

El método para calcular los autovalores consiste en el siguiente procedimiento iterativo: Poniendo $\mathbf{A}_1 = \mathbf{A}$, construimos una matriz ortogonal $\mathbf{Q}_1 = \mathbf{Q}$ y una matriz triangular superior $\mathbf{R}_1 = \mathbf{R}$ tales que la matriz $\mathbf{A}_1 = \mathbf{A}$ se factoriza como

$$(34) \quad \mathbf{A}_1 = \mathbf{Q}_1 \mathbf{R}_1.$$

Entonces se calcula el producto

$$(35) \quad \mathbf{A}_2 = \mathbf{R}_1 \mathbf{Q}_1.$$

Notemos que, como \mathbf{Q}_1 es ortogonal, entonces la relación (34) nos permite escribir

$$(36) \quad \mathbf{Q}'_1 \mathbf{A}_1 = \mathbf{Q}'_1 \mathbf{Q}_1 \mathbf{R}_1 = \mathbf{R}_1;$$

así que podemos expresar \mathbf{A}_2 como

$$(37) \quad \mathbf{A}_2 = \mathbf{Q}'_1 \mathbf{A}_1 \mathbf{Q}_1.$$

Puesto que $\mathbf{Q}'_1 = \mathbf{Q}_1^{-1}$, deducimos que \mathbf{A}_2 es semejante a la matriz inicial \mathbf{A}_1 y, en consecuencia, tiene los mismos autovalores (véase los comentarios hechos en la página 629 y siguientes). En el paso general, se construyen una matriz ortogonal \mathbf{Q}_k y otra triangular superior \mathbf{R}_k tales que

$$(38) \quad \mathbf{A}_k = \mathbf{Q}_k \mathbf{R}_k.$$

A continuación, se define

$$(39) \quad \mathbf{A}_{k+1} = \mathbf{R}_k \mathbf{Q}_k = \mathbf{Q}'_k \mathbf{A}_k \mathbf{Q}_k$$

y, de nuevo, tenemos que $\mathbf{Q}'_k = \mathbf{Q}_k^{-1}$, lo que implica que \mathbf{A}_{k+1} y \mathbf{A}_k son semejantes. Una consecuencia importante de esto es que \mathbf{A}_k y \mathbf{A} son semejantes y, por tanto, tienen los mismos autovalores. También se verifica, como veremos, que si \mathbf{A} es tridiagonal, entonces \mathbf{A}_k es tridiagonal para todo k . El hecho crucial es que, bajo ciertas hipótesis no demasiado restrictivas, la sucesión $\{\mathbf{A}_k\}$ converge a una matriz diagonal semejante a la matriz de partida \mathbf{A} . Los autovalores de \mathbf{A} son, entonces, los elementos de la diagonal de la matriz límite (véase la Referencia [66]).

La factorización QR

Veamos a continuación cómo puede calcularse la factorización QR de una matriz cualquiera \mathbf{A} mediante matrices de Householder. El procedimiento es similar al de la reducción a la forma tridiagonal, por lo que haremos un simple esbozo, dejando los detalles como ejercicio. Supongamos que escribimos

$$(40) \quad \mathbf{A} = \begin{bmatrix} * & * & * & \cdots & * & * & * \\ * & * & * & \cdots & * & * & * \\ * & * & * & \cdots & * & * & * \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ * & * & * & \cdots & * & * & * \\ * & * & * & \cdots & * & * & * \\ * & * & * & \cdots & * & * & * \end{bmatrix}$$

donde $*$ representa los elementos genéricos de \mathbf{A} . Sea \mathbf{X} la primera columna de \mathbf{A} . De acuerdo con el Corolario 11.3, podemos encontrar una matriz de Householder \mathbf{P}_1 de manera que todas las componentes del vector $\mathbf{P}_1\mathbf{X}$ desde la segunda hasta la última son iguales a cero. Entonces

$$(41) \quad \mathbf{P}_1\mathbf{A} = \begin{bmatrix} * & * & * & \cdots & * & * & * \\ 0 & * & * & \cdots & * & * & * \\ 0 & * & * & \cdots & * & * & * \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & * & * & \cdots & * & * & * \\ 0 & * & * & \cdots & * & * & * \\ 0 & * & * & \cdots & * & * & * \end{bmatrix}$$

(hagamos notar que algunos elementos por encima de la superdiagonal pueden ser ahora distintos de cero). De forma similar, podemos encontrar una matriz de Householder \mathbf{P}_2 de manera que los elementos de las dos primeras columnas de la matriz producto $\mathbf{P}_2\mathbf{P}_1\mathbf{A}$ que están por debajo de la diagonal principal son iguales a cero, sin que se destruyan, ya que \mathbf{P}_2 tiene una estructura como la de la expresión (17), los ceros que ya tenemos en la primera columna. Después de $n - 1$ pasos obtenemos

$$(42) \quad \mathbf{P}_{n-1} \cdots \mathbf{P}_1 \mathbf{A} = \begin{bmatrix} * & * & * & \cdots & * & * & * \\ 0 & * & * & \cdots & * & * & * \\ 0 & 0 & * & \cdots & * & * & * \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & * & * & * \\ 0 & 0 & 0 & \cdots & 0 & * & * \\ 0 & 0 & 0 & \cdots & 0 & 0 & * \end{bmatrix} = \mathbf{R}.$$

Puesto que cada matriz de Householder es ortogonal, la relación (42) implica que

$$(43) \quad \mathbf{Q} = \mathbf{P}'_1 \mathbf{P}'_2 \cdots \mathbf{P}'_{n-1}$$

es la matriz ortogonal buscada.

Dejamos como ejercicio la comprobación, usando la estructura de las matrices \mathbf{P}_k , de que la matriz

$$(44) \quad \mathbf{A}_2 = \mathbf{Q}' \mathbf{R} \mathbf{Q} = \mathbf{R} \mathbf{P}'_1 \mathbf{P}'_2 \cdots \mathbf{P}'_{n-1}$$

es también tridiagonal y simétrica (supuesto que \mathbf{A} lo es).

Aceleración mediante traslaciones

Tal y como lo acabamos de describir, el método QR funciona, pero la convergencia es lenta incluso para matrices de orden bajo; por ello, vamos a describir ahora una técnica de traslación que permite acelerar la velocidad de convergencia. Recordemos que si λ_j es un autovalor de \mathbf{A} entonces $\lambda_j - s$ es un autovalor de $\mathbf{B} = \mathbf{A} - s\mathbf{I}$. Partiendo, como antes, de la matriz simétrica y tridiagonal $\mathbf{A}_1 = \mathbf{A}$, esta idea se incorpora al método mediante la siguiente modificación del paso general: Primero se calcula la factorización

$$(45) \quad \mathbf{A}_i - s_i \mathbf{I} = \mathbf{Q}_i \mathbf{R}_i$$

y luego se calcula

$$(46) \quad \mathbf{A}_{i+1} = \mathbf{R}_i \mathbf{Q}_i \quad \text{para } i = 1, 2, \dots, k,$$

donde $\{s_i\}$ es una sucesión cuya suma aproxima el autovalor λ_j ; es decir, $\lambda_j \approx \sigma_j := s_1 + s_2 + \cdots + s_{k_j}$. Observemos que la matriz \mathbf{A}_{k_j+1} así obtenida es semejante a la matriz $\mathbf{A} - \sigma_j \mathbf{I}$.

El valor adecuado de la traslación se calcula en cada paso usando los elementos de la esquina inferior derecha de la matriz. Para el primer autovalor, digamos λ_1 , se calculan los autovalores de la submatriz de orden 2×2 extraída de la esquina inferior derecha

$$(47) \quad \begin{bmatrix} d_{n-1} & e_{n-1} \\ e_{n-1} & d_n \end{bmatrix}.$$

Los autovalores de esta matriz son las raíces x_1 y x_2 de la ecuación de segundo grado

$$(48) \quad x^2 - (d_{n-1} + d_n)x + d_{n-1}d_n - e_{n-1}e_{n-1} = 0.$$

Entonces tomamos como valor s_1 del primer desplazamiento que se usa en la relación (45), la raíz de la ecuación (48) que está más cerca de d_n .

La iteración QR con traslación se repite hasta que tengamos $e_{n-1} \approx 0$, digamos que esto ocurre en el paso k_1 , momento en el que tomamos como aproximación al primer autovalor $\lambda_1 \approx s_1 + s_2 + \dots + s_{k_1}$. Aplicamos el mismo proceso, pero ahora a la submatriz que resulta de eliminar la última fila y la última columna, hasta que obtengamos $e_{n-2} \approx 0$, lo que nos proporciona la aproximación al segundo autovalor λ_2 . Seguimos con las iteraciones, que se van aplicando a matrices cada vez más pequeñas, hasta que obtengamos $e_2 \approx 0$ y la aproximación al autovalor λ_{n-2} . Finalmente, los autovalores de la matriz de orden 2×2 restante se calculan usando la fórmula para la resolución de la ecuación de segundo grado. En el programa que damos al final pueden verse los detalles concretos.

Ejemplo 11.9. Vamos a calcular los autovalores de la matriz

$$M = \begin{bmatrix} 4 & 2 & 2 & 1 \\ 2 & -3 & 1 & 1 \\ 2 & 1 & 3 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix}.$$

En el Ejemplo 11.8 construimos una matriz tridiagonal A_1 que es semejante a M , así que aplicamos el proceso de diagonalización descrito a esta matriz:

$$A_1 = \begin{bmatrix} 4 & -3 & 0 & 0 \\ -3 & 2 & 3.16228 & 0 \\ 0 & 3.16228 & -1.4 & -0.2 \\ 0 & 0 & -0.2 & 1.4 \end{bmatrix}.$$

Los cuatro elementos de la esquina inferior derecha son $d_3 = -1.4$, $d_4 = 1.4$ y $e_3 = -0.2$ y con ellos formamos la ecuación de segundo grado

$$x^2 - (-1.4 + 1.4)x + (-1.4)(1.4) - (-0.2)(-0.2) = x^2 - 2 = 0.$$

Las raíces de esta ecuación son $x_1 = -1.41421$ y $x_2 = 1.41421$, de las cuales elegimos la más cercana a d_4 como el valor del primer desplazamiento $s_1 = 1.41421$. Entonces la primera matriz trasladada es

$$A_1 - s_1 I = \begin{bmatrix} 2.58579 & -3 & 0 & 0 \\ -3 & 0.58579 & 3.16228 & 0 \\ 0 & 3.16228 & -2.81421 & -0.2 \\ 0 & 0 & -0.2 & -0.01421 \end{bmatrix}.$$

Ahora calculamos la factorización $\mathbf{A}_1 - s_1 \mathbf{I} = \mathbf{Q}_1 \mathbf{R}_1$:

$$\mathbf{Q}_1 \mathbf{R}_1 = \begin{bmatrix} -0.65288 & -0.38859 & -0.55535 & 0.33814 \\ 0.75746 & -0.33494 & -0.47867 & 0.29145 \\ 0 & 0.85838 & -0.43818 & 0.26610 \\ 0 & 0 & 0.52006 & 0.85413 \end{bmatrix} \times \begin{bmatrix} -3.96059 & 2.40235 & 2.39531 & 0 \\ 0 & 3.68400 & -3.47483 & -0.17168 \\ 0 & 0 & -0.38457 & 0.08024 \\ 0 & 0 & 0 & -0.06550 \end{bmatrix}.$$

La matriz que resulta de multiplicar éstas en orden inverso es

$$\mathbf{A}_2 = \mathbf{R}_1 \mathbf{Q}_1 = \begin{bmatrix} 4.40547 & 2.79049 & 0 & 0 \\ 2.79049 & -4.21663 & -0.33011 & 0 \\ 0 & -0.33011 & 0.21024 & -0.03406 \\ 0 & 0 & -0.03406 & -0.05595 \end{bmatrix}.$$

El segundo desplazamiento es $s_2 = -0.06024$; calculando la segunda matriz trasladada y su factorización $\mathbf{A}_2 - s_2 \mathbf{I} = \mathbf{Q}_2 \mathbf{R}_2$, resulta

$$\mathbf{A}_3 = \mathbf{R}_2 \mathbf{Q}_2 = \begin{bmatrix} 4.55257 & -2.65725 & 0 & 0 \\ -2.65725 & -4.26047 & 0.01911 & 0 \\ 0 & 0.01911 & 0.29171 & 0.00003 \\ 0 & 0 & 0.00003 & 0.00027 \end{bmatrix}.$$

El valor de la tercera traslación es $s_3 = 0.00027$, calculando la tercera matriz trasladada y su factorización $\mathbf{A}_3 - s_3 \mathbf{I} = \mathbf{Q}_3 \mathbf{R}_3$, resulta

$$\mathbf{A}_4 = \mathbf{R}_3 \mathbf{Q}_3 = \begin{bmatrix} 4.62640 & 2.53033 & 0 & 0 \\ 2.53033 & -4.33489 & -0.00111 & 0 \\ 0 & -0.00111 & 0.29150 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

En consecuencia, el primer autovalor, redondeado con cinco cifras decimales, viene dado por

$$\lambda_1 = s_1 + s_2 + s_3 = 1.41421 - 0.06023 + 0.00027 = 1.35425.$$

Almacenamos λ_1 en la última posición diagonal de \mathbf{A}_4 y repetimos el proceso pero con la submatriz de orden 3×3 que aparece en la esquina superior izquierda de

$$\mathbf{A}_4 = \begin{bmatrix} 4.62640 & 2.53033 & 0 & 0 \\ 2.53033 & -4.33489 & -0.00111 & 0 \\ 0 & -0.00111 & 0.29150 & 0 \\ 0 & 0 & 0 & 1.35425 \end{bmatrix}.$$

Trabajando como antes, el siguiente desplazamiento reduce (con diez decimales) a cero el elemento que ocupa la posición $(3, 2)$, lo que nos proporciona

$$s_4 = 0.29150, \quad \mathbf{A}_4 - s_4 \mathbf{I} = \mathbf{Q}_4 \mathbf{R}_4, \quad \mathbf{A}_5 = \mathbf{R}_4 \mathbf{Q}_4.$$

Por consiguiente, el segundo autovalor es

$$\lambda_2 = \lambda_1 + s_4 = 1.35425 + 0.29150 = 1.64575,$$

que se almacena en la tercera posición diagonal de \mathbf{A}_5 , matriz que viene dada por

$$\mathbf{A}_5 = \begin{bmatrix} 4.26081 & -2.65724 & 0 & 0 \\ -2.65724 & -4.55232 & 0 & 0 \\ 0 & 0 & 1.64575 & 0 \\ 0 & 0 & 0 & 1.35425 \end{bmatrix}.$$

En el último paso, hay que calcular los autovalores de la matriz de orden 2×2 que ocupa la esquina superior izquierda de \mathbf{A}_5 . La ecuación característica es

$$x^2 + (4.26081 - 4.55232)x + (4.26081)(-4.55232) - (2.65724)(2.65724) = 0,$$

o sea,

$$x^2 + 0.29151x - 26.45749 = 0.$$

Las raíces de esta ecuación son $x_1 = 5.00000$ y $x_2 = -5.29150$, así que los dos últimos autovalores vienen dados por

$$\lambda_3 = \lambda_2 + x_1 = 1.64575 + 5.00000 = 6.64575$$

y

$$\lambda_4 = \lambda_2 + x_2 = 1.64575 - 5.29150 = -3.64575. \blacksquare$$

MATLAB

El Programa 11.5 que damos a continuación puede usarse para aproximar los autovalores de una matriz tridiagonal y simétrica. El programa se ha implementado siguiendo los pasos descritos previamente, salvo tres notables excepciones: Primero, que las raíces de la ecuación característica (48) de cada submatriz de orden 2×2 dada en (47) se calculan usando la instrucción `eig` del paquete MATLAB; segundo, que la factorización $\mathbf{A}_i - s_i \mathbf{I} = \mathbf{Q}_i \mathbf{R}_i$ dada en (45) se calcula usando la instrucción `[Q,R]=qr(B)` del paquete MATLAB, que proporciona una matriz ortogonal \mathbf{Q} y una matriz triangular superior \mathbf{R} , tales que $\mathbf{B}=\mathbf{Q}*\mathbf{R}$ y, tercero, que cada traslación mediante un s_i se deshace inmediatamente, no al final; o sea, en vez del paso descrito en (46) lo que se hace es

$$\mathbf{A}_{i+1} = \mathbf{R}_i \mathbf{Q}_i + s_i \mathbf{I} \quad \text{para } i = 1, 2, \dots, k_j,$$

con lo que no hace falta sumar todos los desplazamientos para calcular la aproximación a λ_j , esta aproximación es, directamente, el elemento que está en la diagonal.

Programa 11.5 (El método QR con traslaciones). Construcción de aproximaciones a los autovalores de una matriz tridiagonal y simétrica A mediante el método QR con traslaciones sucesivas.

```

function D=qr2(A,epsilon)
% Datos
% - A es una matriz simétrica y tridiagonal de orden n x n
% - epsilon es la tolerancia
% Resultado
% - D es un vector de orden n x 1 que contiene
%   los autovalores
% Inicialización de los parámetros
[n,n]=size(A);
m=n;
D=zeros(n,1);
B=A;
while (m>1)
    while (abs(B(m,m-1))>=epsilon)
        % Cálculo del valor de desplazamiento
        S=eig(B(m-1:m,m-1:m));
        [j,k]=min([abs(B(m,m)*[1 1]'-S)]);
        % Factorización B=QR
        [Q,R]=qr(B-S(k)*eye(m));
        % Siguiente matriz B
        B=R*Q+S(k)*eye(m);
    end
    % Colocamos el m-ésimo autovalor en la posición A(m,m)
    A(1:m,1:m)=B;
    % Repetición del proceso en la submatriz m-1 x m-1 de A
    m=m-1;
    B=A(1:m,1:m);
end
D=diag(A);

```

Ejercicios

- Explique con cuidado por qué, en la demostración del Teorema 11.23, los vectores Z y W son perpendiculares.
- Pruebe que si X es un vector cualquiera, entonces la matriz $P = I - 2XX'$ es simétrica.

3. Dado un vector \mathbf{X} cualquiera, sea $\mathbf{P} = \mathbf{I} - 2\mathbf{XX}'$.

(a) Calcule $\mathbf{P}'\mathbf{P}$.

(b) ¿Qué condición adicional nos garantiza que \mathbf{P} es una matriz ortogonal?

Algoritmos y programas

En los Problemas 1 a 6 use:

(a) El Programa 11.4 para reducir la matriz dada a forma tridiagonal.

(b) El Programa 11.5 para calcular los autovalores de la matriz dada.

1. $\begin{bmatrix} 3 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 3 \end{bmatrix}$

2. $\begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}$

3. $\begin{bmatrix} 2.75 & -0.25 & -0.75 & 1.25 \\ -0.25 & 2.75 & 1.25 & -0.75 \\ -0.75 & 1.25 & 2.75 & -0.25 \\ 1.25 & -0.75 & -0.25 & 2.75 \end{bmatrix}$

4. $\begin{bmatrix} 3.6 & 4.4 & 0.8 & -1.6 & -2.8 \\ 4.4 & 2.6 & 1.2 & -0.4 & 0.8 \\ 0.8 & 1.2 & 0.8 & -4.0 & -2.8 \\ -1.6 & -0.4 & -4.0 & 1.2 & 2.0 \\ -2.8 & 0.8 & -2.8 & 2.0 & 1.8 \end{bmatrix}$

5. $\mathbf{A} = [a_{ij}]$ con $a_{ij} = \begin{cases} i + j & \text{si } i = j, \\ ij & \text{si } i \neq j, \end{cases} \quad i, j = 1, 2, \dots, 30.$

6. $\mathbf{A} = [a_{ij}]$ con $a_{ij} = \begin{cases} \cos(\operatorname{sen}(i + j)) & \text{si } i = j, \\ i + ij + j & \text{si } i \neq j, \end{cases} \quad i, j = 1, 2, \dots, 40.$

7. Escriba un programa que lleve a cabo la factorización QR de una matriz simétrica.

8. Modifique el Programa 11.5 de manera que utilice su programa del Problema 7 como subprograma. Use su programa modificado para hallar los autovalores de las matrices de los Problemas 1 a 6.

Apéndice: MATLAB

Este apéndice es una introducción a las técnicas de programación con el paquete de programas MATLAB. Suponemos que quien lee este libro tiene ya alguna experiencia de programación con lenguajes de alto nivel y, en particular, conoce algunas técnicas esenciales como los bucles, la ramificación mediante relaciones lógicas, el empleo de subprogramas y la edición de archivos con un procesador de textos. Estas técnicas son directamente aplicables cuando se usa el paquete MATLAB.

El paquete MATLAB es un conjunto de programas matemáticos que se basa en el empleo de matrices. El paquete consta de una amplia colección de programas numéricos y programas gráficos para dibujos bi- y tridimensionales e incluye la posibilidad de realizar programas adicionales usando un lenguaje de alto nivel. También es posible desarrollar y modificar los programas de manera muy sencilla. Todo esto hace que el paquete MATLAB sea un banco de trabajo ideal para explorar y trabajar con los algoritmos que se explican en este libro.

Lo que sigue es una introducción guiada a la programación con el paquete MATLAB y nuestra sugerencia es que se trabaje a lo largo de las líneas que aquí se explican (las instrucciones del paquete MATLAB se escriben con el tipo de letra **máquina de escribir**). Los ejemplos ilustran lo que nos muestra la ventana de trabajo (en inglés, *command window*) del paquete MATLAB en una sesión típica. Lo que aparece a continuación de **>>** es lo que se introduce como dato o instrucción. Una vez que se escribe lo que se desea, hay que pulsar la tecla de retorno; entonces el computador realiza la operación y muestra la respuesta **ans =**. En las guías de uso y de referencia que acompañan el programa, así

como en el programa de ayuda que puede consultarse directamente en la ventana de trabajo, se puede hallar información adicional sobre las instrucciones y sus opciones y, también, varios ejemplos.

Operaciones aritméticas

+	Sumar
-	Restar
*	Multiplicar
/	Dividir
ⁿ	Elevar a una potencia
pi, e, i	Constantes

Ejemplo. `>>(2+3*pi)/2`

```
ans =
      5.7124
```

Funciones incorporadas al programa

Damos a continuación una breve lista de algunas de las funciones disponibles en el paquete MATLAB; el programa de ayuda proporciona información sobre qué otras funciones hay disponibles. El ejemplo ilustra cómo se usan y se combinan las operaciones aritméticas y las funciones.

<code>abs(#)</code>	<code>cos(#)</code>	<code>exp(#)</code>	<code>log(#)</code>	<code>log10(#)</code>	<code>cosh(#)</code>
<code>sin(#)</code>	<code>tan(#)</code>	<code>sqrt(#)</code>	<code>floor(#)</code>	<code>acos(#)</code>	<code>tanh(#)</code>

Ejemplo. `>>3*cos(sqrt(4.7))`

```
ans =
      -1.6869
```

En la respuesta se muestran, habitualmente, cinco cifras decimales significativas; la instrucción `format long` nos permite obtener hasta 15 cifras decimales significativas.

Ejemplo. `>>format long`

```
3*cos(sqrt(4.7))
ans =
      -1.68686892236893
```

InSTRUCCIONES DE ASIGNACIÓN

Mediante el signo de igualdad podemos asignar un nombre al resultado de la evaluación de una expresión.

Ejemplo. `>>a=3-floor(exp(2.9))`

```
a=
-15
```

Cuando se escribe un punto y coma al final de una expresión, el computador realiza las operaciones correspondientes y almacena su resultado bajo el nombre que le hayamos asignado (para su uso en cálculos posteriores) pero no muestra el resultado en la pantalla.

Ejemplo. `>>b=sin(a);` Nota: b no se muestra.
`>>2*b^2`
`ans=`
`0.8457`

Definición de nuevas funciones

Es posible definir nuevas funciones que se pueden usar con el paquete MATLAB escribiendo con un editor un archivo de texto cuya extensión sea `m`. Una vez definido, puede utilizarse como cualquier otra función.

Ejemplo. Vamos a definir la función $f(x) = 1 + x - x^2/4$ en un archivo `fun.m`. Para ello, con el editor de textos, escribimos:

```
function y=fun(x)
y=1+x-x.^2/4;
```

Explicaremos el uso de “`.`” un poco más adelante. Podemos usar letras distintas para las variables y podemos darle un nombre distinto a la función, pero el formato debe ser el mismo. Una vez que almacenamos esta función en un archivo llamado `fun.m`, podemos usarla como cualquier otra función del paquete MATLAB.

```
>>cos(fun(3))
ans=
-0.1782
```

La instrucción `feval` nos permite evaluar funciones de una manera útil y eficiente. Cuando se usa, la función se invoca como una cadena de caracteres.

Ejemplo. `>>feval('fun',4)`
`ans=`
`1`

Matrices

En el paquete MATLAB todas las variables son matrices. Las matrices se introducen de una manera directa:

Ejemplo. `>>A=[1 2 3;4 5 6;7 8 9]`

`A=`

```
1 2 3
4 5 6
7 8 9
```

Los puntos y comas separan las filas de la matriz, mientras que los elementos de la misma fila deben separarse mediante un espacio en blanco (o una coma). Alternativamente, podemos introducir las matrices fila a fila:

Ejemplo. `>>A=[1 2 3`

```
4 5 6
7 8 9]
```

`A =`

```
1 2 3
4 5 6
7 8 9
```

Podemos generar algunas matrices especiales usando funciones ya incorporadas:

Ejemplo. `>>Z=zeros(3,5);`

crea una matriz de ceros de orden 3×5

`>>X=ones(3,5);`

crea una matriz de unos de orden 3×5

`>>Y=0:0.5:2`

crea la matriz de orden 1×5 siguiente

`Y=`

```
0 0.5000 1.0000 1.5000 2.0000
```

`>>sin(Y)`

crea una matriz de orden 1×5 tomando el seno de cada elemento de Y

`ans=`

```
1.0000 0.8776 0.5403 0.0707 -0.4161
```

Podemos trabajar con los elementos de una matriz de diversas maneras.

Ejemplo. `>>A(2,3)`

selecciona una entrada concreta de A

`ans=`

6

`>>A(1:2,2:3)`

selecciona una submatriz de A

`ans=`

2 3

5 6

`>>A([1 3],[1 3])`

otra forma de seleccionar una submatriz de A

`ans=`

1 3

7 9

`>>A(2,2)=tan(7.8);` asigna un nuevo valor a una entrada concreta de A

El programa de ayuda y la documentación que acompañan el paquete proporcionan información sobre otras funciones matriciales disponibles.

Operaciones con matrices

+	Sumar
-	Restar
*	Multiplicar
[^]	Elevar a una potencia
,	Traspuesta conjugada

Ejemplo. `>>B=[1 2;3 4];`

`>>C=B'` C es la traspuesta de B

`C=`

1 3

2 4

`>>3*(B*C)^3` $3(BC)^3$

`ans=`

13080 29568

29568 66840

Operaciones que se realizan elemento a elemento

Una de las características más útiles del paquete MATLAB es que dispone de un gran número de funciones que operan sobre una matriz elemento a elemento; vimos antes un ejemplo de esto cuando tomamos el seno de cada elemento de una matriz de orden 1×5 . Las operaciones matriciales de suma, resta y producto por un escalar se realizan elemento a elemento, lo que no ocurre con las operaciones matriciales de multiplicación, división y potenciación. Estas tres operaciones pueden realizarse elemento a elemento si anteponemos un punto al símbolo correspondiente: `.*`, `./` y `.^`. Es importante el entender cómo y cuándo deben usarse estas operaciones ya que las operaciones elemento a elemento son cruciales a la hora de diseñar e implementar eficientemente programas numéricos y gráficos con el paquete MATLAB.

Ejemplo. `>>A=[1 2;3 4];A^2`

calcula el producto AA

`ans=`

7 10

15 22

`>>A.^2`

eleva al cuadrado cada elemento de A

`ans=`

1 4

9 16

```
>>cos(A./2)           divide cada elemento de A entre 2 y,
                           después, calcula el coseno
ans=
0.8776  0.5403
0.0707 -0.4161
```

Gráficos

El paquete MATLAB puede producir dibujos bi- y tridimensionales de curvas y superficies. Las diversas opciones y aspectos adicionales de las instrucciones gráficas básicas pueden consultarse en el paquete de ayuda o en la documentación.

La instrucción `plot` permite generar gráficas de curvas planas. En el siguiente ejemplo se muestra cómo podemos obtener las gráficas de las funciones $y = \cos(x)$ e $y = \cos^2(x)$ en el intervalo $[0, \pi]$.

Ejemplo. `>>x=0:0.1:pi;`
`>>y=cos(x);`
`>>z=cos(x).^2;`
`>>plot(x,y,x,z, 'o')`

En la primera línea se especifican el dominio y el tamaño de paso 0.1. En las dos líneas siguientes se definen las funciones. Hagamos notar que las tres primeras líneas terminan con un punto y coma; este punto y coma se escribe para evitar que aparezcan en la pantalla los treinta y tantos elementos de cada una de las matrices `x`, `y` y `z`. La cuarta línea contiene la instrucción de dibujo que produce las gráficas. Los dos primeros términos, `x` e `y`, dibujan la función $y = \cos(x)$. Los términos tercero y cuarto, `x` y `z`, dibujan la función $y = \cos^2(x)$. El último término, 'o', hace que se dibuje una 'o' en cada punto (x_k, z_k) con $z_k = \cos^2(x_k)$.

El uso en la tercera fila del indicador de operación elemento a elemento "`.^`" es esencial: primero se calcula el coseno de cada elemento de la matriz `x` y, después, cada elemento de la matriz `cos(x)` se eleva al cuadrado usando la instrucción `.^`.

La instrucción de dibujo `fplot` es una alternativa útil a la instrucción `plot`. La sintaxis de esta instrucción es `fplot('nombre',[a,b],n)`, que produce la gráfica de la función `nombre.m` determinando su valor en `n` puntos del intervalo $[a, b]$. Si no se especifica otra cosa, el valor de `n` es 25.

Ejemplo. `>>fplot('tanh', [-2, 2])` dibuja $y = \tanh(x)$ en $[-2, 2]$

Las instrucciones `plot` y `plot3` se utilizan para dibujar curvas parametrizadas en el espacio bi- y tridimensional, respectivamente. Estas instrucciones son especialmente útiles para visualizar las soluciones de una ecuación diferencial en dimensión dos y tres.

Ejemplo. El dibujo de la elipse $c(t) = (2 \cos(t), 3 \sin(t))$, con $0 \leq t \leq 2\pi$, se obtiene con las siguientes instrucciones:

```
>>t=0:0.2:2*pi;
>>plot(2*cos(t),3*sin(t))
```

Ejemplo. El dibujo de la curva $c(t) = (2 \cos(t), t^2, 1/t)$, con $0.1 \leq t \leq 4\pi$, se obtiene con las siguientes instrucciones:

```
>>t=0.1:0.1:4*pi;
>>plot3(2*cos(t),t.^2,1./t)
```

Para obtener dibujos tridimensionales de superficies hay que especificar un rectángulo del dominio de la función, mediante la instrucción `meshgrid`, y luego las instrucciones `mesh` o `surf` para obtener la gráfica. Estas instrucciones son útiles para visualizar la solución de una ecuación en derivadas parciales.

Ejemplo.

```
>>x=-pi:0.1:pi;
>>y=x;
>>[x,y]=meshgrid(x,y);
>>z=sin(cos(x+y));
>>mesh(z)
```

Bucles y ramificaciones

Operadores de relación

<code>==</code>	Igual que
<code>~=</code>	No igual que
<code><</code>	Menor que
<code>></code>	Mayor que
<code><=</code>	Menor o igual que
<code>>=</code>	Mayor o igual que

Operadores lógicos

<code>~</code>	No	(Verdadero si, y sólo si, la proposición es falsa)
<code>&</code>	Y	(Verdadero si las dos proposiciones son verdaderas)
<code> </code>	O	(Verdadero si alguna de las dos proposiciones es verdadera)

Valores booleanos

<code>1</code>	Verdadero
<code>0</code>	Falso

Las instrucciones `for`, `if` y `while` del paquete MATLAB operan de manera similar a sus homólogas en otros lenguajes de programación. Estas instrucciones adoptan la sintaxis básica siguiente:

```

for (variable del bucle = rango del bucle)
    instrucciones ejecutables
end

if (premisa)
    instrucciones ejecutables
else
    instrucciones ejecutables
end

while (premisa)
    instrucciones ejecutables
end

```

En el siguiente ejemplo mostramos cómo se pueden encajar varios bucles para generar una matriz. Guardando las líneas de texto en un archivo llamado `nido.m`, entonces cada vez que escribamos `nido` en la ventana de trabajo del paquete MATLAB obtendremos la matriz `A`. Hagamos notar que los elementos de la matriz `A` forman, empezando en la esquina superior izquierda, el triángulo de Pascal.

Ejemplo.

```

for i=1:5
    A(i,1)=1;A(1,i)=1;
end
for i=2:5
    for j=2:5
        A(i,j)=A(i,j-1)+A(i-1,j);
    end
end
A

```

Para salir de un bucle antes de que se complete, se usa la instrucción `break`.

Ejemplo.

```

for k=1:100
    x=sqrt(k);
    if ((k>10)&(x-floor(x)==0))
        break
    end
end
k

```

Para mostrar una línea de texto o una matriz se utiliza la instrucción `disp`.

Ejemplo.

```

n=10;
k=0;
while k<=n
    x=k/3; disp([x x^2 x^3]), k=k+1;
end

```

Programas

Una forma eficiente de construir programas es crear nuevas funciones que se almacenan como archivos cuya extensión es `m`. Estos programas nos permiten especificar los datos que deben introducirse y los resultados que deben mostrarse y pueden ser llamados como subprogramas desde otros programas. El siguiente ejemplo nos permite visualizar el efecto de calcular los elementos del triángulo de Pascal previa modulación con un número primo. Para ello, hay que escribir las siguientes líneas con el editor de textos del paquete y almacenarlas en un archivo llamado `pasc.m`.

Ejemplo.

```
function P=pasc(n,m)
    % Datos      - n es la cantidad de filas
    %             - m es el número primo
    % Resultado - P es el triángulo de Pascal

    for j=1:n
        P(j,1)=1;P(1,j)=1;
    end
    for k=2:n
        for j=2:n
            P(k,j)=rem(P(k,j-1),m)+rem(P(k-1,j),m);
        end
    end
```

Ahora, escribiendo en la ventana de trabajo la instrucción `P=pasc(5,3)`, veremos las cinco primeras filas del triángulo de Pascal módulo 3. Podemos intentar también `P=pasc(175,3)`; (use el punto y coma) y luego `spy(P)` (que genera una matriz dispersa para valores grandes de `n`).

Conclusión

Alcanzado este punto, usted debería ser capaz de crear y modificar programas basados en los algoritmos dados en este libro. Para obtener información adicional sobre las funciones del paquete MATLAB o sobre cómo se utiliza en su computador particular, debe usted utilizar el programa de ayuda o la documentación que acompaña el paquete.

Referencias temáticas

Ofrecemos aquí una lista de referencias que pueden servir de guía para que los estudiantes realicen trabajos de profundización en algunos aspectos específicos del análisis numérico.

Algoritmo de Remes [9, 19, 56, 88, 128, 149, 152, 153]

Algoritmo del cociente de las diferencias [3, 29, 62, 78, 79, 86, 112, 152, 200]

Algoritmo *QR* [3, 9, 10, 19, 29, 40, 41, 74, 85, 92, 97, 104, 128, 152, 153, 169, 175, 192, 203]

Aproximación de funciones [34, 44, 114, 149, 157, 161, 182]

Aritmética de coma flotante [8, 9, 35, 40, 41, 51, 57, 62, 90, 101, 103, 128, 129, 142, 153, 181, 184, 208]

Cerchas Básicas (en inglés, *B-splines*) [35, 96, 101, 149, 160]

Computación científica [5, 71, 98, 103, 150, 151, 152, 158, 159, 160]

Computadores en la enseñanza del cálculo infinitesimal [13, 18, 36, 55, 110, 111, 120, 122, 134, 162, 176, 179]

Construcción de modelos matemáticos [15, 17, 22, 23, 32, 39, 42, 64, 72, 83, 95, 98, 102, 104, 107, 113, 115, 116, 131, 135, 136, 190]

Control del tamaño de paso en ecuaciones diferenciales [29, 40, 60, 75, 101, 117, 160]

Corrección iterativa de errores [8, 9, 19, 29, 40, 41, 49, 51, 58, 72, 90, 94, 96, 97, 117, 137, 152, 153, 160]

- Economización de series de potencias [3, 9, 29, 41, 51, 62, 76, 85, 88, 117, 153, 184]
- Ecuaciones diferenciales [7, 31, 33, 39, 42, 99, 104, 136, 138, 152, 171, 173]
- Ecuaciones diferenciales rígidas [9, 29, 40, 57, 60, 98, 117, 152, 153, 160, 173]
- Errores de redondeo [4, 9, 29, 35, 41, 51, 76, 79, 81, 90, 94, 101, 117, 128, 146, 153, 160, 181, 184, 186, 204]
- Estabilidad de ecuaciones diferenciales [3, 8, 9, 29, 40, 60, 76, 78, 79, 96, 101, 128, 146, 152, 153, 160]
- Estrategias de pivoteo [9, 29, 35, 40, 41, 58, 79, 96, 101, 117, 128, 145, 146, 152, 153, 160]
- Extrapolación [19, 29, 35, 40, 41, 78, 117, 153]
- Factorización de Choleski [9, 29, 40, 41, 51, 90, 97, 152, 153, 160]
- Fórmulas de diferencias progresivas [9, 29, 40, 41, 51, 76, 78, 81, 85, 90, 94, 105, 117, 128, 143, 145, 153, 181, 184]
- Fórmulas de Newton-Cotes [9, 29, 62, 76, 78, 81, 90, 94, 97, 105, 117, 126, 128, 152, 153, 154, 160, 175, 193, 208]
- Integrales múltiples [29, 62, 67, 85, 96, 112, 117, 152, 153]
- Interpolación de Hermite [9, 29, 40, 41, 79, 81, 90, 92, 128, 153, 191, 193, 208]
- Interpolación inversa [9, 19, 29, 35, 41, 62, 81, 128, 153, 166, 181, 191]
- Interpolación iterada [29, 78, 81, 90, 126, 128, 129, 181, 184, 208]
- Matrices mal condicionadas [9, 19, 29, 40, 41, 47, 49, 62, 94, 101, 128, 145, 153, 192, 197]
- Método de Gauss-Jordan [29, 44, 51, 62, 79, 85, 90, 117, 152]
- Método de la secante (convergencia) [9, 35, 40, 41, 153, 160]
- Método de Monte Carlo [35, 41, 57, 76, 83, 87, 98, 112, 115, 135, 152, 154]
- Métodos cuasi-Newton [29, 96, 97, 139, 152, 153]
- Métodos de relajación [19, 29, 40, 41, 62, 90, 139, 152, 199, 207]
- Métodos de sobrerrelajación sucesiva [10, 29, 40, 41, 49, 137, 139, 152, 160, 175, 199, 207]
- Métodos numéricos en la ingeniería [6, 17, 20, 31, 33, 39, 54, 59, 71, 88, 93, 104, 131, 136, 141, 163, 174, 183, 190, 195]
- Mínimos cuadrados [39, 92, 109, 112, 152]
- Normas de vectores y matrices [9, 19, 29, 40, 49, 62, 90, 94, 96, 101, 117, 128, 145, 153, 192]
- Número de condición de una matriz [9, 19, 29, 40, 41, 57, 62, 74, 94, 96, 98, 101, 117, 128, 145, 152, 153, 160, 192]
- Números hexadecimales [8, 35, 51, 101, 142]

- Paquetes de programas para el análisis numérico [32, 52, 82, 84, 95, 97, 98, 124, 125, 150, 151, 152, 158, 159, 160, 178]
- Pérdida de cifras significativas (cancelación) [3, 8, 35, 40, 79, 142]
- Polinomios de Legendre [9, 29, 40, 41, 75, 152, 153]
- Polinomios ortogonales [9, 19, 29, 34, 40, 41, 44, 76, 81, 90, 96, 126, 128, 143, 145, 149, 152, 153, 169]
- Programación [12, 103, 119, 150, 151, 152]
- Programación lineal (método del simplex) [19, 27, 35, 37, 41, 44, 50, 53, 79, 83, 94, 104, 115, 135, 152, 153, 154, 165, 169]
- Propagación de errores [4, 9, 40, 41, 49, 51, 78, 79, 81, 133, 142, 145, 153, 204]
- Sistemas de ecuaciones con estructura de banda [29, 35, 41, 128, 160, 192]
- Sistemas dinámicos [2, 17, 48, 164]
- Sistemas lineales [61, 66, 74, 82, 152, 159]
- Transformada rápida de Fourier [25, 29, 33, 40, 51, 62, 79, 96, 98, 112, 136, 141, 145, 149, 150, 152, 153, 155, 169, 210]

Bibliografía y referencias

1. Aberth, Oliver (1988). *Precise Numerical Analysis*, Wm. C. Brown, Dubuque, Ia.
2. Aburdene, Maurice F. (1988). *Computer Simulation of Dynamic Systems*, Wm. C. Brown, Dubuque, Ia.
3. Acton, Forman S. (1970). *Numerical Methods That Work*, Harper & Row, Nueva York.
4. Adby, P. R. y M. A. H. Dempster (1974). *Introduction to Optimization Methods*, Halsted Press, Nueva York.
5. Aho, Alfred V., John E. Hopcroft y Jeffrey D. Ullman (1974). *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, Mass.
6. Al-Khafaji, Amir Wadi y John R. Tooley (1986). *Numerical Methods in Engineering Practice*, Holt, Rinehart and Winston, Nueva York.
7. Ascher, Uri M., Robert M. M. Mattheij y Robert D. Russell (1988). *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Prentice Hall, Englewood Cliffs, N. J.
8. Atkinson, Kendall E. (1985). *Elementary Numerical Analysis*, John Wiley, Nueva York.
9. Atkinson, Kendall E. (1988). *An Introduction to Numerical Analysis*, 2^a ed., John Wiley, Nueva York.
10. Atkinson, Laurence V. y P. J. Harley (1983). *An Introduction to Numerical Methods with Pascal*, Addison-Wesley, Reading, Mass.

11. Bailey, Paul B., Lawrence F. Shampine y Paul E. Waltman (1968). *Non-linear Two Point Boundary Value Problems*, Academic Press, Nueva York.
12. Barnard, David T. y Robert G. Crawford (1982). *Pascal Programming Problems and Applications*, Reston, Va.
13. Beckman, Charlene E. y Ted Sundstrom (1990). *Graphing Calculator Laboratory Manual for Calculus*, Addison-Wesley, Reading, Mass.
14. Bender, Carl M. y Steven A. Orszag (1978). *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill, Nueva York.
15. Bender, Edward A. (1978). *An Introduction to Mathematical Modeling*, John Wiley, Nueva York.
16. Bennett, William Ralph (1976). *Introduction to Computer Applications for Nonscience Students*, Prentice Hall, Englewood Cliffs, N. J.
17. Bennett, William Ralph (1976). *Scientific and Engineering Problem-Solving with the Computer*, Prentice Hall, Englewood Cliffs, N. J.
18. Bitter, Gary G. (1983). *Microcomputer Applications for Calculus*, Prindle, Weber & Schmidt, Boston.
19. Blum, E. K. (1972). *Numerical Analysis and Computation: Theory and Practice*, Addison-Wesley, Reading, Mass.
20. Borse, G. J. (1985). *FORTRAN 77 and Numerical Methods for Engineers*, Prindle, Weber & Schmidt, Boston.
21. Brainerd, Walter S. y Lawrence H. Landweber (1974). *Theory of Computation*, John Wiley, Nueva York.
22. Brams, Steven J., William F. Lucas y Philip D. Straffin, eds. (1983). *Political and Related Models*, Springer-Verlag, Nueva York.
23. Braun, Martin, Courtney S. Coleman y Donald A. Drew, eds. (1983). *Differential Equation Models*, Springer-Verlag, Nueva York.
24. Brent, Richard P. (1973). *Algorithms for Minimization without Derivatives*, Prentice Hall, Englewood Cliffs, N. J.
25. Brigham, E. Oran (1988). *The Fast Fourier Transform and Its Applications*, Prentice Hall, Englewood Cliffs, N. J.
26. Buck, R. Creighton (1978). *Advanced Calculus*, 3^a ed., McGraw-Hill, Nueva York. Versión española (1968): *Cálculo Superior*, Ed. del Castillo, Madrid.
27. Bunday, Brian D. (1984). *Basic Linear Programming*, Edward Arnold, Baltimore, Md.
28. Bunday, Brian D. (1984). *Basic Optimisation Methods*, Edward Arnold, Baltimore, Md.

29. Burden, Richard L. y J. Douglas Faires (1985). *Numerical Analysis*, 3^a ed., Prindle, Weber & Schmidt, Boston. Versiones españolas (1996 y 1998): *Análisis Numérico*, Grupo Editorial Iberoamericana e Int. Thompson, México.
30. Burnett, David S. (1987). *Finite Element Analysis: From Concepts to Applications*, Addison-Wesley, Reading, Mass.
31. Carnahan, Brice, H. A. Luther y James O. Wilkes (1969). *Applied Numerical Methods*, John Wiley, Nueva York.
32. Carroll, John M. (1987). *Simulation Using Personal Computers*, Prentice Hall, Englewood Cliffs, N. J.
33. Chapra, Steven C. (1985). *Numerical Methods for Engineers: With Personal Computer Applications*, McGraw-Hill, Nueva York.
34. Cheney, Ward (1966). *Introduction to Approximation Theory*, McGraw-Hill, Nueva York.
35. Cheney, Ward y David Kincaid (1985). *Numerical Mathematics and Computing*, 2^a ed. Brooks/Cole, Monterey, Calif. Versión española (1994): *Análisis Numérico*, Addison-Wesley Iberoamericana, Wilmington, De.
36. Christensen, Mark J. (1981). *Computing for Calculus*, Academic Press, Nueva York.
37. Chvatal, Vasek (1980). *Linear Programming*, W. H. Freeman, Nueva York.
38. Coddington, Earl A. y Norman Levinson (1955). *Theory of Ordinary Differential Equations*, McGraw-Hill, Nueva York.
39. Constantinides, Alkis (1987). *Applied Numerical Methods with Personal Computers*, McGraw-Hill, Nueva York.
40. Conte, S. D. y Carl de Boor (1980). *Elementary Numerical Analysis: An Algorithmic Approach*, McGraw-Hill, Nueva York.
41. Dahlquist, Germund y Ake Bjorck (1974). *Numerical Methods*, Prentice Hall, Englewood Cliffs, N. J.
42. Danby, J. M. A. (1985). *Computing Applications to Differential Equations: Modelling in the Physical and Social Sciences*, Reston, Va.
43. Daniels, Richard W. (1978). *An Introduction to Numerical Methods and Optimization Techniques*, North-Holland, Nueva York.
44. Davis, Philip J. (1963). *Interpolation and Approximation*, Blaisdell, Nueva York.
45. Davis, Philip J. y Philip Rabinowitz (1984). *Methods of Numerical Integration*, 2^a ed., Academic Press, Nueva York.
46. deBoor, Carl (1978). *A Practical Guide to Splines*, Springer-Verlag, Nueva York.

47. Deif, Assem S. (1986). *Sensitivity Analysis in Linear Systems*, Springer-Verlag, Nueva York.
48. Devaney, Robert L. (1990). *Chaos, Fractals and Dynamics: Computer Experiments in Mathematics*, Addison-Wesley, Reading, Mass.
49. Dew, P. M. y K. R. James (1983). *Introduction to Numerical Computation in Pascal*, Springer-Verlag, Nueva York.
50. Dixon, L. C. W. (1972). *Nonlinear Optimisation*, Crane, Russak & Co., Nueva York.
51. Dodes, Irving Allen (1978). *Numerical Analysis for Computer Science*, North-Holland, Nueva York.
52. Dongarra, J. J. (1979). *LINPACK: Users' Guide*, SIAM, Filadelfia.
53. Dorn, William S. y Daniel D. McCracken (1976). *Introductory Finite Mathematics with Computing*, John Wiley, Nueva York.
54. Dorn, William S. y Daniel D. McCracken (1972). *Numerical Methods with Fortran IV Case Studies*, John Wiley, Nueva York.
55. Edwards, C. H. (1986). *Calculus and the Personal Computer*, Prentice Hall, Englewood Cliffs, N. J.
56. Fike, C. T. (1968). *Computer Evaluation of Mathematical Functions*, Prentice Hall, Englewood Cliffs, N. J.
57. Forsythe, George E., Michael A. Malcolm y Cleve B. Moler (1977). *Computer Methods for Mathematical Computations*, Prentice Hall, Englewood Cliffs, N. J.
58. Forsythe, George E. y Cleve B. Moler (1967). *Computer Solution of Linear Algebraic Systems*, Prentice Hall, Englewood Cliffs, N. J.
59. Fox, L. y D. F. Mayers (1968). *Computing Methods for Scientists and Engineers*, Oxford University Press, Nueva York.
60. Gear, C. William (1971). *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice Hall, Englewood Cliffs, N. J.
61. George, Alan y Joseph W. H. Liu (1981). *Computer Solution of Large Sparse Positive Definite Systems*, Prentice Hall, Englewood Cliffs, N. J.
62. Gerald, Curtis F. y Patrick O. Wheatley (1984). *Applied Numerical Analysis*, 3^a ed., Addison-Wesley, Reading, Mass.
63. Gill, Philip E., Walter Murray y Margaret H. Wright (1981). *Practical Optimization*, Academic Press, Nueva York.
64. Giordano, Frank R. y Maurice D. Weir (1985). *A First Course in Mathematical Modeling*, Brooks/Cole, Monterey, Calif.
65. Goldstine, Herman H. (1977). *A History of Numerical Analysis from the 16th through the 19th Century*, Springer-Verlag, Nueva York.

66. Golub, Gene H. y Charles F. VanLoan (1989). *Matrix Computations*, The Johns Hopkins University Press, Baltimore, Md.
67. Gordon, Sheldon P. (1986). Simpson's Rule for Double Integrals, *UMAP Journal*, Vol. 7, N. 4, pp. 319–328.
68. Gourlay, A. R. y G. A. Watson (1973). *Computational Methods for Matrix Eigenproblems*, John Wiley, Nueva York.
69. Greenspan, Donald y Vincenzo Casulli (1988). *Numerical Analysis for Applied Mathematics, Science and Engineering*, Addison-Wesley, Reading, Mass.
70. Grove, Wendell E. (1966). *Brief Numerical Methods*, Prentice Hall, Englewood Cliffs, N. J.
71. Guggenheim, H. (1987). *BASIC Mathematical Programs for Engineers and Scientists*, Petrocelli Books, West Hempstead, N.Y.
72. Haberman, Richard (1977). *Mathematical Models: Mechanical Vibrations, Population Dynamics and Traffic Flow*, Prentice Hall, Englewood Cliffs, N. J.
73. Hageman, Louis A. y David M. Young (1981). *Applied Iterative Methods*, Academic Press, Nueva York.
74. Hager, William W. (1988). *Applied Numerical Linear Algebra*, Prentice Hall, Englewood Cliffs, N. J.
75. Hamming, Richard W. (1971). *Introduction to Applied Numerical Analysis*, McGraw-Hill, Nueva York.
76. Hamming, Richard W. (1973). *Numerical Methods for Scientists and Engineers*, 2^a ed., McGraw-Hill, Nueva York.
77. Henrici, Peter (1974). *Applied and Computational Complex Analysis*, Vol. 1, John Wiley, Nueva York.
78. Henrici, Peter (1964). *Elements of Numerical Analysis*, John Wiley, Nueva York.
79. Henrici, Peter (1982). *Essentials of Numerical Analysis with Pocket Calculator Demonstrations*, John Wiley, Nueva York.
80. Hildebrand, Francis B. (1976). *Advanced Calculus for Applications*, 2^a ed., Prentice Hall, Englewood Cliffs, N. J.
81. Hildebrand, Francis B. (1974). *Introduction to Numerical Analysis*, 2^a ed., McGraw-Hill, Nueva York.
82. Hill, David R. y Cleve B. Moler (1989). *Experiments in Computational Matrix Algebra*, Random House, Nueva York.
83. Hillier, Frederick S. y Gerald Lieberman (1974). *Operations Research*, 2^a ed., Holden-Day, San Francisco.

84. Hopkins, Tim y Chris Philips (1988). *Numerical Methods in Practice Using the NAG Library*, Addison-Wesley, Reading, Mass.
85. Hornbeck, Robert W. (1975). *Numerical Methods*, Quantum, Nueva York.
86. Householder, Alston S. (1970). *The Numerical Treatment of a Single Nonlinear Equation*, McGraw-Hill, Nueva York.
87. Householder, Alston S. (1953). *Principles of Numerical Analysis*, McGraw-Hill, Nueva York.
88. Hultquist, Paul F. (1988). *Numerical Methods for Engineers and Computer Scientists*, Benjamin/Cummings, Menlo Park, Calif.
89. Hundhausen, Joan R. y Robert A. Walsh (1985). Unconstrained Optimization, *UMAP Journal*, Vol. 6, No. 4, pp. 57-90.
90. Isaacson, Eugene y Herbert Bishop Keller (1966). *Analysis of Numerical Methods*, John Wiley, Nueva York.
91. Jacobs, D., ed. (1977). *The State of the Art in Numerical Analysis*, Academic Press, Nueva York.
92. Jacques, Ian y Colin Judd (1987). *Numerical Analysis*, Chapman and Hall, Nueva York.
93. James, M. L., G. M. Smith y J. C. Wolford (1985). *Applied Numerical Methods for Digital Computation*, 3^a ed., Harper & Row, Nueva York.
94. Jensen, Jens A. y John H. Rowland (1975). *Methods of Computation: The Linear Space Approach to Numerical Analysis*, Scott, Foresman, Glenview, Ill.
95. Jepsen, Charles H. y Eugene Herman (1988). *The Matrix Algebra Calculator: Linear Algebra Problems for Computer Solution*, Brooks/Cole, Pacific Grove, Calif.
96. Johnson, Lee W. y R. Dean Riess (1982). *Numerical Analysis*, 2^a ed., Addison-Wesley, Reading, Mass.
97. Johnston, R. L. (1982). *Numerical Methods: A Software Approach*, John Wiley, Nueva York.
98. Kahaner, David, Cleve Moler y Stephen Nash (1989). *Numerical Methods and Software*, Prentice Hall, Englewood Cliffs, N. J.
99. Keller, Herbert Bishop (1976). *Numerical Solution of Two Point Boundary Value Problems*, SIAM, Filadelfia.
100. Kincaid, David y Ward Cheney (1991). *Numerical Analysis Mathematics of Scientific Computing*, Brooks/Cole, Pacific Grove, Calif. Versión española (1994): *Análisis Numérico*, Addison-Wesley Iberoamericana, Wilmington, De.
101. King, J. Thomas (1984). *Introduction to Numerical Computation*, McGraw-Hill, Nueva York.

102. Klamkin, Murray S. (1987). *Mathematical Modeling: Classroom Notes in Applied Mathematics*, SIAM, Filadelfia.
103. Knuth, Donald E. (1981). *The Art of Computer Programming*, Vol. 2, Seminumerical Algorithms, 2^a ed., Addison-Wesley, Reading, Mass. Versión española (1987): *El Arte de Programar Ordenadores*, Reverté, Barcelona.
104. Kreyszig, Erwin (1983). *Advanced Engineering Mathematics*, 5^a ed., John Wiley, Nueva York. Versión española: *Matemáticas Avanzadas para Ingeniería*, Vols. 1 y 2, Limusa, México.
105. Kunz, Kaiser S. (1957). *Numerical Analysis*, McGraw-Hill, Nueva York.
106. Lambert, J. D. (1973). *Computational Methods in Ordinary Differential Equations*, John Wiley, Nueva York.
107. Lancaster, Peter (1976). *Mathematics: Models of the Real World*, Prentice Hall, Englewood Cliffs, N. J.
108. Lapidus, L. y J. H. Seinfeld (1971). *Numerical Solution of Ordinary Differential Equations*, Academic Press, Nueva York.
109. Lawson, C. L. y R. J. Hanson (1974). *Solving Least-Squares Problems*, Prentice Hall, Englewood Cliffs, N. J.
110. Lax, Peter, Samuel Burstein y Anneli Lax (1976). *Calculus with Applications and Computing*, Springer-Verlag, Nueva York.
111. Leinbach, L. Carl (1991). *Calculus Laboratories Using DERIVE*, Wadsworth, Belmont, Calif.
112. Lindfield, G. R. y J. E. T. Penny (1989). *Microcomputers in Numerical Analysis*, Halsted Press, Nueva York.
113. Lucas, William F., Fred S. Roberts y Robert M. Thrall, eds. (1983). *Discrete and System Models*, Springer-Verlag, Nueva York.
114. Luke, Yudell L. (1975). *Mathematical Functions and Their Applications*, Academic Press, Nueva York.
115. Maki, Daniel P. y Maynard Thompson (1973). *Mathematical Models and Applications*, Prentice Hall, Englewood Cliffs, N. J.
116. Marcus-Roberts, Helen y Maynard Thompson, eds. (1983). *Life Science Models*, Springer-Verlag, Nueva York.
117. Maron, Melvin J. y Robert J. Lopez (1991). *Numerical Analysis: A Practical Approach*, 3^a ed., Wadsworth, Belmont, Calif.
118. Mathews, John H. (1988). *Complex Variables for Mathematics and Engineering*, Wm. C. Brown, Dubuque, Ia.
119. McCalla, Thomas Richard (1967). *Introduction to Numerical Methods and FORTRAN Programming*, John Wiley, Nueva York.

120. McCarty, George (1975). *Calculator Calculus*, Page-Ficklin Publications, Palo Alto, Calif.
121. McCormick, John M. y Mario G. Salvadori (1964). *Numerical Methods in FORTRAN*, Prentice Hall, Englewood Cliffs, N. J.
122. McNeary, Samuel S. (1973). *Introduction to Computational Methods for Students of Calculus*, Prentice Hall, Englewood Cliffs, N. J.
123. Miel, George J. (1981). Calculator Demonstrations of Numerical Stability, *UMAP Journal*, Vol. 2, No. 2, pp. 3-7.
124. Miller, Webb (1984). *The Engineering of Numerical Software*, Prentice Hall, Englewood Cliffs, N. J.
125. Miller, Webb (1987). *A Software Tools Sampler*, Prentice Hall, Englewood Cliffs, N. J.
126. Milne, William Edmund (1949). *Numerical Calculus*, Princeton University Press, Princeton, N. J.
127. Moore, Ramon E. (1966). *Interval Analysis*, Prentice Hall, Englewood Cliffs, N. J.
128. Morris, John L. (1983). *Computational Methods in Elementary Numerical Analysis*, John Wiley, Nueva York.
129. Moursund, David G. y Charles S. Duris (1967). *Elementary Theory and Application of Numerical Analysis*, McGraw-Hill, Nueva York.
130. Murphy, J., D. Ridout y Brigid McShane (1988). *Numerical Analysis, Algorithms and Computation*, Halsted Press, Nueva York.
131. Noble, Ben (1967). *Applications of Undergraduate Mathematics in Engineering*, Macmillan, Nueva York.
132. Noble, Ben y James W. Daniel (1977). *Applied Linear Algebra*, 2^a ed., Prentice Hall, Englewood Cliffs, N. J. Versión española (1989): *Álgebra Lineal Aplicada*, Prentice Hall Hispanoamericana, México.
133. Nonweiler, T. R. F. (1984). *Computational Mathematics: An Introduction to Numerical Approximation*, Halsted Press, Nueva York.
134. Oldknow, Adrian y Derek Smith (1983). *Learning Mathematics with Micros*, Halsted Press, Nueva York.
135. Olinick, Michael (1978). *An Introduction to Mathematical Models in the Social and Life Sciences*, Addison-Wesley, Reading, Mass.
136. O'Neil, Peter V. (1991). *Advanced Engineering Mathematics*, 3^a ed., Wadsworth, Belmont, Calif.
137. Ortega, James M. (1972). *Numerical Analysis*, Academic Press, Nueva York.
138. Ortega, James M. y William G. Poole (1981). *An Introduction to Numerical Methods for Differential Equations*, Pitman, Marshfield, Mass.

139. Ortega, James M. y W. C. Rheinboldt (1970). *Iterative Solution of Non-linear Equations in Several Variables*, Academic Press, Nueva York.
140. Parlett, Beresford N. (1980). *The Symmetric Eigenvalue Problem*, Prentice Hall, Englewood Cliffs, N. J.
141. Pearson, Carl E. (1986). *Numerical Methods in Engineering and Science*, Van Nostrand Reinhold, Nueva York.
142. Pennington, Ralph H. (1970). *Introductory Computer Methods and Numerical Analysis*, 2^a ed., Macmillan, Nueva York.
143. Pettofrezzo, Anthony J. (1984). *Introductory Numerical Analysis*, Orange Publishers, Winter Park, Fla.
144. Phillips, G. M. y P. J. Taylor (1974). *Theory and Applications of Numerical Analysis*, Academic Press, Nueva York.
145. Pizer, Stephen M. (1975). *Numerical Computing and Mathematical Analysis*, Science Research Associates, Chicago.
146. Pizer, Stephen M. y Victor L. Wallace (1983). *To Compute Numerically. Concepts and Strategies*, Little, Brown, Boston.
147. Pokorny, Cornel K. y Curtis F. Gerald (1989). *Computer Graphics: The Principles behind the Art and Science*, Franklin, Beedle & Associates, Irvine, Calif.
148. Potts, J. Frank y J. Walter Oler (1989). *Finite Element Applications with Microcomputers*, Prentice Hall, Englewood Cliffs, N. J.
149. Powell, Michael James David (1981). *Approximation Theory and Methods*, Cambridge University Press, Nueva York.
150. Press, William H., Brian P. Flannery, Saul A. Teukolsky y William T. Vetterling (1988). *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Nueva York.
151. Press, William H., Brian P. Flannery, Saul A. Teukolsky y William T. Vetterling (1989). *Numerical Recipes in Pascal: The Art of Scientific Computing*, Cambridge University Press, Nueva York.
152. Press, William H., Brian P. Flannery, Saul A. Teukolsky y William T. Vetterling (1986). *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Nueva York.
153. Ralston, Anthony y Philip Rabinowitz (1978). *A First Course in Numerical Analysis*, 2^a ed., McGraw-Hill, Nueva York.
154. Ralston, Anthony y Herbert S. Wilf (1960). *Mathematical Methods for Digital Computers*, John Wiley, Nueva York.
155. Ramirez, Robert W. (1985). *The FFT, Fundamentals and Concepts*, Prentice Hall, Englewood Cliffs, N. J.

156. Rheinboldt, Werner C. (1981). Algorithms for Finding Zeros of Functions, *The UMAP Journal*, Vol. 2, No. 1, pp. 43–72.
157. Rice, John R. (1969). *The Approximation of Functions*, Addison-Wesley, Reading, Mass.
158. Rice, John R. (1980). *Mathematical Aspects of Scientific Software*, Springer-Verlag, Nueva York.
159. Rice, John R. (1981). *Matrix Computations and Mathematical Software*, McGraw-Hill, Nueva York.
160. Rice, John Rischard (1983). *Numerical Methods, Software and Analysis: IMSL Reference Edition*, McGraw-Hill, Nueva York.
161. Rivlin, Theodore J. (1969). *An Introduction to the Approximation of Functions*, Blaisdell, Waltham, Mass.
162. Rosser, J. Barkley y Carl de Boor (1979). *Pocket Calculator Supplement for Calculus*, Addison-Wesley, Reading, Mass.
163. Salvadori, Mario G. y Melvin L. Baron (1961). *Numerical Methods in Engineering*, 2^a ed., Prentice Hall, Englewood Cliffs, N. J.
164. Sandefur, James T. (1990). *Discrete Dynamical Systems: Theory and Applications*, Oxford University Press, Nueva York.
165. Scalzo, Frank y Rowland Hughes (1977). *A Computer Approach to Introductory College Mathematics*, Mason/Charter, Nueva York.
166. Scarborough, James B. (1966). *Numerical Mathematical Analysis*, The Johns Hopkins University Press, Baltimore, Md.
167. Scheid, Francis (1968). *Theory and Problems of Numerical Analysis*, McGraw-Hill, Nueva York.
168. Schultz, M. H. (1966). *Spline Analysis*, Prentice Hall, Englewood Cliffs, N. J.
169. Schwarz, Hans Rudolf y J. Waldvogel (1989). *Numerical Analysis: A Comprehensive Introduction*, John Wiley, Nueva York.
170. Scraton, R. E. (1984). *Basic Numerical Methods: An Introduction to Numerical Mathematics on a Microcomputer*, Edward Arnold, Baltimore, Md.
171. Sewell, Granville (1988). *The Numerical Solution to Ordinary and Partial Differential Equations*, Harcourt Brace Jovanovich, Nueva York.
172. Shampine, Lawrence F. y Richard C. Allen (1973). *Numerical Computing: An Introduction*, Saunders, Filadelfia.
173. Shampine, Lawrence F. y M. K. Gordon (1975). *Computer Solution of Ordinary Differential Equations: The Initial Value Problem*, W. H. Freeman, San Francisco.

174. Shoup, Terry E. (1979). *A Practical Guide to Computer Methods for Engineers*, Prentice Hall, Englewood Cliffs, N. J.
175. Shoup, Terry E. (1983). *Numerical Methods for the Personal Computer*, Prentice Hall, Englewood Cliffs, N. J.
176. Sicks, Jon L. (1985). *Investigating Secondary Mathematics with Computers*, Prentice Hall, Englewood Cliffs, N. J.
177. Simmons, George F. (1972). *Differential Equations: With Applications and Historical Notes*, McGraw-Hill, Nueva York. Versión española (1993): *Ecuaciones Diferenciales con Aplicaciones y Notas Históricas*, McGraw-Hill, Madrid.
178. Smith, B. T., J. M. Boyle, J. Dongarra, B. Garbow, Y. Ikebe, V. C. Klema y C. B. Moler (1976). *Matrix Eigensystem Routines: EISPACK Guide*, 2^a ed., Vol. 6 of *Lecture Notes in Computer Science*, Springer-Verlag, Nueva York.
179. Smith, David A. (1976). *INTERFACE: Calculus and the Computer*, Houghton Mifflin, Boston.
180. Smith, G. D. (1978). *The Numerical Solution of Partial Differential Equations*, 2^a ed., Oxford University Press, Nueva York.
181. Smith, W. Allen (1986). *Elementary Numerical Analysis*, Prentice Hall, Englewood Cliffs, N. J.
182. Snyder, Martin Avery (1966). *Chebyshev Methods in Numerical Approximation*, Prentice Hall, Englewood Cliffs, N. J.
183. Stanton, Ralph G. (1961). *Numerical Methods for Science and Engineering*, Prentice Hall, Englewood Cliffs, N. J.
184. Stark, Peter A. (1970). *Introduction to Numerical Methods*, Macmillan, Toronto, Ontario.
185. Strang, G. y G. Fix (1973). *An Analysis of the Finite Element Method*, Prentice Hall, Englewood Cliffs, N. J.
186. Strecker, George E. (1982). Round Numbers: An Introduction to Numerical Expression, *UMAP Journal*, Vol. 3, No. 4, pp. 425–454.
187. Stroud, A. H. (1971). *Approximate Calculation of Multiple Integrals*, Prentice Hall, Englewood Cliffs, N. J.
188. Stroud, A. H. y Don Secrest (1966). *Gaussian Quadrature Formulas*, Prentice Hall, Englewood Cliffs, N. J.
189. Szidarovszky, Ferenc y Sidney Yakowitz (1978). *Principles and Procedures of Numerical Analysis*, Plenum Press, Nueva York.
190. Thompson, William J. (1984). *Computing in Applied Science*, John Wiley, Nueva York.
191. Todd, John (1979). *Basic Numerical Mathematics*, Vol. 1: *Numerical Analysis*, Academic Press, Nueva York.

192. Todd, John (1977). *Basic Numerical Mathematics*, Vol. 2: *Numerical Algebra*, Academic Press, Nueva York.
193. Tompkins, Charles B. y Walter L. Wilson (1969). *Elementary Numerical Analysis*, Prentice Hall, Englewood Cliffs, N. J.
194. Traub, J. F. (1964). *Iterative Methods for the Solution of Equations*, Prentice Hall, Englewood Cliffs, N. J.
195. Tuma, Jan J. (1989). *Handbook of Numerical Calculations in Engineering*, McGraw-Hill, Nueva York.
196. Turner, Peter R. (1989). *Guide to Numerical Analysis*, CRC Press, Boca Raton, Fla.
197. Vandergraft, James S. (1983). *Introduction to Numerical Computations*, Academic Press, Nueva York.
198. VanIwaarden, John L. (1985). *Ordinary Differential Equations with Numerical Techniques*, Harcourt Brace Jovanovich, Nueva York.
199. Varga, Richard S. (1962). *Matrix Iterative Analysis*, Prentice Hall, Englewood Cliffs, N. J.
200. Wachspress, Eugene L. (1966). *Iterative Solution of Elliptic Systems*, Prentice Hall, Englewood Cliffs, N. J.
201. Wendroff, Burton (1966). *Theoretical Numerical Analysis*, Academic Press, Nueva York.
202. Wilkes, Maurice Vincent (1966). *A Short Introduction to Numerical Analysis*, Cambridge University Press, Nueva York.
203. Wilkinson, J. H. (1965). *The Algebraic Eigenvalue Problem*, Oxford University Press, Nueva York.
204. Wilkinson, J. H. (1963). *Rounding Errors in Algebraic Processes*, Prentice Hall, Englewood Cliffs, N. J.
205. Wilkinson, J. H. y C. Reinsch (1971). *Handbook for Automatic Computation*, Vols. 1 y 2, Springer-Verlag, Nueva York.
206. Yakowitz, Sidney y Ferenc Szidarovszky (1989). *An Introduction to Numerical Computations*, 2^a ed., Macmillan, Nueva York.
207. Young, David M. (1971). *Iterative Solution of Large Linear Systems*, Academic Press, Nueva York.
208. Young, David M. y Robert Todd Gregory (1972). *A Survey of Numerical Mathematics*, Vols. 1 y 2, Addison-Wesley, Reading, Mass.
209. Zienkiewicz, O. C. y R. L. Taylor (1989). *The Finite Element Method*, 4^a ed., McGraw-Hill, Nueva York.
210. Zohar, Shalhav (1979). *Faster Fourier Transformation: The Algorithm of S. Winograd*, Jet Propulsion Laboratory, Pasadena, Calif.

Soluciones de algunos ejercicios

Sección 1.1 Un repaso al cálculo infinitesimal

1. (a) $L = 2$, $\{\varepsilon_n\} = \left\{ \frac{1}{2n+1} \right\}$, $\lim_{n \rightarrow \infty} \varepsilon_n = 0$.
3. (a) $c = 1 - \sqrt{2}$.
4. (a) $M_1 = -5/4$, $M_2 = 5$.
5. (a) $c = 0$.
6. (a) $c = 1$.
7. $c = 4/3$.
9. (a) $x^2 \cos(x)$.
10. (a) $c = \pm \sqrt{13/3}$.
11. (a) 2 (b) 1.
15. $13\pi/3$, aplicando el teorema del valor medio para integrales.
16. Sean x_0, x_1, \dots, x_{n-1} las n raíces de $P(x)$. Comprobando que las hipótesis del teorema de Rolle generalizado se cumplen, se sigue que existe $c \in (a, b)$ tal que $P^{(n-1)}(c) = 0$.

Sección 1.2 Números binarios

1. (a) La respuesta de la calculadora no es 0 porque el desarrollo binario de 0.1 tiene infinitas cifras iguales a 1.
(b) 0 (exactamente).

2. (a) 21 (c) 254.
3. (a) 0.84375 (c) 0.6640625.
4. (a) 1.4140625.
5. (a) $\sqrt{2} - 1.4140625 = 0.000151062 \dots$
6. (a) 10111_{dos} (c) 101111010_{dos}
7. (a) 0.0111_{dos} (c) 0.10111_{dos}
8. (a) $0.\overline{0011}_{\text{dos}}$ (c) $0.\overline{001}_{\text{dos}}$
9. (a) $0.006250000 \dots$
11. Usando $c = \frac{3}{16}$ y $r = \frac{1}{16}$ se obtiene $S = \frac{\frac{3}{16}}{1 - \frac{1}{16}} = \frac{1}{5}$.
13. (a) $\frac{\frac{1}{3}}{\frac{8}{15}} \approx \frac{0.1011_{\text{dos}} \times 2^{-1}}{0.100011_{\text{dos}} \times 2^{-0}} = \frac{0.1011_{\text{dos}}}{0.100011_{\text{dos}}} \times 2^{-1}$
 $\frac{\frac{1}{5}}{\frac{1}{6}} \approx \frac{0.1101_{\text{dos}} \times 2^{-2}}{0.001011_{\text{dos}} \times 2^{-0}} = \frac{0.1101_{\text{dos}}}{0.001011_{\text{dos}}} \times 2^{-2}$
 $\frac{\frac{8}{15}}{\frac{7}{10}} \approx \frac{0.1001_{\text{dos}} \times 2^{-0}}{0.101111_{\text{dos}} \times 2^{-0}} = \frac{0.1001_{\text{dos}}}{0.101111_{\text{dos}}} \times 2^0$
 $\frac{\frac{1}{6}}{\frac{7}{10}} \approx \frac{0.1011_{\text{dos}} \times 2^{-2}}{0.101111_{\text{dos}} \times 2^{-0}} = \frac{0.1011_{\text{dos}}}{0.101111_{\text{dos}}} \times 2^{-2}$ $\approx [0.1100_{\text{dos}}]$
14. (a) $10 = 101_{\text{tres}}$ (c) $421 = 120121_{\text{tres}}$
15. (a) $\frac{1}{3} = 0.1_{\text{tres}}$ (b) $\frac{1}{2} = 0.\overline{1}_{\text{tres}}$
16. (a) $10 = 20_{\text{cinco}}$ (c) $721 = 10341_{\text{cinco}}$
17. (b) $\frac{1}{2} = 0.\overline{2}_{\text{cinco}}$

Sección 1.3 Análisis del error

1. (a) $x = 2.71828182$, $\hat{x} = 2.7182$, $(x - \hat{x}) = 0.00008182$,
 $(x - \hat{x})/x = 0.00003010$, cuatro cifras significativas.
2. $\frac{1}{4} + \frac{1}{4^3 3} + \frac{1}{4^5 5(2!)} + \frac{1}{4^7 7(3!)} = \frac{292807}{1146880} = 0.2553074428 = \hat{p}$,
 $p - \hat{p} = 0.0000000178$, $(p - \hat{p})/p = 0.0000000699$.
3. (a) $p_1 + p_2 = 1.414 + 0.09125 = 1.505$, $p_1 p_2 = (1.414)(0.09125) = 0.1290$.
4. Hay una pérdida de cifras significativas.
(a) $\frac{0.70711385222 - 0.70710678119}{0.00001} = \frac{0.00000707103}{0.00001} = 0.707103$
5. (a) $\ln((x+2)/x)$ o $\ln(1+1/x)$. (c) $\cos(2x)$.

$$\begin{aligned}
 6. \text{ (a)} \quad P(2.72) &= (2.72)^3 - 3(2.72)^2 + 3(2.72) - 1 \\
 &= 20.12 - 22.19 + 8.16 - 1 = -2.07 + 8.16 - 1 \\
 &= 6.09 - 1 = 5.09,
 \end{aligned}$$

$$\begin{aligned}
 Q(2.72) &= ((2.72 - 3)2.72 + 3)2.72 - 1 \\
 &= ((-0.28)2.72 + 3)2.72 - 1 = (-0.7616 + 3)2.72 - 1 \\
 &= (2.238)2.72 - 1 = 6.087 - 1 = 5.087, \\
 R(2.72) &= (2.72 - 1)^3 = (1.72)^3 = 5.088.
 \end{aligned}$$

$$7. \text{ (a)} \quad 0.498. \quad \text{(b)} \quad 0.499.$$

$$9. \text{ (a)} \quad \frac{1}{1-h} + \cos(h) = 2 + h + \frac{h^2}{2}h^3 + O(h^4).$$

$$\text{(b)} \quad \frac{1}{1-h} \cos(h) = 1 + h + \frac{h^2}{2} + \frac{h^3}{2} + O(h^4).$$

Sección 2.1 Métodos iterativos para resolver $x = g(x)$

1. (a) La función $g \in C[0, 1]$ aplica $[0, 1]$ sobre $[3/4, 1] \subseteq [0, 1]$ y

$$|g'(x)| = |-x/2| = x/2 \leq 1/2 < 1$$

en $[0, 1]$. Por tanto, las hipótesis del Teorema 2.2 se cumplen, así que g tiene un único punto fijo en $[0, 1]$.

$$2. \text{ (a)} \quad g(2) = -4 + 8 - 2 = 2, \quad g(4) = -4 + 16 - 8 = 4.$$

$$\begin{array}{lll}
 \text{(b)} \quad p_0 = 1.9 & E_0 = 0.1 & R_0 = 0.05, \\
 p_1 = 1.795 & E_1 = 0.205 & R_1 = 0.1025, \\
 p_2 = 1.5689875 & E_2 = 0.4310125 & R_2 = 0.21550625, \\
 p_3 = 1.04508911 & E_3 = 0.95491089 & R_3 = 0.477455444.
 \end{array}$$

(e) La sucesión del apartado (b) no converge a $P = 2$. La sucesión del apartado (c) converge a $P = 4$.

$$4. \quad P = 2, \quad g'(2) = 5, \quad \text{la iteración no convergerá a } P = 2.$$

$$5. \quad P = 2n\pi \text{ donde } n \text{ es cualquier entero, } g'(P) = 1; \text{ el Teorema 2.3 no da información sobre la convergencia.}$$

$$9. \text{ (a)} \quad g(3) = 0.5(3) + 1.5 = 3.$$

(c) Por inducción matemática: Si $n = 1$, entonces $|P - p_1| = |P - p_0|/2^1$, por el apartado (b). Hipótesis de inducción: Supongamos que $|P - p_k| = |P - p_0|/2^k$ y probemos que esto es cierto para $n = k + 1$:

$$\begin{aligned}
 |P - p_{k+1}| &= |P - p_k|/2 && \text{(por el apartado (b))} \\
 &= (|P - p_0|/2^k)/2 && \text{(hipótesis de inducción)} \\
 &= |P - p_0|/2^{k+1}.
 \end{aligned}$$

10. (a) $\frac{|p_{k+1} - p_k|}{|p_{k+1}|} = \left| \frac{\frac{p_k}{2} - p_k}{\frac{p_k}{2}} \right| = 1.$

Sección 2.2 Los métodos de localización de raíces

1. $I_0 = (0.11 + 0.12)/2 = 0.115 \quad A(0.115) = 254\,403,$
 $I_1 = (0.11 + 0.115)/2 = 0.1125 \quad A(0.1125) = 246\,072,$
 $I_2 = (0.1125 + 0.115)/2 = 0.11375 \quad A(0.11375) = 250\,198.$

3. Pueden escogerse muchos intervalos $[a, b]$ en los que $f(a)$ y $f(b)$ tienen signos opuestos. Por ejemplo, podemos escoger los siguientes.

(a) $f(1) < 0$ y $f(2) > 0$, así que hay una raíz en $[1, 2]$; también se cumple $f(-1) < 0$ y $f(-2) > 0$, así que hay una raíz en $[-2, -1]$.

(c) $f(3) < 0$ y $f(4) > 0$, así que hay una raíz en $[3, 4]$.

4. $c_0 = -1.8300782, c_1 = -1.8409252, c_2 = -1.8413854, c_3 = -1.8414048$

6. $c_0 = 3.6979549, c_1 = 3.6935108, c_2 = 3.6934424, c_3 = 3.6934414$

11. Determine N de manera que $\frac{7-2}{2^{N+1}} < 5 \times 10^{-9}$.

14. El método de bisección nunca convergerá (suponiendo que $c_n \neq 2$) a $x = 2$.

Sección 2.3 Aproximación inicial y criterios de convergencia

1. Hay una raíz cerca de $x = -0.7$, así que podríamos usar el intervalo $[-1, 0]$.
3. Hay una raíz cerca de $x = 1$, así que podríamos usar el intervalo $[-2, 2]$.
5. Hay una raíz cerca de $x = 1.4$, así que podríamos usar el intervalo $[1, 2]$. Hay una segunda raíz cerca de $x = 3$, así que para ésta podríamos usar el intervalo $[2, 4]$.

Sección 2.4 Los métodos de Newton-Raphson y de la secante

1. (a) $p_k = g(p_{k-1}) = \frac{p_{k-1}^2 - 2}{2p_{k-1} - 1}.$
(b) $p_0 = -1.5, p_1 = 0.125, p_2 = 2.6458, p_3 = 1.1651.$
3. (a) $p_k = g(p_{k-1}) = \frac{3}{4}p_{k-1} + \frac{1}{2}.$
(b) $p_0 = 2.1, p_1 = 2.075, p_2 = 2.0561, p_3 = 2.0421, p_4 = 2.0316.$
5. (a) $p_k = g(p_{k-1}) = p_{k-1} + \cos(p_{k-1}).$

7. (a) $g(p_{k-1}) = p_{k-1}^2 / (p_{k-1} - 1)$.

(b) $p_0 = 0.20$ (c) $p_0 = 20.0$

$$p_1 = -0.05$$

$$p_1 = 21.05263158$$

$$p_2 = -0.002380953$$

$$p_2 = 22.10250034$$

$$p_3 = -0.000005655$$

$$p_3 = 23.14988809$$

$$p_4 = -0.0000000000$$

$$p_4 = 24.19503505$$

$$\lim_{n \rightarrow \infty} p_k = 0.0$$

$$\lim_{n \rightarrow \infty} p_k = \infty$$

8. $p_0 = 2.6, p_1 = 2.5, p_2 = 2.41935484, p_3 = 2.41436464$

14. No, porque $f'(x)$ no es continua en la raíz $p = 0$. Podría usar también $g(p_{k-1}) = -2p_{k-1}$ y comprobar que esta sucesión diverge.

22. (a) $g(x) = x - \frac{x^2 - a}{2x} \left(1 - \frac{(x^2 - a)2}{2(2x)^2}\right)^{-1} = \frac{x(x^2 + 3a)}{3x^2 + a}$

$$g(x) = \frac{15x + x^3}{5 + 3x^2}$$

$$p_1 = 2.2352941176, p_2 = 2.2360679775, p_3 = 2.2360679775$$

(b) $g(x) = \frac{2 + 4x + 2x^2 + x^3}{3 + 4x + 2x^2}$

$$p_1 = -2.0130081301, p_2 = -2.0000007211, p_3 = -2.0000000000$$

Sección 2.5 Métodos de Aitken, Steffensen y Muller

2. (a) $\Delta^2 p_n = \Delta(\Delta p_n) = \Delta(p_{n+1} - p_n) = (p_{n+2} - p_{n+1}) - (p_{n+1} - p_n)$
 $= p_{n+2} - 2p_{n+1} + p_n = 2(n+2)^2 + 1 - 2(2(n+1)^2 + 1)$
 $+ 2n^2 + 1 = 4$

6. $p_n = 1/(4^n + 4^{-n})$

n	p_n	q_n de Aitken
0	0.5	-0.26437542
1	0.23529412	-0.00158492
2	0.06225681	-0.00002390
3	0.01562119	-0.00000037
4	0.00390619	
5	0.00097656	

7. $g(x) = (6 + x)^{1/2}$

n	p_n	q_n de Aitken
0	2.5	3.00024351
1	2.91547595	3.00000667
2	2.98587943	3.00000018
3	2.99764565	3.00000001
4	2.99960758	
5	2.99993460	

9. Solución de $\cos(x) - 1 = 0$.

n	p_n de Steffensen
0	0.5
1	0.24465808
2	0.12171517
3	0.00755300
4	0.00377648
5	0.00188824
6	0.00000003

11. La suma de la serie es $S = 99$.

n	S_n	T_n
1	0.99	98.9999988
2	1.9701	99.0000017
3	2.940399	98.9999988
4	3.90099501	98.9999992
5	4.85198506	
6	5.79346521	

13. La suma de la serie es $S = 4$.

15. Método de Muller para $f(x) = x^3 - x - 2$.

n	p_n	$f(p_n)$
0	1.0	-2.0
1	1.2	-1.472
2	1.4	-0.656
3	1.52495614	0.02131598
4	1.52135609	-0.00014040
5	1.52137971	-0.00000001

Sección 3.1 Vectores y matrices

1. (i) (a) $(1, 4)$ (b) $(5, -12)$ (c) $(9, -12)$ (d) 5

(e) $(-26, 72)$ (f) -38 (g) $2\sqrt{1465}$

2. $\theta = \arccos(-16/21) \approx 2.437045$ radianes

3. (a) Supongamos que $\mathbf{X}, \mathbf{Y} \neq \mathbf{0}$. $\mathbf{X} \cdot \mathbf{Y} = 0$ si, y sólo si, $\cos(\theta) = 0$ si, y sólo si, $\theta = (2n + 1)\frac{\pi}{2}$ si, y sólo si, \mathbf{X} e \mathbf{Y} son ortogonales.

6. (c) $a_{ji} = \begin{cases} ji & j = i \\ j - ji + i & j \neq i \end{cases} = \begin{cases} ij & i = j \\ i - ij + j & i \neq j \end{cases} = a_{ij}$

Sección 3.2 Multiplicación de matrices

1. $\mathbf{AB} = \begin{bmatrix} -11 & -12 \\ 13 & -24 \end{bmatrix}, \quad \mathbf{BA} = \begin{bmatrix} -15 & 10 \\ -12 & -20 \end{bmatrix}$

3. (a) $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}) = \begin{bmatrix} 2 & -5 \\ -88 & -56 \end{bmatrix}$

5. (a) 33 (c) El determinante no existe porque la matriz no es cuadrada.

8. $(\mathbf{AB})(\mathbf{B}^{-1}\mathbf{A}^{-1}) = \mathbf{A}(\mathbf{B}\mathbf{B}^{-1})\mathbf{A}^{-1} = (\mathbf{AI})\mathbf{A}^{-1} = \mathbf{AA}^{-1} = \mathbf{I}$. Análogamente, $(\mathbf{B}^{-1}\mathbf{A}^{-1})(\mathbf{AB}) = \mathbf{I}$. Por tanto, $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

10. (a) MN (b) $M(N-1)$

14. $\mathbf{XX}' = [6], \quad \mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & -1 & 2 \\ -1 & 1 & -2 \\ 2 & -2 & 4 \end{bmatrix}$

Sección 3.3 Sistemas lineales triangulares

1. $x_1 = 2, x_2 = -2, x_3 = 1, x_4 = 3$ y $\det \mathbf{A} = 120$

5. $x_1 = 3, x_2 = 2, x_3 = 1, x_4 = -1$ y $\det \mathbf{A} = -24$

Sección 3.4 Eliminación gaussiana y pivoteo

1. $x_1 = -3, x_2 = 2, x_3 = 1$

5. $y = 5 - 3x + 2x^2$

10. $x_1 = 1, x_2 = 3, x_3 = 2, x_4 = -2$

15. (a) Solución para la matriz de Hilbert \mathbf{A} :

$x_1 = 25, x_2 = -300, x_3 = 1050, x_4 = -1400, x_5 = 630$

(b) Solución para la otra matriz \mathbf{A} :

$x_1 = 28.02304, x_2 = -348.5887, x_3 = 1239.781$

$x_4 = -1666.785, x_5 = 753.5564$

Sección 3.5 Factorización triangular

1. (a) $\mathbf{Y}' = [-4 \ 12 \ 3], \quad \mathbf{X}' = [-3 \ 2 \ 1]$

(b) $\mathbf{Y}' = [20 \ 39 \ 9], \quad \mathbf{X}' = [5 \ 7 \ 3]$

3. (a) $\begin{bmatrix} -5 & 2 & -1 \\ 1 & 0 & 3 \\ 3 & 1 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -0.2 & 1 & 0 \\ -0.6 & 5.5 & 1 \end{bmatrix} \begin{bmatrix} -5 & 2 & -1 \\ 0 & 0.4 & 2.8 \\ 0 & 0 & -10 \end{bmatrix}$

5. (a) $\mathbf{Y}' = [8 \ -6 \ 12 \ 2]$, $\mathbf{X}' = [3 \ -1 \ 1 \ 2]$

(b) $\mathbf{Y}' = [28 \ 6 \ 12 \ 1]$, $\mathbf{X}' = [3 \ 1 \ 2 \ 1]$

6. La factorización triangular $\mathbf{A} = \mathbf{LU}$ es

$$\mathbf{LU} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 5 & 1 & 1 & 0 \\ -3 & -1 & -1.75 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 4 \\ 0 & -3 & 5 & -8 \\ 0 & 0 & -4 & -10 \\ 0 & 0 & 0 & -7.5 \end{bmatrix}$$

Sección 3.6 Métodos iterativos para sistemas lineales

1. (a) Método de Jacobi

$P_1 = (3.75, 1.8)$

$P_2 = (4.2, 1.05)$

$P_3 = (4.0125, 0.96)$

La sucesión converge al punto $(4, 1)$.

(b) Método de Gauss-Seidel

$P_1 = (3.75, 1.05)$

$P_2 = (4.0125, 0.9975)$

$P_3 = (3.999375, 1.000125)$

La sucesión converge al punto $(4, 1)$.

3. (a) Método de Jacobi

$P_1 = (-1, -1)$

$P_2 = (-4, -4)$

$P_3 = (-13, -13)$

La sucesión diverge y se aleja de la solución

$P = (0.5, 0.5).$

(b) Método de Gauss-Seidel

$P_1 = (-1, -4)$

$P_2 = (-13, -40)$

$P_3 = (-121, -361)$

La sucesión diverge y se aleja de la solución

$P = (0.5, 0.5).$

5. (a) Método de Jacobi

$P_1 = (2, 1.375, 0.75)$

$P_2 = (2.125, 0.96875, 0.90625)$

$P_3 = (2.0125, 0.95703125, 1.0390625)$

La sucesión converge a $P = (2, 1, 1)$.

(b) Método de Gauss-Seidel

$P_1 = (2, 0.875, 1.03125)$

$P_2 = (1.96875, 1.01171875, 0.989257813)$

$P_3 = (2.00449219, 0.99753418, 1.0017395)$

La sucesión converge a $P = (2, 1, 1)$.

9. (15): $\|X\|_1 = \sum_{k=1}^N |x_k| = 0$ si, y sólo si, $|x_k| = 0$ para $k = 0, 1, \dots, N$ si, y sólo si, $X = \mathbf{0}$
(16): $\|cX\|_1 = \sum_{k=1}^N |cx_k| = \sum_{k=1}^N |c||x_k| = |c| \sum_{k=1}^N |x_k| = |c| \|X\|_1$

Sección 3.7 Métodos iterativos para sistemas no lineales

1. (a) $x = 0, y = 0$ (c) $x = 0, y = 2n\pi$

2. (a) $x = 4, y = -2$ (c) $x = 0, y = (2n + 1)\pi/2$

5. $\mathbf{J}(x, y) = \begin{bmatrix} 1-x & y/4 \\ (1-x)/2 & (2-y)/2 \end{bmatrix}, \quad \mathbf{J}(1.1, 2.0) = \begin{bmatrix} -0.1 & 0.5 \\ -0.05 & 0.0 \end{bmatrix}$

k	Iteración de punto fijo		Iteración de Seidel	
	p_k	q_k	p_k	q_k
0	1.1	2.0	1.1	2.0
1	1.12	1.9975	1.12	1.9964
2	1.1165508	1.9963984	1.1160016	1.9966327
∞	1.1165151	1.9966032	1.1165151	1.9966032

7. $0 = x^2 - y - 0.2, 0 = y^2 - x - 0.3$

\mathbf{P}_k	Solución del sistema lineal: $\mathbf{J}(\mathbf{P}_k) d\mathbf{P} = -\mathbf{F}(\mathbf{P}_k)$	$\mathbf{P}_k + d\mathbf{P}$
$\begin{bmatrix} 1.2 \\ 1.2 \end{bmatrix}$	$\begin{bmatrix} 2.4 & -1.0 \\ -1.0 & 2.4 \end{bmatrix} \begin{bmatrix} -0.0075630 \\ 0.0218487 \end{bmatrix} = \begin{bmatrix} 0.04 \\ -0.06 \end{bmatrix}$	$\begin{bmatrix} 1.192437 \\ 1.221849 \end{bmatrix}$
$\begin{bmatrix} 1.192437 \\ 1.221849 \end{bmatrix}$	$\begin{bmatrix} 2.384874 & -1.0 \\ -1.0 & 2.443697 \end{bmatrix} \begin{bmatrix} -0.0001278 \\ -0.0002476 \end{bmatrix} = \begin{bmatrix} 0.0000572 \\ 0.0004774 \end{bmatrix}$	$\begin{bmatrix} 1.192309 \\ 1.221601 \end{bmatrix}$

(a) Por tanto, $(p_1, q_1) = (1.192437, 1.221849)$ y $(p_2, q_2) = (1.192309, 1.221601)$.

\mathbf{P}_k	Solución del sistema lineal: $\mathbf{J}(\mathbf{P}_k) d\mathbf{P} = -\mathbf{F}(\mathbf{P}_k)$	$\mathbf{P}_k + d\mathbf{P}$
$\begin{bmatrix} -0.2 \\ -0.2 \end{bmatrix}$	$\begin{bmatrix} -0.4 & -1.0 \\ -1.0 & -0.4 \end{bmatrix} \begin{bmatrix} -0.0904762 \\ 0.0761905 \end{bmatrix} = \begin{bmatrix} 0.04 \\ -0.06 \end{bmatrix}$	$\begin{bmatrix} -0.2904762 \\ -0.1238095 \end{bmatrix}$
$\begin{bmatrix} -0.2904762 \\ -0.1238095 \end{bmatrix}$	$\begin{bmatrix} -0.5809524 & -1.0 \\ -1.0 & -0.2476190 \end{bmatrix} \begin{bmatrix} 0.0044128 \\ 0.0056223 \end{bmatrix} = \begin{bmatrix} 0.0081859 \\ 0.0058050 \end{bmatrix}$	$\begin{bmatrix} -0.2860634 \\ -0.1181872 \end{bmatrix}$

(b) Por tanto, $(p_1, q_1) = (-0.2904762, -0.1238095)$ y $(p_2, q_2) = (-0.2860634, -0.1181872)$.

8. (b) Los valores del determinante jacobiano en las soluciones son, respectivamente, $|\mathbf{J}(1, 1)| = 0$ y $|\mathbf{J}(-1, -1)| = 0$. El método de Newton se basa en la posibilidad de resolver un sistema lineal en el que la matriz es $\mathbf{J}(p_n, q_n)$ cuando (p_n, q_n) está cerca de la solución. En este ejemplo, el sistema de

ecuaciones está mal condicionado así que es difícil resolverlo con precisión. De hecho, para algunos puntos cercanos a la solución se tiene $\mathbf{J}(x_0, y_0) = 0$, por ejemplo, $\mathbf{J}(1.0001, 1.0001) = 0$.

- 12. (a)** Para las derivadas se tiene $\frac{\partial}{\partial x}(cf(x, y)) = c\frac{\partial}{\partial x}f(x, y)$. Dado que $\mathbf{F}(\mathbf{X})$ se define como $\mathbf{F}(\mathbf{X}) = [f_1(x_1, \dots, x_n) \cdots f_m(x_1, \dots, x_n)]'$; entonces, por las propiedades del producto por escalares,

$$c\mathbf{F}(\mathbf{X}) = [cf_1(x_1, \dots, x_n) \cdots cf_m(x_1, \dots, x_n)]'.$$

Por otro lado, $\mathbf{J}(c\mathbf{F}(\mathbf{X})) = [j_{ik}]_{m \times n}$, donde

$$j_{ik} = \frac{\partial}{\partial x_k}(cf_i(x_1, \dots, x_n)) = c\frac{\partial}{\partial x_k}f_i(x_1, \dots, x_n).$$

En consecuencia, se tiene $\mathbf{J}(c\mathbf{F}(\mathbf{X})) = c\mathbf{J}(\mathbf{F}(\mathbf{X}))$.

Sección 4.1 Series de Taylor y cálculo de los valores de una función

- 1. (a)** $P_5(x) = x - x^3/3! + x^5/5!$

$$P_7(x) = x - x^3/3! + x^5/5! - x^7/7!$$

$$P_9(x) = x - x^3/3! + x^5/5! - x^7/7! + x^9/9!$$

- (b)** $|E_9(x)| = |\operatorname{sen}(c)x^{10}/10!| \leq (1)(1)^{10}/10! = 0.0000002755$

- (c)** $P_5(x) = 2^{-1/2}(1 + (x - \pi/4) - (x - \pi/4)^2/2 - (x - \pi/4)^3/6 + (x - \pi/4)^4/24 + (x - \pi/4)^5/120)$

- 3.** En $x_0 = 0$ las derivadas de $f(x)$ no están definidas, mientras que en $x_0 = 1$ sí lo están.

- 5.** $P_3(x) = 1 + 0x - x^2/2 + 0x^3 = 1 - x^2/2$

- 8. (a)** $f(2) = 2$, $f'(2) = \frac{1}{4}$, $f''(2) = -\frac{1}{32}$, $f^{(3)}(2) = \frac{3}{256}$
 $P_3(x) = 2 + (x - 2)/4 - (x - 2)^2/64 + (x - 2)^3/512$

- (b)** $P_3(1) = 1.732421875$; compárese con $3^{1/2} = 1.732050808$

- (c)** $f^{(4)}(x) = -15(2 + x)^{-7/2}/16$; el máximo de $|f^{(4)}(x)|$ en el intervalo $1 \leq x \leq 3$ se alcanza en $x = 1$, luego $|f^{(4)}(x)| \leq |f^{(4)}(1)| \leq 3^{-7/2}(15/16) \approx 0.020046$. Por tanto, $|E_3(x)| \leq \frac{(0.020046)(1)^4}{4!} = 0.00083529$

- 13. (d)** $P_3(0.5) = 0.41666667$

$$P_6(0.5) = 0.40468750$$

$$P_9(0.5) = 0.40553230$$

$$\ln(1.5) = 0.40546511$$

- 14. (d)** $P_2(0.5) = 1.21875000$

$$P_4(0.5) = 1.22607422$$

$$P_6(0.5) = 1.22660828$$

$$(1.5)^{1/2} = 1.22474487$$

Sección 4.2 Introducción a la interpolación

1. (a) Use $x = 4$ para obtener $b_3 = -0.02$, $b_2 = 0.02$, $b_1 = -0.12$, $b_0 = 1.18$; con lo que, $P(4) = 1.18$.
 - (b) Use $x = 4$ para obtener $d_2 = -0.06$, $d_1 = -0.04$, $d_0 = -0.36$; con lo que, $P'(4) = -0.36$.
 - (c) Use $x = 4$ para obtener $i_4 = -0.005$, $i_3 = 0.01333333$, $i_2 = -0.04666667$, $i_1 = 1.47333333$, $i_0 = 5.89333333$; con lo que, $I(4) = 5.89333333$. Análogamente, use $x = 1$ para obtener $I(1) = 1.58833333$.
- $\int_1^4 P(x) dx = I(4) - I(1) = 5.89333333 - 1.58833333 = 4.305$
- (d) Use $x = 5.5$ para obtener $b_3 = -0.02$, $b_2 = -0.01$, $b_1 = -0.255$, $b_0 = 0.2575$; con lo que, $P(5.5) = 0.2575$.

Sección 4.3 Interpolación de Lagrange

1. (a) $P_1(x) = -1(x - 0)/(-1 - 0) + 0 = x + 0 = x$
 - (b) $P_2(x) = -1 \frac{(x - 0)(x - 1)}{(-1 - 0)(-1 - 1)} + 0 + \frac{(x + 1)(x - 0)}{(1 + 1)(1 - 0)}$
 $= -0.5(x)(x - 1) + 0.5(x)(x + 1) = 0x^2 + x + 0 = x$
 - (c) $P_3(x) = -1 \frac{(x)(x - 1)(x - 2)}{(-1)(-2)(-3)} + 0 + \frac{(x + 1)(x)(x - 2)}{(2)(1)(-1)}$
 $+ 8 \frac{(x + 1)(x)(x - 1)}{(3)(2)(1)} = x^3 + 0x^2 + 0x + 0 = x^3$
 - (d) $P_1(x) = 1(x - 2)/(1 - 2) + 8(x - 1)/(2 - 1) = 7x - 6$
 - (e) $P_2(x) = 0 + \frac{(x)(x - 2)}{(1)(-1)} + 8 \frac{(x)(x - 1)}{(2)(1)} = 3x^2 - 2x$
5. (c) $f^{(4)}(c) = 120(c - 1)$ para todo c ; así que

$$E_3(x) = 5(x + 1)(x)(x - 3)(x - 4)(c - 1)$$

10. $|f^{(2)}(c)| \leq |- \operatorname{sen}(1)| = 0.84147098 = M_2$
(a) $h^2 M_2 / 8 = h^2 (0.84147098) / 8 < 5 \times 10^{-7}$

12. (a) $z = 3 - 2x + 4y$

Sección 4.4 Polinomio interpolador de Newton

1. $P_1(x) = 4 - (x - 1)$
 $P_2(x) = 4 - (x - 1) + 0.4(x - 1)(x - 3)$
 $P_3(x) = P_2(x) + 0.01(x - 1)(x - 3)(x - 4)$
 $P_4(x) = P_3(x) - 0.002(x - 1)(x - 3)(x - 4)(x - 4.5)$
 $P_1(2.5) = 2.5, P_2(2.5) = 2.2, P_3(2.5) = 2.21125, P_4(2.2) = 2.21575$

5. $f(x) = 3(2)^x$

$$P_4(x) = 1.5 + 1.5(x+1) + 0.75(x+1)(x) + 0.25(x+1)(x)(x-1) \\ + 0.0625(x+1)(x)(x-1)(x-2)$$

$$P_1(1.5) = 5.25, P_2(1.5) = 8.0625,$$

$$P_3(1.5) = 8.53125, P_4(1.5) = 8.47265625$$

7. $f(x) = 3.6/x$

$$P_4(x) = 3.6 - 1.8(x-1) + 0.6(x-1)(x-2) \\ - 0.15(x-1)(x-2)(x-3) \\ + 0.03(x-1)(x-2)(x-3)(x-4)$$

$$P_1(2.5) = 0.9, P_2(2.5) = 1.35,$$

$$P_3(2.5) = 1.40625, P_4(2.5) = 1.423125$$

Sección 4.5 Polinomios de Chebyshev

9. (a) $\ln(x+2) \approx 0.69549038 + 0.49905042x - 0.14334605x^2 + 0.04909073x^3$

$$(b) |f^{(4)}(x)|/(2^3(4!)) \leq |-6|/(2^3(4!)) = 0.03125000$$

11. (a) $\cos(x) \approx 1 - 0.46952087x^2$

$$(b) |f^{(3)}(x)|/(2^2(3!)) \leq |\sin(1)|/(2^2(3!)) = 0.03506129$$

13. La cota del error para el polinomio de Taylor es

$$\frac{|f^{(8)}(x)|}{8!} \leq \frac{|\sin(1)|}{8!} = 0.00002087.$$

La cota del error para la aproximación minimax es

$$\frac{|f^{(8)}(x)|}{2^7(8!)} \leq \frac{|\sin(1)|}{2^7(8!)} = 0.00000016.$$

Sección 4.6 Aproximaciones de Padé

1. $1 = p_0, 1 + q_1 = p_1, \frac{1}{2} + q_1 = 0, q_1 = -\frac{1}{2}, p_1 = \frac{1}{2}$
 $e^x \approx R_{1,1}(x) = (2+x)/(2-x)$

3. $1 = p_0, \frac{1}{3} + 2q_1/15 = p_1, \frac{2}{15} + q_1/3 = 0, q_1 = -\frac{2}{5}, p_1 = -\frac{1}{15}$

5. $1 = p_0, 1 + q_1 = p_1, \frac{1}{2} + q_1 + q_2 = p_2.$

Primero se resuelve el sistema $\begin{cases} \frac{1}{6} + \frac{q_1}{2} + q_2 = 0 \\ \frac{1}{24} + \frac{q_1}{6} + \frac{q_2}{2} = 0. \end{cases}$

Entonces $q_1 = -\frac{1}{2}, q_2 = \frac{1}{12}, p_1 = \frac{1}{2}, p_2 = \frac{1}{12}.$

7. (a) $1 = p_0, \frac{1}{3} + q_1 = p_1, \frac{2}{15} + q_1/3 + q_2 = p_2.$

Primero se resuelve el sistema $\begin{cases} \frac{17}{315} + \frac{2q_1}{15} + \frac{q_2}{3} = 0 \\ \frac{62}{2835} + \frac{17q_1}{315} + \frac{2q_2}{15} = 0. \end{cases}$

Entonces $q_1 = -\frac{4}{9}, q_2 = \frac{1}{63}, p_1 = -\frac{1}{9}, p_2 = \frac{1}{945}.$

Sección 5.1 Rectas de regresión en mínimos cuadrados

1. (a) $10A + 0B = 7$

$0A + 5B = 13$

$y = 0.70x + 2.60, E_2(f) \approx 0.2449$

2. (a) $40A + 0B = 58$

$0A + 5B = 31.2$

$y = 1.45x + 6.24, E_2(f) \approx 0.8958$

3. (c) $\sum_{k=1}^5 x_k y_k / \sum_{k=1}^5 x_k^2 = 86.9/55 = 1.58$

$y = 1.58x, E_2(f) \approx 0.1720$

11. (a) $y = 1.6866x^2, E_2(f) \approx 1.3$

$y = 0.5902x^3, E_2(f) \approx 0.29.$ Éste es el mejor ajuste.

Sección 5.2 Ajuste de curvas

1. (a) $164A + 20C = 186$

$20B = -34$

$20A + 4C = 26$

$y = 0.875x^2 - 1.70x + 2.125 = 7/8x^2 - 17/10x + 17/8$

3. (a) $15A + 5B = -0.8647$

$5A + 5B = 4.2196$

$y = 3.8665e^{-0.5084x}, E_1(f) \approx 0.10$

6.

	Linealizando	Por mínimos cuadrados
(a)	$\frac{1000}{1 + 4.3018e^{-1.0802t}}$	$\frac{1000}{1 + 4.2131e^{-1.0456t}}$
(b)	$\frac{5000}{1 + 8.9991e^{-0.81138t}}$	$\frac{5000}{1 + 8.9987e^{-0.81157t}}$

18. (a) $14A + 15B + 8C = 82$

$$15A + 19B + 9C = 93$$

$$8A + 9B + 5C = 49$$

$A = 2.4, B = 1.2, C = 3.8$ con lo cual $z = 2.4x + 1.2y + 3.8$.

Sección 5.3 Interpolación polinomial a trozos

4. $h_0 = 1 \quad d_0 = -2$

$$h_1 = 3 \quad d_1 = 1 \quad u_1 = 18$$

$$h_2 = 3 \quad d_2 = -2/3 \quad u_2 = -10$$

Se resuelve el sistema $\begin{cases} \frac{15}{2}m_1 + m_2 = 21 \\ 3m_1 + \frac{21}{2}m_2 = -15 \end{cases}$ y se obtienen los valores

$$m_1 = \frac{314}{101} \text{ y } m_2 = -\frac{234}{101}.$$

Entonces $m_0 = -\frac{460}{101}$ y $m_3 = \frac{856}{303}$. La cercha cúbica es

$$S_0(x) = \frac{129}{101}(x+3)^3 - \frac{230}{101}(x+3)^2 - (x+3) + 2 \quad -3 \leq x \leq -2$$

$$S_1(x) = -\frac{274}{909}(x+2)^3 + \frac{157}{101}(x+2)^2 - \frac{96}{101}(x+2) \quad -2 \leq x \leq 1$$

$$S_2(x) = \frac{779}{2727}(x-1)^3 - \frac{117}{101}(x-1)^2 + \frac{72}{303}(x-1) + 3 \quad 1 \leq x \leq 4$$

5. $h_0 = 1 \quad d_0 = -2$

$$h_1 = 3 \quad d_1 = 1 \quad u_1 = 18$$

$$h_2 = 3 \quad d_2 = -2/3 \quad u_2 = -10$$

Se resuelve el sistema $\begin{cases} 8m_1 + 3m_2 = 18 \\ 3m_1 + 12m_2 = -10 \end{cases}$ y se obtiene $m_1 = \frac{82}{29}$ y

$$m_2 = -\frac{134}{87}.$$

Poniendo $m_0 = 0 = m_3$, la cercha cúbica es

$$S_0(x) = \frac{41}{87}(x+3)^3 - \frac{215}{87}(x+3) + 2 \quad -3 \leq x \leq -2$$

$$S_1(x) = -\frac{190}{783}(x+2)^3 + \frac{41}{29}(x+2)^2 - \frac{92}{87}(x+2) \quad -2 \leq x \leq 1$$

$$S_2(x) = \frac{67}{783}(x-1)^3 - \frac{67}{87}(x-1)^2 + \frac{76}{87}(x-1) + 3 \quad 1 \leq x \leq 4$$

6. $h_0 = 1 \quad d_0 = -2$

$$h_1 = 3 \quad d_1 = 1 \quad u_1 = 18$$

$$h_2 = 3 \quad d_2 = -2/3 \quad u_2 = -10$$

Se resuelve el sistema $\begin{cases} \frac{28}{3}m_1 + \frac{8}{3}m_2 = 18 \\ 0m_1 + 18m_2 = -10 \end{cases}$ y se obtiene $m_1 = \frac{263}{126}$ y $m_2 = -\frac{5}{9}$.

Entonces $m_0 = \frac{187}{63}$ y $m_3 = -\frac{403}{126}$ y la cerca cónica es

$$S_0(x) = -\frac{37}{252}(x+3)^3 + \frac{187}{126}(x+3)^2 - \frac{841}{252}(x+3) \quad -3 \leq x \leq -2$$

$$S_1(x) = -\frac{37}{252}(x+2)^3 + \frac{263}{252}(x+2)^2 - \frac{17}{21}(x+2) \quad -2 \leq x \leq 1$$

$$S_2(x) = -\frac{37}{252}(x-1)^3 - \frac{5}{18}(x-1)^2 + \frac{125}{84}(x-1) + 3 \quad 1 \leq x \leq 4$$

Sección 5.4 Series de Fourier y polinomios trigonométricos

$$1. f(x) = \frac{4}{\pi} \left(\operatorname{sen}(x) + \frac{\operatorname{sen}(3x)}{3} + \frac{\operatorname{sen}(5x)}{5} + \frac{\operatorname{sen}(7x)}{7} + \dots \right)$$

$$3. f(x) = \frac{\pi}{4} + \sum_{j=1}^{\infty} \left(\frac{(-1)^{j-1}}{\pi j^2} \right) \cos(jx) - \sum_{j=1}^{\infty} \left(\frac{(-1)^j}{j} \right) \operatorname{sen}(jx)$$

$$5. f(x) = \frac{4}{\pi} \left(\operatorname{sen}(x) - \frac{\operatorname{sen}(3x)}{3^2} + \frac{\operatorname{sen}(5x)}{5^2} - \frac{\operatorname{sen}(7x)}{7^2} + \dots \right)$$

$$12. f(x) = 6 + \frac{36}{\pi^2} \sum_{j=1}^{\infty} \left(\frac{(-1)^{j+1}}{j^2} \right) \cos \left(\frac{j\pi x}{3} \right)$$

Sección 6.1 Aproximaciones a la derivada

$$1. f(x) = \operatorname{sen}(x)$$

h	Aproximación a $f'(x)$, fórmula (3)	Error de la aproximación	Cota del error de truncamiento
0.1	0.695546112	0.001160597	0.001274737
0.01	0.696695100	0.000011609	0.000012747
0.001	0.696706600	0.000000109	0.000000127

$$3. f(x) = \operatorname{sen}(x)$$

h	Aproximación a $f'(x)$, fórmula (10)	Error de la aproximación	Cota del error de truncamiento
0.1	0.696704390	0.000002320	0.000002322
0.01	0.696706710	-0.000000001	0.000000000

$$5. f(x) = x^3 \quad (\text{a}) \quad f'(2) \approx 12.0025000 \quad (\text{b}) \quad f'(2) \approx 12.0000000$$

(c) Para el apartado (a): $O(h^2) = -(0.05)^2 f^{(3)}(c)/6 = -0.0025$. Para el apartado (b): $O(h^4) = -(0.05)^4 f^{(3)}(c)/30 = -0.0000$

7. $f(x, y) = xy/(x + y)$

(a) $f_x(x, y) = (y/(x + y))^2$, $f_x(2, 3) = 0.36$

h	Aproximación a $f_x(2, 3)$	Error de la aproximación
0.1	0.360144060	-0.000144060
0.01	0.360001400	-0.000001400
0.001	0.360000000	0.000000000

$f_y(x, y) = (x/(x + y))^2$, $f_y(2, 3) = 0.16$

h	Aproximación a $f_y(2, 3)$	Error de la aproximación
0.1	0.160064030	-0.000064030
0.01	0.160000600	-0.000000600
0.001	0.160000000	0.000000000

10. (a) La fórmula (3) da $I'(1.2) \approx -13.5840$ y $E(1.2) \approx 11.3024$. La fórmula (10) da $I'(1.2) \approx -13.6824$ y $E(1.2) \approx 11.2975$.

(b) Usando las reglas de derivación del cálculo infinitesimal, obtenemos $I'(1.2) \approx -13.6793$ y $E(1.2) \approx 11.2976$.

12.

h	Aprox. $f'(x)$, fórmula (17)	Error de la aproximación	Fórmula (19), cota del error total redondeo + trunc.
0.1	-0.93050	-0.00154	0.00005 + 0.00161 = 0.00166
0.01	-0.93200	-0.00004	0.00050 + 0.00002 = 0.00052
0.001	-0.93000	-0.00204	0.00500 + 0.00000 = 0.00500

15. $f(x) = \cos(x)$, $f^{(5)}(x) = -\operatorname{sen}(x)$

Use la cota $|f^{(5)}(x)| \leq \operatorname{sen}(1.4) \approx 0.98545$.

h	Aprox. $f'(x)$, fórmula (22)	Error de la aproximación	Fórmula (24), cota del error total redondeo + trunc.
0.1	-0.93206	0.00002	0.00008 + 0.00000 = 0.00008
0.01	-0.93208	0.00004	0.00075 + 0.00000 = 0.00075
0.001	-0.92917	-0.00287	0.00750 + 0.00000 = 0.00750

Sección 6.2 Fórmulas de derivación numérica

1. $f(x) = \ln(x)$

(a) $f''(5) \approx -0.040001600$ (b) $f''(5) \approx -0.040007900$

(c) $f''(5) \approx -0.039999833$ (d) $f''(5) = -0.040000000 = -1/5^2$

La respuesta más precisa es la del apartado (b).

3. $f(x) = \ln(x)$

(a) $f''(5) \approx 0.0000$ (b) $f''(5) \approx -0.0400$

(c) $f''(5) \approx 0.0133$ (d) $f''(5) = -0.0400 = -1/5^2$

La respuesta más precisa es la del apartado (b).

5. (a) $f(x) = x^2, f''(1) \approx 2.0000$

(b) $f(x) = x^4, f''(1) \approx 12.0002$

9. (a)

x	$f'(x)$
0.0	0.141345
0.1	0.041515
0.2	-0.058275
0.3	-0.158025

Sección 7.1 Introducción a la integración numérica

1. (a) $f(x) = \sin(\pi x)$	regla del trapecio	0.0
	regla de Simpson	0.666667
	regla $\frac{3}{8}$ de Simpson	0.649519
	regla de Boole	0.636165
(c) $f(x) = \sin(\sqrt{x})$	regla del trapecio	0.420735
	regla de Simpson	0.573336
	regla $\frac{3}{8}$ de Simpson	0.583143
	regla de Boole	0.593376
2. (a) $f(x) = \sin(\pi x)$	regla compuesta del trapecio	0.603553
	regla compuesta de Simpson	0.638071
	regla de Boole	0.636165
(b) $f(x) = \sin(\sqrt{x})$	regla compuesta del trapecio	0.577889
	regla compuesta de Simpson	0.592124
	regla de Boole	0.593376

Sección 7.2 Las reglas compuestas del trapecio y de Simpson

1. (a) $F(x) = \arctan(x), F(1) - F(-1) = \pi/2 \approx 1.57079632679$

(i): $M = 10, h = 0.2, T(f, h) = 1.56746305691,$

$E_T(f, h) = 0.00333326989$

(ii): $M = 5, h = 0.2, S(f, h) = 1.57079538809,$

$E_S(f, h) = 0.00000093870$

(c) $F(x) = 2\sqrt{x}, F(4) - F(\frac{1}{2}) = 3$

- (i): $M = 10, h = 0.375, T(f, h) = 3.04191993765, E_T(f, h) = -0.04191993765$
(ii): $M = 5, h = 0.375, S(f, h) = 3.00762208163, E_S(f, h) = -0.00762208163$
2. (a) $\int_0^1 \sqrt{1 + 9x^4} dx = 1.54786565469019$
(i): $M = 10, T(f, 1/10) = 1.55260945$
(ii): $M = 5, S(f, 1/10) = 1.54786419$
3. (a) $2\pi \int_0^1 x^3 \sqrt{1 + 9x^4} dx = 3.5631218520124$
(i): $M = 10, T(f, 1/10) = 3.64244664$
(ii): $M = 5, S(f, 1/10) = 3.56372816$
8. (a) Use la cota $|f^{(2)}(x)| = |- \cos(x)| \leq |\cos(0)| = 1$ para obtener $((\pi/3 - 0)h^2)/12 \leq 5 \times 10^{-9}$; ahora sustituya $h = \pi/(3M)$ y deduzca que $\pi^3/162 \times 10^8 \leq M^2$. Resuelva y obtenga $4374.89 \leq M$; puesto que M debe ser un número entero, $M = 4375$ y $h = 0.000239359$.
9. (a) Use la cota $|f^{(4)}(x)| = |\cos(x)| \leq |\cos(0)| = 1$ para obtener $((\pi/3 - 0)h^4)/180 \leq 5 \times 10^{-9}$; ahora sustituya $h = \pi/(6M)$ y deduzca que $\pi^5/34,992 \times 10^7 \leq M^4$; puesto que M debe ser un número entero, $M = 18$ y $h = 0.029088821$.
- 10.
- | M | h | $T(f, h)$ | $E_T(f, h) = O(h^2)$ |
|-----|--------|-----------|----------------------|
| 1 | 0.2 | 0.1990008 | 0.0006660 |
| 2 | 0.1 | 0.1995004 | 0.0001664 |
| 4 | 0.05 | 0.1996252 | 0.0000416 |
| 8 | 0.025 | 0.1996564 | 0.0000104 |
| 16 | 0.0125 | 0.1996642 | 0.0000026 |

Sección 7.3 Reglas recursivas y método de Romberg

1. (a)

J	$R(J, 0)$	$R(J, 1)$	$R(J, 2)$
0	-0.00171772		
1	0.02377300	0.03226990	
2	0.60402717	0.79744521	0.84845691

(c)

J	$R(J, 0)$	$R(J, 1)$	$R(J, 2)$
0	2.88		
1	2.10564024	1.84752031	
2	1.78167637	1.67368841	1.66209962

10. (ii) Para $\int_0^1 \sqrt{x} dx$, el método de integración de Romberg converge lentamente porque las derivadas superiores del integrando $f(x) = \sqrt{x}$ no están acotadas cerca de $x = 0$.

Sección 7.5 El método de integración de Gauss-Legendre (opcional)

1. $\int_0^2 6t^5 dt = 64$ (b) $G(f, 2) = 58.6666667$

3. $\int_0^1 \sin(t)/t dt \approx 0.9460831$ (b) $G(f, 2) = 0.9460411$

6. (a) $N = 4$ (b) $N = 6$

8. Si la derivada cuarta no cambia mucho, entonces $\left| \frac{f^{(4)}(c_1)}{135} \right| < \left| \frac{-f^{(4)}(c_2)}{90} \right|$.

El término del error de truncamiento de la regla de Gauss-Legendre es menor que el de la regla de Simpson.

Sección 8.1 Minimización de una función

3. (a) $f(x) = 4x^3 - 8x^2 - 11x + 5$; $f'(x) = 12x^2 - 16x - 11$;

mínimo local en $x = \frac{11}{6}$

(d) $f(x) = e^x/x^2$; $f'(x) = e^x(x-2)/x^3$; mínimo local en $x = 2$

7. (a) $f(x, y) = x^3 + y^3 - 3x - 3y + 5$

$$f_x(x, y) = 3x^2 - 3, f_y(x, y) = 3y^2 - 3$$

Puntos críticos: $(1, 1), (1, -1), (-1, 1), (-1, -1)$

Mínimo local en $(1, 1)$

(c) $f(x, y) = x^2y + xy^2 - 3xy$

$$f_x(x, y) = 2xy + y^2 - 3y, f_y(x, y) = x^2 + 2xy - 3x$$

Puntos críticos: $(0, 0), (0, 3), (3, 0), (1, 1)$

Mínimo local en $(1, 1)$

11. Como “reflejamos” el triángulo respecto del lado \overline{BG} , los extremos de los vectores \mathbf{W} , \mathbf{M} y \mathbf{R} están todos sobre un mismo segmento rectilíneo. En consecuencia, usando las propiedades de la suma de vectores y del producto por escalares, tenemos $\mathbf{R} - \mathbf{W} = 2(\mathbf{M} - \mathbf{W})$, o sea, $\mathbf{R} = 2\mathbf{M} - \mathbf{W}$.

Sección 9.1 Introducción a las ecuaciones diferenciales

1. (b) $L = 1$

3. (b) $L = 3$

5. (b) $L = 60$

10. (c) No, porque $f_y(t, y) = \frac{1}{2}y^{-2/3}$ no es continua cuando $t = 0$ y $\lim_{y \rightarrow 0} f_y(t, y) = \infty$.

13. $y(t) = t^3 - \cos(t) + 3$

15. $y(t) = \int_0^t e^{-s^2/2} ds$

17. (b) $y(t) = y_0 e^{-0.000120968t}$ (c) 2808 años (d) 6.9237 segundos

Sección 9.2 El método de Euler

1. (a)

t_k	$y_k (h = 0.1)$	$y_k (h = 0.2)$
0.0	1	1
0.1	0.90000	
0.2	0.81100	0.80000
0.3	0.73390	
0.4	0.66951	0.64800

3. (a)

t_k	$y_k (h = 0.1)$	$y_k (h = 0.2)$
0.0	1	1
0.1	1.00000	
0.2	0.99000	1.00000
0.3	0.97020	
0.4	0.94109	0.96000

6. $P_{k+1} = P_k + (0.02P_k - 0.00004P_k^2)10$ para $k = 1, 2, \dots, 8$.

Año	t_k	Población en el año t_k , $P(t_k)$	P_k	
			Método de Euler (redondeado)	Método de Euler (con más cifras)
1900	0.0	76.1	76.1	76.1
1910	10.0	92.4	89.0	89.0035
1920	20.0	106.5	103.6	103.6356
1930	30.0	123.1	120.0	120.0666
1940	40.0	132.6	138.2	138.3135
1950	50.0	152.3	158.2	158.3239
1960	60.0	180.7	179.8	179.9621
1970	70.0	204.9	202.8	203.0000
1980	80.0	226.5	226.9	227.1164

9. No. Cualquiera que sea M , el método de Euler produce $0 < y_1 < y_2 < \dots < y_M$. La solución exacta es $y(t) = \tan(t)$ e $y(3) < 0$.

Sección 9.3 El método de Heun

1. (a)

t_k	$y_k (h = 0.1)$	$y_k (h = 0.2)$
0	1	1
0.1	0.90550	
0.2	0.82193	0.82400
0.3	0.75014	
0.4	0.69093	0.69488

3. (a)

t_k	$y_k (h = 0.1)$	$y_k (h = 0.2)$
0	1	1
0.1	0.99500	
0.2	0.98107	0.98000
0.3	0.95596	
0.4	0.92308	0.92277

7. Método de mejora de Richardson para resolver $y' = (t - y)/2$ en $[0, 3]$ con $y(0) = 1$. Los elementos de la tabla son aproximaciones a $y(3)$.

k	y_k	$(4y_k - y_{2k})/3$
1	1.732422	
$1/2$	1.682121	1.665354
$1/4$	1.672269	1.668985
$1/8$	1.670076	1.669345
$1/16$	1.669558	1.669385
$1/32$	1.669432	1.669390
$1/64$	1.669401	1.669391

8. $y' = f(t, y) = 1.5y^{1/3}$, $f_y(t, t) = 0.5y^{-2/3}$. Como la derivada parcial $f_y(0, 0)$ no existe, el problema de valor inicial no está bien planteado (en el sentido del teorema de existencia y unicidad) en ningún rectángulo que contenga el punto $(0, 0)$.

Sección 9.4 El método de la serie de Taylor

1. (a)

t_k	$y_k (h = 0.1)$	$y_k (h = 0.2)$
0	1	1
0.1	0.90516	
0.2	0.82127	0.82127
0.3	0.74918	
0.4	0.68968	0.68968

3. (a)

t_k	$y_k (h = 0.1)$	$y_k (h = 0.2)$
0	1	1
0.1	0.99501	
0.2	0.98020	0.98020
0.3	0.96000	
0.4	0.92312	0.92313

6. Método de mejora de Richardson para las aproximaciones obtenidas con el método Taylor a la solución de $y' = (t - y)/2$ en $[0, 3]$ con $y(0) = 1$. Los elementos de la tabla son aproximaciones a $y(3)$.

h	y_k	$(16y_h - y_{2h})/15$
1	1.6701860	
1/2	1.6694308	1.6693805
1/4	1.6693928	1.6693903
1/8	1.6693906	1.6693905

Sección 9.5 Los métodos de Runge-Kutta

1. (a)

t_k	$y_k (h = 0.1)$	$y_k = (h = 0.2)$
0	1	1
0.1	0.90516	
0.2	0.82127	0.82127
0.3	0.74918	
0.4	0.68968	0.68969

3. (a)

t_k	$y_k(h = 0.1)$	$y_k = (h = 0.2)$
0	1	1
0.1	0.99501	
0.2	0.98020	0.98020
0.3	0.95600	
0.4	0.92312	0.92312

Sección 9.6 Métodos de predicción y corrección

1. $y_4 = 0.82126825, y_5 = 0.78369923$
 3. $y_4 = 0.74832050, y_5 = 0.66139979$
 4. $y_4 = 0.98247692, y_5 = 0.97350099$
 7. $y_4 = 1.1542232, y_5 = 1.2225213$

Sección 9.7 Sistemas de ecuaciones diferenciales

1. (a) $(x_1, y_1) = (-2.5500000, 2.6700000)$
 $(x_2, y_2) = (-2.4040735, 2.5485015)$
 (b) $(x_1, y_1) = (-2.5521092, 2.6742492)$
 5. (b) $x' = y$
 $y' = 1.5x + 2.5y + 22.5e^{2t}$
 (c) $x_1 = 2.05, x_2 = 2.17$
 (d) $x_1 = 2.0875384$

Sección 9.8 Problemas de contorno

2. No; $q(t) = -1/t^2 < 0$ para todo $t \in [0.5, 4.5]$.

Sección 9.9 El método de las diferencias finitas

1. (a) $h_1 = 0.5, x_1 = 7.2857149$
 $h_2 = 0.25, x_1 = 6.0771913, x_2 = 7.2827443$
 2. (a) $h_1 = 0.5, x_1 = 0.85414295$
 $h_2 = 0.25, x_1 = 0.93524622, x_2 = 0.83762911$

Sección 10.1 Ecuaciones hiperbólicas**4.**

t_j	x_2	x_3	x_4	x_5
0.0	0.587785	0.951057	0.951057	0.587785
0.1	0.475528	0.769421	0.769421	0.475528
0.2	0.181636	0.293893	0.293893	0.181636

5.

t_j	x_2	x_3	x_4	x_5
0.0	0.500	1.000	1.500	0.750
0.1	0.500	1.000	0.875	0.800
0.2	0.500	0.375	0.300	0.125

Sección 10.2 Ecuaciones parabólicas

3.

$x_1 = 0.0$	$x_2 = 0.2$	$x_3 = 0.4$	$x_4 = 0.6$	$x_5 = 0.8$	$x_6 = 1.0$
0.0	0.587785	0.951057	0.951057	0.587785	0.0
0.0	0.475528	0.769421	0.769421	0.475528	0.0
0.0	0.384710	0.622475	0.622475	0.384710	0.0

Sección 10.3 Ecuaciones elípticas

1. (a) $-4p_1 + p_2 + p_3 = -80$

$$p_1 - 4p_2 + p_4 = -10$$

$$p_1 - 4p_3 + p_4 = -160$$

$$p_2 + p_3 - 4p_4 = -90$$

(b) $p_1 = 41.25, p_2 = 23.75, p_3 = 61.25, p_4 = 43.75$

5. (a) $u_{xx} + u_{yy} = 2a + 2c = 0$, if $a = -c$

6. Determine si $u(x, y) = \cos(2x) + \sin(2y)$ es una solución (pues está definida en el interior de R), esto es, $u_{xx} + u_{yy} = -4\cos(2x) - 4\sin(2y) = -4u$.

Sección 11.1 El problema de los autovalores

1. (a) $|A - \lambda I| = \lambda^2 - 3\lambda - 4 = 0$ implica que $\lambda_1 = -1$ y $\lambda_2 = 4$. Sustituyendo cada autovalor en $|A - \lambda I| = 0$ y resolviendo los sistemas, obtenemos $V_1 = [-1 \ 1]'$ y $V_2 = [2/3 \ 1]'$, respectivamente.

10. Si $\lambda = 2$ es un autovalor de A con autovector V , entonces $AV = 2V$. Multiplicando por la izquierda en ambos miembros por A^{-1} , obtenemos: $A^{-1}AV = A^{-1}(2V)$, o sea, $V = 2A^{-1}V$. Así pues, $A^{-1}V = 1/2V$.

Sección 11.2 Los métodos de las potencias

1. $(A - \alpha I)V = AV - \alpha IV = AV - \alpha V = \lambda V - \alpha V = (\lambda - \alpha)V$. Por tanto, $(\lambda - \alpha, V)$ es una pareja autovalor-autovector de $A - \alpha I$.

5. (a) $|A - 1I| = \begin{vmatrix} -0.2 & 0.3 \\ 0.2 & -0.3 \end{vmatrix} = 0$

(b) $\begin{bmatrix} -0.2 & 0.3 & 0 \\ 0.2 & -0.3 & 0 \end{bmatrix}$ equivale a $\begin{bmatrix} -0.2 & 0.3 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, luego $0.2x = 0.3y$.

Tomemos $y = t$, entonces $x = 3/2$. Así que los autovectores asociados con $\lambda = 1$ son $\{t[3/2 \ 1]': t \in \mathbb{R}, t \neq 0\}$.

- (c) Usando el autovector del apartado (b) deducimos que, a la larga, las 50 000 personas estarán divididas, en sus preferencias por las marcas \mathbf{X} e \mathbf{Y} , en una proporción 3 a 2, respectivamente. Es decir, $[30\,000 \ 20\,000]'$.

Sección 11.3 El método de Jacobi

3. (a) Las dos parejas autovalor-autovector de la matriz $\mathbf{A} = \begin{bmatrix} 4 & 2 \\ 3 & -1 \end{bmatrix}$ son $(5, [2 \ 1]')$ y $(-2, [-1/3 \ 1]')$. Por tanto, la solución general es $\mathbf{X}(t) = c_1 e^{5t} [2 \ 1]' + c_2 e^{-2t} [-1/3 \ 1]'$. Poniendo $t = 0$ determinamos c_1 y c_2 , esto es, $[1 \ 2]' = c_1 [2 \ 1]' + c_2 [-1/3 \ 1]'$. Así que $c_1 = 0.7143$ y $c_2 = 1.2857$.

Sección 11.4 Autovalores de matrices simétricas

1. A partir de (3) deducimos que $\mathbf{W} = \frac{\mathbf{X}-\mathbf{Y}}{\|\mathbf{X}-\mathbf{Y}\|_2}$ y, de acuerdo con la Figura 11.4, $\mathbf{Z} = \frac{1}{2}(\mathbf{X} + \mathbf{Y})$.
Haciendo el producto escalar

$$\begin{aligned} \frac{\mathbf{X} - \mathbf{Y}}{\|\mathbf{X} - \mathbf{Y}\|_2} \cdot \frac{1}{2}(\mathbf{X} + \mathbf{Y}) &= \frac{(\mathbf{X} - \mathbf{Y}) \cdot (\mathbf{X} + \mathbf{Y})}{2\|\mathbf{X} - \mathbf{Y}\|_2} \\ &= \frac{\mathbf{X} \cdot \mathbf{X} + \mathbf{X} \cdot \mathbf{Y} - \mathbf{Y} \cdot \mathbf{X} - \mathbf{Y} \cdot \mathbf{Y}}{2\|\mathbf{X} - \mathbf{Y}\|_2} = \frac{\|\mathbf{X}\|^2 - \|\mathbf{Y}\|^2}{2\|\mathbf{X} - \mathbf{Y}\|_2} = 0, \end{aligned}$$

ya que \mathbf{X} e \mathbf{Y} tienen la misma norma.

2. $\mathbf{P}' = (\mathbf{I} - 2\mathbf{X}\mathbf{X}')' = \mathbf{I}' - 2(\mathbf{X}\mathbf{X}')' = \mathbf{I} - 2(\mathbf{X}')'\mathbf{X}' = \mathbf{I} - 2\mathbf{X}\mathbf{X}' = \mathbf{P}$

Índice analítico

- Aceleración de la convergencia, 91
 método de Aitken, 99, 109 (#10–#14)
 método de Steffensen, 99, 104
 Newton-Raphson, 78, 91, 96 (#23), 192
- Ajuste de datos en mínimos cuadrados
 ajuste exponencial, 285, 287, 288
 ajuste lineal, 276, 280, 282 (#7), 300 (#17)
 ajuste no lineal, 279, 285, 288, 294
 ajuste plano, 300 (#17, #18)
 ajuste polinomial, 230, 294, 296
 error cuadrático medio, 274
 linealización de los datos, 289
 polinomios trigonométricos, 322, 328, 331
 recta de regresión, 282 (#7), 286
- Análisis gráfico
 iteración de punto fijo, 51–53
 método de la secante, 88
 método de Newton-Raphson, 77, 86, 87
- Aproximación
 lineal, 300 (#17), 302
 orden de, 32, 35
- Aproximación de datos
 ajuste exponencial, 285, 287
 ajuste polinomial, 294, 296
 lineal, 274, 276, 280
 mínimos cuadrados, 279, 300 (#17)
 recta de regresión, 276, 280
- Aproximación de funciones
 aproximaciones de Padé, 263, 267
 cerchas, 302, 304, 309, 317
 funciones racionales, 263
 lineal, 238 (#12)
 minimax, 251, 253, 258
 mínimos cuadrados, 276, 279, 294
 polinomio de Chebyshev, 250, 251, 253, 258, 261
 polinomio de Lagrange, 225, 230, 232, 235, 258
 polinomio de Newton, 239, 243, 247
 polinomio de Taylor, 9, 28, 34, 206
splines, 309, 317
- Área de una superficie de revolución, 395 (#3)
- Autovalores, 605
 algoritmo QR, 651, 658
 dominante, 616
 método de Householder, 643
 método de Jacobi, 629, 639

- método de las potencias, 615, 617, 621
método de las potencias inversas, 620, 623, 624
métodos de las potencias, 625
polinomio característico, 605
Autovector, 605
- Base, 604
- $C[a, b]$, 3
 $C^n[a, b]$, 5
Cancelación de cifras significativas, 30
Cercha
cúbica, 304, 305
cúbica natural, 308, 309
cúbica sujeta, 308, 309, 317
integración numérica, 320 (#12)
lineal, 302
restricciones en los extremos, 308, 309
Ceros
de los polinomios de Chebyshev, 252
de una función, 59, 83
Cifras
binarias, 15, 19, 20
decimales, 15, 21, 24
significativas, 27, 310
Coeficientes
del polinomio trigonométrico, 323, 329
serie de Fourier, 329
Coma flotante, 21, 23, 24
exponente, 21
mantisa, 21
Condición
de contorno de Neumann, 587, 591
de Lipschitz, 467
Convergencia
aceleración, 91, 96 (#21–#23), 99, 101, 104
criterios, 69, 73
cuadrática, 83, 85, 91, 96 (#21, #23)
global (local), 69
lineal, 83, 85, 99
Newton-Raphson, 85, 91, 96 (#21, #23)
orden de, 35, 83
velocidad, 83
Convergente
serie, 8, 109 (#10–#14)
sucesión, 3
Criterio de parada, 68 (#13)
método de bisección, 65
método de la posición falsa, 65, 66
método de la secante, 93
método de Newton-Raphson, 92
método de Romberg, 410
método de Runge-Kutta-Fehlberg, 509
Cuadrático
error, 274, 276
Cuadratura
adaptativa, 415, 421
cerchas cúbicas, 320 (#12)
método de Gauss-Legendre, 423, 425, 427, 428
método de Romberg, 404, 406, 409, 410, 413 (#11)
Newton-Cotes, 374
precisión de una fórmula, 380
regla compuesta de Simpson, 385, 393
regla compuesta del trapecio, 384, 393
regla de Boole, 374, 403, 407, 412 (#3, #4), 423 (#3)
regla de Simpson, 374, 383 (#9), 385, 389, 393, 402, 412 (#6), 420, 421
regla del punto medio, 397 (#12), 413 (#11)
regla del trapecio, 374, 384, 388, 393, 399, 409
regla tres octavos de Simpson, 383 (#9)
reglas compuestas, 378, 379, 384, 385, 388, 389, 393
Deflación de autovalores, 627
Derivación numérica, 355
de orden superior, 360
diferencias centradas, 338, 339, 355, 367 (#7, #8)

- diferencias progresivas, 360, 368 (#13)
- diferencias regresivas, 360
- error, 338, 340, 342, 344
- extrapolación de Richardson, 345, 347
- fórmulas, 338, 347, 355, 360, 548, 560, 571, 583
- orden superior, 355
- Derivada**
- de orden superior, 355, 360, 548
 - definición, 4, 336
 - parcial, 351 (#7), 560, 571, 583
- polinomio de Lagrange, 360
- polinomio de Newton, 363
- polinomios, 222
- Determinante, 124, 126, 135, 166
- triangular, 135
- Diferencias**
- centradas, 338, 339, 355, 367 (#7, #8)
 - divididas, 242, 243
 - finitas, 548, 554, 557, 560, 571, 584, 595
 - progresivas, 360, 368 (#13), 571, 578
 - regresivas, 360
 - tabla, 243
- División**
- entre cero, 82, 85
 - sintética, 10, 217
- Ecuación**
- de Helmholtz, 583, 593
 - de Laplace, 559, 583, 594, 595
 - de ondas, 558, 560, 562
 - de Poisson, 583, 593
 - del calor, 558, 570
 - en diferencias, 548, 560, 571, 576, 584, 595
- Ecuaciones diferenciales ordinarias**
- de orden superior, 531
 - estabilidad, 520, 523
 - existencia y unicidad, 468
 - extrapolación de Richardson, 488 (#7), 495 (#6), 512 (#7)
 - método de Adams-Bashforth-Moulton, 515, 524
 - método de disparo, 541, 546
- método de Euler, 470, 474, 478
 - método de Euler modificado, 505
 - método de Hamming, 526
 - método de Heun, 482, 484, 486, 505
 - método de las diferencias finitas, 548, 554, 595
 - método de Milne-Simpson, 518, 525
 - método de predicción y corrección, 515, 518, 524–526
 - método de Runge-Kutta, 497, 501, 508, 543
 - método de Runge-Kutta-Fehlberg, 505, 509
 - método de Taylor, 490, 491, 494
 - problema de contorno, 540, 541, 546, 548, 554
 - problema de valor inicial, 465, 468, 529
- Ecuaciones en derivadas parciales**, 557
- casi-lineales, 557
 - elípticas, 559, 582
 - hiperbólicas, 557, 560
 - método de Crank-Nicholson, 575, 579
 - método de las diferencias finitas, 557, 560, 571, 584
 - método de las diferencias progresivas, 571, 578
 - parabólicas, 558, 570
- Ecuaciones normales de Gauss**, 276
- Eliminación gaussiana**, 137, 141, 150, 157
- complejidad computacional, 161
 - factorización LU , 155, 157, 165, 166
 - multiplicadores, 139, 142
 - pivoteo, 139, 144
 - pivoteo parcial, 166
 - sistema tridiagonal, 153 (#1), 182 (#3), 308, 548
 - sustitución progresiva, 137 (#2), 158
 - sustitución regresiva, 133, 135, 158
- Error**
- absoluto, 26
 - cancelación, 30

- computador, 23, 28, 148
- cuadrático medio, 274, 276
- datos, 39, 220, 341
- de redondeo, 23, 28, 338, 340, 342, 344
- de truncamiento, 28, 338, 340
- derivación numérica, 338, 340, 342, 344, 358
- ecuaciones diferenciales, 475, 484, 491, 501, 516, 519, 562
- estable, 36
- global final, 475, 484, 491, 501
- inestable, 36
- integración, 373, 388, 389, 409
- iteración de punto fijo, 51
- polinomio de Taylor, 207, 211
- polinomio interpolador, 207, 232, 258
- propagación, 35
- pérdida de cifras significativas, 30
- relativo, 26, 73
- Exponente, 21
- Extremos, 434, 438
- Factorización
 - LU , 155, 157, 164–166
 - $PA = LU$, 164, 165
 - QR , 652
 - triangular, 155, 157, 164–166
- Fenómeno de Runge, 256
- Fórmula
 - de iteración de Newton-Raphson, 91, 92
 - ecuación de segundo grado, 42 (#12)
 - para el valor corrector, 516
 - para el valor predictor, 515
- Fórmulas
 - de Euler, 323
 - de Newton-Cotes, 374
- Fracciones binarias, 19
- Función
 - continua, 3
 - continua a trozos, 323
 - impar, 326
 - lipschitziana, 467
 - par, 325
 - periódica, 322
 - racional, 263
- unimodal, 436
- Gradiente, 455
- Incremento óptimo
 - derivación, 341, 343, 344, 358
 - ecuaciones diferenciales, 506, 517, 521
 - integración, 415
 - integración numérica, 388, 389
- Independencia lineal, 603
- Integración numérica
 - adaptativa, 415, 421
 - cerchas cúbicas, 320 (#12)
 - extrapolación de Richardson, 406
 - método de Gauss-Legendre, 423, 425, 427, 428
 - método de Romberg, 404, 406, 409, 410, 413 (#11)
 - Newton-Cotes, 374
 - precisión de una fórmula, 380
 - regla compuesta de Simpson, 385, 393
 - regla compuesta del trapecio, 384, 393
 - regla de Boole, 374, 407, 412 (#3, #4), 423 (#3)
 - regla de Boole recursiva, 403, 407
 - regla de Simpson, 374, 383 (#9), 385, 389, 393, 412 (#6), 420, 421
 - regla de Simpson recursiva, 402, 407
 - regla del punto medio, 397 (#12), 413 (#11)
 - regla del trapecio, 374, 384, 388, 393, 399
 - regla del trapecio recursiva, 400, 407, 409
 - regla tres octavos de Simpson, 383 (#9)
 - reglas compuestas, 378, 379, 384, 385, 388, 389, 393
 - spline cúbico, 320 (#12)
- Interpolación
 - aproximaciones de Padé, 263, 267
 - cercha cúbica, 305, 309, 310, 317
 - error, 9, 34, 206, 230, 232, 258
 - extrapolación, 217

- fenómeno de Runge, 256
 función cúbica a trozos, 305
 función lineal a trozos, 302
 funciones racionales, 263
 integración numérica, 320 (#12), 373
 lineal, 225, 238 (#12), 276, 300 (#17)
 mínimos cuadrados, 276, 294
 oscilación polinomial, 295
 polinomial, 274
 polinomio de Chebyshev, 250, 253, 258, 261
 polinomio de Lagrange, 225, 230, 232, 235, 258
 polinomio de Newton, 239, 243, 247
 polinomio de Taylor, 9, 28, 34, 206, 339, 356
 polinomios trigonométricos, 322, 328, 331
spline cúbico, 309, 310, 317
 Iteración de punto fijo, 48, 54, 189
 cota del error, 51
- Lineal**
 a trozos, 302
 convergencia, 83, 85, 99
- Localización de raíces**, 57, 59, 75
Longitud de una curva, 394 (#2)
- Límite**
 de una función, 2
 de una sucesión, 3
 suma de una serie, 8
- Mala condición**
 ajuste en mínimos cuadrados, 148
 matriz, 147, 153 (#15)
- Mantisa**, 21
- Matriz**
 ampliada, 138, 141, 158
 autovalor, 605
 autovector, 605
 de diagonal dominante, 175, 177, 179, 180
 de Hilbert, 153 (#15)
 de una rotación, 126
 determinante, 124, 126, 135, 166
 diagonalización, 609
- factorización *LU*, 155, 157, 165, 166
 factorización *QR*, 652
 forma de Hessenberg, 649
 gradiente, 447
 identidad, 123, 157
 igualdad de matrices, 117
 inversa, 123, 124, 126
 invertible, 123
 jacobiana, 186, 192
 mal condicionada, 147, 153 (#15)
 multiplicación, 121, 123, 157, 165
 no invertible, 124
 no singular, 123
 norma, 613
 ortogonal, 612, 643
 permutación, 163, 165
 polinomio característico, 605
 reflexión de Householder, 643
 semejante, 630
 simétrica, 119 (#6), 611, 612, 629, 639, 643
 singular, 124
 suma, 118
 traspuesta, 114, 119 (#5), 292
- triangular, 137 (#2)
 triangular inferior, 132, 137 (#2)
 triangular superior, 132
 tridiagonal, 153 (#1), 182 (#3), 308, 548, 649
- unidad, 123
- Media aritmética**, 281 (#4, #5, #6)
- Método**
 iterativo de punto fijo, 590
 iterativo para ecuaciones en derivadas parciales, 590
 iterativo para la ecuación de Laplace, 590
 de Adams-Bashforth-Moulton, 515, 524
 de Aitken, 99, 109 (#10–#14)
 de bisección, 59, 60, 65
 de Bolzano, 59
 de Crank-Nicholson, 575, 579
 de disparo, 541, 546
 de eliminación de Gauss, 133, 135, 137, 139, 141, 144, 146, 150, 155, 157, 158, 161, 165, 166
 de Euler, 470, 474, 478, 530

- de Euler modificado, 505
- de extrapolación de Richardson, 345, 347, 406, 488 (#7), 495 (#6), 512 (#7)
- de Gauss-Legendre, 423, 425, 427, 428
- de Halley, 96 (#22)
- de Hamming, 519, 526
- de Heun, 482, 484, 486, 505
- de Horner, 10, 217
- de Householder, 643
- de Jacobi para autovalores, 629, 639
- de la posición falsa, 63, 66
- de la secante, 88, 93, 96 (#20)
- de la sección áurea, 435, 448
- de las diferencias finitas, 548, 554, 557, 560, 571, 584, 595
- de las diferencias progresivas, 571
- de las potencias, 615, 617, 625
- de las potencias inversas, 620
- de Milne-Simpson, 518, 525
- de Muller, 101, 106
- de Nelder-Mead, 440, 450
- de Newton-Raphson, 77, 85, 91, 92, 96 (#21, #23), 96 (#23), 192, 196
- de Newton-Raphson para raíces múltiples, 83, 91
- de paso simple, 515
- de predicción y corrección, 515, 518, 524–526
- de Romberg, 404, 406, 409, 410, 413 (#11)
- de Runge-Kutta, 497, 501, 508, 530, 543
- de Runge-Kutta-Fehlberg, 505, 509
- de sobrerelajación sucesiva, 590
- de Steffensen, 101, 104
- de sustitución progresiva, 137 (#2)
- de sustitución regresiva, 133, 135
- de Taylor para ecuaciones diferenciales, 490, 491, 494
- del descenso por la máxima pendiente, 447, 455
- iterativo de Gauss-Seidel, 174, 177, 180
- iterativo de Jacobi, 171, 177, 179
- iterativo de punto fijo, 47, 54, 189
- iterativo de Seidel, 190, 195, 197, 198
- iterativo para ecuaciones en derivadas parciales, 592
- iterativo para la ecuación de Laplace, 592
- multipaso, 515, 518, 524–526
- QR* para autovalores, 651, 658
- tangencial, 77, 92
- Minimización**
- método del descenso por la máxima pendiente, 447, 455
- método del gradiente, 447, 455
- Nelder-Mead, 440, 450
- sección áurea, 435, 448
- Mínimos cuadrados**
- ajuste lineal, 276, 280
- ajuste no lineal, 279, 288, 294
- ajuste polinomial, 294, 296
- error cuadrático medio, 274
- linealización de los datos, 289
- polinomios trigonométricos, 322, 328, 331
- Modelos**
- balística, 81, 480 (#8), 489 (#6)
- cadena de Markov, (#5) 627
- crecimiento de poblaciones, 299 (#6, #7)
- desintegración radiactiva, 470 (#17)
- epidemia, 481 (#9)
- movimiento de un proyectil, 81
- predador-presa, 537 (#13)
- Múltiple**
- raíz, 83, 91, 96 (#21, #23)
- Multiplicación encajada**, 10, 241
- Nodos**, 220, 225, 230, 232, 373, 423 de Chebyshev, 252, 254
- Norma**
- euclídea, 178, 179
- matricial, 613
- Notación científica**, 21
- Números**
- binarios, 14, 19, 20
- del computador, 21, 23
- en coma flotante, 23, 24

- error de redondeo, 23
precisión en coma flotante, 23
- $O(h^n)$, 32, 35, 233, 338, 340, 356, 360, 404, 409, 475, 484, 491, 501, 516, 519, 548, 560, 571, 583
- Optimización
método del descenso por la máxima pendiente, 447, 455
método del gradiente, 447, 455
Nelder-Mead, 440, 450
sección áurea, 435, 448
- Orden
de aproximación, 32, 35, 233, 338, 356, 360, 404, 409
de convergencia, 35, 83
- Oscilaciones polinomiales, 295
- Pérdida de cifras significativas, 30
- Pesos de las fórmulas de integración, 373, 427
- Pivote
elemento, 139
estrategias, 144, 146
fila, 139
parcial, 146
parcial escalado, 146
- Poda, o redondeo por debajo, 29
- Polinomio
a trozos, 304, 305
característico, 605
de Chebyshev, 250, 253, 254, 258, 261
de Lagrange, 225, 230, 232, 256
de Newton, 239, 243, 247, 363
de Taylor, 9, 28, 34, 206, 208, 339, 356
derivada, 222, 360, 363
evaluación, 222
interpolación, 222, 225, 227, 230, 235, 243, 247, 258
ortogonalidad, 258
oscilaciones, 295
trigonométrico, 322, 328, 331
- Precisión
cuadratura, 380
de un computador, 23
doble, 24
simple, 24
- Problema
de contorno, 540, 541, 546, 554
de Dirichlet, 595
de valor inicial, 465, 468, 529
de valores en la frontera, 540, 546, 554
- Progresión geométrica, 18, 57
- Propagación del error, 35
- Punto fijo, 47
- Radio espectral, 613
- Raíces
bisección, 59, 60, 65
cálculo, 45, 57, 77, 83, 99, 183, 191
cúbicas, 94 (#11)
de una ecuación, 59, 83
división sintética, 10, 217
dobles, 83, 85, 96 (#21)
ecuación de segundo grado, 42 (#12)
localización, 57, 59, 75
Muller, 101, 106
múltiples, 83, 91, 96 (#21, #23)
Newton-Raphson, 91, 92, 96 (#23), 192, 196
posición falsa, 63, 66
secante, 88, 93, 96 (#20)
simples, 83, 85, 96 (#22)
Steffensen, 101, 104
- Redondeo, 28
poda, 29
por debajo, 29
- Regla
compuesta de Simpson, 379, 385, 389, 393
compuesta del trapecio, 378, 384, 388, 393
de Barrow, 7
de Boole, 374, 403, 407, 412 (#3, #4), 423 (#3)
de Ruffini, 10, 217
de Simpson, 374, 383 (#9), 389, 402, 420
de tres, 302
del punto medio, 397 (#12), 413 (#11)
del trapecio, 374, 388, 400, 409

- tres octavos de Simpson, 374, 383
 (#9), 412 (#6)
- Rotación, 126, 630
- Sección áurea, 435, 448
- Serie
- binomial, 214 (#14)
 - convergente, 8, 109 (#10–#14), 206, 212
 - de Fourier, 323
 - de Fourier discreta, 328
 - de Maclaurin, 264
 - de Taylor, 9, 28, 34, 206, 339, 356, 490, 491, 494
 - geométrica, 18, 57
 - suma, 8
- Simple
- raíz, 83, 85, 96 (#22)
- Sintética
- división, 10
- Sistema de ecuaciones diferenciales, 529
- método de Euler, 530
 - método de Runge-Kutta, 530
- Sistema lineal, 126, 133, 141, 165, 171, 290
- eliminación gaussiana, 133 (#2), 135 (#2), 137, 137 (#2), 139, 141, 144, 146, 157
 - equivalencia, 137, 139
 - factorización LU , 155, 157, 165, 166
 - factorización $PA = LU$, 166
 - forma matricial, 122, 126, 139, 155, 157, 165, 166
 - homogéneo, 602
 - mal condicionado, 147
 - matriz ampliada, 138, 141, 158
 - métodos iterativos, 171, 174, 177, 179, 180
 - pivoteo, 139, 144, 147
 - solución única, 126
 - sustitución progresiva, 137 (#2)
 - sustitución regresiva, 133, 135, 150
 - triangular, 133, 135, 137 (#2)
 - tridiagonal, 153 (#1), 182 (#3), 308, 548
- Sistema no lineal, 183, 191
- método de Newton-Raphson, 191, 192, 196
- método iterativo de Seidel, 190, 195, 197, 198
- Spline*
- cúbico, 304, 305
 - lineal, 302
 - natural, 308, 309
 - restricciones en los extremos, 308
 - sujeto, 308, 309, 317
- Sucesión, 3
- convergente, 3
 - de errores, 3
 - iteración de punto fijo, 47
- Suma de una serie, 8
- Tamaño de paso óptimo
- derivación, 341, 343, 344, 358
 - ecuaciones diferenciales, 506, 521
 - ecuación diferencial, 517
 - integración numérica, 388, 389, 415
 - interpolación, 254
- Teorema
- de Bolzano, 3, 59
 - de los círculos de Gershgorin, 613
 - de los ejes principales, 612
 - de los valores extremos, 4
 - de Rolle, 5, 6, 215 (#20), 231, 238 (#13)
 - de Rolle generalizado, 6, 215 (#20)
 - de Schur, 610
 - de Taylor, 9, 34
 - de Weierstrass, 4
 - del radio espectral, 613
 - del valor intermedio, 3, 59
 - del valor medio, 50, 467
 - del valor medio de Lagrange, 6
 - del valor medio para derivadas, 6
 - del valor medio para integrales, 7
 - del valor medio ponderado para integrales, 8
 - espectral para matrices simétricas, 612
 - fundamental del cálculo, 7, 482
 - regla de Barrow, 7
- Transformaciones
- de semejanza, 630, 631

- elementales, 137, 139
- Triangularización superior, 150, 165
- Trigonométrico
 - polinomio, 328, 331
- Truncamiento
 - error, 28, 338
- Unimodal
 - función, 436
- Vectores
 - combinación lineal, 113
 - distancia entre dos puntos, 114, 177
 - módulo, 113, 178, 179
 - norma euclídea, 113, 177, 179
 - producto escalar, 113