MÉTODOS NUMÉRICOS
APLICADOS A LA INGENIERÍA

Buscar CD.

MÉTODOS NUMÉRICOS :
APLICADOS A LA INGENIERÍA

ANTONIO NIEVES HURTADO

FEDERICO C. DOMÍNGUEZ SÁNCHEZ

Profesores de la Academia de Matemáticas Aplicadas
ESIQIE - IPN

QUINTA REIMPRESIÓN
MÉXICO, 1999

COMPAÑÍA EDITORIAL CONTINENTAL, S.A. DE C.V.
MÉXICO

Colaboración especial:
Dr. Guillermo Marroquín Suárez
Profesor de la Academia de Matemáticas Aplicadas
ESIQIE-IPN

Revisión Técnica
M.C. José Luis Turriza
Profesor de Matemáticas
ESIME-IPN

Métodos numéricos

Renacimiento 180, Colonia San Juan Tlihuaca,
Delegación Azcapotzalco, Código Postal 02400, México, D.F.

Miembro de la Cámara Nacional de la Industria Editorial.
Registro núm. 43

ISBN 968-26-1260-8

Impreso en México
Printed in Mexico

Primera edición: 1995
Cuarta reimpresión: 1998
Quinta reimpresión: 1999

A los Eggli:
Violet (Mom), Fred,
Josephine, Richard y David.

Gracias

Antonio

A mis hijos Alura,
Alejandra y Federico,
a mis hermanos, y a
la memoria de mis padres.

Federico

PREFACIO

El análisis numérico y sus métodos son una dialéctica entre el análisis matemático cualitativo y el análisis matemático cuantitativo; el primero nos dice por ejemplo que bajo ciertas condiciones algo existe, que es o no único, etc. mientras que el segundo complementa al primero, permitiendo calcular **aproximadamente** el valor de aquéllo que existe. Es pues una reflexión sobre los cursos tradicionales de cálculo, álgebra lineal, ecuaciones diferenciales, etc. desde el punto de vista numérico, concretando en una serie de métodos o algoritmos cuyo estudio y uso en diferentes áreas de ingeniería y ciencias es la finalidad de este libro.

Dado que cada algoritmo implica numerosas operaciones lógicas, aritméticas y en algunos casos graficaciones, la computadora es fundamental para el estudio y uso de éstos. El binomio computadora-lenguaje de alto nivel (Fortran, Basic, Pascal y otros), ha sido utilizado durante muchos años para la enseñanza y el aprendizaje de los métodos numéricos; si bien esta fórmula ha sido exitosa y sigue vigente, también es cierto que la aparición de paquetes comerciales como Graphics Calculus (GC), Math-CAD, Maple, por citar algunos de los más conocidos, han venido a apoyar el trabajo de profesores y alumnos, permitiendo variantes como ilustraciones geométricas de algunos métodos y de las ideas que los sustentan; programación más sencilla y rápida de ciertos algoritmos; uso directo de los métodos; exploración de conjeturas planteadas por el alumno o profesor etc. de modo que esta rama de las matemáticas resulta hoy en día más atractiva y útil para casi todos los estudiantes de ingeniería y ciencias.

El contenido del libro gira alrededor de cuatro ideas matemáticas fundamentales: punto fijo, eliminación de Gauss (ortogonalización), aproximación de funciones con polinomios y aproximación de derivadas con diferencias finitas; y se usan como herramientas de deducción la expansión en serie de Taylor y el teorema del valor medio.

Por la naturaleza del material el libro puede dividirse en tres partes : algebraica, de análisis y de dinámica, sustentándose todas ellas en un primer capítulo de ideas básicas sobre los sistemas numéricos y los errores debidos al manejo de números en la computadora (véase la red de temas e interrelación).

En la parte algebraica los problemas que se resuelven son: una ecuación no lineal en una incógnita, sistemas de ecuaciones lineales y sistemas de ecuaciones no lineales (capítulos 2 a 4).

En la parte de análisis, se resuelven problemas de interpolación, derivación e integración, desarrollándose sus algoritmos de solución a partir de la idea central de aproximación de funciones con polinomios (capítulos 5 y 6).

La tercera parte (capítulos 7 y 8), de dinámica, se refiere a la solución de ecuaciones diferenciales ordinarias y ecuaciones diferenciales parciales, en donde los conceptos de integración y de aproximación de derivadas por diferencias divididas son fundamentales, ya que a partir de ellos, así como de las expansiones multivariables de Taylor, se obtienen los diferentes algoritmos de solución.

Los algoritmos presentados en el libro se sustentan en teoremas que se **describen** o **enuncian** a lo largo de los temas desarrollados, de modo que se tenga fundamentación teórica (y no un conjunto de "recetas" de aplicación), y proporcione al lector recursos para usar con más propiedad, profundidad y racionalidad estos algoritmos.

La aplicación de los diferentes métodos numéricos se lleva a cabo guiando al lector en la visualización de problemas realistas en ingeniería, en los que al aplicarse las leyes básicas que los rigen, se obtienen las ecuaciones matemáticas que los modelan.

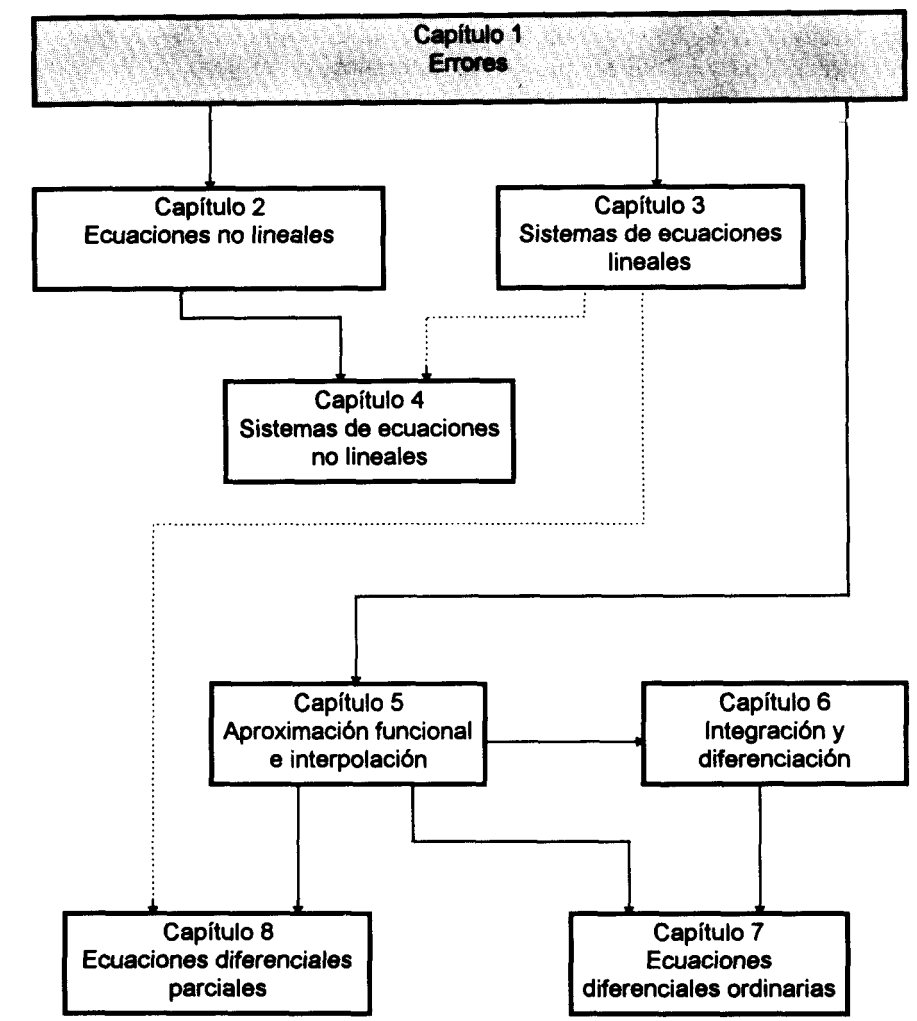
Se aprovechan las características y limitaciones de los distintos algoritmos estudiados, para seleccionar los más adecuados para el modelo matemático a resolver. Debido a la preferencia de los profesores por algún lenguaje en el cual programar, estos algoritmos se dan en pseudocódigo y se codificaron en Fortran (Lahey Personal Fortran), Pascal, Basic y C (Turbo Pascal, Turbo Basic y Turbo C de Borland). La mayoría son programas de propósito general y están documentados para permitir su utilidad en otras aplicaciones de métodos del mismo tipo. Además se desarrollaron algunos de ellos en Math-CAD (un pizarrón electrónico) que permite una variante en la programación tradicional y en las posibilidades de exploración.

Mediante la discusión de los resultados obtenidos con los programas, se motiva en el lector la necesidad de analizar la congruencia entre la interpretación física, la solución y el sistema modelado, así también como entre los resultados, el modelo matemático y el algoritmo de cálculo usado.

El libro viene acompañado con un disco que contiene los programas mencionados anteriormente: 28 en Pascal, Basic y C y 29 en Fortran (el programa que calcula raíces imaginarias con el método de Müller sólo está en Fortran). Además, el disco contiene dos programas tutoriales en métodos numéricos, ambos desarrollados con apoyo institucional del I.P.N. En el primero (TMN1.EXE) se estudia la solución de una ecuación no lineal con el método de Newton-Raphson y la solución de sistemas de ecuaciones lineales y de sistemas de ecuaciones no lineales. En el segundo (TMN2.EXE) se estudia la interpolación, derivación e integración numérica. Ambos tutoriales están divididos en tres secciones: la primera presenta el método con ayuda gráfica y otros efectos visuales y sonoros; en la segunda el usuario deberá resolver, paso a paso, un ejemplo propuesto por el programa; se finaliza con una sección de cálculo directo en donde el usuario puede solicitar la solución de un problema. En la segunda sección (interactiva), el programa muestra las operaciones a realizar y espera la respuesta. Para calcularla, se puede invocar una calculadora con la tecla F1, indicar las operaciones y oprimir la tecla Enter. Una vez que se obtiene el resultado, se oprime la tecla Esc y la calculadora traslada el valor al tutorial; éste lo valida y, en caso de ser correcto, pasa a la siguiente operación. Si el resultado es incorrecto, el programa lo indica y solicita un nuevo intento. Al tercer intento, el tutorial anotará el resultado correcto y pasará a la siguiente operación.

Estos programas trabajan en ambiente gráfico bajo MS-DOS 3.0 o posterior en microcomputadoras personales IBM o compatibles con un mínimo de 512 Kb y con monitor a color CGA o de mayor resolución (es posible ejecutarlos en monitor monocromático también).

RED DE TEMAS E INTERRELACION



AGRADECIMIENTOS

Esta obra tiene su origen en apuntes para los cursos de métodos numéricos en la carrera de Ingeniero Químico Industrial del Instituto Politécnico Nacional, desarrollados durante una estancia de año sabático en el Instituto Tecnológico de Celaya, y posteriormente, a raíz de un certamen organizado por el propio I.P.N., se convirtieron en una propuesta de libro que ganó el primer lugar en el Primer Certamen Editorial Politécnico en 1984. Desde entonces, con actualizaciones continuas, ha sido utilizado como texto para estos cursos. Los autores agradecen al Instituto Politécnico Nacional la facilidad que otorgó para que la Editorial CECSA lo publicara y su autorización para incluir los programas tutoriales con el libro.

Agradecemos también a las muchas personas que en distintas formas colaboraron para la realización de este libro. En especial al Dr. Guillermo Marroquín Suárez por el material, ideas y colaboración intensa que durante un año aportó a la elaboración de los capítulos 5 a 8; su valiosa ayuda permitió darle un enfoque interesante de aplicación al material matemático; sus observaciones a las soluciones enriquecieron el análisis de los ejercicios. Al Ing. Arturo López García por el programa del método de Muller. Al M.C. José Luis Turriza, por su minuciosa revisión técnica. A los ingenieros Eva Zepeda Lobato, Mary Carmen Peláez Acero y Victor M. Martínez Reyes por la captura y corrección de los manuscritos; a los ingenieros Gloria Catalina Valdez Barrón, Jesús García Manríquez, José Hernández Sánchez, Rosa González Cortés y Alejandro Correa Flores que programaron los tutoriales.

Nuestro agradecimiento especial al personal de la Editorial CECSA por su interés constante en lograr una óptima presentación técnica, estética y de estilo en el libro.

Contenido

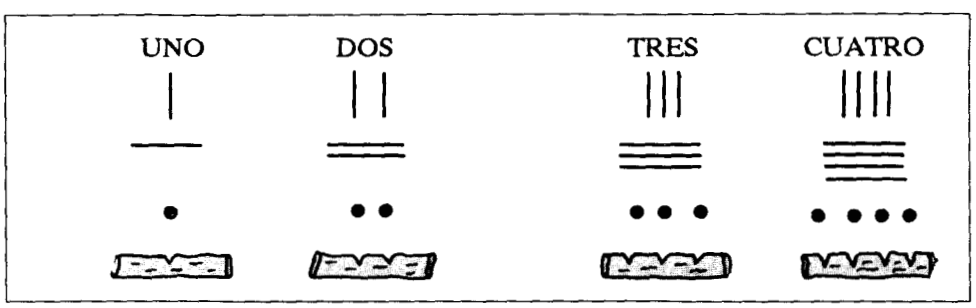
CAPÍTULO 1

ERRORES

EN ESTE CAPÍTULO se revisarán los sistemas numéricos binario, decimal y octal; las conversiones entre ellos, su manejo en una computadora, los diversos errores que ello puede ocasionar y algunas formas de evitarlos.

INTRODUCCIÓN

En la antigüedad, los números naturales se representaban con distintos tipos de símbolos o **numerales**. A continuación se presentan algunos posibles numerales primitivos.



Obsérvese que cada numeral es un conjunto de marcas **sencillas** e **iguales**. ¡Imagínese si así se escribiera el número de páginas del directorio telefónico de la Ciudad de México! No sería práctico por la enorme cantidad de tiempo y de espacio que requeriría tal sucesión de marcas iguales. Más aún, nadie podría reconocer, a primera vista, el número representado. Por ejemplo, ¿podría identificar rápidamente el numeral siguiente?

| | | | | | | | | | | | | | | | | | | | = ?

Los antiguos egipcios evitaron algunos de los inconvenientes de los numerales representados con marcas iguales, usando un sólo jeroglífico o figura. Por ejemplo, en lugar de | | | | | | | | | |, usaron el símbolo ∩. Este jeroglífico representaba el hueso del talón. Abajo se muestran otros numerales egipcios básicos con los del sistema decimal que les corresponden

Números egipcios antiguos

1	10	100	1,000	10,000	100,000	1'000,000
						
Raya	Hueso del talón	Cuerda enrollada	Flor de loto	Dedo señalando	Pez	Hombre sorprendido

SECCIÓN 1.1 SISTEMAS NUMÉRICOS

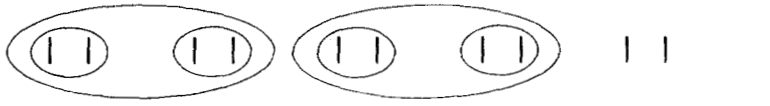
Numeración de base dos (sistema binario)

Dado el siguiente conjunto de marcas simples e iguales

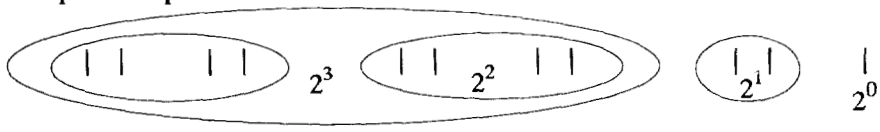
| | | | | | | | | | | | ,

si se encierran en óvalos por parejas, a partir de la izquierda, se tiene

A continuación, también empezando por la izquierda, se encierra cada par de óvalos en otro mayor



Finalmente, se encierra cada par de óvalos en uno mayor todavía, comenzando también por la izquierda.



Nótese que el número de marcas dentro de cualquier óvalo es una potencia de 2.

El número representado por el numeral | | | | | | | | | | | se obtiene así

$$2^3 + 2^1 + 2^0 ,$$

o también

$$(1 \times 2^3) + (1 \times 2^1) + (1 \times 2^0)$$

Obsérvese que en esta suma no aparece 2^2. Como $0 \times 2^2 = 0$, entonces la suma puede escribirse así

$$(1 \times 2^3) + (0 \times 2^2) + (1 \times 2^1) + (1 \times 2^0)$$

Ahora puede formarse un nuevo símbolo para representar esta suma omitiendo los paréntesis, los signos de operación + y × y las potencias de 2, de la siguiente manera:

$$(1 \times 2^3) + (0 \times 2^2) + (1 \times 2^1) + (1 \times 2^0)$$

Nuevo símbolo: 1 0 1 1

Ahora bien, ¿cómo interpretaremos este nuevo símbolo?

El significado de los números 1 en este nuevo símbolo depende del lugar que ocupan en el numeral. Así pues, el primero de derecha a izquierda representa una unidad; el segundo, un grupo de dos (o bien 2^1), el cuarto cuatro grupos de dos (8, o bien 2^3). El cero es un medio de dar a cada "1" su posición correcta. A los números o potencias de 2 que representa el "1" según su posición en el numeral, se les llama **valores de posición**; se dice que un sistema de numeración que usa valores de posición es un **sistema posicional.**

El sistema de este ejemplo es un **sistema de base dos**, o sistema binario, porque usa un grupo básico de dos símbolos: 0 y 1. Los símbolos "1" y "0" utilizados para escribir los numerales se denominan **dígitos binarios** o **bits.**

¿Qué número representa el numeral 101010_{dos}?

(Se lee : "uno, cero, uno, cero, uno, cero, base dos").

Escríbanse los valores de posición debajo de los dígitos:

Dígitos binarios	1	0	1	0	1	0_{dos}
Valores de posición	2^5	2^4	2^3	2^2	2^1	2^0

Multiplicando los valores de posición por los dígitos binarios correspondientes y sumándolos todos, se obtiene el equivalente en decimal.

$$101010_{dos} = (1 \times 2^5) + (0 \times 2^4) + (1 \times 2^3) + (0 \times 2^2) + (1 \times 2^1) + (0 \times 2^0)$$

$$= 42_{diez}$$ (se lee : "cuatro, dos, base diez").

El sistema de numeración más difundido en la actualidad es el **sistema decimal.** Es un sistema posicional que usa un grupo básico de diez (base diez).

Considérese por ejemplo el numeral 582_{diez}

Dígitos decimales	5	8	2
Valores de posición	10^2	10^1	10^0
Forma desarrollada	$(5 \times 10^2) + (8 \times 10^1) + (2 \times 10^0)$		

Al escribir números decimales se omite la palabra "diez" y se establece la convención de que un numeral con valor de posición, es un número decimal, sin necesidad de indicar la base. De ahí que siempre se anote 582 en lugar de 582_{diez}.

El desarrollo y arraigo del sistema decimal, quiza se deba al hecho de tener, siempre a la vista, los diez dedos de las manos. El sistema binario se emplea en las computadoras digitales, debido a que los alambres que forman los circuitos electrónicos presentan sólo dos estados: magnetizados o no magnetizados, ya sea que pase o no corriente por ellos.

Conversión de números enteros del sistema decimal a un sistema de base b y viceversa

Para convertir un número n del sistema decimal a un sistema de base b, se divide el número n entre la base b y se registra el cociente c_1 y el residuo r_1 resultantes; se divide c_1 entre la base b y se anotan el nuevo cociente c_2 y el nuevo residuo r_2. Este procedimiento se repite hasta obtener un cociente c_i igual a cero con residuo r_i. El número equivalente a n en el sistema de base b queda formado así: $r_i \quad r_{i-1} \quad r_{i-2} \quad \ldots r_1$.

Ejemplo 1.1

Convierta 358_{10} al sistema octal.

SOLUCIÓN

La base del sistema octal* es 8, por lo tanto

$$\begin{array}{rcrcrcl} 358 & = & 8 & \times & 44 & + & 6 \\ & & & & c_1 & & r_1 \\ 44 & = & 8 & \times & 5 & + & 4 \\ & & & & c_2 & & r_2 \\ 5 & = & 8 & \times & 0 & + & 5 \\ & & & & c_3 & & r_3 \end{array}$$

Así que el número equivalente en octal es 546

Ejemplo 1.2

Convierta 358_{10} a binario (base 2).

SOLUCIÓN

$$358 = 2 \times 179 + 0$$

*El sistema octal usa un grupo básico de ocho símbolos: 0, 1, 2, 3, 4, 5, 6, 7.

$$
\begin{aligned}
179 &= 2 \times 89 + 1 \\
89 &= 2 \times 44 + 1 \\
44 &= 2 \times 22 + 0 \\
22 &= 2 \times 11 + 0 \\
11 &= 2 \times 5 + 1 \\
5 &= 2 \times 2 + 1 \\
2 &= 2 \times 1 + 0 \\
1 &= 2 \times 0 + 1
\end{aligned}
$$

Por lo tanto $358_{10} = 101100110_2$

Para convertir un entero m de un sistema de base b al sistema decimal, se multiplica cada dígito de m por la base b elevada a una potencia igual a la posición del dígito, tomando como posición cero la del dígito más a la derecha. La suma da el equivalente en decimal. Así

$$276_8 = 2 \times 8^2 + 7 \times 8^1 + 6 \times 8^0 = 190_{10}$$

$$1010001_2 = 1 \times 2^6 + 0 \times 2^5 + 1 \times 2^4 + 0 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 = 81_{10}$$

Conversión de números enteros del sistema octal al binario y viceversa

Dado un número del sistema octal, su equivalente en binario se obtiene sustituyendo cada dígito del número octal con los tres dígitos equivalentes del sistema binario.

BASE OCTAL	EQUIVALENTE BINARIO EN TRES DIGITOS
0	000
1	001
2	010
3	011
4	100
5	101
6	110
7	111

Ejemplo 1.3

Convierta 546_8 a binario.

SOLUCIÓN

5	4	6
101	100	110

Así que $546_8 = 101100110_2$

Dado un número en binario, su equivalente en octal se obtiene formando ternas de dígitos, contando de derecha a izquierda y sustituyendo cada terna por su equivalente en octal. Así

Convertir 10011001_2 a octal

010	011	001_2	Por lo tanto $10011001_2 = 231_8$
2	3	1	

Dado que la conversión de octal a binario es simple y la de decimal a binario resulta muy tediosa, se recomienda usar la conversión a octal como paso intermedio al convertir un número decimal a binario.

Las flechas tienen dos sentidos porque en ambas direcciones es válido lo dicho.

Ejemplo 1.4

Convierta 101100110_2 a decimal.

SOLUCIÓN

a) Conversión directa

$$101100110_2 = 1 \times 2^8 + 0 \times 2^7 + 1 \times 2^6 + 1 \times 2^5 + 0 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 0 \times 2^0 = 358_{10}$$

b) Usando la conversión a octal como paso intermedio:

1) Conversión a octal	101	100	110
	5	4	6

Por tanto $101100110_2 = 546_8$

2) Conversión de octal a decimal

$$546_8 = 5 \times 8^2 + 4 \times 8^1 + 6 \times 8^0 = 358_{10}$$

Conversión de números fraccionarios del sistema decimal a un sistema de base *b*

Para convertir un número x_{10} fraccionario a un número en base b, se multiplica dicho número por la base b; el resultado tiene una parte entera e_1 y una parte fraccionaria f_1. Se multiplica ahora f_1 por b y se obtiene un nuevo producto con parte entera e_2 y fraccionaria f_2. Este procedimiento se repite un número suficiente de veces o hasta que se presenta $f_i = 0$. El equivalente de x_{10} en base b queda así $0.e_1\ e_2\ e_3\ e_4\ \ldots$

Ejemplo 1.5

Convierta 0.2_{10} a octal y binario.

SOLUCIÓN

a) Conversión a octal

0.2	0.6	0.8	0.4	0.2
× 8	× 8	× 8	× 8	× 8
1.6	4.8	6.4	3.2	1.6
$e_1 f_1$	$e_2 f_2$	$e_3 f_3$	$e_4 f_4$	$e_5 f_5$

Después de e_4 se van a repetir $e_1\ e_2\ e_3\ e_4$ indefinidamente, por lo que $0.2_{10} = 0.14631463\ldots_8$

b) Conversión a binario

0.2	0.4	0.8	0.6	0.2
× 2	× 2	× 2	× 2	× 2
0.4	0.8	1.6	1.2	0.4
$e_1 f_1$	$e_2 f_2$	$e_3 f_3$	$e_4 f_4$	$e_5 f_5$

Igual que en el inciso a), después de e_4 se repite $e_1\ e_2\ e_3\ e_4$ indefinidamente, por lo que $0.2_{10} = 0.001100110011\ \ldots_2$

Obsérvese que 0.2_{10} pudo convertirse en binario simplemente tomando su equivalente en octal, y sustituyendo cada número con su terna equivalente en binario. Así

$$\begin{array}{lccccccccc} 0.2_{10} = & 0.1 & & 4 & & 6 & & 3 & & 1 & & 4 & & 6 & & 3 \\ & 0.001 & & 100 & & 110 & & 011 & & 001 & & 100 & & 110 & & 011 \end{array}$$

y

$$0.2_{10} = 0.00110011001100110011001100110011\ldots_2$$

De lo anterior puede observarse que

$$358.2_{10} = 101100110.00110011001100110011001100110011\ldots_2$$

y cualquier número con parte entera y fraccionaria puede pasarse a otro sistema, cambiando su parte entera y fraccionaria independientemente, y al final integrarse.

Conversión de un número fraccionario en sistema binario a sistema decimal

El procedimiento es similar al caso de números enteros, sólo hay que tomar en cuenta que la posición inicia con –1, a partir del punto.

Ejemplo 1.6

Convierta 0.010101110_2 a decimal

SOLUCIÓN

$$\begin{aligned} 0.010101110 &= 0 \times 2^{-1} + 1 \times 2^{-2} + 0 \times 2^{-3} + 1 \times 2^{-4} \\ &+ 0 \times 2^{-5} + 1 \times 2^{-6} + 1 \times 2^{-7} + 1 \times 2^{-8} + 0 \times 2^{-9} \\ &= 0.33984375_{10} \end{aligned}$$

Ejemplo 1.7

Convierta 0.010101110_2 a decimal.

SOLUCIÓN

a) Conversión a octal

$$\begin{array}{ccc} 0.010 & 101 & 110 \\ 2 & 5 & 6 \end{array}$$

y

$$0.010101110_2 = 0.256_8$$

b) Conversión a decimal

$$0.256_8 = 2 \times 8^{-1} + 5 \times 8^{-2} + 6 \times 8^{-3} = 0.33984375_{10}$$

SECCIÓN 1.2 MANEJO DE NÚMEROS EN LA COMPUTADORA

Por razones prácticas, sólo puede manejarse una cantidad finita de bits para cada número en una computadora y esta cantidad o longitud varía de una máquina a otra. Por ejemplo, cuando se realizan cálculos de ingeniería y ciencias, es deseable una longitud grande; por otro lado, una longitud pequeña es más económica y útil en una computadora empleada en cálculos y procesamientos administrativos.

Para una computadora dada, el número de bits generalmente se llama palabra. Las palabras van desde ocho bits hasta 64 bits. Para facilitar su manejo, la palabra se divide en partes más cortas denominadas bytes; por ejemplo, una palabra de 32 bits puede dividirse en cuatro bytes (ocho bits cada uno).

Números enteros

Cada palabra, cualquiera que sea su longitud, almacena un número, aunque en ciertas circunstancias se usan varias para contener un número. Por ejemplo, considérese una **palabra** de 16 bits para almacenar números enteros. De los 16 bits, el primero representa el signo del número; un cero es signo más y un uno un signo menos. Los 15 bits restantes pueden usarse para guardar números binarios desde 000000000000000 hasta 111111111111111 (véase figura 1.1). Al convertir este número en decimal se obtiene

$$(1 \times 2^{14}) + (1 \times 2^{13}) + (1 \times 2^{12}) + \ldots + (1 \times 2^{1}) + (1 \times 2^{0})$$

que es igual a 32767 ($2^{15} - 1$). Por tanto cada palabra de 16 bits puede contener un número cualquiera del intervalo –32768 a + 32767 (véase Prob. 1.10).

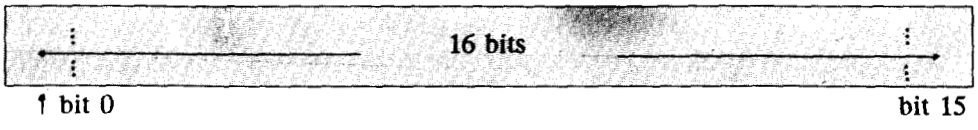


Figura 1.1 Esquema de una palabra de 16 bits para un número entero.

Ejemplo 1.8

Represente el número –26 en una palabra de 16 bits.

SOLUCIÓN

$-26_{10} = -11010_2$ y su almacenamiento en una palabra de 16 bits quedaría así

1	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0

Ejemplo 1.9

Represente el número 525_{10} en una palabra de 16 bits.

SOLUCIÓN

$525_{10} = 1015_8 = 1000001101_2$ y su almacenamiento quedaría así

0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	1

Números reales (punto flotante)

Cuando se desea almacenar un número real, se emplea en su representación binaria, llamada de punto flotante, la notación

$$0.d_1d_2d_3d_4d_5d_6d_7d_8 \times 2^{d_1'\,d_2'\,d_3'\,d_4'\,d_5'\,d_6'\,d_7'}$$

donde $d_1 \approx 0$ y d_i y d_j' con i = 2, ..., 8 y j = 1,2, ..., 7 pueden ser ceros ounos, y se guarda en una palabra como se muestra en la figura 1.2

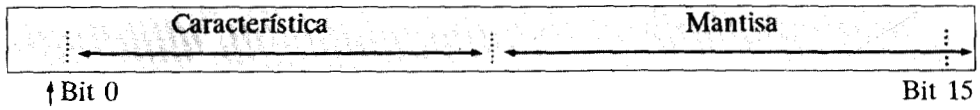
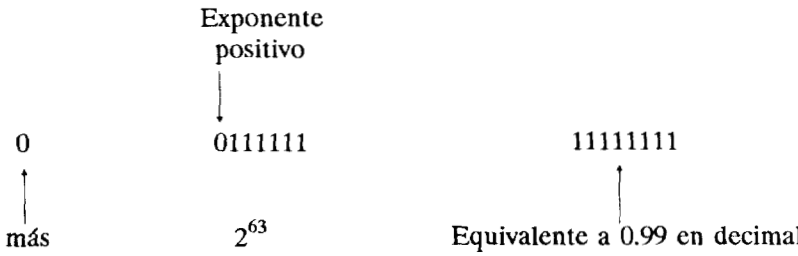


Figura 1.2 Esquema de una palabra de 16 bits para un número de punto flotante.

Igual que antes, el bit cero se usa para guardar el signo del número. En los bits del uno al siete se almacenan el exponente de la base 2 y los ocho bits restantes para la fracción*. Según el lenguaje de los logaritmos, la fracción es llamada **mantisa** y el exponente **característica.** El número mayor que puede guardarse en una palabra de 16 bits usando la notación de punto flotante es



*El exponente es un número binario de sies dígitos, ya que el bit uno se emplea para su signo. En algunas computadoras el exponente se almacena en base ocho (octal) o base 16 (hexadecimal) en lugar de base 2.

y los números que se pueden guardar en punto flotante binario van de alrededor de 2^{-64} (si la característica es negativa) a cerca de 2^{63}; en decimal, de 10^{-19} a cerca de 10^{18} en magnitud (incluyendo números positivos, negativos y cero).

Ejemplo 1.10

El número decimal –125.32 que en binario es

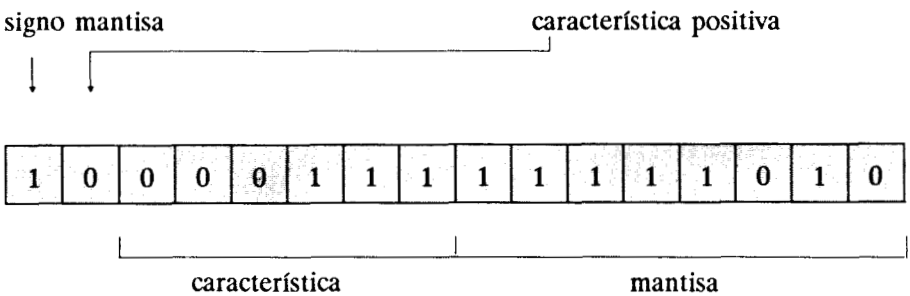
$$-1111101.0101000111110101,$$

normalizado queda así

$$-.11111010101000111110101 \times 2^{+111}$$

bits truncados en el almacenamiento

y la palabra de memoria de 16 bits donde se almacena este valor quedaría como



Nótese que primero se normaliza el número, después se almacenan los primeros ocho bits y se truncan los restantes.

El número decimal + 0.2, que en binario es

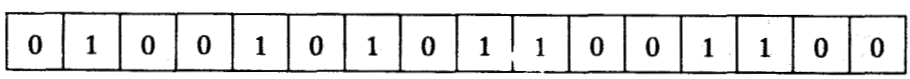
$$0.0011001100110011...$$

y que normalizado queda

$$.1100110011001100... \times 2^{-10}$$

bits truncados

se almacena así



Doble precisión

La doble precisión es un esfuerzo para aumentar la exactitud de los cálculos adicionando más bits a la mantisa. Esto se hace al utilizar dos palabras, la primera en la forma expuesta anteriormente, y los bits de la segunda para aumentar la mantisa de la primera. Entonces, con una palabra de 16 bits puede usarse en doble precisión una mantisa de 8 + 16 = 24 bits. Los 24 bits de la mantisa permiten expresar alrededor de 7 dígitos de exactitud en un número decimal, en lugar de 3 de la precisión sencilla.

La desventaja del uso de la doble precisión es que se emplean más palabras, con lo cual se consume más memoria para un programa.

Error de redondeo

Para finalizar esta sección, se analizarán brevemente algunas consecuencias de utilizar el sistema binario y una longitud de palabra finita.

Como no es posible guardar un número binario de longitud infinita o un número de más dígitos de los que posee la mantisa de la computadora que se está empleando, se almacena sólo un número finito de estos dígitos; como consecuencia, se comete automáticamente un pequeño error, conocido como error de redondeo, que al repetirse muchas veces puede llegar a ser considerable. Por ejemplo, si se desea guardar la fracción decimal 0.0001 que en binario es la fracción infinita

$$0.0000000000000011010001101101110001011101011000...,$$

quedaría, después de normalizarse, almacenado en una palabra de 16 bits como

$$.11010001 \times 2^{-1101}$$

Si se desea sumar el número 0.0001 con él mismo diez mil veces, usando una computadora, naturalmente que no se esperará obtener 1 como resultado, ya que los números que se adicionen no serían realmente 0.0001 sino valores aproximados a él (véase Prob. 1.16).

SECCIÓN 1.3 ERRORES

Error absoluto, error relativo y error en porciento

Si p^* es una aproximación a p, el error se define como

$$\boxed{E = p^* - p}$$

Sin embargo, para facilitar el manejo y el análisis se emplea el error absoluto definido como

$$\boxed{EA = | p^* - p |}$$

y el error relativo como

$$ER = \frac{|\, p^* - p \,|}{p}, \text{ si } p \neq 0$$

y como porciento de error a

$$ERP = ER \times 100$$

En otros libros las definiciones pueden ser diferentes; por ejemplo algunos autores definen el error E como $p - p^*$; por lo tanto, sugerimos que al consultar las distintas bibliografías se vean las definiciones de error dadas.

Ejemplo 1.11

Suponga que el valor para un cálculo debería ser

$p = 0.10 \times 10^2$ pero se obtuvo el resultado $p^* = 0.08 \times 10^2$, entonces

$$EA = |\, 0.08 \times 10^2 - 0.10 \times 10^2 \,| = 0.2 \times 10^1$$

$$ER = \frac{|\, 0.08 \times 10^2 - 0.10 \times 10^2 \,|}{0.10 \times 10^2} = 0.2 \times 10^0$$

$$ERP = ER \times 100 = 20\%$$

Por lo general, interesa el error absoluto y no el error relativo; pero cuando el valor exacto de una cantidad es "muy pequeño" o "muy grande", los errores relativos son más significativos.

Por ejemplo si

$$p = 0.24 \times 10^{-4} \text{ y } \quad p^* = 0.12 \times 10^{-4},$$

entonces

$$EA = |\, 0.12 \times 10^{-4} - 0.24 \times 10^{-4} \,| = 0.12 \times 10^{-4}$$

Sin reparar en las cantidades que se comparan, puede pensarse que el error absoluto es muy pequeño y, lo más grave, aceptar p^* como una buena aproximación a p.

Si, por otro lado, se calcula el error relativo

$$ER = \frac{|\, 0.12 \times 10^{-4} - 0.24 \times 10^{-4} \,|}{0.24 \times 10^{-4}} = 0.5 \times 10^0$$

se observa que la "aproximación" es tan sólo la mitad del valor verdadero y por tanto, está muy lejos de ser aceptable como aproximación a p. Finalmente

$$ERP = 50\%$$

De igual manera puede verse que si

$$p = 0.46826564 \times 10^6 \text{ y} \qquad p^* = 0.46830000 \times 10^6,$$

entonces

$$EA = 0.3436 \times 10^2,$$

y si de nueva cuenta no se toman en consideración las cantidades en cuestión, puede creerse que el EA es muy grande y que se tiene una mala aproximación a p. Sin embargo, al calcular el error relativo

$$ER = 0.7337715404 \times 10^{-4},$$

se advierte que el error es muy pequeño, como en realidad ocurre.

Advertencia

Cuando se manejan cantidades "muy grandes" o "muy pequeñas", el error absoluto puede ser engañoso, mientras que el error relativo es más significativo en esos casos.

Definición

Se dice que el número p^* aproxima a p con t dígitos significativos si t es el entero más grande no negativo para el cual se cumple

$$\frac{|p^* - p|}{p} < 5 \times 10^{-t}$$

Supóngase por ejemplo el número 10. Para que p^* aproxime a 10 con dos cifras significativas, usando la definición, p^* debe cumplir con

$$\frac{|p^* - 10|}{10} < 5 \times 10^{-2}$$

$$-5 \times 10^{-2} < \frac{p^* - 10}{10} < 5 \times 10^{-2}$$

$$10 - 5 \times 10^{-1} < p^* < +5 \times 10^{-1} + 10$$

$$9.5 < p^* < 10.5$$

esto es, cualquier valor de p^* en el intervalo (9.5, 10.5) cumple la condición.

En general para t dígitos significativos

$$\frac{| p^* - p |}{p} < 5 \times 10^{-t} \qquad \text{si } p > 0$$

$$| p^* - p | < 5\, p \times 10^{-t}$$

$$p - 5\, p \times 10^{-t} < p^* < p + 5\, p \times 10^{-t}$$

Si, por ejemplo, $p = 1000$ y $t = 4$

$$1000 - 5 \times 1000 \times 10^{-4} < p^* < 1000 + 5 \times 1000 \times 10^{-4}$$

$$999.5 < p^* < 1000.5$$

Causas de errores graves en computación

Existen muchas causas de errores en la ejecución de un programa de cómputo, de las cuales se discutirán ahora algunas de las más serias. Para esto, **piense** en una computadora imaginaria que trabaja con números en el sistema decimal, en forma tal que se tiene una mantisa de cuatro dígitos decimales, y una característica de dos dígitos decimales, el primero de los cuales es usado para el signo. Sumados estos seis al bit empleado para el signo del número, se tendrá una longitud de palabra de siete bits. Los números que se van a guardar deben normalizarse primero en la siguiente forma

$$3.0 = .3000 \times 10^{1}$$

$$7956000 = .7956 \times 10^{7}$$

$$-0.0000025211 = -.2521 \times 10^{-5}$$

Valiéndose de esta computadora imaginaria, pueden estudiarse algunos de los errores más serios que se cometen en su empleo.

a) Suma de números muy distintos en magnitud

Supóngase que se trata de sumar 0.002 a 600 en la computadora decimal imaginaria.

$$0.002 = .2000 \times 10^{-2}$$

$$600 = .6000 \times 10^{3}$$

Estos números **normalizados** no pueden sumarse directamente y, por tanto, la computadora debe desnormalizarlos antes de efectuar la suma.

$$\begin{array}{rr} & .000002 \times 10^{3} \\ + & .600000 \times 10^{3} \\ \hline & .600002 \times 10^{3} \end{array}$$

Como sólo puede manejar cuatro dígitos, los últimos dos son eliminados y la respuesta es $.6000 \times 10^3$ ó 600. Por el resultado, la suma nunca se realizó.

Este tipo de errores cuyo origen es el redondeo es muy común y se recomienda, de ser posible, no sumar o restar dos números muy diferentes (véase ejercicio 1.2).

b) Resta de números casi iguales

Supóngase que la computadora decimal va a restar 0.2144 de 0.2145.

$$\begin{array}{rr} & .2145 \times 10^0 \\ - & .2144 \times 10^0 \\ \hline & .0001 \times 10^0 \end{array}$$

Como la mantisa de la respuesta está desnormalizada, la computadora automáticamente la normaliza y el resultado se almacena como $.1000 \times 10^{-3}$.

Hasta aquí no hay error, pero en la respuesta sólo hay un dígito significativo; por lo tanto, se sugiere no confiar en su exactitud, ya que un pequeño error en alguno de los números originales produciría un error relativo muy grande en la respuesta de un problema que involucrara este error, como se ve a continuación.

Supóngase que la siguiente expresión aritmética es parte de un programa

$$X = (A - B) * C$$

Considérese ahora que los valores de A, B y C son

$$A = 0.2145 \times 10^0, \qquad B = 0.2144 \times 10^0, \qquad C = 0.1000 \times 10^5$$

Al efectuarse la operación se obtiene el valor de $X = 1$, que es correcto. Sin embargo, supóngase que A fue calculada en el programa con un valor de 0.2146×10^0 (error absoluto 0.0001, error relativo 0.00046 y $ERP = 0.046\%$). Usando este valor de A en el cálculo de X, se obtiene como respuesta $X = 2$. Un error de 0.046% de pronto provoca un error del 100%. Aun más, este error puede pasar desapercibido.

c) Overflow y Underflow

Con frecuencia una operación aritmética con dos números válidos da como resultado un número tan grande o tan pequeño que la computadora no puede manejarlo; como consecuencia se tiene un overflow o un underflow, respectivamente.

Por ejemplo al multiplicar 0.5000×10^8 por 0.2000×10^9 se tiene

$$\begin{array}{rr} & 0.5000 \times 10^8 \\ \times & 0.2000 \times 10^9 \\ \hline & 0.1000 \times 10^{17} \end{array}$$

Cada uno de los números que se multiplican puede guardarse en la palabra de la computadora imaginaria; sin embargo, su producto es muy grande y no puede almacenarse en ella porque la característica requiere tres dígitos. Entonces se dice que hay *overflow.*

Otro caso de overflow puede ocurrir en la división; por ejemplo

$$\frac{2000000}{0.000005} = \frac{0.2000 \times 10^{7}}{0.5000 \times 10^{-5}} = 0.4000 \times 10^{12}$$

Las computadoras comúnmente reportan esta circunstancia con un mensaje que varía con la máquina.

El *underflow* puede aparecer en la multiplicación o división, y generalmente no es tan serio como el *overflow*; las computadoras casi nunca envían mensaje de *underflow*. Por ejemplo

$$(0.3000 \times 10^{-5}) \times (0.02000 \times 10^{-3}) = 0.006 \times 10^{-8} = 0.6000 \times 10^{-10}$$

Como el exponente -10 está excedido en un dígito, no puede guardarse en la computadora y este resultado se expresa como valor cero. Este error expresado como error relativo es muy pequeño y a menudo no es serio. No obstante, puede ocurrir, por ejemplo

$$A = 0.3000 \times 10^{-5}, \qquad B = 0.0200 \times 10^{-3}, \qquad C = 0.4000 \times 10^{7},$$

y que se desee en algún punto del programa calcular el producto de A, B y C

$$X = A \times B \times C$$

Se multiplican primero A y B. El resultado parcial es cero. La multiplicación de este resultado por C da también cero. Si, en cambio, se arregla la expresión como

$$X = A \times C \times B$$

se multiplica A por C y se obtiene 0.1200×10^{2}. La multiplicación siguiente da la respuesta correta: 0.2400×10^{-3}. De igual manera, un arreglo en una divisón puede evitar underflow.

d) División entre un número muy pequeño

Como se dijo, la división entre un número muy pequeño puede causar *overflow*.

Supóngase que se realiza en la computadora una división válida y que no se comete error alguno en la operación; pero considérese que ocurrió un pequeño error de redondeo previamente en el programa, cuando se calculó el denominador. Si el numerador es grande y el denominador pequeño, puede presentarse un error absoluto considerable en el cociente. Si éste se resta después, de otro número del mismo tamaño relativo, puede presentarse un error mayor en la respuesta final.

Como ejemplo considérese la siguiente instrucción en un programa

$$X = A - B / C$$

donde

$$A = 0.1120 \times 10^{9} = 112000000$$

$$B = 0.1000 \times 10^{6} = 100000$$

$$C = 0.900 \times 10^{-3} = 0.0009$$

Si el cálculo se realiza en la computadora decimal de cuatro dígitos, el cociente B / C es 0.1111×10^9, y X es 0.0009×10^9 o, después de ser normalizado, $X = 0.9000 \times 10^6$. Nótese que sólo hay un dígito significativo.

Imagínese ahora que se cometió un pequeño error de redondeo al calcular C en algún paso previo y resultó un valor $C^* = 0.9001 \times 10^{-3}$ ($EA = 0.0001 \times 10^{-3}$; $ER = 10^{-4}$ y $ERP = 0.01\%$).

Si se calcula B / C^* se obtiene como cociente 0.1110×10^9 y $X^* = 0.1000 \times 10^7$. El valor correcto de X es 0.9000×10^6.

Entonces

$$EA = | \ 1000000 - 900000 \ | = 100000$$

$$ER = \frac{| \ 1000000 - 900000 \ |}{900000} = 0.11$$

$$ERP = 0.11 \times 100 = 11\%$$

El error relativo se ha multiplicado cerca de 1100 veces. Como se dijo ya, estos cálculos pueden conducir a un resultado final sin significado o relación con la respuesta verdadera.

e) Error de discretización

Dado que un número específico no se puede almacenar exactamente como número binario de punto flotante, el error generado se conoce como error de **discretización** (error de cuantificación), ya que los números expresados exactamente por la máquina (números de máquina) no forman un conjunto continuo sino discreto.

Ejemplo 1.12

Cuando se suma 10000 veces 0.0001 con él mismo, debe resultar 1; sin embargo, el número 0.0001 en binario resulta en una sucesión infinita de ceros y unos que se trunca al ser almacenada en una palabra de memoria, con lo que se perderá información y el resultado de la suma ya no será 1. Se obtuvieron los siguientes resultados que corroboran lo anterior, utilizando una PC, precisión sencilla y Quick-Basic.

SOLUCIÓN

a) $\sum_{i=1}^{10000} 0.0001 = 1.000054$

b) $1 + \sum_{i=1}^{10000} 0.0001 = 2.000166$

c) $1000 + \sum_{i=1}^{10000} 0.0001 = 1001.221$

d) $10000 + \sum_{i=1}^{10000} 0.0001 = 10000$

Nótese que en los tres últimos incisos, además del error de discretización, se generó el error de sumar un número muy grande con un número muy pequeño (véase Prob. 1.16 y 1.17).

f) Errores de salida

Aún cuando no se haya cometido error alguno durante la fase de cálculos de un programa, puede presentarse un error al imprimir resultados.

Por ejemplo, supóngase que la respuesta de un cálculo particular es exactamente 0.015625. Cuando este número se imprime con un formato tal como F10.6 ó E14.6 (de FORTRAN), se obtiene la respuesta correcta. Si, por el contrario, se decide usar F8.3, se imprimirá el número 0.016 (si la computadora redondea), o bien 0.015 (si la computadora trunca), con lo cual se presenta un error.

Propagación de errores

Una vez que se sabe como se producen los errores en un programa de cómputo, podría pensarse en tratar de determinar el error cometido en cada paso, y conocer de esa manera el error total en la respuesta final. Sin embargo, esto no es práctico. Resulta más adecuado analizar las operaciones individuales realizadas por la computadora para ver cómo se propagan los errores de dichas operaciones.

a) Suma

Se espera que al sumar a y b, se obtenga el valor correcto de $c = a + b$; sin embargo, se tiene en general un valor de c incorrecto debido a la longitud finita de palabra que se emplea. Puede considerarse que este error fue causado por una operación incorrecta de la computadora $\dot{+}$ (el punto indica que es suma con error). Entonces el error es

$$\text{Error} = (a \dot{+} b) - (a + b)$$

La magnitud de este error depende de las magnitudes relativas, de los signos de a y b, y de la forma binaria en que a y b son almacenados en la computadora. Esto último varía de computadora en computadora, y por tanto es un error muy difícil de analizar y no se discutirá aquí.

Si por otro lado a y b de entrada son inexactos, hay un segundo error posible. Por ejemplo, considérese que en lugar del valor verdadero de a, la computadora tiene el valor a^* el cual presenta un error $\in_a$

$$a^* = a + \in_a$$

y similarmente para b

$$b^* = b + \in_b$$

Como consecuencia se tendría, aun si no se cometiera error en la adición, un error en el resultado

$$\begin{aligned}\text{Error} &= (a^* + b^*) - (a + b) \\ &= (a + \in_a + b + \in_b) - (a + b) = \in_a + \in_b = \in_c\end{aligned}$$

o sea $c^* = c + \in_c$

El error absoluto es

$$| (a^* + b^*) - (a + b) | = | \in_a + \in_b | \leq | \in_a | + | \in_b |$$

o bien

$$| \in_c | \leq | \in_a | + | \in_b |$$

Se dice que los errores $\in_a$ y $\in_b$ se han extendido a c y $\in_c$ se conoce como el error de propagación.

Dicho error es causado por valores inexactos de los valores iniciales y se propaga en los cómputos siguientes, con lo cual causa un error en el resultado final.

b) Resta

El error de propagación ocasionado por valores inexactos iniciales a^* y b^*, puede darse en igual forma que en la adición, con un simple cambio de signo (véase prob. 1.24).

c) Multiplicación

Si se multiplican los números a^* y b^*, se obtiene (ignorando el error causado por la operación misma)

$$(a^* \times b^*) = (a + \in_a) \times (b + \in_b)$$

$$= (a \times b) + (a \times \in_b) + (b \times \in_a) + (\in_a \times \in_b)$$

Si $\in_a$ y $\in_b$ son suficientemente pequeños, puede considerarse que su producto es muy pequeño en comparación con los otros términos, y, por tanto, eliminar el último término. Se obtiene entonces el error del resultado final

$$(a^* \times b^*) - (a \times b) \approx (a \times \in_b) + (b \times \in_a)$$

Esto hace posible encontrar el valor absoluto del error relativo del resultado dividiendo ambos lados entre $a \times b$.

$$\left| \frac{(a^* \times b^*) - (a \times b)}{(a \times b)} \right| \approx \left| \frac{\in_b}{b} + \frac{\in_a}{a} \right| \leq \left| \frac{\in_b}{b} \right| + \left| \frac{\in_a}{a} \right|$$

El error de propagación relativo en valor absoluto en la multiplicación es aproximadamente igual o menor a la suma de los errores relativos de a y b en valor absoluto.

d) División

Puede considerarse la división de a^* y b^* como sigue:

$$a^*/b^* = (a + \in_a) / (b + \in_b)$$

$$= (a + \in_a) \frac{1}{(b + \in_b)}$$

Multiplicando numerador y denominador por $b - \in_b$

$$a^* / b^* = \frac{(a + \in_a)(b - \in_b)}{(b + \in_b)(b - \in_b)}$$

$$= \frac{ab - a\in_b + \in_a b - \in_a \in_b}{b^2 - \in_b^2}$$

Si, como en la multiplicación, se considera el producto $\in_a \in_b$ muy pequeño y, por las mismas razones a $\in_b^2$ y se desprecian se tiene.

$$a^* / b^* \approx \frac{ab}{b^2} + \frac{\in_a b}{b^2} - \frac{a\in_b}{b^2}$$

$$\approx \frac{a}{b} + \frac{\in_a}{b} - \frac{a\in_b}{b^2}$$

El error es entonces

$$a^*/b^* - \frac{a}{b} \approx \frac{\in_a}{b} - \frac{a\in_b}{b^2}$$

Dividiendo entre a/b se obtiene el error relativo. Al tomar el valor absoluto del error relativo, se tiene

$$\left| \frac{\frac{a^*}{b^*} - \frac{a}{b}}{a/b} \right| \approx \left| \frac{\frac{\in_a}{b} - \frac{a\in_b}{b^2}}{a/b} \right| \approx \left| \frac{\in_a}{a} - \frac{\in_b}{b} \right| \leq \left| \frac{\in_a}{a} \right| + \left| \frac{\in_b}{b} \right|$$

Se concluye que, el error de propagación relativo del cociente en valor absoluto es aproximadamente igual o menor a la suma de los errores relativos en valor absoluto de a y b.

e) Evaluación de funciones

Por último, se estudiará la propagación del error (asumiendo operaciones básicas $+, -, \times$ y $/$ ideales o sin errores), cuando se evalúa una función $f(x)$ en un punto

$x = a$. En general, se dispone de un valor de a aproximado: a^*; la intención es determinar el error resultante

$$\in_f = f(a^*) - f(a)$$

La figura 1.3 muestra la gráfica de la función $f(x)$ en las cercanías de $x = a$. A continuación se determina la relación entre $\in_a$ y $\in_f$.

Si $\in_a$ es pequeño, puede aproximarse la curva $f(x)$ por su tangente en $x = a$. Se sabe que la pendiente de esta tangente es $f'(a)$ o aproximadamente $\in_f / \in_a$; esto es

$$\in_f / \in_a \approx f'(a)$$

y

$$\in_f \approx \in_a f'(a) \approx \in_a f'(a^*)$$

En valor absoluto

$$| \in_f | \approx | \in_a f'(a^*) | \approx | \in_a | \, | f'(a^*) |$$

El error al evaluar una función en un argumento inexacto es proporcional a la primera derivada de la función en el punto donde se ha evaluado.

SECCIÓN 1.4 ALGORITMOS Y ESTABILIDAD

El tema fundamental de este libro es el estudio, selección y aplicación de algoritmos, que se definen como secuencias de operaciones algebraicas y lógicas para obtener la solución de un problema. Generalmente, se dispone de varios algoritmos para resolver un problema particular; uno de los criterios de selección es la estabilidad del algoritmo; esto es que a pequeños errores de los valores manejados se obtengan pequeños errores en los resultados finales.

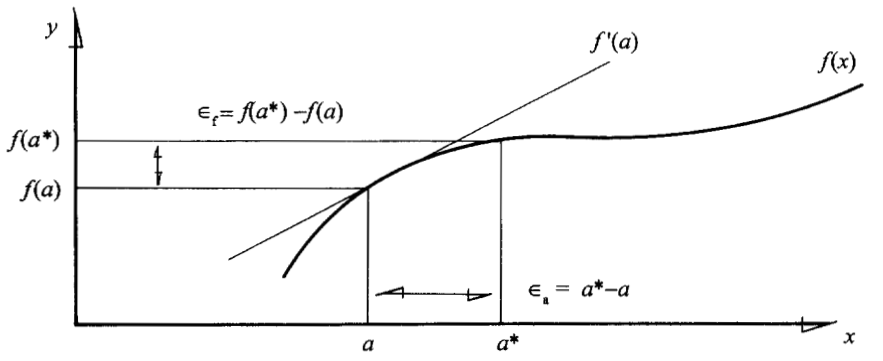


Figura 1.3. Gráfica de una función y su primera derivada en a.

Supóngase que un error $\in$ se introduce en algún paso en los cálculos y que el error de propagación de n operaciones subsiguientes se denote por $\in_n$. En la práctica son generalmente dos los casos que se presentan

a) $| \in_n | \approx n\,c \in$, donde c es una constane independiente de n; se dice entonces que la propagación del error es lineal.

b) $| \in_n | \approx k^n \in$, para $k > 1$; se dice entonces que la propagación del error es exponencial.

La propagación lineal de los errores suele ser inevitable; cuando c y $\in$ son pequeños, los resultados finales normalmente son aceptables. Por otro lado la propagación exponencial debe evitarse, ya que el término k^n crece con rapidez para valores relativamente pequeños de n. Esto conduce a resultados finales muy poco exactos, sea cual sea el tamaño de $\in$. Como consecuencia, se dice que un algoritmo con crecimiento lineal del error es estable, mientras que un algoritmo con una propagación exponencial es inestable (véase Fig. 1.4).

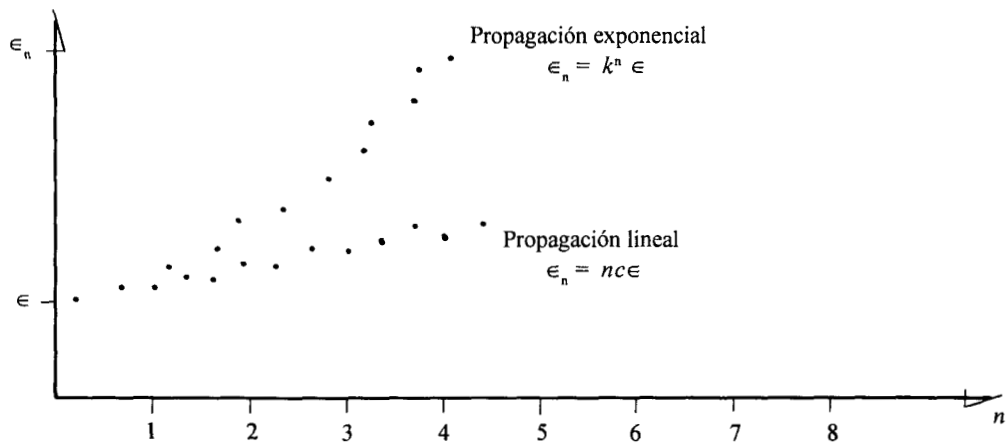


Figura 1.4. Propagación lineal y propagación exponencial de errores.

Ejercicios

1.1 Error de redondeo al restar dos números casi iguales

Considere las ecuaciones

$$31.69x + 14.31y = 45.00 \qquad (1)$$

$$13.05x + 5.89y = 18.53 \qquad (2)$$

La única solución de este sistema de ecuaciones es (redondeando a cinco cifras decimales) $x = 1.25055$, $y = 0.37527$. Un método para resolver este tipo de problemas es multiplicar la ecuación (1) por el coeficiente de x de la ecuación (2), multiplicar la ecuación (2) por el coeficiente de x de la ecuación (1) y después restar

las ecuaciones resultantes. Para este sistema se obtendría (como los coeficientes tienen dos cifras decimales, todas las operaciones intermedias se efectúan redondeando a dos cifras decimales)

$$[\ 13.05\ (14.31)\ -\ 31.69\ (\ 5.89)\ ]\ y\ =13.05\ (\ 45.00\ )\ -\ 31.69\ (\ 18.53\ )$$

$$(\ 186.75\ -\ 186.65\ )\ y\ =587.25\ -\ 587.22$$

$$0.10\ y\ =\ 0.03$$

de donde $y = 0.3$, luego

$$x = \frac{(\ 18.53\ ) - 5.89\ (\ 0.3\ )}{13.05} = \frac{18.53 - 1.77}{13.05} = \frac{16.76}{13.05} = 1.28$$

Para la variable x

$$EA = |\ 1.28 - 1.25\ | = 0.03;\quad ER = 0.03/1.25 = 0.024;\quad ERP = 2.4\%$$

Para la variable y

$$EA = |\ 0.3 - 0.38\ | = 0.08;\quad ER = 0.08/0.38 = 0.21;\quad ERP = 21\%$$

1.2 Error de redondeo al sumar un número grande y uno pequeño

Considere la sumatoria infinita

$$s = \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{1}{1} + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \ldots + \frac{1}{100} + \ldots$$

resulta (usando precisión simple y 5000 como valor final de n) 1.644725 si se suma de izquierda a derecha, pero resulta 1.644834 si se suma de derecha a izquierda, a partir de $n = 5000$.

Debe notarse que el resultado de sumar de derecha a izquierda es más correcto ya que en todos los términos se suman valores de igual magnitud.

Por el contrario, al sumar de izquierda a derecha, una vez que se avanza en la sumatoria, se sumarán números cada vez más grandes con números más pequeños.

Lo anterior se corrobora si se realiza la suma en ambos sentidos, pero ahora con doble precisión. El resultado obtenido es 1.6448340718406.

1.3 Reducción de errores

Para resolver la ecuación cuadrática

$$100\ x^2\ -\ 10011\ x\ +\ 10.011\ =\ 0,$$

el método común sería usar la fórmula

$$x = \frac{-\ b\ \pm\sqrt{b^2\ -\ 4\ a\ c}}{2\ a},$$

después de dividir la ecuación entre 100.

$$x^2 - 100.11\,x + 0.10011 = 0$$

$$x = \frac{100.11 \pm \sqrt{100.11^2 - 4(0.10011)}}{2}$$

Trabajando con aritmética de cinco dígitos

$$x = \frac{100.11 \pm \sqrt{10022 - 0.40044}}{2} = \frac{100.11 \pm \sqrt{10022}}{2}$$

$$= \frac{100.11 \pm 100.11}{2} = \begin{cases} \frac{200.22}{2} = 100.11 \\ 0 \end{cases}$$

Las soluciones verdaderas, redondeadas a cinco dígitos decimales son 100.11 y 0.00100.

El método empleado fue adecuado para la solución mayor, pero no del todo para la solución menor. Si las soluciones fueran divisores de otras expresiones, la solución $x = 0$ hubiese causado problemas serios.

Se restaron dos números "casi iguales" (números iguales en aritmética de cinco dígitos) y sufrieron pérdida de exactitud.

¿Cómo evitar esto? Una forma sería reescribir la expresión para la solución de una ecuación cuadrática a fin de evitar la resta de números "casi iguales".

El problema, en este caso, se da en el signo negativo asignado a la raíz cuadrada; esto es

$$\frac{-b - \sqrt{b^2 - 4\,a\,c}}{2\,a}$$

Multiplicando numerador y denominador por $-b + \sqrt{b^2 - 4a\,c}$, queda

$$\frac{\left(-b - \sqrt{b^2-4\,a\,c}\right)\left(-b + \sqrt{b^2-4\,a\,c}\right)}{2\,a\left(-b + \sqrt{b^2-4\,a\,c}\right)} = \frac{(-b)^2 - (b^2 - 4\,a\,c)}{2a\left(-b + \sqrt{b^2-4\,a\,c}\right)}$$

$$= \frac{4\,a\,c}{2a\left(-b + \sqrt{b^2-4\,a\,c}\right)} = \frac{2\,c}{-b + \sqrt{b^2-4\,a\,c}}$$

Usando esta expresión con $a = 1$, $b = -100.11$, y $c = 0.10011$, se obtiene

$$\frac{2\,(0.10011)}{100.11 + \sqrt{10022}} = \frac{0.20022}{200.22} = 0.001 \text{ (en aritmética de cinco dígitos)}$$

que es el valor verdadero, redondeado a cinco dígitos decimales.

Esta forma alternativa para calcular una raíz pequeña de una ecuación cuadrática, casi siempre produce una respuesta más exacta que la de la fórmula usual (véase Prob. 2.12).

1.4 Más sobre reducción de errores

Se desea evaluar la expresión $A / (1 - \text{sen } x)$, en $x = 89° 41'$. En tablas con cinco cifras decimales, sen 89° 41' = 0.99998. Con aritmética de cinco dígitos y redondeando se tiene

$$\text{sen } x = 0.99998 \text{ y } 1 - \text{sen } x = 0.00002$$

La función sen x sólo tiene cuatro dígitos exactos (confiables). Por otro lado, el único dígito no cero en $1 - \text{sen } x$ se ha calculado con el dígito no confiable de sen x, por lo que se pudo perder la exactitud en la resta.

Esta situación de arriba puede mejorarse observando que

$$1 - \text{sen } x = \frac{(1 - \text{sen } x)(1 + \text{sen } x)}{1 + \text{sen } x} = \frac{1 - \text{sen}^2 x}{1 + \text{sen } x} = \frac{\cos^2 x}{1 + \text{sen } x}$$

Por esto, es posible escribir 1 - sen x de una forma que no incluye la resta de dos números casi iguales.

1.5 Comparaciones seguras

En los métodos numéricos, a menudo la comparación de igualdad de dos números en notación de punto flotante permitirá terminar la repetición de un conjunto de cálculos (proceso cíclico o iterativo). En vista de los errores observados, es recomendable comparar la diferencia de los dos números en valor absoluto contra una tolerancia ε apropiada, usando por ejemplo el operador de relación menor o igual ($\leq$). Se ilustra esto enseguida.

En lugar de

SI $X = Y$ ALTO; En caso contrario REPETIR las instrucciones 5 a 9

Deberá usarse

SI ABS $(X - Y) \leq \varepsilon$ ALTO; en caso contrario REPETIR las instrucciones 5 a 9

En lugar de

REPETIR

{ pasos de un ciclo }
HASTA QUE $X = Y$

Deberá usarse

REPETIR
{pasos de un ciclo}
HASTA QUE ABS $(X - Y) \leq \varepsilon$

donde ε es un número pequeño (generalmente menor que uno, pero puede ser mayor dependiendo el contexto en que se trabaje) e indicará la cercanía de X con Y que se aceptará como "igualdad" de X y Y.

1.6 Análisis de resultados

Codifique las siguientes instrucciones en QUICK-BASIC

```
Y = 1000.2
A = Y - 1000.0
PRINT A
```

Se obtiene 0.2000122

En precisión sencilla pueden manejarse alrededor de siete dígitos decimales de exactitud, de modo que la resta de arriba se representa

$$1000.200 - 1000.000$$

La computadora convierte Y a binario dando un número infinito de ceros y unos, y almacena un número distinto a 1000.2 (véase Prob. 1.6 b).

Por otro lado, 1000 sí se puede almacenar o representar exactamente en la computadora en binario en punto flotante (los números con esta característica se llaman **números de máquina**). Al efectuarse la resta se obtiene un número diferene de 0.2, 0.2000122. Esto muestra por qué deberá examinarse siempre un resultado de un dispositivo digital antes de aceptarlo.

1.7 Más sobre análisis de resultados

El método de posición falsa (véase sección 2.4) obtiene su algoritmo al encontrar el punto de corte de la línea recta que pasa por los puntos (x_D, y_D), (x_I, y_I) y el eje x. Pueden obtenerse dos expresiones para encontrar el punto de corte x_M

$$\text{i)}\quad x_M = \frac{x_I\, y_D - x_D\, y_I}{y_D - y_I} \qquad \text{ii)}\quad x_M = x_D - \frac{(x_D - x_I)\, y_D}{y_D - y_I}$$

Si (x_D , y_D) = (2.13, 4.19) y (x_I , y_I) = (1.96, 6.87) y usando aritmética de tres dígitos y redondeando, ¿cuál es la mejor expresión y por qué?

SOLUCIÓN

Sustituyendo en i) y en ii)

$$\text{i)}\quad x_M = \frac{1.96\,(4.19) - 2.13\,(6.87)}{4.19 - 6.87} = 2.38$$

$$\text{ii)}\quad x_M = 2.13 - \frac{(2.13 - 1.96)\,4.19}{4.19 - 6.87} = 2.40$$

Al calcular los errores absoluto y relativo y tomando como valor verdadero a 2.395783582, el cual se calculó con aritmética de 13 dígitos, se tiene

i) $EA = 2.395783582 - 2.38 = 0.015783582$

$$ER = \frac{0.015783582}{2.395783582} = 0.006588066$$

ii) $EA = 2.395783582 - 2.40 = 0.004216418$

$$ER = \frac{0.004216418}{2.395783582} = 0.001759932$$

de donde es evidente que la forma ii) es mejor. El por qué se deja como ejercicio al lector.

Problemas

1.1 Averigüe los símbolos o numerales romanos correspondientes a los siguientes símbolos arábigos

10, 100, 1000, 10000, 100000, 1000000

1.2 Convierta los siguientes números decimales a los sistemas de base 2 y base 8 y viceversa

a) 536 *b*) 923 *c*) 1536 *d*) 8 *e*) 2 *f*) 10 *g*) 0

1.3 Convierta los siguientes números enteros del sistema octal a binario y viceversa

a) 777 *b*) 573 *c*) 7 *d*) 2 *e*) 10 *f*) 0

1.4 Resuelva las siguientes preguntas.

a) ¿El número 101121 pertenece al sistema binario?
b) ¿El número 3852 pertenece al sistema octal?

Si su respuesta es NO en alguno de los incisos, explique por qué; si es SI, conviértalo(s) a decimal.

1.5 Convierta los siguientes números dados en binario a decimal y viceversa, usando la conversión a octal como paso intermedio

a) 1000 *b*) 10101 *c*) 111111

1.6 Convierta los siguientes números fraccionarios dados en decimal, a binario y octal

a) 0.8 *b*) 0.2 *c*) 0.973 *d*) 0.356 *e*) 0.713 *f*) 0.10

1.7 Convierta los siguientes números fraccionarios, dados en binario, a decimal

a) 0.1 *b*)0.010101 *c*)0.0001 *d*) 0.11111 *e*) 0.00110011 *f*) 0.0110111

1.8 Repita los incisos (*a*) a (*f*) del problema 1.7, pero pasando a octal como paso intermedio.

1.9 Convierta los siguientes números, dados en decimal, a octal y binario.

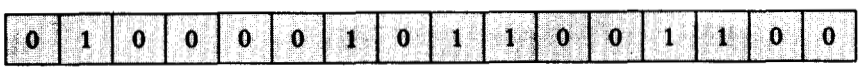
a) 985.34 *b*) 10.1 *c*) 888.222 *d*) 3.57 *e*) 977.93 *f*) 0.357 *g*) 0.9389 *h*) –0.9389

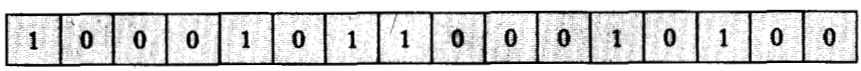
1.10 Se dijo en la sección 1.2 que cada palabra de 16 bits puede contener un número entero cualquiera del intervalo –32768 a +32767. Investigue por qué se incluye al –32768, o bien por qué el intervalo no va de –32767.

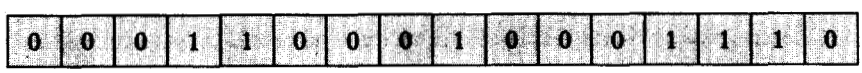
1.11 Considere una computadora con una palabra de 8 bits. ¿Qué rango de números enteros puede contener dicha palabra?

1.12 Represente el número –26 en una palabra de 8 bits.

1.13 Dados los siguientes números de máquina en una palabra de 16 bits

a) 

b) 

c) 

¿Qué decimales representan?

1.14 Normalice los siguientes números

a) 723.5578 *b*) –15.324 *c*) 0.003485 *d*) 8×10^3

Sugerencia: Pase los números a binario y después normalícelos.

1.15 Represente en doble precisión el número decimal del ejemplo 1.10

1.16 Elabore un programa para su calculadora o el dispositivo de cálculo con que cuente, de modo que el número 0.0001 se sume diez mil veces consigo mismo

$$\underset{1}{0.0001} + \underset{2}{0.0001} + \ldots + \underset{10000}{0.0001}$$

El resultado deberá imprimirse. Interprete este resultado de acuerdo con los siguientes lineamientos

a) Si es 1, ¿cómo es posible si se sumaron diez mil valores que no son realmente 0.0001?
b) En caso de obtener 1, explore con el valor 0.00001, 0.000001, etc., hasta obtener un resultado diferente de 1.
c) ¿Es posible obtener un resultado menor de 1? ¿Por qué?

1.17 Con el programa del problema 1.16 efectúe los cálculos de los incisos (*a*) a (*d*) del ejemplo 1.12 y obtenga los resultados de la siguiente manera

a) Inicialice la variable SUMA con 0, 1, 1000 y 10000 en los incisos (*a*), (*b*), (*c*) y (*d*), respectivamente, y luego en un cilo súmese a ese valor diez mil veces el 0.0001. Anote sus resultados.

b) Inicialice la variable SUMA con 0 para los cuatro incisos y al final del ciclo donde se habrá sumado 0.0001 consigo mismo 10000 veces, sume a ese resultado los números 0, 1, 1000 y 10000 e imprima los resultados.

Interprete las diferencias de los resultados.

1.18 La mayoría de las calculadoras científicas almacenan dos o tres dígitos de seguridad más de los que despliegan. Por ejemplo, una calculadora que despliega ocho dígitos puede almacenar realmente diez (dos dígitos de seguridad); por tanto, será un dispositivo de diez dígitos. Para encontrar la exactitud real de su calculadora, efectúe las siguientes operaciones.

Divida 10 entre 3, al resultado réstele 3.

Divida 100 entre 3, al resultado réstele 33.

Divida 1000 entre 3, al resultado réstele 333.

Divida 10000 entre 3, al resultado réstele 3333.

Notará que la cantidad de los números 3 desplegados se va reduciendo.

La cantidad de 3 desplegada en cualquiera de las operaciones anteriores, sumada al número de ceros utilizados con el 1, indica el número de cifras significativas que maneja su calculadora. Por ejemplo, si con la segunda operación despliega 0.3333333 la calculadora manjea nueve cifras significativas de exactitud (7 + 2 ceros que tiene 100).

NOTA: Si su calculadora es del tipo intérprete BASIC, no realice las operaciones como 1000/3-333 porque obtendrá otros resultados.

1.19 Evalúe la expresión A / (1-cos x), en un valor de x cercao a 0°. ¿Cómo podría evitar la resta de dos números casi iguales en el denominador?

1.20 Determine en su calculadora o microcomputadora si muestra un mensaje de *overflow* o no.

1.21 Deduzca las expresiones para x_M dadas en el ejercicio 1.7.

1.22 Un número de máquina para una calculadora o computadora es un número real que se almacena exactamente (en forma binaria de punto flotante). El número –125.32 del ejemplo 1.10, evidentemente no es un número de máquina (si el dispositivo de cálculo tiene una palabra de 16 bits). Por otro lado, el número –26 del ejemplo 1.8 si lo es, empleando una palabra de 16 bits. Determine 10 números de máquina en el intervalo $[10^{-19}, 10^{18}]$ cuando se emplea una palbra de 16 bits.

1.23 Investigue cuántos números de máquina positivos es posible representar en una palabra de 16 bits.

1.24 Haga el análisis de la propagación de errores para la resta (véase análisis de la suma, en la sección 1.3).

1.25 Se desea evaluar la función e^{5x} en el punto $x = 1.0$; sin embargo, si el valor de x se calculó en un paso previo con un pequeño error y se tiene $x^* = 1.01$; determine $\in_f$ con las expresiones dadas en la evaluación de funciones de la sección 1.3. Luego determine $\in_f$ como $f(1) - f(1.01)$ y compare los resultados.

1.26 Resuelva el siguiente sistema de ecuaciones, usando dos cifras decimales para guardar los resultados intermedios y finales.

$$21.76x + 24.34y = 1.24 \qquad 14.16x + 15.84y = 1.15$$

y determine el error cometido. La solución exacta (redondeada a 5 cifras decimales es) $x = -347.89167, y = 311.06667$.

1.27 Escriba el siguiente programa BASIC en su microcomputadora

```
INPUT A
WHILE A > 0
    PRINT LOG (EXP (A) ) - A,     EXP (LOG (A) ) - A
    INPUT A
WEND
END
```

utilice QUICK-BASIC; TURBO-BASIC o GWBASIC (en este último caso necesitará usar un número para cada línea). Ejecútelo con diferentes valores para A, tales como 1, 1.5, 1.8, 2.5, 3.1416, 0.008205, etc. y observe los resultados.

1.28 Modifique el programa del problema 1.27 agregándole al principio la instrucción

```
DEFDBL A
```

y compare los resultados

1.29 Modifique las instrucciones PRINT del programa del problema 1.27 para que queden así

```
PRINT SQR (A^2)- A,SQR(A)^2 - A
```

y vuelva a ejecutarlo con los mismos valores.

1.30 Realice la modificación indicada en el problema 1.29 al programa del problema 1.28. Compare los resultados.

1.31 Repita los problemas 1.27 a 1.30 con lenguaje PASCAL (puede usar TURBO PASCAL por ejemplo), con lenguaje C (TURBO C) y compare los resultados con los obtenidos en BASIC.

Programa en PASCAL

```
Program Errores;
Var a: Real;
Begin
    Readln (a);
    While a > 0 Do
    Begin
        Writeln (Exp (LN (a)) –a, Ln (Exp (a) ) –a;
        Readln (a);
    End;
End.
```

Programa en C;

```
# include <stdio.h >
# include < math.h >
main ( )
{
    float a;
    scanf ( "%g", &a);
    while ( a > 0 )
        {
            printf ("%g %g\n", log (exp (a ))–a, exp (log (a) )–a);
            scanf ("%g", &a);
        }
}
```

Las modificaciones para doble precisión son: En Pascal cambiar la instrucción Var a: Real; por Var a: Double;. En C cambiar la instrucción float a; por double a;. En las instrucciones scanf y printf cambiar "%g" por "%lg".

CAPÍTULO 2

SOLUCIÓN DE ECUACIONES NO LINEALES

EN ESTE CAPÍTULO se presenta un estudio minucioso de métodos muy diversos, desde perspectivas analíticas y geométricas; como prototipo de todos se tiene el método de punto fijo.

INTRODUCCIÓN

Uno de los problemas que se presenta con frecuencia en ingeniería es encontrar las raíces de ecuaciones de la forma $f(x) = 0$, donde $f(x)$ es una función real de una variable x, como un polinomio en x

$$f(x) = 4x^5 + x^3 - 8x + 2$$

o una función trascendente*

$$f(x) = e^x \operatorname{sen} x + \ln 3x + x^3$$

Existen distintos algoritmos para encontrar las raíces o ceros de $f(x) = 0$, pero ninguno es general; es decir, no hay un algoritmo que funcione con todas las ecuaciones; por ejemplo, se puede tener un algoritmo que funciona perfectamente para encontrar las raíces de $f_1(x) = 0$, pero al aplicarlo no se pueden encontrar los ceros de una ecuación distinta $f_2(x) = 0$.

Sólo en muy pocos casos será posible obtener las raíces exactas de $f(x) = 0$, como cuando $f(x)$ es un polinomio factorizable, tal como

$$f(x) = (x - \bar{x}_1)(x - \bar{x}_2) \ldots (x - \bar{x}_n),$$

*Las funciones trascendentes contienen términos trigonométricos, exponenciales o logarítmicos o ambos de la variable independiente.

donde $\bar{x}_i$, $1 \leq i \leq n$ denota la i-ésima raíz de $f(x) = 0$. Sin embargo, se pueden obtener soluciones aproximadas al utilizar algunos de los métodos numéricos de este capítulo. Se empezará con el método de punto fijo (también conocido como de aproximaciones sucesivas, de iteración funcional, etc.), por ser el prototipo de todos ellos.

SECCIÓN 2.1 MÉTODO DE PUNTO FIJO

Sea el inicio la ecuación general

$$f(x) = 0, \tag{2.1}$$

de la cual se desea encontrar una raíz real* $\bar{x}$.

El primer paso consiste en transformar algebraicamente la ecuación 2.1 a la forma equivalente

$$x = g(x) \tag{2.2}$$

Por ejemplo para la ecuación

$$f(x) = 2x^2 - x - 5 = 0, \tag{2.3}$$

cuyas raíces son 1.850781059 y –1.350781059, algunas posibilidades de $x = g(x)$ son

$$\begin{array}{lll} a) & x = 2x^2 - 5, & \text{"despejando" el segundo término.} \\ b) & x = \sqrt{\dfrac{x+5}{2}} & \text{"despejando" } x \text{ del primer término.} \\ c) & x = \dfrac{5}{2x-1} & \text{factorizando } x \text{ y "despejándola".} \\ d) & x = 2x^2 - 5 & \text{sumando } x \text{ a cada lado.} \\ e) & x = x - \dfrac{2x^2 - x - 5}{4x - 1} & \text{véase sección 2.2} \end{array} \tag{2.4}$$

Una vez que se ha determinado una forma equivalente (Ec. 2.2), el siguiente paso es **tantear** una raíz; esto puede hacerse por observación directa de la ecuación (por ejemplo en la Ec. 2.3 se ve directamente que $x = 2$ es un valor cercano a una raíz). Se denota el valor de tanteo o valor de inicio como x_0. Otros métodos de tanteo se estudiarán en la sección 2.8.

Una vez que se tiene x_0, se evalúa $g(x)$ en x_0, denotándose el resultado de esta evaluación como x_1; esto es

$$g(x_0) = x_1$$

*En las secciones 2.9 y 2.10 se discutirá el caso de raíces complejas.

El valor de x_1 comparado con x_0 presenta los dos siguientes casos

Caso 1. Que $x_1 = x_0$

Esto indica que se ha elegido como valor inicial una raíz y el problema queda concluido. Para aclararlo, recuérdese que si $\bar{x}$ es raíz de la ecuación 2.1, se cumple que

$$f(\bar{x}) = 0,$$

y como la ecuación 2.2 es sólo un rearreglo de la ecuación 2.1, también es cierto que

$$g(\bar{x}) = \bar{x}.$$

Si se hubiese elegido como $x_0 = 1.850781059$ para la ecuación 2.3, el lector puede verificar que cualquiera que sea la $g(x)$ seleccionada, $g(1.850781059) = 1.850781059$; esto se debe a que 1.850781059 es una raíz de la ecuación 2.3. Esta característica de $g(x)$ de fijar su valor en una raíz $\bar{x}$ ha dado a este método el nombre que lleva.

Caso 2. Que $x_1 \neq x_0$

Es el caso más frecuente e indica que x_1 y x_0 son distintos de $\bar{x}$. Esto es fácil de explicar, ya que si $\dot{x}$ no es una raíz de 2.1, se tiene que

$$f(\dot{x}) \neq 0,$$

y por otro lado, evaluando $g(x)$ en $\dot{x}$, se tiene

$$g(\dot{x}) \neq \dot{x}.$$

En estas circunstancias se procede a una segunda evaluación de $g(x)$, ahora en x_1, denotándose el resultado como x_2

$$g(x_1) = x_2$$

Este proceso se repite y se obtiene el siguiente esquema iterativo

Valor inicial:	x_0	$f(x_0)$	
Primera iteración	$x_1 = g(x_0)$	$f(x_1)$	
Segunda iteración	$x_2 = g(x_1)$	$f(x_2)$	
Tercera iteración	$x_3 = g(x_2)$	$f(x_3)$	
.	.	.	(2.5)
.	.	.	
.	.	.	
i-ésima iteración	$x_i = g(x_{i-1})$	$f(x_i)$	
$i+1$-ésima iteración	$x_{i+1} = g(x_i)$	$f(x_{i+1})$	
.	.	.	
.	.	.	
.	.	.	

Aunque hay excepciones, generalmente se encuentra que los valores $x_0, x_1, x_2, \ldots$ se van acercando a $\bar{x}$ de manera que x_i está más cerca de $\bar{x}$ que x_{i-1}, o bien se van alejando de $\bar{x}$ de modo que cualquiera está más lejos que el valor anterior.

Si para la ecuación 2.3 se emplea $x_0 = 2.0$ como valor inicial y las $g(x)$ de los incisos (a) y (b) de la ecuación 2.4 se obtiene, respectivamente

$x_0 = 2 \; ; \; g(x) = 2x - 5$

i	x_i	$g(x_i)$
0	2	3
1	3	13
2	13	333
3	333	221773

$x_0 = 2; \; g(x) = \sqrt{\frac{x+5}{2}}$

i	x_i	$g(x_i)$
0	2.00000	1.87083
1	1.87083	1.85349
2	1.85349	1.85115
3	1.85115	1.85083

Puede apreciarse que la sucesión diverge con la $g(x)$ del inciso (a) y converge a la raíz 1.850781059 con la $g(x)$ del inciso (b).

Finalmente, para determinar si la sucesión $x_0, x_1, x_2, \ldots$ está convergiendo o divergiendo de una raíz $\bar{x}$, cuyo valor se desconoce, puede calcularse en el proceso 2.5 la sucesión $f(x_0), f(x_1), f(x_2), \ldots$ Si dicha sucesión tiende a cero, el proceso 2.5 converge a $\bar{x}$ y dicho proceso se continuará hasta que $| f(x_i) | < \varepsilon_1$, donde ε_1 es un valor pequeño e indicativo de la exactitud o cercanía de x_i con $\bar{x}$. Se toma a x_i como la raíz y el problema de encontrar una raíz real queda concluido. Si por el contrario $f(x_0), f(x_1), f(x_2), \ldots$ no tiende a cero, la sucesión $x_0, x_1, x_2, \ldots$ diverge de $\bar{x}$ y el proceso deberá detenerse y ensayarse uno nuevo con una $g(x)$ diferente.

Ejemplo 2.1

Encuentre una aproximación a una raíz real de la ecuación

$$\cos x - 3x = 0$$

SOLUCIÓN

Dos posibilidades de $g(x) = x$ son

$$a) \; x = \cos x - 2x \qquad b) \; x = \cos x / 3$$

Graficando por separado las funciones cos x y $3x$, se obtiene la figura 2.1 (para graficar puede usar software comercial).

De donde un valor cercano a $\bar{x}$ es $x_0 = (\pi/2)/4$*. Iterando se obtiene para la forma del inciso (a)

*En el caso de funciones trigonométricas x debe estar en radianes.

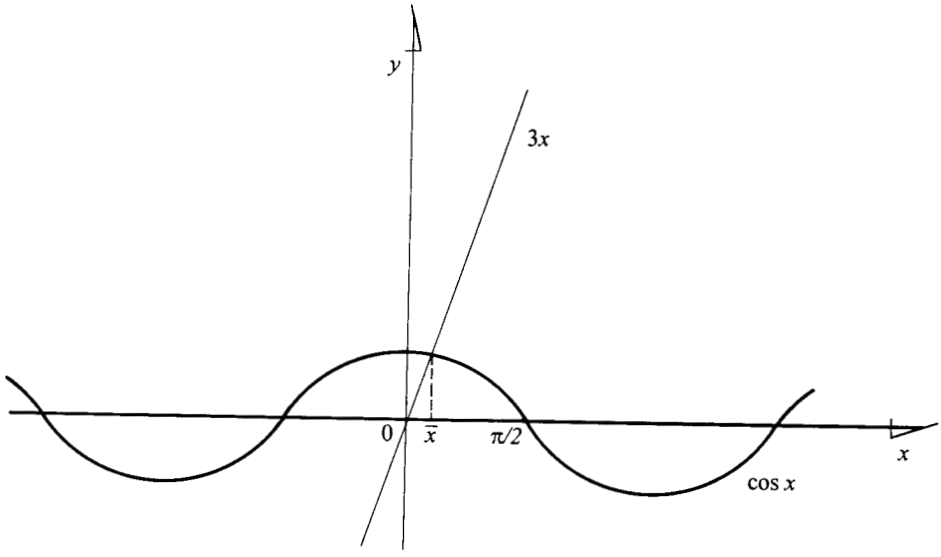


Figura 2.1. Gráfica de cos x y de $3x$

i	x_i	$g(x_i)$	$\lvert f(x_i) \rvert$
0	$\pi/8$	0.21578	0.35626
1	0.214578	0.57084	0.71256
2	0.57084	–0.14172	1.42516
3	–0.14172	1.28344	2.85057
4	1.28344	–1.56713	5.70102

Se detiene el proceso en la cuarta interación, porque $f(x_0)$, $f(x_1)$, $f(x_2)$, ... no tiende a cero. Se emplea el valor absoluto de $f(x)$ para manejar la idea de distancia.

Se inicia un nuevo proceso con $x_0 = (\pi/2)/4$ y la forma equivalente del inciso (b)

i	x_i	$g(x_i)$	$\lvert f(x_i) \rvert$
0	$\pi/8$	0.30796	0.25422
1	0.30796	0.31765	0.02907
2	0.31765	0.31666	0.00298
3	0.31666	0.31676	0.00031
4	0.31676	0.31675	0.00003

y la aproximación de la raíz es

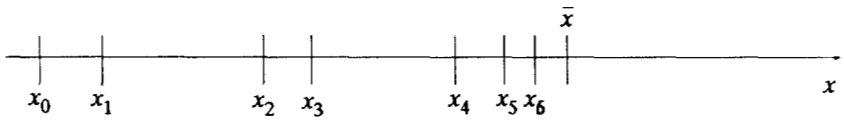
$$\bar{x} \approx x_4 = 0.31675$$

Criterio de convergencia

Se estudiará un criterio más de convergencia del proceso iterativo 2.5, basado en que

$$g(\bar{x}) = \bar{x},$$

por lo cual puede suponerse que si la sucesión $x_0, x_1, x_2, \ldots$ converge a $\bar{x}$, los valores consecutivos x_i y x_{i+1} irán acercándose entre sí conforme el proceso iterativo avanza, como puede verse enseguida



Un modo práctico de saber si los valores consecutivos se acercan es ir calculando la distancia entre ellos

$$d_i = | x_{i+1} - x_i |$$

Si la sucesión $d_1, d_2, d_3, \ldots$ tiende a cero, puede pensarse que el proceso 2.5 está convergiendo a una raíz $\bar{x}$ y debe continuarse hasta que $d_i < \varepsilon$, y tomar a x_{i+1} como la raíz buscada. Si $d_1, d_2, d_3, \ldots$ no converge para un número "grande" de iteraciones (llámense MAXIT), entonces $x_0, x_1, x_2, \ldots$ diverge de $\bar{x}$, y se detiene el proceso para iniciar uno nuevo, modificando la función $g(x)$, el valor inicial o ambos.

Este criterio de convergencia se utiliza ampliamente en el análisis numérico y resulta más sencillo de calcular que el que emplea la sucesión $f(x_0), f(x_1), f(x_2), \ldots$ pero también es menos seguro, como se verá más adelante.

Para finalizar esta sección se da un algoritmo del método de punto fijo en forma propia para lenguajes de programación.

ALGORITMO 2.6 Método de punto fijo

Para encontrar una raíz real de la ecuación $g(x) = x$ proporcionar la función G (X) y los

DATOS: Valor inicial X0, criterio de convergencia EPS y número máximo de iteraciones MAXIT.
RESULTADOS: La raíz aproximada X o un mensaje de falla.

PASO 1. Hacer I = 1
PASO 2. Mientras I <MAXIT, realizar los pasos 3 a 6.
PASO 3. Hacer X = G(X0) (calcular (x_i))

PASO 4. Si ABS (X - X0) ≤ EPS entonces IMPRIMIR X y TERMINAR. De otro modo CONTINUAR
PASO 5. Hacer I = I + 1
PASO 6. Hacer X0 = (actualiza X0)
PASO 7. IMPRIMIR mensaje de falla: "EL MÉTODO NO CONVERGE A UNA RAÍZ" y TERMINAR.

El criterio $| g'(x) | < 1$

Es importante analizar por qué algunas formas equivalentes $x = g(x)$ de $f(x) = 0$ conducen a una raíz en el método de punto fijo y otras no, aun empleando el mismo valor inicial en ambos casos.

Se inicia el análisis aplicando el teorema del punto medio* a la función $g(x)$ en el intervalo comprendido entre x_{i-1} y x_i.

$$g(x_i) - g(x_{i-1}) = g'(\xi_i)(x_i - x_{i-1}) \tag{2.6}$$

donde

$$\xi_i \in (x_i, x_{i-1}).$$

Como

$$g(x_i) = x_{i+1} \text{ y } g(x_{i-1}) = x_i$$

sustituyendo se obtiene

$$x_{i+1} - x_i = g'(\xi_i)(x_i - x_{i-1})$$

Tomando valor absoluto en ambos miembros

$$| x_{i+1} - x_i | = | g'(\xi_i) | \, | x_i - x_{i-1} |$$

Para $i = 1, 2, 3, \ldots$ la ecuación 2.7 queda así

$$\begin{aligned} | x_2 - x_1 | &= | g'(\xi_1) | \, | x_1 - x_0 | \qquad \xi_1 \in (x_1, x_0) \\ | x_3 - x_2 | &= | g'(\xi_2) | \, | x_2 - x_1 | \qquad \xi_2 \in (x_2, x_1) \\ | x_4 - x_3 | &= | g'(\xi_3) | \, | x_3 - x_2 | \qquad \xi_3 \in (x_3, x_2) \end{aligned} \tag{2.8}$$

. . .

Supóngase ahora que en la región que comprende a $x_0, x_1, \ldots$ y en $\bar{x}$ misma, la función $g'(x)$ está acotada; esto es

$$| g'(x) | \leq M,$$

*Se supone que $g(x)$ satisface las condiciones de aplicabilidad de este teorema.

para algún número M. Entonces

$$\begin{aligned} | x_2 - x_1 | &\leq M | x_1 - x_0 | \\ | x_3 - x_2 | &\leq M | x_2 - x_1 | \\ | x_4 - x_3 | &\leq M | x_3 - x_2 | \\ &\vdots \end{aligned} \tag{2.9}$$

Si se sustituye la primera desigualdad en la segunda, se tiene

$$| x_3 - x_2 | \leq M | x_2 - x_1 | \leq MM | x_1 - x_0 |$$

o bien

$$| x_3 - x_2 | \leq M^2 | x_1 - x_0 |$$

Si se sustituye este resultado en la tercera desigualdad de la ecuación 2.9 se tiene

$$| x_4 - x_3 | \leq M | x_3 - x_2 | \leq MM^2 | x_1 - x_0 |$$

o

$$| x_4 - x_3 | \leq M^3 | x_1 - x_0 |$$

Procediendo de igual manera se llega a

$$\boxed{| x_{i+1} - x_i | \leq M^i | x_1 - x_0 |} \tag{2.10}$$

El proceso 2.5 puede converger por razones muy diversas, pero es evidente que si $M < 1$, dicho proceso convergirá, ya que M^i tenderá a cero al tender i a un número grande.

En conclusión, el proceso 2.5 puede converger si M es grande y convergirá si $M < 1$ en un entorno de x que incluya $x_0, x_1, x_2, \ldots$ Entonces $M < 1$ es una condición suficiente, pero no necesaria para la convergencia.

Un método práctico de emplear este resultado es obtener distintas formas $x = g(x)$ de $f(x) = 0$, y calcular $| g'(x) |$; las que satisfagan el criterio $| g'(x_0) | < 1$ prometerán convergencia al aplicar el proceso 2.5.

Ejemplo 2.2

Calcule una raíz real de la ecuación*

$$f(x) = x^3 + 2x^2 + 10x - 20 = 0,$$

empleando como valor inicial $x_0 = 1$.

*Resuelta por Leonardo de Pisa en 1225.

SOLUCIÓN

Dos formas $x = g(x)$ de esta ecuación son

a) $x = \dfrac{20}{x^2 + 2x + 10}$ y b) $x = x^3 + 2x^2 + 11x - 20$

de donde

$$g'(x) = \frac{-20(2x + 2)}{(x^2 + 2x + 10)^2} \quad \text{y} \quad g'(x) = 3x^2 + 4x + 11$$

Sustituyendo $x_0 = 1$.

$$|g'(1)| = \left|\frac{-80}{169}\right| = 0.47 \quad \text{y} \quad |g'(1)| = 8$$

De donde la forma (*a*) promete convergencia y la forma (*b*) no.

Aplicando el proceso 2.5 y el criterio $\varepsilon = 10^{-3}$ a $|x_{i+1} - x_i|$ en caso de convergencia, se tiene

i	x_i	$\lvert x_{i+1} - x_i \rvert$	$\lvert g'(x_i) \rvert$
0	1.00000		0.47337
1	1.53846	0.53846	0.42572
2	1.29502	0.24344	0.45100
3	1.40183	1.10681	0.44047
4	1.35421	0.04762	0.44529
5	1.37009	0.02101	0.44317
6	1.36593	0.00937	0.44412
7	1.37009	0.00416	0.44370
8	1.36824	0.00185	0.44389
9	1.36906	0.00082	0.44386

Obsérvese que $|g'(x_i)|$ se mantiene menor de uno. Una vez que $|x_{i+1} - x_i| < 10^{-3}$, se detiene el proceso y se toma como raíz a x_9

$$\bar{x} \approx 1.36906$$

Si se hubiese tomado la forma equivalente

$$x = \frac{-x^3 - 2x^2 + 20}{10}$$

para la cual, se tiene

$$g'(x) = \frac{-3x^2 - 4x}{10}$$

y con $x_0 = 1$

$$|g'(1)| = \left|\frac{-7}{10}\right| = 0.7,$$

lo cual indica posibilidad de convergencia, pero al aplicar el proceso 2.5 se tiene

i	x_i	$\lvert x_{i+1}-x_i \rvert$	$\lvert g'(x_i) \rvert$
0	1.00000		0.70000
1	1.70000	0.70000	1.54700
2	0.93070	0.76930	0.63214
3	1.74614	0.81544	1.61316
4	0.85780	0.88835	0.56386
5	1.78972	0.93192	1.67682

Una divergencia lenta, ya que $| g'(x_i) |$ toma valores mayores de 1 en algunos puntos.

La condición de que el valor absoluto de $g'(x)$ sea menor que 1 en la región que comprende la raíz buscada $\bar{x}$ y los valores x_i, se interpreta geométricamente a continuación.

En caso de contar con software comercial pueden graficarse las funciones $g'(x)$ correspondientes a los incisos (*a*) y (*b*) y la recta $y = x$, y observar los valores de $g'(x)$ en las x_i del proceso iterativo.

Interpretación geométrica de $| g'(x) | < 1$

Al graficar los dos miembros de la ecuación 2.2 como las funciones $y = x$ y $y = g(x)$, la raíz buscada $\bar{x}$ es la abscisa del punto de cruce de dichas funciones (véase Fig. 2.2).

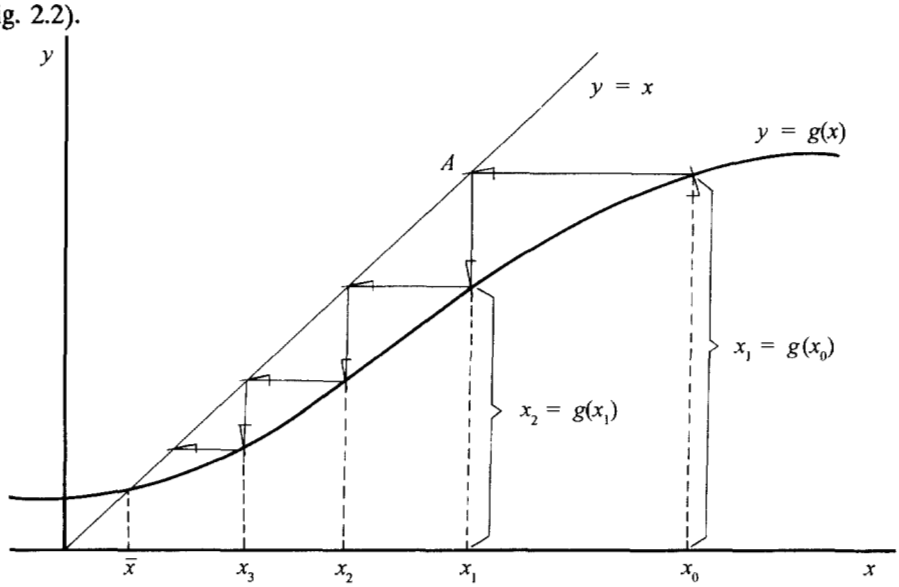
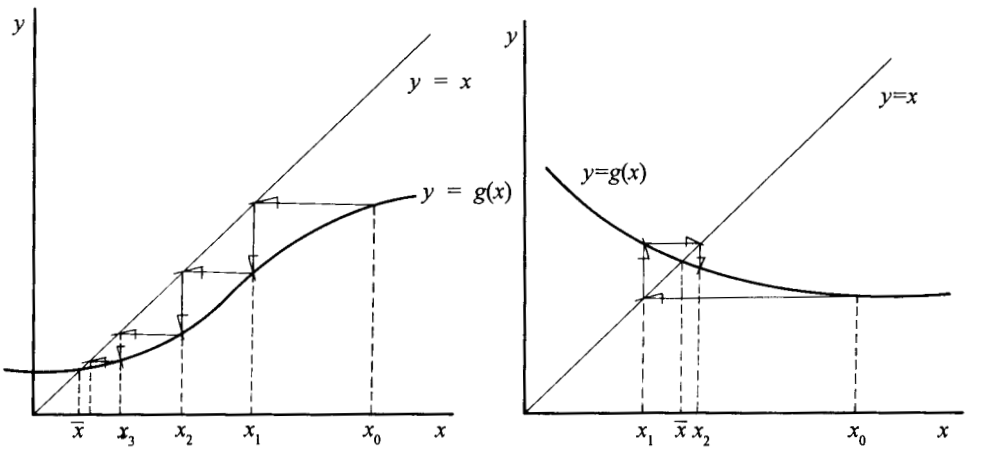


Fig. 2.2 Interpretación geométrica de $| g'(x) | < 1$.

El proceso 2.5 queda geométricamente representado en la figura 2.2, la cual muestra un caso de convergencia, ya que $g'(x)$ es menor que 1 en $x_0, x_1, \ldots \bar{x}$.

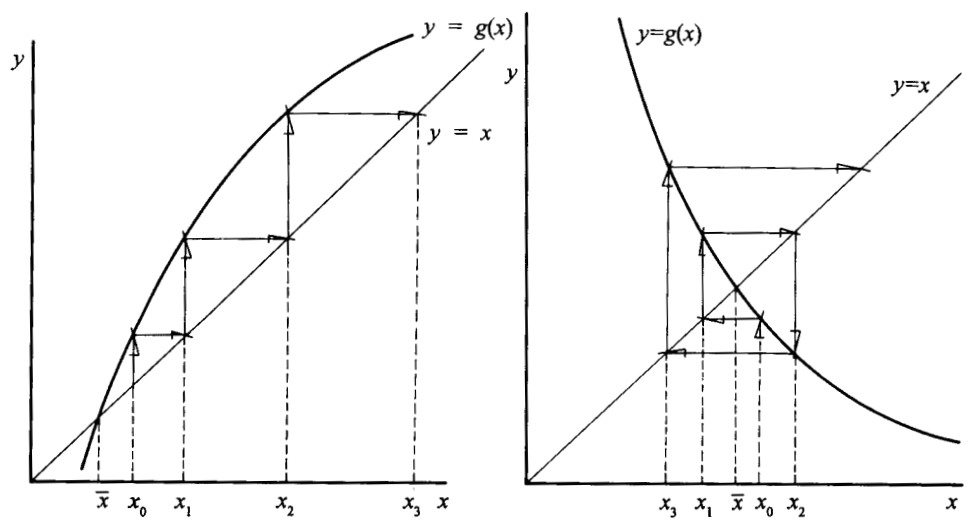
Para ver esto se trazan las tangentes a $g(x)$ en (x_0, x_1), (x_1, x_2), ... y se observa que todas tienen un ángulo de inclinación menor que la función $y = x$ cuya pendiente es 1.

A continuación se presentan geométricamente los casos posibles de convergencia y divergencia.



a. Convergencia monotónica

b. Convergencia oscilatoria



c. Divergencia monotónica

d. Divergencia oscilatoria

Figura 2.3 Cuatro casos posibles de convergencia y divergencia en la iteración $x = g(x)$.

Caso 1. La figura 2.3a ilustra qué ocurre si $g'(x)$ se encuentra entre 0 y 1. Incluso si x_0 está lejos de la raíz $\bar{x}$ —que se encuentra en el cruce de las curvas $y = x, y = g(x)$— los valores sucesivos de x_i se acercan a la raíz por un solo lado. Esto se conoce como convergencia monotónica.

Caso 2. La figura 2.3b muestra la situación en que $g'(x)$ está entre –1 y 0. Aun si x_0 está alejada de la raíz $\bar{x}$, los valores sucesivos de x_i se aproximan por el lado derecho e izquierdo de la raíz. Esto se conoce como convergencia oscilatoria.

Caso 3. En la figura 2.3c se ve la divergencia cuando $g'(x)$ es mayor que 1. Los valores sucesivos de x_i se alejan de la raíz por un solo lado. Esto se conoce como divergencia monotómica.

Caso 4. La figura 2.3d presenta la divergencia cuando $g'(x)$ es menor que –1. Los valores sucesivos de x_i se alejan de la raíz oscilando alrededor de ella. Esto se conoce como divergencia oscilatoria.

Nuevamente se recomienda usar un graficador para encontrar diversas $g(x)$ que cumplan los cuatro posibles casos mostrados.

Orden de convergencia

Se verá ahora que la magnitud de $g'(x)$ no sólo indica si el proceso converge o no, sino que además puede usarse como indicador de cuán rápida es la convergencia.

Sea $\in_i$ el error en la i-ésima iteración; esto es

$$\in_i = x_i - \bar{x}$$

Si se conoce el valor de la función $g(x)$ y sus derivadas en $\bar{x}$, puede expanderse $g(x)$ alrededor de $\bar{x}$ en serie de Taylor y encontrar así el valor de $g(x)$ en x_i

$$g(x_i) = g(\bar{x}) + g'(\bar{x})(x_i - \bar{x}) + g''(x)\frac{(x_i - \bar{x})^2}{2!} + g'''(\bar{x})\frac{(x_i - \bar{x})^3}{3!} + \ldots$$

o bien

$$g(x_i) - g(x) = g'(\bar{x})(x_i - \bar{x}) + g''(\bar{x})\frac{(x_i - \bar{x})^2}{2!} + g'''(\bar{x})\frac{(x_i - \bar{x})^3}{3!} + \ldots$$

Como

$$x_{i+1} = g(x_i)$$

y

$$\bar{x} = g(\bar{x}),$$

también puede escribirse la última ecuación como

$$x_{i+1} - \bar{x} = g'(\bar{x})\in_i + g''(\bar{x})\frac{\in_i^2}{2!} + g'''(\bar{x})\frac{\in_i^3}{3!} + \ldots$$

El miembro de la izquierda es el error en la $(i + 1)$ -ésima iteración y, por tanto, se expresa como $\in_{i+1}$ de modo que

$$\in_{i+1} = g'(\bar{x})\, \in_i + g''(\bar{x}) \frac{\in_i^2}{2!} + g'''(\bar{x}) \frac{\in_i^3}{3!} + \ldots \qquad (2.11)$$

donde puede observarse que si después de las primeras iteraciones $\in_i$ tiene un valor pequeño ($| \in_i | < 1$), entonces $\in_i^2$, $| \in_i^3 |$, $\in_i^4$, ... serán valores más pequeños que $| \in_i |$, de modo que si $g'(\bar{x}) \neq 0$, la magnitud del primer término de la ecuación 2.11 generalmente domina las de los demás términos y $\in_{i+1}$ es proporcional a $\in_i$; en cambio si $g'(\bar{x}) = 0$ y $g''(\bar{x}) \neq 0$, la magnitud del segundo término de la ecuación 2.11 predomina sobre la de los términos restantes y $\in_{i+1}$ es proporcional a $\in_i^2$. Si $g'(\bar{x}) = g''(\bar{x}) = 0$ y $g'''(\bar{x}) \neq 0$, $\in_{i+1}$ es proporcional a $\in_i^3$, etcétera.

Se dice entonces que en caso de convergencia, el proceso 2.5 tiene orden uno si $g'(\bar{x}) \neq 0$, orden dos si $g'(\bar{x}) = 0$ y $g''(\bar{x}) \neq 0$, orden tres si $g'(\bar{x}) = g''(\bar{x}) = 0$ y $g'''(\bar{x}) \neq 0$, etc. Una vez determinado el orden n se tiene que $\in_{i+1} \propto \in_i^n$ y el error $\in_{i+1}$ será más pequeño que $\in_i$ entre más grande sea n y la convergencia por tanto más rápida.

Obsérvese que en los ejemplos resueltos $g'(x) \neq 0$, y el orden ha sido uno. Como al iniciar el proceso sólo se cuenta con x_0 y algunas formas $g(x)$, puede obtenerse $g'(x)$ para cada forma y las que satisfagan la condición $| g'(x_0) | < 1$ prometerán convergencia. Dicha convergencia será más rápida para aquéllas donde $| g'(x_0) |$ sea más cercano a cero y más lenta entre más próximo esté dicho valor a 1. Así pues, para la ecuación 2.3, las formas 2.4 y el valor inicial $x_0 = 2$ se obtiene respectivamente

a) $g'(x) = 4x$ y $| g'(2) | = 4$

b) $g'(x) = \dfrac{1}{\dfrac{4(x+5)^{1/2}}{2}}$ y $| g'(2) | = 0.1336$

c) $g'(x) = \dfrac{-10}{(2x-1)^2}$ y $| g'(2) | = 1.111$

d) $g'(x) = 4x$ y $| g'(2) | = 8$

e) $g'(x) = 1 - \dfrac{(4x-1)(4x-1) - (2x^2 - x - 5)4}{(4x-1)^2}$ y $| g'(2) | = 0.08163$

Las formas de los incisos (b) y (e) quedan con posibilidades de convergencia, y la (e) como la mejor opción porque su valor está más cercano a cero.

Se deja al lector encontrar una raíz real de la ecuación 2.3 con el método de punto fijo, con la forma (e) y detener la iteración una vez que $| f(x_i) | \leq 10^{-4}$, en caso de convergencia, o desde un principio si observa divergencia en las primeras iteraciones.

SECCIÓN 2.2 MÉTODO DE NEWTON-RAPHSON

Ahora se estudiará un método de segundo orden de convergencia cuando se trata de raíces reales no repetidas. Consiste en un procedimiento que lleva la ecuación $f(x) = 0$ a la forma $x = g(x)$, de modo que $g'(\bar{x}) = 0$. Su deducción se presenta enseguida.

En la figura 2.4 se tiene la gráfica de $f(x)$ cuyo cruce con el eje x es una raíz real $\bar{x}$.

Supóngase que se escoge un valor inicial x_0 que se sitúa en el eje horizontal. Trácese una tangente a la curva en el punto $(x_0, f(x_0))$ y a partir de ese punto sígase por la tangente hasta su intersección con el eje x; el punto de corte x_1 es una nueva aproximación a $\bar{x}$ (nótese que se ha reemplazado la curva $f(x)$ con su tangente en $(x_0, f(x_0))$. El proceso se repite comenzando con x_1, se obtiene una nueva aproximación x_2 y así sucesivamente, hasta que un valor x_i satisfaga $| f(x_i) | \leq \varepsilon_1$, $| x_{i+1} - x_i | < \varepsilon$ o ambos. Si lo anterior no se cumpliera en un máximo de iteraciones (MAXIT), debe reiniciarse con un nuevo valor x_0.

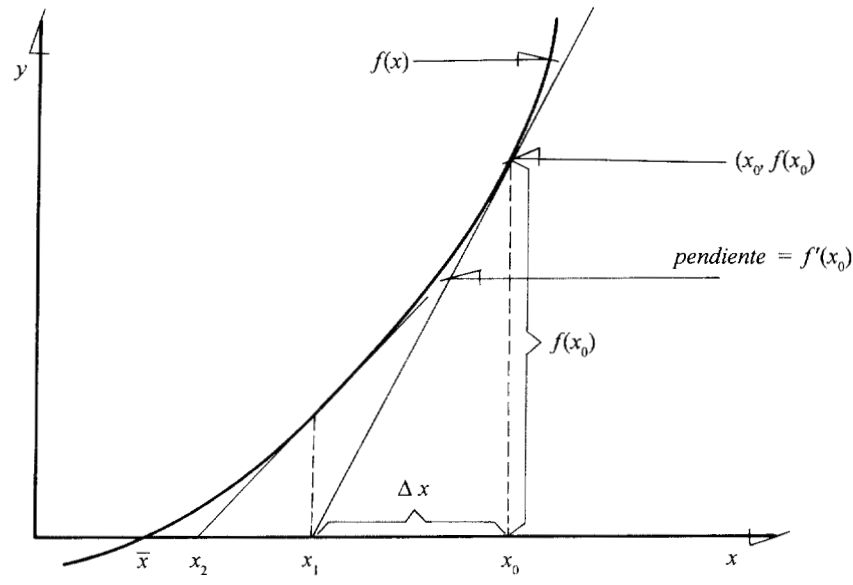


Figura 2.4 Derivación del método de Newton-Raphson.

La ecuación central del algoritmo se obtiene así

$$x_1 = x_0 - \Delta x$$

La pendiente de la tangente a la curva en el punto $(x_0, f(x_0))$ es

$$f'(x_0) = \frac{f(x_0)}{\Delta x},$$

así que

$$\Delta x = \frac{f(x_0)}{f'(x_0)}$$

y sustituyendo

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

o en general

$$\boxed{x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} = g(x_i)} \qquad (2.12)$$

Este método es de orden 2, porque $g'(\bar{x}) = 0$ y $g''(\bar{x}) \neq 0$ (véase Probl. 2.11).

Ejemplo 2.3

Encuentre una raíz real de la ecuación

$$f(x) = x^3 + 2x^2 + 10x - 20$$

mediante el método de Newton-Raphson, $x_0 = 1$,

con $\varepsilon = 10^{-3}$ aplicado a $| x_{i+1} - x_i |$

SOLUCIÓN

Se sustituyen $f(x)$ y $f'(x)$ en (2.12)

$$x_{i+1} = x_i - \frac{x_i^3 + 2x_i^2 + 10x_i - 20}{3x_i^2 + 4x_i + 10}$$

Primera iteración

$$x_1 = 1 - \frac{(1)^3 + 2(1)^2 + 10(1) - 20}{3(1)^2 + 4(1) + 10} = 1.41176$$

Como $x_1 \neq x_0$, se calcule x_2

Segunda iteración

$$x_2 = 1.41176 - \frac{(1.41176)^3 + 2(1.41176)^2 + 10(1.41176) - 20}{3(1.41176)^2 + 4(1.41176) + 10} = 1.36934$$

Con este proceso se obtiene la tabla 2.1

i	x_i	$\lvert x_{i+1} - x_i \rvert$	$\lvert g'(x_i) \rvert$
0	1.00000		0.24221
1	1.41176	0.41176	0.02446
2	1.36934	0.04243	0.00031
3	1.36881	0.00053	1.09×10^{-6}
4	1.36881	0.00000	1.2714×10^{-6}

Tabla 2.1 Resultados del ejemplo 2.3.

Se requirieron sólo tres iteraciones para satisfacer el criterio de convegencia; además se obtuvo una mejor aproximación a $\bar{x}$ que en el ejemplo 2,2, ya que f (1.36881) está más cercana a cero que f (1.36906), como se ve a continuación

$$f(1.36881) = (1.36881)^3 + 2(1.36881)^2 + 10(1.36881) - 20 = -0.00004$$

$$\lvert f(1.36881) \rvert = 0.00004 \text{ y } \lvert f(1.36906) \rvert = 0.00531$$

Obsérvese que x_4 ya no cambia con respecto a x_3 en cinco cifras decimales y que $g'(x_4)$ es prácticamente cero.

El software del libro presenta este método con diferentes posibilidades, deducción del método, solución paso a paso de un ejemplo y solución de una ecuación propuesta por el usuario.

ALGORITMO 2.2 Método de Newton-Raphson

Para encontrar una raíz real de la ecuación $f(x) = 0$, proporcionar la función F (X) y su derivada DF (X) y los

DATOS: Valor inicial X0, criterio de convergencia EPS, criterio de exactitud EPS1 y número máximo de iteraciones MAXIT.

RESULTADOS: La raíz aproximada X o un mensaje de falla.

PASO 1. Hacer I = 1

PASO 2. Mientras I < MAXIT, repetir los pasos 3 a 7.

PASO 3. Hacer X = X0 – F(X0) / DF (X0) (calcula x_i)
PASO 4. SI ABS (X – X0) < EPS, entonces IMPRIMIR X y TERMINAR. De otro modo CONTINUAR.
PASO 5. Si ABS (F (X)) < EPS1, entonces IMPRIMIR X y TERMINAR. De otro modo CONTINUAR.
PASO 6. Hacer I = I +1
PASO 7. Hacer X0 = X
PASO 8. IMPRIMIR mensaje de falla "EL MÉTODO NO CONVERGE A UNA RAÍZ" y TERMINAR.

Fallas del método de Newton-Raphson

Cuanto el método de Newton-Raphson converge se obtienen los resultados en relativamente pocas iteraciones, ya que para raíces no repetidas este método converge con orden 2 y el error $\in_{i+1}$ es proporcional al cuadrado del error anterior* $\in_i$. Para precisar más, supóngase que el error en una iteración es 10^{-n}, el error siguiente —que es proporcional al cuadrado del error anterior— es entonces aproximadamente 10^{-2n}, el que sigue será aproximadamente 10^{-4n}, etc. De esto puede afirmarse que cada iteración duplica aproximadamente el número de dígitos correctos.

Sin embargo, algunas veces el método de Newton-Raphson no converge sino que oscila. Esto ocurre si no hay raíz real como se ve en la figura 2.5a; si la raíz es un punto de inflexión como en la figura 2.5b, o si el valor inicial está muy alejado de la raíz buscada y alguna otra parte de la función "atrapa" la iteración, como en la figura 2.5c. Esto puede explorarse con el software del libro, con el GC o con un graficador.

El método de Newton- Raphson requiere la evaluación de la primera derivada de $f(x)$. En la mayoría de los problemas de los textos este requisito es trivial, pero éste no es el caso en problemas reales donde, por ejemplo, la función $f(x)$ está dada en forma tabular.

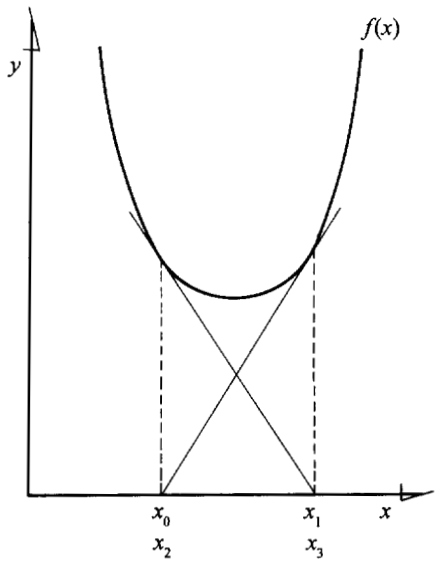
Es importante discutir algunos métodos para resolver $f(x) = 0$ que no requieran el cálculo de $f'(x)$, pero que retengan algunas de las propiedades favorables de convergencia del método de Newton-Raphson. A continuación se estudian algunos métodos que tienen estas características y que se conocen como métodos de dos puntos.

SECCIÓN 2.3 MÉTODO DE LA SECANTE

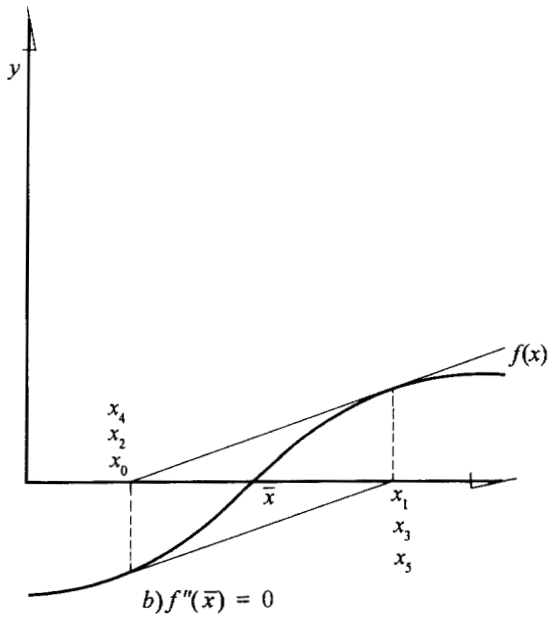
El método de la secante consiste en aproximar la derivada $f'(x_i)$ de la ecuación 2.12 por el cociente**

$$\frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}},$$

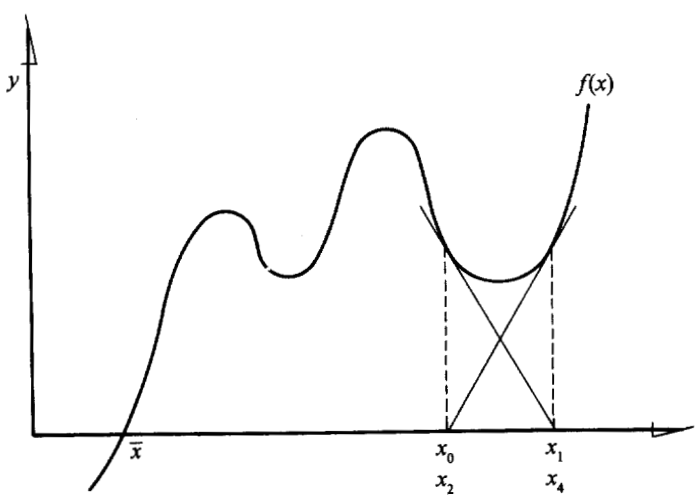
*Véase Probl. 2.13
**Nótese que este cociente es la derivada numérica de $f(x)$.



a) Raíces complejas



b) $f''(\bar{x}) = 0$



c) Valor inicial muy lejos de la raíz

Figura 2.5. Funciones donde falla el método de Newton-Raphson.

formado con los resultados de las dos iteraciones anteriores x_{i-1} y x_i. De esto resulta la fórmula

$$x_{i+1} = x_i - \frac{(x_i - x_{i-1})f(x_i)}{f(x_i) - f(x_{i-1})} = g(x_i) \tag{2.13}$$

Para la primera aplicación de la ecuación 2.13 e iniciar el proceso iterativo, se requerirán dos valores iniciales: x_0 y x_1.* La siguiente aproximación, x_2, está dada por

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

x_3 por

$$x_3 = x_2 - \frac{(x_2 - x_1)f(x_2)}{f(x_2) - f(x_1)},$$

y así sucesivamente hasta que $g(x_i) \approx x_{i+1}$ o una vez que

$$| x_{i+1} - x_i | < \varepsilon$$

o

$$| f(x_{i+1}) | < \varepsilon_1$$

Ejemplo 2.4

Use el método de la secante para encontrar una raíz real de la ecuación polinominal

$$f(x) = x^3 + 2x^2 + 10x - 20 = 0$$

SOLUCIÓN

Con la ecuación 2.13 se obtiene

$$x_{i+1} = x_i - \frac{(x_i - x_{i-1})(x_i^3 + 2x_i^2 + 10x_i - 20)}{(x_i^3 + 2x_1^2 + 10x_i - 20) - (x_{i-1}^3 + 2x_{i-1}^2 + 10x_{i-1} - 20)}$$

Mediante $x_0 = 0$ y $x_1 = 1$ se calcula x_2

$$x_2 = 1 - \frac{(1-0)(1^3 + 2(1)^2 + 10(1) - 20)}{(1^3 + 2(1)^2 + 10(1) - 20) - (0^3 + 2(0)^2 + 10(0) - 20)} = 1.53846$$

*Que pueden obtenerse por el método de punto fijo.

Los valores de las iteraciones subsecuentes se encuentran en la tabla 2.2. Si bien no se convergió a la raíz tan rápido como en el caso del método de Newton-Raphson, la velocidad de convergencia no es tan lenta como en el método de punto fijo (véase ejemplo 2.2); entonces se tiene para este ejemplo una velocidad de convergencia intermedia.

i	x_i	$\lvert x_{i+1} - x_i \rvert$
0	0.00000	
1	1.00000	1.00000
2	1.53846	0.53846
3	1.35031	0.18815
4	1.36792	0.01761
5	1.36881	0.00090
$\lvert x_{i+1} - x_i \rvert \leq \varepsilon = 10^{-3}$		

Tabla 2.2 Resultados del ejemplo 2.4.

ALGORITMO 2.3 Método de la secante

Para encontrar una raíz real de la ecuación $f(x) = 0$, dada $f(x)$ analíticamente, proporcionar la función F (X) y los

DATOS: Valores iniciales X0, X1; criterio de convergencia EPS, criterio de exactitud EPS1 y número máximo de iteraciones MAXIT.

RESULTADOS: La raíz aproximada X o un mensaje de falla.

PASO 1. Hacer I = 1

PASO 2. Mientras I < MAXIT, repetir los pasos 3 a 8.

PASO 3. Hacer
X=X0 – (X1-X0)*F(X0)/(F(X1)–F(X0))

PASO 4. Si ABS (X – X1) < EPS entonces IMPRIMIR X y TERMINAR

PASO 5. Si ABS (F (X)) < EPS1 entonces IMPRIMIR X y TERMINAR

PASO 6. Hacer X0 = X1

PASO 7. Hacer X1 = X

PASO 8. Hacer I = I + 1

PASO 9. IMPRIMIR mensaje de falla "EL MÉTODO NO CONVERGE A UNA RAÍZ" y TERMINAR.

Interpretación geométrica del método de la secante

Los dos miembros de la ecuación $x = g(x)$ se grafican por separado, como se ve en la figura 2.6.

Se eligen dos puntos del eje x: x_0 y x_1 como primeras aproximaciones a $\bar{x}$.

Se evalúa $g(x)$ en x_0 y en x_1 y se obtienen los puntos A y B de coordenadas $(x_0, g(x_0))$ y $(x_1, g(x_1))$, respectivamente.

Los puntos A y B se unen con una línea recta [secante a la curva $y = g(x)$] y se sigue por la secante hasta su intersección con la recta $y = x$. La abscisa correspondiente al punto de intersección es x_2, la nueva aproximación a $\bar{x}$.

Para obtener x_3 se repite el proceso comenzando con x_1 y x_2 en lugar de x_0 y x_1.

Este método **no garantiza** la convergencia a una raíz, lo cual puede lograrse con ciertas modificaciones que dan lugar a los métodos de posición falsa y de bisección.

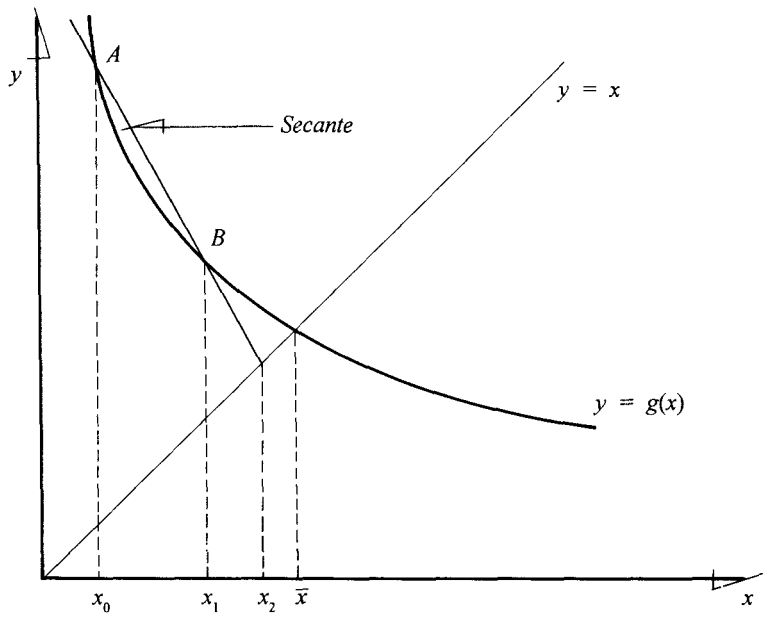


Figura 2.6. Interpretación geométrica del método de la secante.

SECCIÓN 2.4 MÉTODO DE POSICIÓN FALSA

El método de posición falsa, también llamado de Regula-Falsi, al igual que el algoritmo de la secante, aproxima la derivada $f'(x_i)$ de la ecuación 2.12 por el cociente

$$\frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}$$

pero en este caso los valores de x_i y x_{i-1} se encuentran en lados opuestos de la raíz buscada y sus valores funcionales correspondientes tienen signos opuestos; esto es

$$f(x_i) \times f(x_{i-1}) < 0$$

Se denotan x_i y x_{i+1} como x_D y x_I, respectivamente.

Para ilustrar el método se utilizará la figura 2.7 y se partirá del hecho que se tienen dos valores iniciales x_D y x_I definidos arriba y de que la función es continua en (x_I, x_D).

Se traza una línea recta que une los puntos A y B de coordenadas $(x_I, f(x_I))$ y $(x_D, fx_D))$, respectivamente. Se remplaza $f(x)$ en el intervalo (x_I, x_D) con el segmento de recta $\overline{AB}$ y el punto de intersección de este segmento con el eje x, x_M, será la siguiente aproximación a $\bar{x}$.

Se evalúa $f(x_M)$ y se compara su signo con el de $f(x_D)$. Si son iguales, se actualiza x_D sustituyendo su valor con el de x_M; si los signos son diferentes, se actualiza x_I sustituyendo su valor con el de x_M. Nótese que el objetivo es mantener los valores descritos (x_D y x_I) cada vez más cercanos entre sí y la raíz entre ellos.

Se traza una nueva línea secante entre los puntos actuales A y B y se repite el proceso hasta que se satisfaga el criterio de exatitud $| f(x_M) | < \varepsilon_1$ tomándose como aproximación a $\bar{x}$ el valor último de x_M. Para terminar el proceso también puede usarse el criterio $| x_D - x_I | < \varepsilon$. En este caso se toma como aproximación a $\bar{x}$ la media entre x_D y x_I.

Para calcular el valor de x_M se sustituye x_D por x_i y x_I por x_{i-1} en la ecuación 2.13, con lo que se llega a

$$x_M = x_D - \frac{(x_D - x_I) f(x_D)}{f(x_D) - f(x_I)} = \frac{x_I f(x_D) - x_D f(x_I)}{f(x_D) - f(x_I)} \qquad (2.14)$$

el algoritmo de posición falsa.

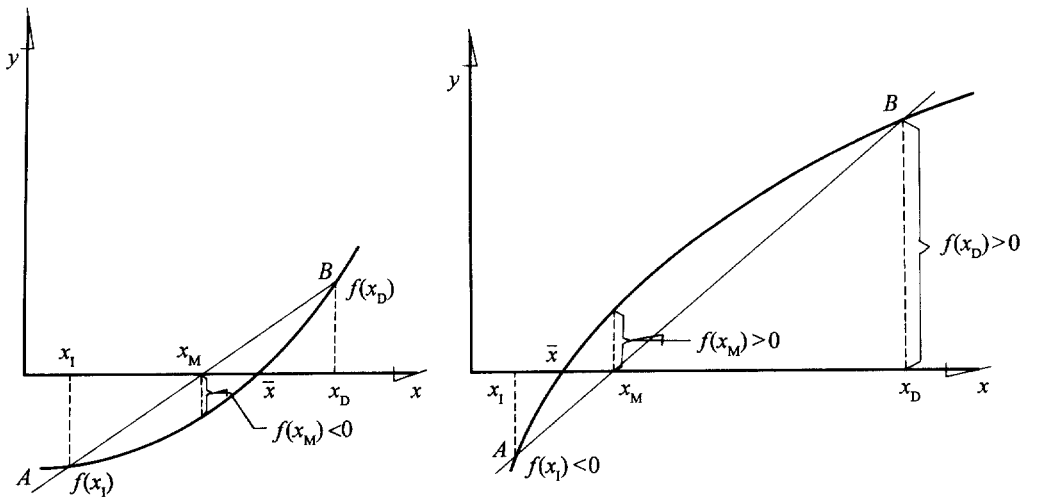


Figura 2.7. Método de posición falsa.

Ejemplo 2.5

Utilice el método de posición falsa para obtener una raíz real del polinomio

$$f(x) = x^3 + 2x^2 + 10x - 20$$

SOLUCIÓN

Para obtener x_I y x_D se puede, por ejemplo, evaluar la función en algunos puntos donde este cálculo sea fácil o bien se grafica. Así

$$f(0) = -20$$
$$f(1) = -7$$
$$f(-1) = -29$$
$$f(2) = 16$$

De acuerdo con el teorema de Bolzano hay una raíz real, por lo menos, en el intervalo (1, 2); por tanto

$$x_D = 1 \; ; f(x_D) = -7$$
$$x_I = 2 \; ; f(x_I) = 16$$

Al aplicar la ecuación 2.14 se obtiene x_M

$$x_M = 1 - \frac{(1-2)(-7)}{-7-16} = 1.30435$$

y

$$f(x_M) = (1.30435)^3 + 2(1.30435)^2 + 10(1.30435) - 20 = -1.33476$$

Como $f(x_M) < 0$, (igual signo que $f(x_D)$), se reemplaza el valor de x_D con el de x_M, con lo cual queda el nuevo intervalo como (1.30435,2). Por tanto

$$x_D = 1.30435 \; ; f(x_D) = -1.33476$$
$$x_I = 2 \; ; f(x_I) = 16$$

Se calcula una nueva x_M

$$x_M = 1.30435 - \frac{(1.30435 - 2)(-1.33476)}{(-1.33476 - 16)} = 1.35791,$$

$$f(x_M) = (1.35791)^3 + 2(1.35791)^2 + 10(1.35791) - 20 = -0.22914$$

Como $f(x_M) < 0$, el valor actual de x_D se remplaza con el último valor de x_M; así el intervalo queda reducido a (1.35791,2). La tabla 2.3 muestra los cálculos llevados a cabo hasta satisfacer el criterio de exactitud

$$| f(x_M) | \leq 10^{-3}$$

i	x_D	x_I	x_M	$\| f(x_M) \|$
0	1.00000	2.00000		
1	1.00000	2.00000	1.30435	1.33476
2	1.30435	2.00000	1.35791	0.22914
3	1.35791	2.00000	1.36698	0.03859
4	1.36698	2.00000	1.36850	0.00648
5	1.36850	2.00000	1.36876	0.00109
6	1.36876	2.00000	1.36880	0.00018

Tabla 2.3 Resultados del ejemplo 2.5.

NOTA: El GC proporciona los métodos de punto fijo, Newton-Raphson, posición falsa y bisección, de modo tal que pueden verse las iteraciones gráfica y numéricamente al resolver una ecuación dada. También hay calculadoras que disponen de algunos de estos métodos con las cuales auxiliarse.

ALGORITMO 2.4 Método de posición falsa

Para encontrar una raíz real de la ecuación $f(x) = 0$, dada $f(x)$ analíticamente, proporcionar la función F (X) y los

DATOS: Valores iniciales XI y XD que forman un intervalo en donde se halla una raíz $\bar{x}$ (F (XI) * F (XD) < 0), criterio de convergencia EPS, criterio de exactitud EPS1 y número máximo de iteraciones MAXIT.

RESULTADOS: La raíz aproximada X o un mensaje de falla.

PASO 1. Hacer I = 1; FI = F (XI); FD = F (XD)

PASO 2. Mientras I < MAXIT, repetir los pasos 3 a 8.

PASO 3. Hacer XM = (XI*FD - XD*FI) / (FD - FI); FM = F (XM).

PASO 4. Si ABS (FM) < EPS1 entonces IMPRIMIR XM y TERMINAR.

PASO 5. Si ABS (XD-XI) < EPS, entonces Hacer XM = (XD+XI) / 2; IMPRIMIR "LA RAÍZ BUSCADA ES", IMPRIMIR XM y TERMINAR.

PASO 6. Si FD * FM >0, hacer XD = XM (actualiza XD).

PASO 7. Si FD * FM <0, hacer XI = XM (actualiza XI).
PASO 8. Hacer I = I + 1.
PASO 9. IMPRIMIR mensaje de falla "EL MÉTODO NO CONVERGE A UNA RAÍZ" y TERMINAR.

SECCIÓN 2.5 MÉTODO DE LA BISECCIÓN

El método de la bisección es muy similar al de posición falsa, aunque algo más simple. Como en el método de posición falsa, también se requieren dos valores iniciales a ambos lados de la raíz y que sus valores funcionales correspondientes sean de signos opuestos.

En este caso el valor de x_M se obtiene como el punto medio entre x_I y x_D.

$$x_M = (x_I + x_D) / 2$$

Dependiendo de la función que se tenga en particular, el método de bisección puede converger ligeramente más rápido o más lentamente que el método de posición falsa. Su gran ventaja sobre el método de posición falsa es que proporciona el tamaño exacto del intervalo en cada iteración (en ausencia de errores de redondeo). Para aclarar esto, nótese que en este método después de cada iteración el tamaño del intervalo se reduce a la mitad; después de n interaciones, el intervalo original se habrá reducido 2^n veces. Por lo anterior, si el intervalo original es de tamaño a y el criterio de convergencia aplicado al valor absoluto de la diferencia de dos x_M consecutivas es ε, entonces se requerirán n iteraciones, donde n se calcula con la igualdad de la expresión

$$\frac{a}{2^n} \leq \varepsilon ,$$

de donde:

$$n = \frac{\ln a - \ln \varepsilon}{\ln 2} \tag{2.15}$$

Por esto se dice que se puede saber de antemano cuántas iteraciones se requieren.

Ejemplo 2.6

Utilice el método de bisección para obtener una raíz real del polinomio

$$f(x) = x^3 + 2x^2 + 10x - 20$$

SOLUCIÓN

Con los valores iniciales obtenidos en el ejemplo 2.5

$$x_D = 1 ; f(x_D) = -7,$$
$$x_I = 2 ; f(x_I) = 16$$

Si $\varepsilon = 10^{-3}$, el número de iteraciones n será

$$n = \frac{\ln a - \ln \varepsilon}{\ln 2} = \frac{\ln (2 - 1) - \ln 10^{-3}}{\ln 2} = 6.64$$

o bien

$$n \approx 7$$

Primera iteración

$$x_M = \frac{1 + 2}{2} = 1.5$$
$$f(1{,}5) = 2.88$$

Como $f(x_M) > 0$ (distinto signo de $f(x_D)$), se remplaza el valor de x_I con el de x_M, con lo cual queda un nuevo intervalo (1,1.5). Entonces

$$x_D = 1 ; f(x_D) = -7$$
$$x_I = 1.5 ; f(x_I) = 2.88$$

Segunda iteración

$$x_M = \frac{1 + 1.5}{2} = 1.25$$

y

$$f(x_M) = -2.42$$

Como ahora $f(x_M) < 0$ (igual signo que $f(x_D)$), se remplaza el valor de x_D con el valor de la nueva x_M; de esta manera queda como intervalo (1.25, 1.5).

La tabla 2.4 muestra los cálculos, llevados a cabo trece veces, con el fin de hacer ciertas observaciones.

El criterio $| x_{i+1} - x_i | \leq 10^{-3}$ se satisface en diez iteraciones; es decir, tres más de las previstas en la ecuación 2.15, debido principalmente a los errores de redondeo involucrados en el método.

Nótese que si ε se hubiese aplicado sobre $| f(x_M) |$, se habrían requerido 13 iteraciones en lugar de 10. En general se necesitarán más iteraciones para satisfacer un valor de ε sobre $| f(x_M) |$ que cuando se aplica a $| x_{i+1} - x_i |$.

i	x_D	x_I	x_M	$\lvert x_{Mi} - x_{Mi+1} \rvert$	$\lvert f(x_M) \rvert$
0	1.00000	2.00000			
1	1.00000	2.00000	1.50000		2.87500
2	1.00000	1.50000	1.25000	0.25000	2.42188
3	1.25000	1,50000	1.37500	0.12500	2.42188
4	1.25000	1.37500	1.31250	0.06250	0.13086
5	1.31250	1.37500	1.34375	0.03125	0.52481
6	1.34375	1.37500	1.35938	0.01563	0.19846
7	1.35938	1.37500	1.36719	0.00781	0.03417
8	1.36719	1.37500	1.37109	0.00391	0.04825
9	1.36719	1.37109	1.36914	0.00195	0.00702
10	1.36719	1.36914	1.36816	0.00098	0.01358
11	1.36826	1.36914	1.36865	0.00049	0.00329
12	1.36865	13.6914	1.36890	0.00025	0.00186
13	1.36865	1.36890	1.36877	0.00013	0.00071

Tabla 2.4 Resultados del ejemplo 2.6

SECCIÓN 2.6 PROBLEMAS DE LOS MÉTODOS DE DOS PUNTOS Y ORDEN DE CONVERGENCIA

A continuación se mencionan algunos problemas que se presentan en la aplicación de los métodos de dos puntos.

1. El hecho de requerir dos valores iniciales. Esto resulta imposible de satisfacer (en bisección y posición falsa) si se tienen raíces repetidas por parejas (x_1 y x_2) o muy difícil si la raíz buscada se encuentra muy cerca de otra (x_3 y x_4) (Véase Fig. 2.8). En el último caso, uno de los valores iniciales debe estar entre las dos raíces o de otra manera no se detectará ninguna de ellas.
2. Debido a los errores de redondeo $f(x_M)$ se calcula con un ligero error. Esto no es un problema sino hasta que x_M está muy cerca de la raíz x y $f(x_M)$ resulta ser positiva cuando debería ser negativa o viceversa, o bien resulta ser cero.
3. En el método de la secante no hay necesidad de tener valores iniciales a ambos lados de la raíz que se busca. Esto constituye una ventaja, pero puede ser peligroso, ya que en la ecuación 2.13

$$x_{i+1} = x_i - \frac{(x_i - x_{i-1})f(x_i)}{f(x_i) - f(x_{i-1})} = \frac{f(x_i)x_{i-1} - f(x_{i-1})x_i}{f(x_i) - f(x_{i+1})}$$

la diferencia

$$f(x_i) - f(x_{i-1})$$

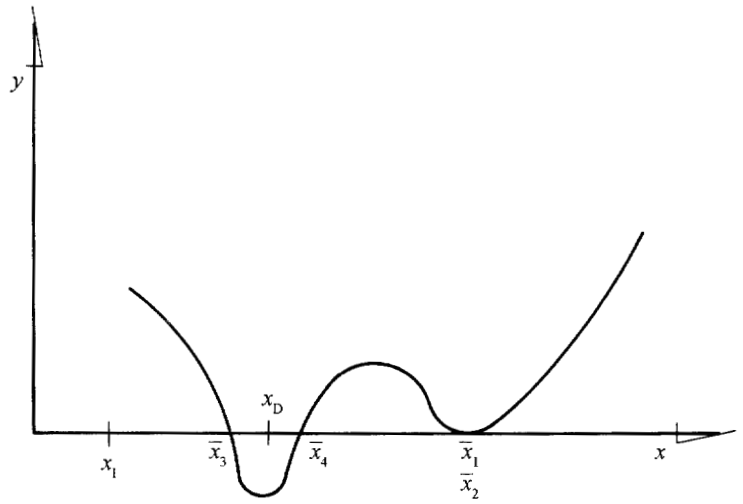


Fig.ura 2.8. Raíces repetidas por parejas y muy cercanas entre sí.

puede causar serios problemas de redondeo al evaluar x_{i+1}, pues $f(x_i)$ y $f(x_{i-1})$ no tienen necesariamente signos opuestos. Por último debe decirse que en el método de la secante no hay certeza de convergencia.

Orden de convergencia

Se determinará el orden de convergencia del método de la secante solamente, ya que para los demás métodos de dos puntos vistos, se siguen las mismas ideas.

Si, como antes, $\in_i$ representa el error en la i-ésima iteración

$$\in_{i-1} = x_{i-1} - \bar{x}$$
$$\in_i = x_i - \bar{x}$$
$$\in_{i+1} = x_{i+1} - \bar{x}$$

Al sustituir en la ecuación 2.13 x_{i+1}, x_i, x_{i-1} despejadas de las ecuaciones de arriba, se tiene

$$\bar{x} + \in_{i+1} = \bar{x} + \in_i - \frac{(\bar{x} + \in_i - \bar{x} - \in_{i-1}) f(\in_i + \bar{x})}{f(\in_i + \bar{x}) - f(\in_{i-1} + \bar{x})}$$

o bien

$$\in_{i+1} = \in_i - \frac{(\in_i - \in_{i-1}) f(\in_i + \bar{x})}{f(\in_i + \bar{x}) - f(\in_{i-1} + \bar{x})} \tag{2.17}$$

Si se expande en serie de Taylor a $f(\in_i + \bar{x})$ y $f(\in_{i-1} + \bar{x})$ alrededor de $\bar{x}$ se tiene

$$f(\in_i + \bar{x}) = f(\bar{x}) + \in_i f'(\bar{x}) + \frac{\in_2}{2!} f''(\bar{x}) + \ldots$$

$$f(\in_{i-1} + \bar{x}) = f(\bar{x}) + \in_{i-1} f'(\bar{x}) + \frac{\in_{i-1}^2}{2!} f''(\bar{x}) + \ldots$$

Sustituyendo estas expansiones en la ecuación 2.17 y como $f(\bar{x}) = 0$, queda

$$\epsilon_{i+1} = \epsilon_i - \frac{(\epsilon_i - \epsilon_{i-1})(\epsilon_i f'(\bar{x}) + \epsilon_i^2 f''(\bar{x})/2! + \ldots)}{(\epsilon_i - \epsilon_{i-1}) f'(\bar{x}) + \frac{1}{2!}(\epsilon_i^2 - \epsilon_{i-1}^2) f''(\bar{x}) + \ldots}$$

Factorizando a $(\epsilon_i - \epsilon_{i-1})$ en el denominador y cancelándolo con el mismo factor del numerador queda

$$\epsilon_{i+1} = \epsilon_i - \frac{(\epsilon_i f'(\bar{x}) + \epsilon_i^2 f''(\bar{x})/2! + \ldots)}{f'(\bar{x}) + \frac{1}{2!}(\epsilon_i + \epsilon_{i-1}) f''(\bar{x}) + \ldots}$$

$$= \epsilon_i - \frac{(\epsilon_i f'(\bar{x}) + \epsilon_i^2 f''(\bar{x})/2! + \ldots)}{f'(x)} \left(1 + \frac{1}{2!}(\epsilon_i + \epsilon_{i-1}) \frac{f''(\bar{x})}{f'(\bar{x})} + \ldots\right)^{-1}$$

Por el teorema binominal

$$\epsilon_{i+1} = \epsilon_i - \frac{1}{f'(\bar{x})}\left(\epsilon_i f'(\bar{x}) + \frac{\epsilon_i^2}{2!} f'(\bar{x}) + \ldots\right)\left(1 - \frac{1}{2!}(\epsilon_i + \epsilon_{i-1}) \frac{f''(\bar{x})}{f'(\bar{x})} + \ldots\right)$$

$$\epsilon_i - \frac{1}{f'(\bar{x})}\left(\epsilon_i f'(\bar{x}) + \frac{1}{2!}\epsilon_i^2 f''(\bar{x}) + \ldots - \frac{-1}{2!}(\epsilon_i + \epsilon_{i-1}) f''(\bar{x}) + \ldots\right)$$

$$= \epsilon_i - \frac{1}{f'(x)}\left(\epsilon_i f'(\bar{x}) - \frac{1}{2!}\epsilon_i \epsilon_{i-1} f''(\bar{x}) + \ldots\right)$$

$$= \frac{1}{2!}\epsilon_i \epsilon_{i-1} \frac{f''(\bar{x})}{f'(\bar{x})} + \ldots$$

o bien

$$\boxed{\epsilon_{i+1} \approx \frac{1}{2!} \frac{f''(\bar{x})}{f'(\bar{x})} \epsilon_i \epsilon_{i-1},}$$

donde se aprecia que el error en la $(i+1)$-ésima iteración es proporcional al producto de los errores de las dos iteraciones previas.

El error en el método de Newton-Raphson está dado así (véase Probl. 2.13)

$$\boxed{\epsilon_{i+1} \approx \frac{f''(\bar{x})}{2! f'(\bar{x})} \epsilon_i^2,}$$

donde por comparación puede observarse que el error en el método de la secante es ligeramente mayor que en el de Newton-Raphson; por tanto, su orden de convergencia será ligeramente menor, pero con la ventaja de que no hay que derivar la función $f(x)$.

Por otro lado en los métodos de primer orden el error en la iteración $(i+1)$-ésima es proporcional al error de la iteración previa solamente, por lo que puede decirse que los métodos de dos puntos son **superlineales** (orden de convergencia mayor de uno pero menor de dos).

SECCIÓN 2.7 ACELERACIÓN DE CONVERGENCIA

Se han visto métodos cuyo orden de convergencia es uno y dos, o bien un valor intermedio (superlineales). Existen métodos de orden 3 (véase Probl. 2.14) y de orden superior; sin embargo, es importante dar otro giro a la búsqueda de raíces reales y averiguar si la convergencia de los métodos vistos se puede acelerar.

Métodos de un punto

Si en alguno de los métodos vistos se tiene que la sucesión , $x_0\, x_1, x_2, \ldots$ converge muy lentamente a la raíz buscada, pueden tomarse, entre otras, las siguientes decisiones

a) Continuar el proceso hasta satisfacer alguno de los criterios de convergencia prestablecidos.
b) Ensayar con una $g(x)$ distinta; es decir, buscar una nueva $g(x)$ en punto fijo o cambiar de método.
c) Utilizar la sucesión de valores $x_0, x_1, x_2, \ldots$ para generar otra sucesión: $x_0', x_1', x_2', \ldots$ que converja más rapidamente a la raíz $\bar{x}$ que se busca.

Los incisos (*a*) y (*b*) son suficientemente claros, mientras que la sucesión $x_o', x_1', x_2', \ldots$ de la parte (c) se basa en que en ciertas condiciones de $g'(x)$*, se tiene que

$$\lim_{i \to \infty} \frac{\in_{i+1}}{\in_i} = g'(\bar{x}) \tag{2.18}$$

donde $\in_i = x_i - \bar{x}$ es el error en la i-ésima iteración.

Para valores finitos de *i*, la ecuación 2.18 puede escribirse como

$$\frac{\in_{i+1}}{\in_i} \approx g'(\bar{x})$$

o

$$x_{i+1} - \bar{x} \approx g'(\bar{x})\,(x_i - \bar{x}) \tag{2.19}$$

o también

$$x_{i+2} - \bar{x} \approx g'(\bar{x})\,(x_{i+1} - \bar{x}) \tag{2.20}$$

Restando la ecuación 2.19 de la 2.20 se tiene

$$x_{i+2} - x_{i+1} \approx g'(\bar{x})\,(x_{i+1} - x_i),$$

de donde

$$g'(\bar{x}) \approx \frac{x_{i+2} - x_{i+1}}{x_{i+1} - x_i} \tag{2.21}$$

*Véase Problema 2.22.

Despejando $\bar{x}$ de la ecuación 2.19

$$\bar{x} \approx \frac{x_{i-1} - g'(\bar{x})x_i}{1-g'(\bar{x})}$$

sustituyendo la ecuación 2.21 en la última ecuación, se llega a

$$\bar{x} \approx x_i - \frac{(x_{i+1} - x_i)^2}{x_{i+2} - 2x_{i+1} + x_i}$$

que da aproximaciones a $\bar{x}$ a partir de los valores ya obtenidos en alguna sucesión. Llámese a esta nueva sucesión $x_0', x_1', x_2', \ldots$

$$\boxed{x_i' = x_i - \frac{(x_{i+1} - x_i)^2}{x_{i+2} - 2x_{i+1} + x_i} \qquad i \geq 0} \tag{2.22}$$

Por ejemplo x_0' requiere de x_0, x_1, x_2, ya que

$$x_0' = x_0 - \frac{(x_1 - x_0)^2}{x_2 - 2x_1 + x_0}$$

y x_1' de x_1, x_2, x_3, pues

$$x_1' = x_1 - \frac{(x_2 - x_1)^2}{x_3 - 2x_2 + x_1}$$

y así sucesivamente.

Este proceso conducirá, en la mayoría de los casos, a la solución buscada $\bar{x}$ más rápido que si se siguiera el inciso (a); asimismo evita la búsqueda de una nueva $g(x)$ y el riesgo de no obtener convergencia con esa nueva $g(x)$. A este proceso se le conoce como aceleración de convergencia y se presenta como algoritmo de Aitken

Algoritmo de Aitken

Dada una sucesión de números $x_0, x_1, x_2, \ldots$ a partir de ella se genera una nueva sucesión $x_0', x_1', x_2', \ldots$ con la ecuación 2.22.

Si se emplea la notación

$$\Delta x_i = x_{i+1} - x_i, \qquad i = 0,1,2, \ldots$$

donde Δ es un operador* de diferencias cuyas potencias se pueden obtener así

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

o

$$\Delta^2 x_i = x_{i+2} - 2x_{i+1} + x_i$$

*Véase capítulo 5

la ecuación 2.22 adquiere la forma simplificada

$$x_i' = x_i - \frac{(\Delta x_i)^2}{\Delta^2 x_i} \tag{2.23}$$

Ejemplo 2.7

Acelerar la convergencia de la sucesión del ejemplo 2.2, mediante el algoritmo de Aitken.

SOLUCIÓN

Con la ecuación 2.22 con $x_0 = 1, x_1 = 1.53846$ y $x_2 = 1.29502$, se tiene

$$x_0' = 1 - \frac{(1.53846 - 1)^2}{1.29502 - 2(1.53846) + 1} = 1.37081$$

Ahora, con la ecuación 2.22 y con $x_1 = 1.53846$, $x_2 = 1.29502$ y $x_3 = 1.40183$, resulta

$$x_1' = 1.53486 - \frac{(1.29502 - 1.53846)^2}{1.40183 - 2(1.29502) + 1.53846} = 1.36926$$

En una tercera iteración se obtiene

$$x_2' = 1.36889$$

Obsérvese que x_1' está prácticamente tan cerca de la raíz real de la ecuación como el valor de x_9 del ejemplo 2.2 y x_2' mejora tanto la aproximación que es preciso comparar este valor con el de x_3 del ejemplo 2.3. La comparación puede establecerse mediante de $| f(x_i') |$ y $| f(x_i) |$.

Se ha encontrado que el método de Aitken es de segundo orden* y se emplea normalmente para acelerar la convergencia de cualquier sucesión de valores que converge linealmente, cualquiera que sea origen. La aplicación del método de Aitken a la iteración de punto fijo da el procedimiento conocido como método de Steffensen, que se ilustra a continuación.

Ejemplo 2.8

Encuentre una raíz real de la ecuación

$$f(x) = x^3 + 2x^2 + 10x - 20 = 0$$

con el método de Steffensen, usando $\varepsilon = 10^{-3}$ aplicado a $| f(x_i') |$.

*Henrici, P., Elements of Numerical Analysis. John Wiley & Sons, Inc. (1964). p 91-92.

SOLUCIÓN

Se pasa primero la ecuación $f(x) = 0$ a la forma $g(x) = x$. Al igual que en el ejemplo 2.2, se factoriza x en la ecuación y luego se "despeja"

$$x = \frac{20}{x^2 + 2x + 10}$$

Primera iteración

Se elige un valor inicial $x_0 = 1$ y se calcula x_1 y x_2

$$x_1 = 1.53846$$

$$x_2 = 1.29502$$

Se aplica ahora la ecuación 2.22 para acelerar la convergencia

$$x_0' = 1 - \frac{(1.53846 - 1)^2}{1.29502 - 2(1.53846) + 1} = 1.37081$$

Como $| f(x_0') | = (1.37081)^3 + 2(1.37081)^2 + 10(1.37081) - 20 =$

$= 0.04234 > 10^{-3}$, se pasa a la

Segunda iteración

Con el valor de x_0' que ahora se denota como x_3 y con la $g(x)$ que se tiene, resulta

$$x_4 = 1.36792$$

$$x_5 = 1.36920$$

Aplicando nuevamente la ecuación 2.22 a x_3, x_4 y x_5 se llega a

$$x_1' = x_6 = 1.37081 - \frac{(1.36792 - 1.37081)^2}{1.36920 - 2(1.36792) + 1.37081}$$

$$= 1.36881$$

Luego, con el criterio de exactitud se tiene

$$| f(x_6) | = 0.0000399 < 10^{-3}$$

y el problema queda resuelto.

A continuación se da el algoritmo de Steffensen.

ALGORITMO 2.5 Método de Steffensen

Para encontrar una raíz real de la ecuación $g(x) = x$, proporcionar la función G(X) y los

DATOS: Valor inicial X0, criterio de convergencia EPS y número máximo de iteraciones MAXIT.

```
    RESULTADOS:   La raíz aproximada X o un mensaje de falla.
PASO 1.  Hacer I = 1
PASO 2.  Mientras I < MAXIT, repetir los pasos 3 a 6
         PASO 3.  Hacer
                  X1 = G( X0 )
                  X2 = G( X1 )
                  X = X0 - ( X1-X0 ) ^2/ ( X2-2*X1+X0 )
         PASO 4.  SI ABS ( X-X0 ) < EPS, IMPRIMIR X y
                  TERMINAR. De otro modo CONTINUAR.
         PASO 5.  Hacer I = I + 1
         PASO 6.  Hacer X0 = X ( actualiza X0 )
PASO 7.  IMPRIMIR mensaje de falla: "EL MÉTODO NO CONVERGE
         A UNA RAÍZ" y TERMINAR.
```

Métodos de dos puntos

Los métodos de dos puntos bisección y posición falsa garantizan convergencia; pero ya que puede ser muy lenta en algunos casos, conviene acelerarla. Se estudia enseguida una modificación de posición falsa que cumple con este cometido.

Método Illinois*

Esta técnica difiere del método de posición falsa (véase algoritmo 2.4) en que los valores (X_I, F_I), (X_D, F_D) de las sucesivas iteraciones se determinan de acuerdo con las siguientes reglas

a) Si $F_D * F_M > 0$, hacer $X_D = X_I$, $F_D = F_I$

b) Si $F_D * F_M < 0$, hacer $F_D = F_D/2$

y en ambos casos se sustituye a X_I con X_M y F_I con F_M.

El empleo de $F_D/2$ en lugar de F_D evita que uno de los extremos X_I o X_D se mantenga fijo (caso frecuente en posición falsa). Esta modificación acelera considerablemente la convergencia del método. Los valores funcionales F_I, F_D empleados conservan sus signos opuestos. El algoritmo correspondiente puede obtenerse sustituyendo los pasos 6 y 7 en el algoritmo 2.4 con los incisos (a) y (b), respectivamente y además un paso donde se sustituye a X_I con X_M y F_I con F_M.

SECCIÓN 2.8 BÚSQUEDA DE VALORES INICIALES

El uso de cualquier algoritmo numérico para encontrar las raíces de $f(x) = 0$, requiere uno o más valores iniciales; además en métodos como el de bisección y el

---

*Dowel M. and Jarrat P., *A modified Regula Falsi Method for Computing the Root of an Equation*. BIT. Vol. 11 P. 168 (1971).

de posición falsa, los dos valores iniciales requeridos deben estar a los lados de la raíz buscada y sus valores funcionales correspondientes tienen que ser de signos opuestos.

A continuación se dan algunos lineamientos generales para obtener valores aproximados a las raíces de $f(x) = 0$.

1. Por lo general, la ecuación cuyas raíces se buscan tiene algún significado físico; entonces a partir de consideraciones físicas pueden estimarse valores aproximados a las raíces. Este razonamiento es particular para cada ecuación. A continuación se presenta un ejemplo para ilustrar esta idea.

Ejemplo 2.9

Determine el valor inicial en la solución de una ecuación de estado.

SOLUCIÓN

El cálculo del volumen molar de un gas dado, a cierta presión y temperatura también dadas, es un problema común en ingeniería química. Para realizar dicho cálculo se emplea alguna de las ecuaciones de estado conocidas. Una de ellas es la ecuación de Beattie-Bridgeman

$$P = \frac{RT}{V} + \frac{\beta}{V^2} + \frac{\gamma}{V^3} + \frac{\delta}{V^4}, \qquad (2.24)$$

donde los parámetros β, γ, y δ quedan determinados al fijar el gas de que se trata, su temperatura T y su presión P.

En las condiciones expuestas, el problema se reduce a encontrar el o los valores de V que satisfagan la ecuación 2.24, o en otros términos, a determinar las raíces del polinomio en V

$$f(V) = P\,V^4 - R\,T\,V^3 - \beta\,V^2 - \gamma\,V - \delta = 0, \qquad (2.25)$$

que resulta de multiplicar por V^4 la ecuación 2.24 y pasar todos sus términos a un solo miembro.

La solución de la ecuación 2.25 tiene como primer problema encontrar cuando menos un valor inicial V_0 cercano al volumen buscado V. Este valor V_0, se obtiene a partir de la ley de los gases ideales; así

$$V_0 = \frac{RT}{P},$$

que generalmente es una primera aproximación razonable.

Como puede verse, el razonamiento es sencillo y se basa en el sentido común y las leyes básicas del fenómeno involucrado.

2. Otra manera de conseguir información sobre la función, que permita determinar "buenos" valores iniciales, consiste en obtener su gráfica aproximada mediante un análisis de $f(x)$, a la manera clásica del cálculo diferencial e

integral, o bien como se ha venido sugiriendo, con algún software comercial y, en el mejor de los casos, empleando ambos. A continuación se presentan los pasos del análisis de la función $f(x)$ y de la construcción de su gráfica en la forma clásica.

a) Determinar el dominio de definición de la función.
b) Determinar un subintervalo de (a), que puede ser (a) mismo. Es un intervalo donde se presupone que es de interés analizar la función. Evalúese la función en los siguientes puntos de ese subintervalo: puntos extremos y aquellos donde sea fácil el cálculo de $f(x)$. En los siguientes pasos todo estará referido a este subintervalo.
c) Encontrar los puntos singulares de la función (puntos en los cuales es infinita o no está definida).
d) La primera y la segunda derivadas dan información muy útil sobre la forma de la función, aún más útil que información de valores computados; por ejemplo, dan los intervalos de crecimiento y decrecimiento de la función. Por esto, obténgase la primera derivada y evalúese en puntos apropiados, en particular en puntos cercanos a aquelllos donde la función ya está evaluada y en los que es fácil esta evaluación.
e) Encontrar los puntos máximo y mínimo, así como los valores de la función en esos puntos.
f) Los dominios de concavidad y convexidad de la curva y los puntos de inflexión es información cualitativa y cuantitativa, que se obtiene a partir de la segunda derivada y es imprescindible para este análisis.
g) Obtener las asíntotas de la función. Éstas, en caso de existir, indican cierta regularidad en los comportamientos de la gráfica de $y = f(x)$ al tender x o y hacia infinito.
h) Descomponer la función en sus partes más sencillas que se sumen o se multipliquen. Graficar cada parte y construir la gráfica de la función original, combinando las gráficas de las partes y la información conseguida en los pasos anteriores.

Ejemplo 2.10

Análisis de una función.
A continuación se presenta el análisis clásico de la función

$$f(x) = x - e^{1-x}\ (1 + \ln x)$$

hecho por Pizer.*

Nótese que $\ln x$ está definida sólo para $x > 0$, así que $f(x)$ está definida sólo en $(0, \infty)$.

En este ejemplo ilustrativo, se analiza la función en todo el dominio de definición; es decir, el intervalo de interés será $(0, \infty)$.

*Stephen, M. Pizer. *Numerical Computing and Mathematical Analysis*. S.R.A., (1975) p. 176-179

Un punto donde es fácil evaluar la función es en $x = 1$, ya que la parte exponencial y la parte logarítmica se determinan fácilmente en ese punto.

$$f(1) = 1 - e^{1-1}(1 + \ln 1) = 0$$

De esta forma se ha encontrado una raíz de la ecuación $x_1 = 1$.

En $x = 10$

$$f(10) = 10 - e^{-9}(1 + \ln 10) = 10$$

En x 100

$$f(100) = 100 - e^{-99}(1 + \ln 100) \approx 100$$

Con esta información puede adelantarse que la función tiene la asíntota $y = x$, la función identidad.

Un punto donde la función no está definida es en el extremo $x = 0$. Al analizarlo se advierte que cuando $x \to 0$, el $\ln x \to -\infty$ y $f(x) \to \infty$ y se encuentra una asíntota más de la función, que es la parte positiva del eje y. Por un lado, $x \to \infty$, $\ln x \to \infty$, pero e^{1-x} se acerca más rápidamente a cero y, por tanto, el producto $e^{1-x}(1+\ln x)$ tiende a cero, dejando como resultado global que $f(x) \to \infty$. Se concluye que $f(x) \to \infty$ cuando $x \to 0$, o cuando $x \to \infty$. Como $f(x)$ no tiene otros puntos singulares, se da por terminado el inciso (c).

Al calcular la primera y segunda derivadas de $f(x)$, se tiene que

$$f'(x) = 1 - e^{1-x}(1/x - 1 - \ln x)$$

y

$$f''(x) = e^{1-x}(2/x + 1/x^2 - 1 - \ln x)$$

Al evaluar $f'(x)$ en $x = 1$, se obtiene $f'(1) = 1$.

Al evaluar $f'(x)$ en $x = \infty$, resulta $f'(\infty) = 1$.

Lo que se sabe hasta aquí de la función, se muestra en la figura 2.9(a). Como $f(x)$ es continua (todas las funciones sencillas que la forman lo son) en $(0, \infty)$, deberá haber por lo menos otra raíz de $f(x)$ en (0,1).

El inciso (e) del análisis de la función no procede en este caso, ya que sería tan complejo como encontrar las raíces de $f(x)$. En su lugar se analiza la forma de la curva con la segunda derivada. Evaluando $f''(x)$ en valores muy grandes de x, se tiene que $f''(x) < 0$, o sea que la función es convexa para valores muy grandes de x (también se dice que la curva gira su convexidad hacia la parte positiva del eje y). Además, se tiene $f''(1) = 2$, lo que indica que la función es cóncava en $x = 1$ (o en otras palabras gira su convexidad hacia la parte negativa del eje y). La información se muestra en la figura 2.9(b).

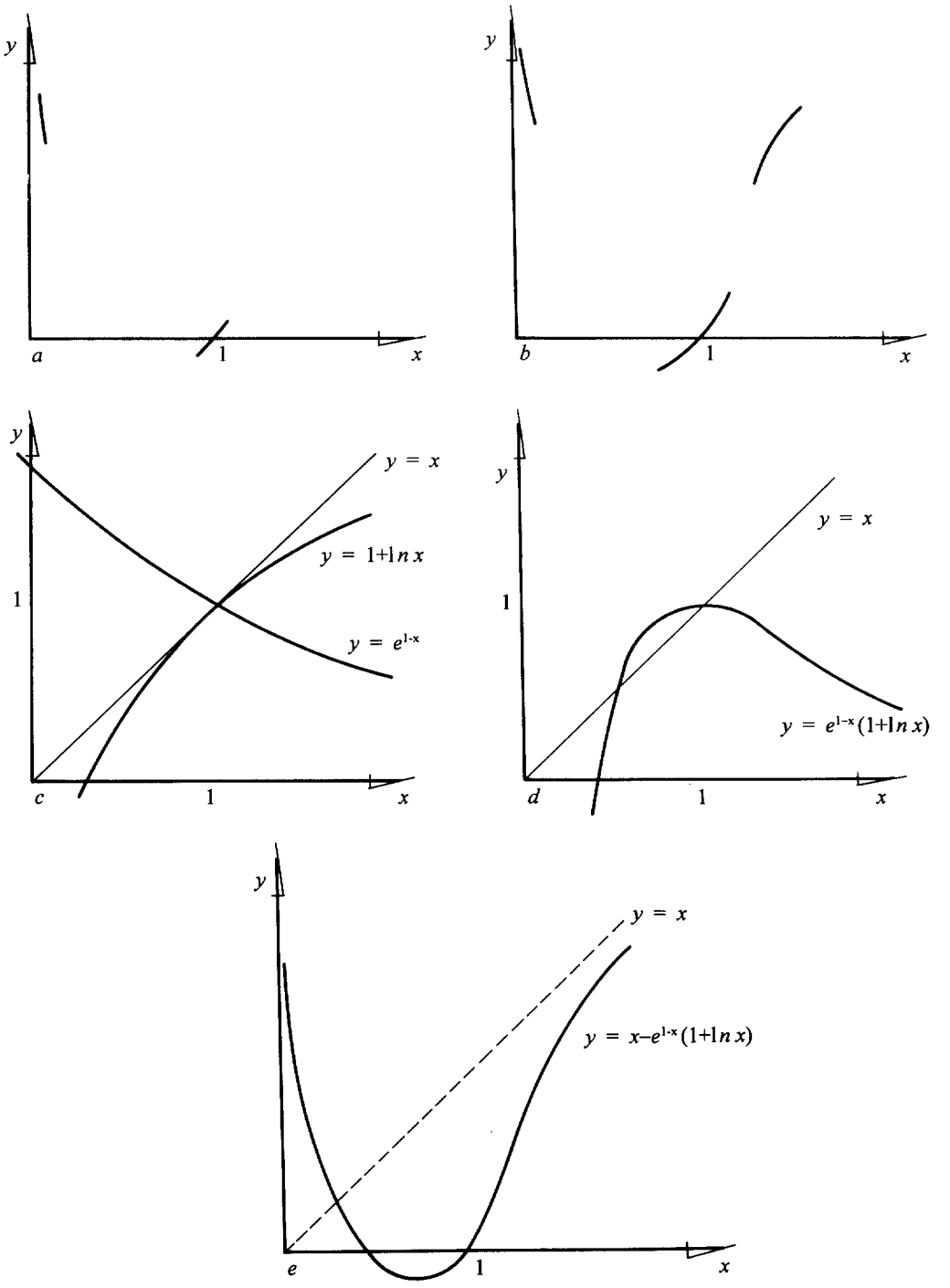


Figura 2.9 Construcción de la gráfica de $f(x) = x - e^{1-x}(1+\ln x)$

Puede obtenerse aún más información de $f(x)$, analizando las funciones elementales que la componen, como x, e^{1-x}, y $1 + \ln x$. La familiaridad con las gráficas de las funciones elementales es útil cuando se consideran funciones más complejas. Las partes en que se puede descomponer $f(x)$ se muestran en la figura 2.9(*c*). Primero nótese que la gráfica de $1 + \ln x$ es la gráfica de $\ln x$ aumentada en una unidad y que la gráfica de e^{1-x} es la gráfica de e^{-x} llevada una unidad a la derecha. Multiplicando e^{1-x} y $1 + \ln x$ entre sí (Fig. 2.9d), se ve que este producto es negativo entre cero y algún valor menor que 1, tiende a cero cuando x aumenta y permanece debajo de $y = x$ para $x > 1$.

Como la derivada del producto es cero en $x = 1$, la curva del producto tiene ahí un máximo y el resto de la gráfica puede obtenerse como se ilustra en la figura 2.9(e).

Nótese que los ceros de $f(x)$ son los puntos donde el producto $e^{1-x}(1 + \ln x)$ y la función identidad $y = x$ se intersecan. Esto significa que sólo hay dos raíces de la función. También puede concluirse que hay una raíz en $x = 1$ y otra cerca de $x = 0.5$, por lo que 0.5 sería un buen valor inicial para calcular esta segunda raíz.

SECCIÓN 2.9 RAÍCES COMPLEJAS

Hasta ahora se han discutido sólo técnicas para encontrar raíces reales de ecuaciones de la forma $f(x) = 0$. Sin embargo, a menudo se presentan ecuaciones polinomiales con coeficientes reales, cuyas raíces son complejas, o bien polinomios complejos, y ecuaciones trascendentes con raíces reales y complejas.

Generalmente, dichas ecuaciones pueden resolverse por el método de Newton-Raphson (Sec. 2.2), pero proponiendo un valor inicial x_0 complejo o bien por algún otro método.

Método de Newton-Raphson

Supóngase que se tiene

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 \qquad (2.26)$$

con todos los coeficiente a_i reales. $f'(x)$ es un polinomio de grado $(n-1)$ y de coeficientes también reales

$$f'(x) = n\,a_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \ldots + 2a_2 x + a_1 \qquad (2.27)$$

Si el valor inicial x_0 es real, entonces

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

también será real y todos los valores x_i siguientes. Consecuentemente no se puede encontrar una raíz compleja de la ecuación 2.26 si se inicia con un valor x_0 real.

Si por el contrario, el valor inicial x_0 es complejo, x_1 entonces será complejo, x_2 también y así sucesivamente. De esta manera, si el proceso converge, puede encontrarse una raíz $\bar{x}$ compleja.

Ejemplo 2.11

Encuentre las raíces complejas de la ecuación

$$f(x) = x^2 + 4 = 0,$$

con el método de Newton-Raphson.

SOLUCIÓN

Al derivar $f(x)$ se tiene

$$f'(x) = 2x$$

Sea $x_0 = j$ el valor inicial propuesto. Aplicando la ecuación 2.12 con este valor inicial, se tiene

$$x_1 = j - \frac{(j^2 + 4)}{2(j)}$$

pero $(j)^2 = -1$, entonces: $x_1 = j - \dfrac{-1 + 4}{2j} = j - \dfrac{3}{2j}$

Multiplicando y dividiendo por j el término $3/(2j)$, se obtiene

$$x_1 = j - (-1.5\, j) = 2.5\, j$$

$$x_2 = 2.5\, j - \frac{(2.5j)^2 + 4}{2(2.5j)} = 2.05\, j$$

$$x_3 = 2.05\, j - \frac{(2.05j)^2 + 4}{2(2.05j)} = 2.001\, j$$

La sucesión de valores complejos $x_0, x_1, \ldots$, va acercándose rápidamente a la raíz $\bar{x}_1 = 2j$

$$f(\bar{x}_1) = f(2\, j) = (2\, j)^2 + 4 = -4 + 4 = 0$$

Para evaluar la distancia entre dos valores complejos consecutivos, se utiliza

$$| x_{i+1} - x_i |,$$

donde las barras representan el módulo del número complejo $x_{i+1} - x_i$. Esto es, si

$$x_{i+1} - x_i = a + b\, j$$

Entonces

$$| x_{i+1} - x_i | = \sqrt{a^2 + b^2}$$

Por lo que se tiene para la sucesión previa

$$| x_1 - x_0 | = | 2.5\,j - j | = \sqrt{0^2 + (1.5)^2} = 1.5$$

$$| x_2 - x_1 | = | 2.05\,j - 2.5\,j | = \sqrt{0^2 + (-0.45)^2} = 0.45$$

$$| x_3 - x_2 | = | 2.001\,j - 2.05\,j | = \sqrt{0^2 + (-0.049)^2} = 0.049$$

y la convergencia es notoria.

Como en general un polinomio con coeficientes reales siempre tiene un número par de raíces complejas, si $x = a + b\,j$ es raíz, también lo será $x = a - b\,j$ (toda vez que al multiplicarlos deben producir los coeficientes reales).

Por esto

$$\bar{x}_2 = -2\,j$$

es la segunda raíz que se busca.

$$f(\bar{x}_2) = f(-2\,j) = (-2\,j)^2 + 4 = -4 + 4 = 0$$

El problema queda terminado.

Si bien se resolvió una ecuación cuadrática que no representa dificultad, el método igualmente puede emplearse para un polinomio de mayor grado, siguiendo los mismos pasos. El lector puede hacer un programa para el algoritmo en algún lenguaje de alto nivel o en un pizarrón electrónico como Math-CAD.

Método de Müller

Un método deducido por Müller*, se ha puesto en práctica en las computadoras con éxito sorprendente. Se puede usar para encontrar cualquier tipo de raíz, real o compleja, de una función arbitraria. Converge casi cuadráticamente en un intervalo cercano a la raíz y, a diferencia del método de Newton-Raphson, no requiere la evaluación de la primera derivada de la función y obtiene raíces reales y complejas aun cuando estas raíces sean repetidas.

*Müller, D.E. "A Method of Solving algebraic Equations Using an Automatic Computer". *Mathematical Tables and Other Aids to Computation* (MTAC), 10. p 208-215 (1956).

La aplicación del método requiere valores iniciales y es una extensión del método de la secante, el cual aproxima la gráfica de la función $f(x)$ por una línea recta que pasa por los puntos $(x_{i-1}, f(x_{i-1}))$ y $(x_i, f(x_i))$. El punto de intersección de esta línea con el eje x da la nueva aproximación x_{i+1}.

En lugar de aproximar $f(x)$ por una función lineal (línea recta o polinomio de grado 1), resulta natural tratar de obtener una convergencia más rápida aproximando $f(x)$ por un polinomio $p(x)$ de grado $n > 1$ que coincida con $f(x)$ en los puntos de abscisas $x_i, x_{i-1}, \ldots, x_{i-n}$ y determinar x_{i+1} como una de las raíces de $p(x)$.

A continuación se describe el caso $n = 2$, en que el estudio detallado de Müller encontró que la eleción de n da resultados satisfactorios.

Se toman tres valores iniciales x_0, x_1, x_2 y se halla el polinomio $p(x)$ de segundo grado que pasa por los puntos $(x_0, f(x_0))$, $(x_1, f(x_1))$ y $(x_2, f(x_2))$ y se toma una de las raíces de $p(x)$, la más cercana a x_2, como la siguiente aproximación x_3. Se repite la operación con los nuevos valores iniciales x_1, x_2, x_3 y se termina el proceso tan pronto como se satisfaga algún criterio de convergencia. La figura 2.10 ilustra este método.

Sean x_i, x_{i-1}, x_{i-2} tres aproximaciones distintas a una raíz de $f(x) = 0$. Usando la siguiente notación.

$$f_i = f(x_i)$$

$$f_{i-1} = f(x_{i-1})$$

$$f_{i-2} = f(x_{i-2})$$

en el capítulo 5, se demostrará que con

$$f[x_i, x_{i-1}] = \frac{f_i - f_{i-1}}{x_i - x_{i-1}}$$

$$f[x_{i-1}, x_{i-2}] = \frac{f_{i-1} - f_{i-2}}{x_{i-1} - x_{i-2}} \tag{2.28}$$

$$f[x_i, x_{i-1}, x_{i-2}] = \frac{f[x_i, x_{i-1}] - f[x_{i-1}, x_{i-2}]}{x_i - x_{i-2}} \tag{2.29}$$

la función

$$p(x) = f_i + f[x_i, x_{i-1}](x - x_i)$$

$$+ f[x_i, x_{i-1}, x_{i-2}](x - x_i)(x - x_{i-1}) \tag{2.30}$$

es la parábola única que pasa por los puntos (x_i, f_i), (x_{i-1}, f_{i-1}) y (x_{i-2}, f_{i-2}). El lector recordará que la manera usual de escribir un polinomio de segundo grado o parábola es

$$p(x) = a_0 + a_1 x + a_2 x^2$$

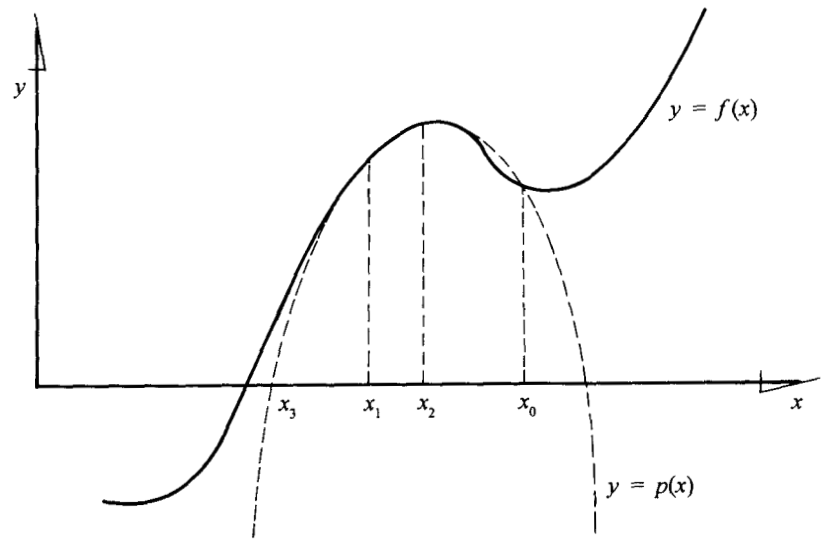


Figura 2.10 Interpretación gráfica del método de Müller.

Al comparar esta última expresión con la ecuación 2.30 se establece la siguiente identificación

$$a_2 = f\,[x_i,\ x_{i-1},\ x_{i-2}]$$

$$a_1 = f\,[x_i,\ x_{i-1}] - (x_i + x_{i-1})a_2$$

$$a_0 = f_i - x_i\,(f\,[x_i,\ x_{i-1}] - x_{i-1}\ a_2)$$

Una vez calculados los valores de a_0, a_1 y a_2, las raíces de $p\,(x)$ se determinan a partir de la fórmula cuadrática

$$x_{i+1} = \frac{2\,a_0}{-a_1 \pm (a_1^2 - 4\,a_0\,a_2)^{1/2}} \tag{2.31}$$

cuya explicación se encuentra en el problema 2.31, y en el ejercicio 1.3 del capítulo 1.

Se selecciona el signo que precede al radical de manera que el denominador sea máximo en magnitud*, y la raíz correspondiente es la siguiente aproximación x_{i+1}. La razón para escribir la fórmula cuadrática de esta manera es obtener mayor exactitud (véase Probl. 2.31), ya disminuida por las diferencias de las ecuaciones 2.28 y 2.29, que se utilizan en el cálculo de a_0, a_1 y a_2 y que son aproximaciones a las derivadas de la función $f\,(x)$.

*Con esto se encuentra el valor más cercano a x_i.

Puede ocurrir que la raíz cuadrada en la ecuación 2.31 sea compleja. Si $f(x)$ no está definida para valores complejos, el algoritmo deberá reiniciarse con nuevos valores iniciales. Si $f(x)$ es un polinomio, la posibilidad de raíces complejas es latente y el valor de x puede considerarse como aproximación a alguna de estas raíces y, por tanto, deberá emplearse en la siguiente iteración.

Ejemplo 2.12

Encuentre una raíz real de la ecuación polinomial

$$f(x) = x^3 + 2x^2 + 10x - 20 = 0,$$

con el método de Müller.

SOLUCIÓN

Primera iteración

Al seleccionar como valores iniciales a

$$x_0 = 0; \qquad x_1 = 1; \qquad x_2 = 2$$

y evaluar la función $f(x)$ en estos puntos, se tiene

$$f_0 = -20; \qquad f_1 = -7; \qquad f_2 = 16$$

Se calculan ahora los coeficientes del polinomio de segundo grado

$$f[x_1, x_0] = \frac{f_1 - f_0}{x_1 - x_0} = \frac{-7 + 20}{1 - 0} = 13$$

$$f[x_2, x_1] = \frac{f_2 - f_1}{x_2 - x_1} = \frac{16 + 7}{2 - 1} = 23$$

$$f[x_2, x_1, x_0] = \frac{f[x_2, x_1] - f[x_1, x_0]}{x_2 - x_0} = \frac{23 - 13}{2 - 0} = 5$$

Por lo tanto

$$a_2 = f[x_2, x_1, x_0] = 5$$

$$a_1 = f[x_2, x_1] - (x_2 + x_1)\, a_2 = 23 - (2 + 1)5 = 8$$

$$a_0 = f_2 - x_2 (f[x_2, x_1] - x_1 a_2) = 16 - 2(23 - 1(5)) = -20$$

Se calculan los denominadores de la ecuación 2.31

$$-a_1 + (a_1^2 - 4a_0a_2)^{1/2} = -8 + (64 + 400)^{1/2} = 13.54066$$

$$-a_1 - (a_1^2 - 4a_0a_2)^{1/2} = -8 - (64 + 400)^{1/2} = -29.54066$$

Como el segundo es mayor en valor absoluto, se usa en la ecuación 2.31, de donde

$$x_3 = \frac{2a_0}{-a_1 - (a_1^2 - 4a_0a_2)^{1/2}} = \frac{2(-20)}{-29.54} = 1.35407$$

Segunda iteración

Recorriendo ahora los subíndices de x, se tiene

$$x_0 = 1; \qquad x_1 = 2; \qquad x_2 = 1.35407$$

$$f_0 = -7; \qquad f_1 = 16; \qquad f_2 = -0.30968$$

En consecuencia

$$f[x_1, x_0] = \frac{16 + 7}{2 - 1} = 23$$

$$f[x_2, x_1] = \frac{-0.30968 - 16}{1.35407 - 2} = 25.24999$$

$$f[x_2, x_1, x_0] = \frac{25.24999 - 23}{1.35407 - 1} = 6.35407$$

De donde

$$a_2 = f[x_2, x_1, x_0] = 6.35407$$

$$a_1 = f[x_2, x_1] - (x_2 + x_1)\, a_2 =$$

$$25.24999 - (1.35407 + 2)\, 6.35407 = 3.9378$$

$$a_0 = f_2 - x_2\,(f[x_2 - x_1] - x_1a_2) =$$

$$-0.30968 - 1.35407\,(25.24999 - 2\,(6.35407)) = -17.29187$$

Calculando los denominadores de la ecuación 2.31

$$-a_1 + (a_1^2 - 4a_0a_2)^{1/2} = 17.39295$$

$$-a_1 - (a_1^2 - 4a_0a_2)^{1/2} = -25.26855$$

Como el segundo es mayor en valor absoluto, se usa en la ecuación 2.31, de donde

$$x_3 = \frac{2a_0}{-a_1 - (a_1^2 - 4a_0a_2)^{1/2}} = 1.36865$$

La tabla 2.5 se obtiene repitiendo el procedimiento.

i	x_i	$\lvert x_{i+1} - x_i \rvert$
0	0	
1	1	1.00000
2	2	1.00000
3	1.35407	0.64593
4	1.36865	0.01458
5	1.36881	0.00016

Tabla 2.5

Ejemplo 2.13

Encuentre las raíces complejas de la ecuación polinomial del ejemplo 2.11

$$f(x) = x^2 + 4 = 0,$$

con el método de Müller.

SOLUCIÓN

Primera iteración

Al elegir como valores iniciales

$$x_0 = 0; \qquad x_1 = 1; \qquad x_2 = -1$$

y evaluar la función en estos puntos, se tiene

$$f_0 = 4; \qquad f_1 = 5; \qquad f_2 = 5.$$

Se calculan ahora los coeficientes del polinomio de segundo grado

$$f[x_1, x_0] = \frac{f_1 - f_0}{x_1 - x_0} = \frac{5 - 4}{1 - 0} = 1$$

$$f[x_2, x_1] = \frac{f_2 - f_1}{x_2 - x_1} = \frac{5 - 5}{-1 - 1} = 0$$

$$f[x_2, x_1, x_0] = \frac{f[x_2, x_1] - f[x_1, x_0]}{x_2 - x_0} = \frac{0 - 1}{-1 - 0} = 1$$

Por lo tanto

$$a_2 = f[x_2, x_1, x_0] = 1$$

$$a_1 = f[x_2, x_1] - (x_2 + x_1)a_2 = 0 - (-1 + 1)(1) = 0$$

$$a_0 = f_2 - x_2(f[x_2, x_1] - x_1 a_2) = 5 - (-1)(0 - 1(1)) = 4$$

Calculando los denominadores de la ecuación 2.31

$$-a_1 + (a_1^2 - 4a_0a_2)^{1/2} = 0 + (0 - 4(4)(1))^{1/2} = (-16)^{1/2} = 4j$$

$$-a_1 - (a_1^2 - 4a_0a_2)^{1/2} = 0 - (0 - 4(4)(1))^{1/2} = -(-16)^{1/2} = -4j$$

Como son de igual magnitud se usa cualquiera, por ejemplo $4j$. Entonces

$$x_3 = \frac{2a_0}{-a_1 + (a_1^2 - 4a_0a_2)^{1/2}} = \frac{2(4)}{4j} = \frac{2}{j}$$

al multiplicar numerador y denominador por j, queda

$$x_3 = \frac{2}{j}\,\frac{j}{j} = \frac{2j}{-1} = -2j$$

Nótese que aun cuando x_0, x_1 y x_2 son números reales, x_3 ha resultado número complejo y además es la raíz buscada, lo cual resulta lógico, ya que la ecuación polinomial

$$f(x) = x^2 + 4 = 0$$

es una parábola y el método de Müller consiste, en el caso $n = 2$, en usar una parábola para sustituir la función.

La otra raíz es el complejo conjugado de x_3, o sea $2j$.

A continuación se da el algoritmo del método de Müller para el caso $n = 2$.

ALGORITMO 2.6 Método de Müller

Para encontrar una raíz real o compleja de la ecuación $f(x) = 0$, proporcionar la función F (X) y los

DATOS: Valores iniciales X0, X1, X2; criterio de convergencia EPS, criterio de exactitud EPS1 y número máximo de iteraciones MAXIT.

RESULTADOS: La raíz aproximada X o un mensaje de falla.

PASO 1. Hacer I = 1

PASO 2. Mientras I < MAXIT, repetir los pasos 3 a 7.

PASO 3. Hacer F10 = (F(X1)–F(X0)) / (X1–X0)
F21 = (F(X2)–F(X1)) / (X2–X1)
F210 = (F21–F10) / (X2–X0).
A2 = F210
A1 = F21–(X2+X1)*A2
A0 = F (X2)–X2*(F21–X1*A2)
D1 = –A1+(A1**2–4*A0*A2) **0.5
D2 = –A1–(A1**2–4*A0*A2)**0.5

PASO 4. Si ABS (X3–X0) > ABS (D2) hacer X3 = 2*A0/D1 En caso contrario hacer X3 = 2*A0/D2

PASO 5. Si ABS (X3–X0) < EPS o ABS (F (X3)) < EPS1 IMPRIMIR X3 y TERMINAR.
De otro modo, continuar.

PASO 6. Hacer X0 =X1
X1 = X2 (actualización de valores iniciales).
X2 = X3

PASO 7. Hacer I = I + 1

PASO 8. IMPRIMIR mensaje de falla: "EL MÉTODO NO CONVERGE A UNA RAÍZ" y TERMINAR.

La siguiente sección puede omitirse sin pérdida de continuidad en el resto del material.

SECCIÓN 2.10 POLINOMIOS Y SUS ECUACIONES

Evaluación de polinomios

Método de Horner

Se desea evaluar un polinomio $p(x)$ en un valor particular de x. Por ejemplo, sea el polinomio

$$p(x) = 4x^4 + 3x^3 - 2x^2 + 4x - 8, \tag{2.32}$$

que se desea evaluar en $x = 2$.

Factorícese x en los primeros cuatro términos

$$p\ (x) = (4\ x^3 + 3\ x^2 - 2\ x + 4)\ x - 8$$

Dentro de los paréntesis, factorizar x en los primeros tres términos

$$p\ (x) = ((4\ x^2 + 3\ x - 2)\ x + 4)\ x - 8$$

Dentro de los paréntesis interiores, factorícese x en los primeros dos términos

$$p\ (x) = (((4\ x + 3)\ x - 2)\ x + 4)\ x - 8$$

El método de Horner consiste en evaluar, secuencialmente, los paréntesis en esta expresión

Paso 1. Evaluar $(4\ x + 3)$ en $x = 2$: $4(2) + 3 = 11$
Paso 2. Evaluar $((11)\ x - 2)$ en $x = 2$: $(11)\ 2 - 2 = 20$
Paso 3. Evaluar $((20)\ x + 4)$ en $x = 2$: $(20)\ 2 + 4 = 44$
Paso 4. Evaluar $(44)\ x - 8$ en $x = 2$: $(44)\ 2 - 8 = 80$

De esto, $p\ (2) = 80$.

Este proceso puede llevarse a cabo sin las factorizaciones.

Escríbase $p_4\ (x) = a_4\ x^4 + a_3\ x^3 + a_2\ x^2 + a_1\ x + a_0$

Para la ecuación 2.32, $a_4 = 4$, $a_3 = 3$, $a_2 = -2$, $a_1 = 4$ y $a_0 = -8$

Conviene almacenar los valores intermedios de la evaluación de esta ecuación 11, 20, 44 y 80, como b_3, b_2, b_1 y b_0, respectivamente. Sea además, por conveniencia, $b_4 = a_4 (= 4)$.

Ahora dispónganse los coeficientes, el valor de x donde se desea evaluar el polinomio y b_4 en la siguiente forma

	a_4	a_3	a_2	a_1	a_0
$x = 2$	4	3	−2	4	−8
	$b_4 = 4$				

En la columna de a_3, se desarrolla el paso $1{:}4(2) + 3 = 11$. Esto puede verse como multiplicar b_4 por el valor de x $(= 2)$ y sumar el producto a a_3. Llámese este resultado b_3. Esto es

	a_4	a_3	a_2	a_1	a_0
	4	3	−2	4	−8
$x = 2$		+			
		$4\ (2) = 8$			
	$b_4 = 4$	$b_3 = 11$			

En la columna de a_2, se desarrolla el paso 2:(11)2 − 2 = 20. Esto es, multiplíquese b_3 por el valor de x (= 2) y súmese el producto a a_2. Llámese este resultado b_2. Lo anterior se ilustra así

	a_4	a_3	a_2	a_1	a_0
$x = 2$	4	3	−2	4	−8
			+		
			11(2) = 22		
	$b_4 = 4$	$b_3 = 11$	$b_2 = 20$		

Repitiendo este proceso hasta calcular b_0 se tiene

	a_4	a_3	a_2	a_1	a_0
$x = 2$	4	3	−2	4	−8
				+	+
				20(2) = 40	44(2) = 88
	$b_4 = 4$	$b_3 = 11$	$b_2 = 20$	$b_1 = 44$	$b_0 = 80$

El valor $p(2)$ resulta en b_0.

Ejemplo 2.14

Evalúe el polinomio

$$x^5 - 4x^3 + 2x + 3 \text{ en } x = 3,$$

mediante el método de Horner.

SOLUCIÓN

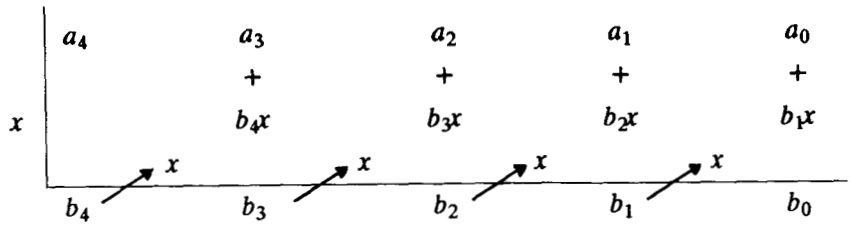
La no aparición de los términos en x^4 y en x^3 del polinomio significa que sus coeficientes son cero; para fines del método en estudio, dichos ceros deben aparecer en el arreglo

Coeficientes de:

	x^5	x^4	x^3	x^2	x	Término independiente
	$a_5=1$	$a_4=0$	$a_3=-4$	$a_2=0$	$a_1=2$	$a_0=3$
$x = 3$		+	+	+	+	+
		1(3)=3	3(3)=9	5(3)=15	15(3)=45	47(3)=141
	$b_5=1$	$b_4=3$	$b_3=5$	$b_2=15$	$b_1=47$	$b_0=144$

De aquí $p(3) = 144$.

Al generalizar este método con polinomios de cuarto grado, la extensión a cualquier grado, es inmediata



donde puede verse que

$$b_4 = a_4,$$

$$b_3 = a_3 + b_4x,\ b_2 = a_2 + b_3x,\ b_1 = a_1 + b_2x,\ b_0 = a_0 + b_1x;$$

esto es

$$\boxed{b_4 = a_4 \text{ y } b_k = a_k + b_{k+1}x, \text{ para } k = 3, 2, 1, 0} \qquad (2.33)$$

Mediante una sustitución regresiva puede verse con claridad por qué $p(x) = b_0$:

Sustituyendo en $b_0 = a_0 + b_1 x$ a b_1 por $a_1 + b_2 x$, se tiene

$$b_0 = a_0 + (a_1 + b_2 x)x$$

y ahora se reemplaza en la última expresión b_2 con $a_2 + b_3x$ y así sucesivamente, con lo cual se obtiene

$$b_0 = (((a_4 x + a_3)x + a_2)x + a_1)x + a_0 = p(x)$$

Las ecuaciones 2.33 representan un algoritmo programable y, como se verá más adelante, de elevada eficiencia para evaluar un polinomio $p(x)$ en algún valor particular de x.

Se describe enseguida el algoritmo de Horner.

ALGORITMO 2.7 Método de Horner

Para evaluar el polinomio

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$$

proporcionar los

DATOS: n: Grado del polinomio
$a_n, a_{n-1}, \ldots, a_0$: Coeficientes del polinomio.
t: Valor de x en donde se desee evaluar $p(x)$

RESULTADO: $p\,(t)$ en b_0
PASO 1. Hacer $b_n = a_n$
PASO 2. Para k = n–1, n–2, ..., 0 realizar el paso 3.
PASO 3. Hacer $b_k = b_{k+1}\,t + a_k$
PASO 4. IMPRIMIR b_0

Método de Horner iterado

El método de Horner tiene otras características, que se verán enseguida.

Tómese de nuevo el polinomio general de cuarto grado $p_4\,(x)$ y divídase entre $(x–t)$, donde t es un valor particular de x, lo que se expresa como

$$p\,(x) = (x–t)\,q\,(x) + R, \tag{2.34}$$

donde $q\,(x)$ es el polinomio cociente (en este caso de tercer grado) y R una constante llamada residuo.

Sustituyendo x con t se obtiene $p\,(t) = R$, de modo que el polinomio evaluado en un valor particular de x es igual al residuo R de la división, $R = b_0$.

Al derivar la ecuación 2.34 con respecto a x (recuérdese que t y R son constantes), se tiene

$$p'\,(x) = (x–t)\,q'\,(x) + q\,(x)$$

Haciendo $x = t$ resulta

$$p'\,(t) = q\,(t), \tag{2.35}$$

esto es, la derivada del polinomio $p\,(x)$ evaluada en $x = t$ es el cociente $q\,(x)$ evaluado en t, toda vez que

$$p'\,(t) = q\,(t) = b_4t^3 + b_3t^2 + b_2t + b_1$$

y en general

$$q\,(x) = b_4x^3 + b_3x^2 + b_2x + b_1 \tag{2.36}$$

donde b_4, b_3, b_2 y b_1 son los valores intermedios que resultan en la evaluación de $p\,(x)$ en t por el método de Horner (véase Ej. 2.14). Así pues, si después de evaluar $p\,(x)$ en t se desea evaluar también $p'\,(x)$ en t, puede aplicarse una vez más el método de Horner a los valores intermedios b_4, b_3, b_2 y b_1, como se ilustra enseguida.

Ejemplo 2.15

Sea $p\,(x) = 3x^3 - 4x - 1$. Evalúe

a) $p\,(2)$ b) $p'\,(2)$

SOLUCIÓN

a) Para evaluar p (2), se tiene

	a_3	a_2	a_1	a_0
	3	0	− 4	−1
$x = 2$		+	+	+
		3(2)=6	6(2)=12	8(2)=16
	$b_3 = 3$	$b_2 = 6$	$b_1 = 8$	$b_0 = 15$

y p (2) = 15.

b) Como se dijo

$$p'(t) = b_3t^2 + b_2t + b_1$$

Para evaluar p' (2) se emplea de nuevo el método de Horner. Esto se logra eficientemente, repitiendo los pasos de los cálculos descritos; esto es bajo b_3, b_2 y b_1 del arreglo anterior. Para almacenar los nuevos valores intermedios de esta evaluación se emplean c_3, c_2 y c_1. Nótese que como b_1 es el término independiente de p' (x), el proceso de evaluación termina una vez que se obtuvo c_1, y éste es el valor buscado de p' (2).

	a_3	a_2	a_1	a_0
	3	0	−4	−1
$x = 2$		+	+	+
		3 (2) = 6	6 (2) = 12	8 (2) = 16
$x = 2$	$b_3 = 3$	$b_2 = 6$	$b_1 = 8$	$b_0 = 15$
		+	+	
		3 (2) = 6	12 (2) = 24	
	$c_3 = 3$	$c_2 = 12$	$c_1 = 32$	

De esto, p' (2) = 32. El lector puede verificar el resultado derivando p (x) y evaluando la derivada en $x = 2$.

En la práctica, los cálculos suelen disponerse sin tantos comentarios.

Ejemplo 2.16

Evalúe $5x^3 - 2x^2 + 10$ y su primera derivada en $x = 0.5$

SOLUCIÓN

	a_3	a_2	a_1	a_0
	5	−2	0	10
		+	+	+
0.5		2.5	0.25	0.125
	b_3	b_2	b_1	b_0
	5	0.5	0.25	10.125
		+	+	
0.5		2.5	1.50	
	c_3	c_2	c_1	
	5	3	1.75	

De esto $p\ (0.5) = 10.125$ y $p'\ (0.5) = 1.75$.

En este punto conviene presentar el algoritmo de Horner iterado para evaluar un polinomio y su primera derivada en un valor t.

ALGORITMO 2.8 Método de Horner iterado

Para evaluar el polinomio

$$p\ (x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$$

y su primera derivada $p'\ (x)$ en $x = t$, proporcionar los

DATOS: n: Grado del polinomio,
$a_n, a_{n-1}, \ldots, a_0$: Coeficientes del polinomio.
t: Valor de x en donde se desea evaluar $p\ (x)$ y $p'\ (x)$.

RESULTADOS: $p\ (t)$ en b_0 y $p'\ (t)$ en c_1.

PASO 1. Hacer $b_n = a_n$ y $c_n = b_n$

PASO 2. Para $k = n-1, n-2, \ldots, 1$ realizar los pasos 3 y 4.

PASO 3. Hacer $b_k = b_{k+1}\, t + a_k$

PASO 4. Hacer $c_k = c_{k+1}\, t + b_k$

PASO 5. Hacer $b_0 = b_1 t + a_0$

PASO 6. IMPRIMIR b_0 y c_1

Cuenta de operaciones.

Si bien la implementación del método de Horner en una computadora es una de sus ventajas, no lo es menos su eficiencia, que se verá a continuación, contando las operaciones en el método de evaluación usual y comparando su número con el del método de Horner.

Tomando de nuevo el polinomio general de cuarto grado

$$p_4(x) = a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$$

a) **Método usual**

a_4x^4 requiere cuatro multiplicaciones

a_3x^3 requiere tres multiplicaciones

a_2x^2 requiere dos multiplicaciones

a_1x requiere una multiplicación

$a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$ necesitas cuatro sumas/restas.

En total se realizan 10 multiplicaciones y cuatro sumas/restas.

b) Método de Horner

$$\left.\begin{aligned} b_4 &= a_4 \\ b_3 &= b_4x + a_3 \\ b_2 &= b_3x + a_2 \\ b_1 &= b_2x + a_1 \\ b_0 &= b_1x + a_0 \end{aligned}\right\}$$ Se requiere una multiplicación y una suma para cada b.

En total cuatro multiplicaciones y cuatro sumas/restas.

Hay una reducción del 60% en el número de multiplicaciones requeridas y, consecuentemente un error de redondeo menor.

A continuación se verá una aplicación del método de Horner en la búsqueda de raíces reales de ecuaciones de la forma $f(x) = 0$, donde $f(x)$ es un polinomio de grado n.

Combinando las ecuaciones 2.34 y 2.36 y el resultado $R = b_0$.

$$f(x) = (x-t)(b_4x^3 + b_3x^2 + b_2x + b_1) + b_0$$

y como $f(t) = b_0$,

$$f(x) = (x-t)(b_4x^3 + b_3x^2 + b_2x + b_1) + f(t)$$

Si t es una raíz de $f(x) = 0$, se tiene $f(t) = 0$ y la expresión resultante

$$f(x) = (x-t)(b_4x^3 + b_3x^2 + b_2x + b_1)$$

indica que $x = t$ es una raíz (lo cual ya se sabía), pero lo más importante es que las raíces restantes de $f(x) = 0$ son las raíces de

$$b_4x^3 + b_3x^2 + b_2x + b_1 = 0, \tag{2.37}$$

una ecuación polinomial de tercer grado y, por tanto, más fácil de manejar que la ecuación original; además, sus coeficientes son los valores ya citados b_4, b_3, b_2 y b_1.

Si se sospecha que la raíz t se repite (es decir t es raíz de la ecuación 2.37), véase el valor de c_1 del método de Horner iterado ya que éste será muy cercano a cero si así fuera; esto es $p'(t) = 0$ en ese caso.

Ahora, desarróllese el método de Newton-Raphson con el método de Horner iterado.

Raíces de una ecuación polinomial $P_n(x) = 0$

Método de Newton-Raphson-Horner

De los métodos vistos para encontrar raíces, el de Newton-Raphson resulta el más adecuado para usarse en conjunción con el método de Horner iterado. Se resuelve a continuación un ejemplo con esta combinación.

Ejemplo 2.17

Aproxime las raíces reales del polinomio

$$p(x) = 4x^4 + 3x^3 - 2x^2 + 4x - 8$$

SOLUCIÓN

PASO 1. Al analizar gráficamente la función se advierte que tiene dos raíces reales, una alrededor de 1, y la otra alrededor de –2.

PASO 2. Se elige 0 como valor inicial para encontrar la primera raíz.

PASO 3. Con el método de Horner $b_0 = -8$ y $c_1 = 4$

PASO 4. Con el método de Newton-Raphson $t_1 = t_0 - b_0/c_1 = 0 - (-8)/4 = 2$

PASO 5. Al repetir los pasos 3 y 4 se tiene

$t_2 = t_1 - b_0/c_1 = 2 - 8/16 = 1.5$

$t_3 = t_2 - b_0/c_1 = 1.5 - 23.875/72.25 = 1.1696$

$t_4 = t_3 - b_0/c_1 = 1.1696 - 6.2258/37.2287 = 1.0023$

Este proceso converge al valor 0.9579

PASO 6. Se toma 0.9579 como primera raíz del polinomio.

PASO 7. El polinomio de menor grado que se obtiene con esta raíz conduce a $p(x) = 4x^3 + 6.831315x^2 + 4.5435x + 8.3518$

PASO 2. Se elige nuevamente 0 como valor inicial para encontrar la segunda raíz.

PASO 3. Con el método de Horner $b_0 = 8.3518$ y $c_1 = 4.5435$

PASO 4. Con el método de Newton-Raphson
$t_1 = t_0 - b_0/c_1 = 0 - 8.3518/4.5435 = -1.8382$

PASO 5. Al repetir los pasos 3 y 4 se tiene
$t_2 = t_1 - b_0/c_1 = -1.8382 - (-1.7623) / 19.9772 = -1.7500$
$t_3 = t_2 - b_0/c_1 = -1.7500 - (-1.1575) / 17.3841 = -1.7434$

Este proceso converge al valor -1.7433

PASO 6. Se toma -1.7433 como segunda raíz del polinomio.

PASO 7. El polinomio disminuido con esta raíz conduce a
$p(x) = 4x^2 - 0.141885x + 4.79085$

Obsérvese que tiene dos raíces imaginarias.

En el disco se encuentra el programa 2.1 para este algoritmo.

En cada etapa se ha calculado una aproximación a cada una de las raíces reales de $p(x) = 0$; conforme se avanza en las etapas, los coeficientes $b_1, b_2, \ldots, b_n$ de cada etapa se alejan de los valores verdaderos, debido a la propagación de errores, y las aproximaciones a las raíces correspondientes también son más inexactas. Para disminuir la pérdida de exactitud se ha sugerido trabajar primero con la raíz más pequeña en valor absoluto, luego con la raíz real restante más pequeña en magnitud y así sucesivamente.

Método de Lin.

En 1941 S.N. Lin publicó un procedimiento que se fundamenta en el resultado

$$R = f(t) = b_0 = b_1 t + a_0$$

y en que si t es una raíz de $p_n(x) = 0$, entonces

$$R = 0 = b_1 t + a_0$$

o

$$t = -a_0 / b_1 \quad (t) \tag{2.38}$$

Se ha escrito $b_1(t)$ en lugar de b_1 para hacer énfasis en que el valor de b_1 (y de las demas b) depende del valor t donde se evalúa $f(x)$ y así ver el lado derecho de la ecuación 2.38 como una función de t. Lo que puede escribirse como

$$t = -a_0 / b_1(t) = g(t) \tag{2.39}$$

y se le puede aplicar el método de punto fijo, empezando con un valor inicial t_0 cercano a la raíz t, de modo que

$$t_1 = -a_0 / b_1(t_0) = g(t_0)$$

Restando en ambos lados t_0

$$t_1 - t_0 = -\frac{a_0 + t_0 b_1(t_0)}{b_1(t_0)} = -\frac{a_0 + t_0 b_1}{b_1}$$

o

$$\boxed{t_1 = t_0 - \frac{R(t_0)}{b_1(t_0)}} \tag{2.40}$$

y se obtiene el algoritmo de Lin. Este método no requiere el cálculo de las c como en el de Newton-Raphson-Horner, por lo que el trabajo por iteración se reduce a la mitad. Esta reducción contrasta con un orden bajo de convergencia y la inestabilidad propia del método de punto fijo.

Ejemplo 2.18

Encuentre una raíz real de la ecuación

$$x^4 - 3x^3 + 2x - 1 = 0,$$

con el método de Lin y un valor inicial $t_0 = 2.8$

SOLUCIÓN

Primera iteración

$$R(2.8) = 0.2096; \qquad b_1(2.8) = 0.432;$$

$$t_1 = t_0 - R(t_0) / b_1(t_0) = 2.8 - 0.2096 / 0.432 = 2.3148$$

Segunda iteración

$$R(2{,}3148) = -4.8692; \; b_1(2.3148) = -1.6715;$$

$$t_2 = t_1 - R(t_1) / b_1(t_1) = 2.3148 - (-4.8692) / (-1.6715) = -0.5983$$

Al continuar las iteraciones se advierte que el método es inestable y no llega a la raíz 2.78897.

La estabilidad* del método puede mejorarse en una raíz $\bar{x}_k$, si se conoce una buena aproximación a $\bar{x}_k$. Para esto se incorpora el parámetro λ a la ecuación 2.40 de Lin y queda

$$t_1 = t_0 - \lambda \frac{R}{b_1}$$

donde

$$\lambda = - \frac{f(0)}{t_0 f'(t_0)}$$

Con $t_0 = 2.8$, $\lambda = 0.018555$ y la fórmula modificada de Lin, queda en general

$$\boxed{t = t - \lambda \frac{R}{b_1}} \qquad (2.41)$$

Ejemplo 2.19

Con la fórmula modificada de Lin, aproxime una raíz real de la ecuación

$$f(x) = x^4 - 3x^3 + 2x - 1 = 0,$$

use como valor inicial $t_0 = 2.8$

SOLUCIÓN

$$f(0) = -1; \quad f'(2.8) = 19.248$$

$$\lambda = - (-1) / 2.8 / 19.248 = 0.018555$$

Primera iteración

$$R(2.8) = 0.2096; \qquad b_1(2.8) = 0.432;$$

$$t_1 = t_0 - \lambda R(t_0) / b_1(t_0) = 2.8 - 0.018555 (0.2096) / 0.432$$

$$= 2.791$$

Segunda iteración

$$R(2.791) = 0.03808; \qquad b_1(2.791) = 0.37194;$$

$$t_2 = t_1 - \lambda R(t_1) / b_1(t_1)$$

$$= 2.791 - 0.018555 (0.03808) / (0.37194) = 2.7891$$

Al continuar las iteraciones se encuentra la raíz 2.78897

Los métodos anteriores son válidos para raíces reales y complejas. Sin embargo, para las segundas deberá inicializarse con un número complejo y llevar a cabo las

*Hildebrand. Introduction to Numerical Analysis. McGraw Hill, Second Edition. p. 591-595.

operaciones complejas correspondientes. Cuando los coeficientes de $p_n(x) = 0$ son reales, las raíces complejas son conjugadas

$$\bar{x}_k = a + b\ \mathrm{i},\ \bar{x}_{k+1} = a - b\ \mathrm{i},$$

lo que se puede aprovechar buscando en $p_n(x) = 0$ el factor cuadrático

$$(x - \bar{x}_k)(x - \bar{x}_{k+1}) = x^2 - 2ax + (a^2 + b^2)$$

de coeficientes reales que genera $\bar{x}_k$ y $\bar{x}_{k+1}$.

Factores cuadráticos. Método de Lin

Sea el polinomio

$$f(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_2x^2 + a_1x + a_0 \tag{2.42}$$

Si a_n no es uno, $f(x)$ puede dividirse entre a_n para obtener la ecuación 2.42.

Al dividir la ecuación 2.42 entre la expresión cuadrática

$$x^2 + px + q \tag{2.43}$$

$$f(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_2x^2 + a_1x + a_0$$

$$= (x^2 + px + q)(x^{n-2} + b_{n-3}x^{n-3} + \ldots + b_1x + b_0) + \mathrm{R}x + \mathrm{S}, \tag{2.44}$$

donde $\mathrm{R}x + \mathrm{S}$ es el residuo lineal de la división y R y S dependen de p y q.

Para que la ecuación 2.43 sea un factor cuadrático de la 2.42 (es decir, que la divida exactamente) es necesario que

$$\mathrm{R} = 0 \text{ y } \mathrm{S} = 0 \tag{2.45}$$

Conviene tener un método que permita calcular R y S sin verificar la división de la 2.42 por la 2.43. Para obtenerlo, se igualan los coeficientes de las mismas potencias de x en los dos miembros de la ecuación 2.44

$$\left.\begin{aligned}
a_{n-1} &= b_{n-3} + p \\
a_{n-2} &= b_{n-4} + p\,b_{n-3} + q \\
a_{n-3} &= b_{n-5} + p\,b_{n-4} + q\,b_{n-3} \\
&\ \ \vdots \\
a_k &= b_{k-2} + p\,b_{k-1} + q\,b_k \\
&\ \ \vdots \\
a_1 &= p\,b_0 + q\,b_1 + \mathrm{R} \\
a_0 &= q\,b_0 + \mathrm{S}
\end{aligned}\right\} \tag{2.46}$$

Despejando b_k de la expresión general (usando para ello el término b_{k-2}), se obtiene

$$b_k = a_{k+2} - p\,b_{k+1} - q\,b_{k+2} \text{ para } k = n-3, n-4, \ldots, 0 \tag{2.47}$$

con

$$b_{n-1} = 0 \quad \text{y} \quad b_{n-2} = 1 \tag{2.48}$$

el algoritmo buscado para obtener los coeficientes del polinomio cociente de la 2.44 y además

$$R = a_1 - p\,b_0 - q\,b_1 \tag{2.49}$$

$$s = a_0 - qb_0 \tag{2.50}$$

Al emplear las condiciones de la ecuación 2.45

$$\begin{aligned} a_1 - p\,b_0 - q\,b_1 &= 0 \\ a_0 - q\,b_0 &= 0 \end{aligned} \tag{2.51}$$

El método de Lin consiste en

PASO 1. Proponer aproximaciones iniciales de los valores desconocidos p y q (pueden llamarse p_0 y q_0).

PASO 2. Emplear las ecuaciones 2.47 para obtener aproximaciones de b_{n-3}, b_{n-4}, ... , b_1, b_0.

PASO 3. Calcular R y S. Si son cero o suficientemente cercanas a éste, el problema está terminado. En caso contrario, se estiman nuevos valores de p y q (pueden llarmarse p_1 y q_1)

$$p_1 = \frac{a_1 - q_0 b_1}{b_0} \quad \text{y} \quad q_1 = \frac{a_0}{b_0}$$

para volver al paso 2.

Ejemplo 2.20

Encuentre los factores cuadráticos de la ecuación polinomial de grado cuatro

$$f(x) = x^4 - 8x^3 + 39x^2 - 62x + 50 = 0$$

SOLUCIÓN

PASO 1. Se propone $p = 0$ y $q = 0$

PASO 2. $b_3 = 0$; $b_2 = 1$;

$$b_1 = a_3 - p\,b_2 - q\,b_3 = -8; \qquad b_0 = a_2 - p\,b_1 - q\,b_2 = 39$$

PASO 3. $R = a_1 - p\,b_0 - q\,b_1 = -62 \qquad S = a_0 - q\,b_0 = 50$

$$p_1 = \frac{a_1 - q\,b_1}{b_0} = \frac{-62}{39} = -1.5897;$$

$$q_1 = a_0 / b_0 = 50 / 39 = 1.2821$$

Al repetir los pasos 2 y 3 se encuentra la siguiente sucesión de valores:

p	q	R	S
-1.9358	1.8164	-10.0204	14.7086
-2.0109	1.9708	-1.4494	3.9171
-2.0090	2.0011	0.0469	0.7586
-2.0034	2.0030	0.1396	0.0458
-2.0009	2.0013	0.0632	-0.0410
-2.0001	2.0004	0.0187	-0.0235

Por lo que el factor cuadrático es

$$x^2 - 2x + 2$$

Ejercicios

2.1 La ecuación de estado de Van der Walls para un gas real es

$$\left(P + \frac{a}{V^2}\right)(V - b) = RT \qquad (1)$$

donde

P = presión en *atm*
T = temperatura en K
R = constante universal de los gases en *atm*-l / (gmol K) = 0.08205
V = volumen molar del gas en l/gmol
a, b = constantes particulares para cada gas

Para los siguientes gases, calcule V a 80 °C para presiones de 10, 20, 30 y 100 *atm*.

Gas	a	b
CO_2	3.599	0.04267
Dimetilamina	37.49	0.19700
He	0.03412	0.02370
Óxido nítrico	1.34	0.02789

SOLUCIÓN

La ecuación 1 también puede escribirse como

$$PV^3 - bPV^2 - RTV^2 + aV - ab = 0 \qquad (2)$$

que es un polinomio cúbico en el volumen molar V; entonces, para una P y una T dadas, puede escribirse como una función de la variable V

$$f(V) = p V^3 - (P b + R T) V^2 + a V - a b = 0 \qquad (3)$$

Esta ecuación se resuelve con el método de posición falsa para encontrar el volumen molar.

Valores iniciales

El programa 2.2 del apéndice realiza los cálculos necesarios para resolver esta ecuación, usando como intervalo inicial: $V_I = 0.8\ v$ y $V_D = 1.2\ v$, donde $v = R\ T\ /\ P$, el volumen molar ideal. (Se resuelve sólo el caso del CO_2 a 10 atm y 80°C, dejando como ejercicio para el lector los demás casos.)

Los valores obtenidos para las diferentes iteraciones son los siguientes

iteración	V_M (l / gmol)	$\mid f(V_M) \mid$
1	2.603856	0.1362×10^{2}
2	2.734767	0.5711×10^{1}
3	2.785884	0.2141×10^{1}
4	2.804528	0.7685
5	2.811156	0.2716
6	2.813489	0.9546×10^{-1}
7	2.814309	0.3348×10^{-1}
8	2.814596	0.1173×10^{-1}
9	2,814697	0.4113×10^{-2}
10	2.814732	0.1441×10^{-2}
11	2.814744	0.5050×10^{-3}
12	2.814749	0.1769×10^{-3}
13	2.814750	0.6200×10^{-4}

Se utilizó el criterio de exactitud

$$\mid f(V) \mid < 10^{-4}$$

aunque puede verse que desde la iteración 7, el cambio en los valores de V_M son solamente en la cuarta cifra decimal, que en este caso representan décimas de mililitro.

Resultado: El volumen molar del CO_2 a una presión de 10 atm y una temperatura de 80°C (= 353.2 K) es 2.81475 l/gmol.

2.2 La fórmula de Bazin para la velocidad de un fluido en canales abiertos está dada por

$$v = c\ (re)^{1/2}$$

con

$$c = \frac{87}{0.552 + \frac{m}{(r)^{1/2}}}$$

donde

m = coeficiente de rugosidad
r = radio hidráulico en pies (área dividida entre el perímetro mojado)
e = pendiente de la superficie del fluido
v = velocidad del fluido en pies/segundos

Calcule el radio hidráulico correspondiente a los siguientes datos (dados en unidades consistentes) por el método de Steffensen

$$m = 1.1; \qquad e = 0.001; \qquad v = 5$$

SOLUCIÓN

Sustituyendo c en v: $v = \dfrac{87\,(re)^{1/2}}{0.552 + \dfrac{m}{(r)^{1/2}}}$

o bien

$$(0.552 + \frac{m}{(r)^{1/2}})\, v = 87\,(r)^{1/2}\,(e)^{1/2}$$

multiplicando ambos lados por $(r)^{1/2}$

$$[0.552\,(r)^{1/2} + m]\, v = 87\,(e)^{1/2} r$$

y despejando r se llega a

$$\frac{[0.552\,(r)^{1/2} + m]\, v}{87\,(e)^{1/2}}$$

una de las formas de $g\,(r) = r$, necesaria para el método de Steffensen. Sin embargo, antes de usar el método, conviene averiguar el comportamiento de $g'\,(r)$

$$g'(r) = \frac{0.552\,v}{174\,(r)^{1/2}\,(e)^{1/2}}$$

sustituyendo valores

$$g'(r) = \frac{0.5}{(r)^{1/2}}$$

Como el radio hidráulico debe ser mayor de cero, ya que un valor negativo o cero no tendría significado físico y como $|\,g'\,(r)\,| < 1$ para $(r)^{1/2} > 0.5$, o $r > 0.7$, se selecciona como valor inicial de r a 1.0. Con esto

$$g'(1) = 0.5$$

y el método puede aplicarse con cierta garantía de convergencia

Primera iteración

$$r_0 = 1$$

$$r_1 = g(r_0) = \frac{[0.552(1)^{1/2} + 1.1](5)}{87(0.001)^{1/2}} = 3.00235$$

$$r_2 = g(r_1) = \frac{[0.552(3.00235)^{1/2} + 1.1](5)}{87(0.001)^{1/2}} = 3.73742$$

$$r_3 = r_0 - \frac{(r_1 - r_0)^2}{r_2 - 2r_1 + r_0} = 1 - \frac{(3.00235 - 3.73742)^2}{3.73742 - 2(3.00235) + 1} = 4.16380$$

Segunda iteración

Tomando ahora como nuevo valor inicial $r_3 = 4.16380$, se tiene

$$r_0 = 4.16380$$

$$r_1 = g(r_0) = \frac{[0.552(4.16380)^{1/2} + 1.1](5)}{87(0.001)^{1/2}} = 4.04622$$

$$r_2 = g(r_1) = \frac{[0.552(4.04622)^{1/2} + 1.1](5)}{87(0.001)^{1/2}} = 4.01711$$

$$r_3 = 4.16380 - \frac{(4.04622 - 4.16380)}{4.01711 - 2(4.04622) + 4.16380} = 4.00753$$

Dado que la sucesión es convergente y que se trata del radio de un canal abierto, donde la exactitud después del primer decimal no es necesaria, se toma como valor a $r = 4$ pies.

2.3 La siguiente fórmula es atribuida a Francis* y se aplica a un vertedor con contracciones

$$Q = 3.33(B - 0.2H)(H^3)^{1/2},$$

donde

Q = cantidad de agua que pasa por el vertedor en *pies*^{3}*/s*
B = ancho del vertedor en *pies*
H = carga sobre la cresta del vertedor en *pies*

*J. Lipka. Computaciones gráficas y mécanicas. CECSA (1972) p 139-141.

Si se sabe que B varía de 0 a 5 y Q de 0 a 33, calcule los valores de H correspondientes a las siguientes parejas de valores de B y Q (las unidades son consistentes), con el método de Newton-Raphson.

B	3	2	4	3.6
Q	12	20	13	30

SOLUCIÓN

Se escribe la ecuación en la forma

$$f(H) = 3.33(B - 0.2H)(H^3)^{1/2} - Q = 0$$

Se deriva

$$f'(H) = 3.33(B - 0.2H)(1.5)H^{1/2} + (H^3)^{1/2}(3.33)(-0.2)$$

y sustituyendo en la fórmula de Newton-Raphson a $f(H)$ y $f'(H)$, se tiene

$$H_{i+1} = H_i - \frac{3.33(B - 0.2H_i)(H_i^3)^{1/2} - Q}{4.995(B - 0.2H_i)H_i^{1/2} - 0.666(H_i^3)^{1/2}}$$

Para elegir un valor inicial de H en cada caso, se considera que por cuestiones de diseño H debe ser menor que B. Por lo anterior, se sugiere utilizar como valor inicial $H_0 = B/2$.

Para la pareja $B = 3$, $Q = 12$

Primera iteración

$$H_0 = B/2 = 3/2 = 1.5$$

$$H_1 = 1.5 - \frac{3.33[3-0.2(1.5)](1.5^3)^{1/2} - 12}{4.995[3-0.2(1.5)](1.5)^{1/2} - 0.666(1.5^3)^{1/2}} = 1.2046$$

Segunda iteración

$$H_2 = 1.2046 - \frac{3.33[3-0.2(1.2046)](1.2046^3)^{1/2} - 12}{4.995[3-0.02(1.2046)](1.2046)^{1/2} - 0.666(1.2046^3)^{1/2}}$$

$$= 1.1942$$

Tercera iteración

$$H_3 = 1.1942 - \frac{3.33[3-0.2(1.1942)](1.1942^3)^{1/2} - 12}{4.995[3-0.2(1.1942)](1.1942)^{1/2} - 0.666(1.1942^3)^{1/2}}$$

$$= 1.1942$$

El método ha convergido al valor 1.1942 y se toma como carga sobre la cresta del vertedor 1.2 pies; las demás cifras significativas no interesan, por el sentido físico que tiene H.

Los resultados para las siguientes parejas de B y Q se dan a continuación

B	2	4	3.6
Q	20	13	30
H	2,5	1.0	2.0

2.4 En la solución de problemas de valor inicial en ecuaciones diferenciales por transformadas de Laplace* se presentan funciones racionales del tipo

$$F(s) = \frac{p_1(s)}{p_2(s)}$$

donde p_1 y p_2 son polinomios con: grado $p_1 \leq$ grado p_2.

La expresión de $F(s)$ en fracciones parciales es parte importante del proceso de solución y se efectúa descomponiendo primero $p_2(s)$ en sus factores más sencillos posibles.

En la solución de un problema de valor inicial (PVI) que modela un sistema de control lineal**, la función de transferencia es (obtenida al aplicar la transformada de Laplace al PVI)

$$F(s) = \frac{C(s)}{R(s)} = \frac{24040(s+25)}{s^4 + 125s^3 + 5100s^2 + 65000s + 598800}$$

Para encontrar los factores más sencillos de $F(s)$ se resuelve primero la ecuación polinomial

$$s^4 + 125s^3 + 5100s^2 + 65000s + 598800 = 0$$

Con el método de Müller (programa 2.3 del disco), se obtiene

$$s_1 = -6.6 + 11.4i$$
$$s_2 = -6.6 - 11.4i$$
$$s_3 = -55.9 + 18i$$
$$s_4 = -55.9 - 18i$$

y los factores buscados son

$$(s+6.6-11.4i)(s+6.6+11.4i)(s+55.9-18i)(s+55.9+18i)$$

---

*Spiegel, Murray R., Applied Differential Equations. 2nd Ed Prentice Hall, Inc (1967) p 263-270.
**Véase capítulo 7.

con lo que $F(s)$ queda

$$F(s) = \frac{24040(s + 25)}{F_1 F_2 F_3 F_4}$$

donde

$$F_1 = s - s_1; \qquad F_2 = s - s_2; \qquad F_3 = s - s_3; \qquad F_4 = s - s_4$$

El segundo paso que completa la descomposición pedida es encontrar los valores de A_1, A_2, A_3, A_4 que satisfagan la ecuación

$$F(s) = \frac{A_1}{F_1} + \frac{A_2}{F_2} + \frac{A_3}{F_3} + \frac{A_4}{F_4}$$

Esto se logra pasando el denominador de $F(s)$ al lado derecho

$$24040(s + 25) = A_1F_2F_3F_4 + A_2F_1F_3F_4 + A_3F_1F_2F_4 + A_4F_1F_2F_3$$

y dando valores a s, por ejemplo $s = s_1$. Así

$$24040(-6.6 + 11.4i + 25) =$$

$$A_1(-6.6 + 11.4i + 6.6 + 11.4i)(-6.6 + 11.4i + 55.9 - 18i)(-6.6 + 11.4i + 55.9 + 18i),$$

ya que: $A_2F_1F_3F_4 = A_3F_1F_2F_4 = A_4F_1F_2F_3 = 0$.

Al despejar A_1 y realizar operaciones, se encuentra su valor

$$A_1 = 1.195 - 7.904i$$

Procediendo de igual manera, se calcula A_2, A_3, y A_4 con $s = s_2$, $s = s_3$ y $s = s_4$, respectivamente.

Esto se deja como ejercicio para el lector.

2.5 Una vez descompuesto $F(s) = C(s) / R(s)$ en fracciones parciales (véase ejercicio anterior), se les aplica el proceso de "transformación" inversa de Laplace, que da como resultado la solución del problema de valor inicial.

Sea esta solución

$$F(t) = 1.21e^{-6.6t}\, sen\,(11.4t - 111.7°) + 0.28e^{-55.9t}\, sen\,(18t + 26.1°)$$

La solución obtenida debe analizarse matemáticamente e interpretarse físicamente si procede.

Breve análisis clásico

Si t es el tiempo, el intervalo de interés es $t > 0$.

En los términos primero y segundo de $F(t)$ aparece la función seno, que es oscilatoria, afectada de la función exponencial. Esta tiende a cero aproximadamente

cuando t tiene valores superiores a 1; se lleva tanto sus factores como la función $F(t)$ a dicho valor, con lo cual la gráfica $F(t)$ se confunde con el eje t para $t \geq 1$.

Estas funciones son conocidas como oscilatorias amortiguadas y sus gráficas son del tipo mostrado en la figura 2.11

Si, por el contrario, el exponente de e es positivo, al tender t a infinito, la función es creciente y tiende rápidamente a infinito; lo cual se conoce como función oscilatoria no amortiguada.

Por otro lado, obsérvese que la contribución numérica del segundo término de $F(t)$ es despreciable y que el análisis y la gráfica de $F(t)$ pueden obtenerse sin menoscabo de exactitud con el primer término.

Si se dan algunos valores particulares a t se obtiene

t	0.0	0.2	0.4	0.6	0.8	1.0
$F(t)$	-1.001	0.105	0.044	-0.023	0.005	-4.22×10^{-5}

Estos valores señalan claramente la presencia de raíces reales en los intervalos (0,0.2), (0.4,0.6), (0.6,0.8), (0.8,1.0). Utilizando como valores iniciales 0.1, 0.5, 0.7, y 0.9 y el método de Newton-Raphson se obtiene, respectivamente

$$\bar{t}_1 = 0.171013; \qquad \bar{t}_2 = 0.44659; \qquad \bar{t}_3 = 0.72217; \qquad \bar{t}_4 = 0.99775$$

Los posibles máximos y mínimos de esta función se consiguen resolviendo la ecuación que resulta de igualar con cero la primera derivada de $F(t)$

$$F'(t) = 13.794e^{-6.6t}\cos(11.4t - 111.7°) - 7.986e^{-6.6t}\,sen(11.4t - 111.7°)$$
$$+ 5.04e^{-55.9t}\cos(18t + 26.1°) - 15.652e^{-55.9t}\,sen(18t + 26.1°) = 0$$

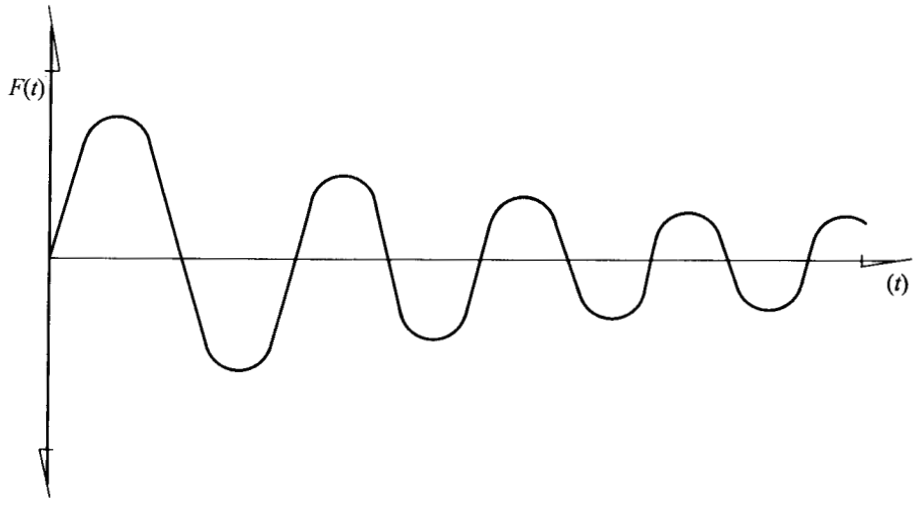


Figura 2.11 Comportamiento de una función oscilatoria amortiguada.

Aprovechando las evaluaciones que se hicieron de $F'(t)$ en el método de Newton-Raphson, se tiene

t	0.0	0.1	0.3	0.6	0.9	1.0
$F'(t)$	–0.040	7.849	–0.900	0.196	–0.035	–0.00184

Con los valores iniciales dados a la izquierda, se obtuvieron las raíces anotadas a la derecha

$$t_0 = 0 \qquad t_1 = 0.00175$$
$$t_0 = 0.2 \qquad \bar{t}_2 = 0.26277$$
$$t_0 = 0.45 \qquad \bar{t}_3 = 0.53834$$
$$t_0 = 0.75 \qquad \bar{t}_4 = 0.81399$$

Con los valores de la función en diferentes puntos, sus raíces y puntos máximos y mínimos, la gráfica aproximada de $F(t)$ se muestra en la figura 2.12

Este análisis se puede comprobar con el software del libro o el G.C., por ejemplo.

2.6 Determine la cantidad de vapor V (moles/hr) y la cantidad de líquido L (moles/hr) que se generan en una vaporización instantánea continua a una presión de 1600 psia y una temperatura de 120 F de la siguiente mezcla.

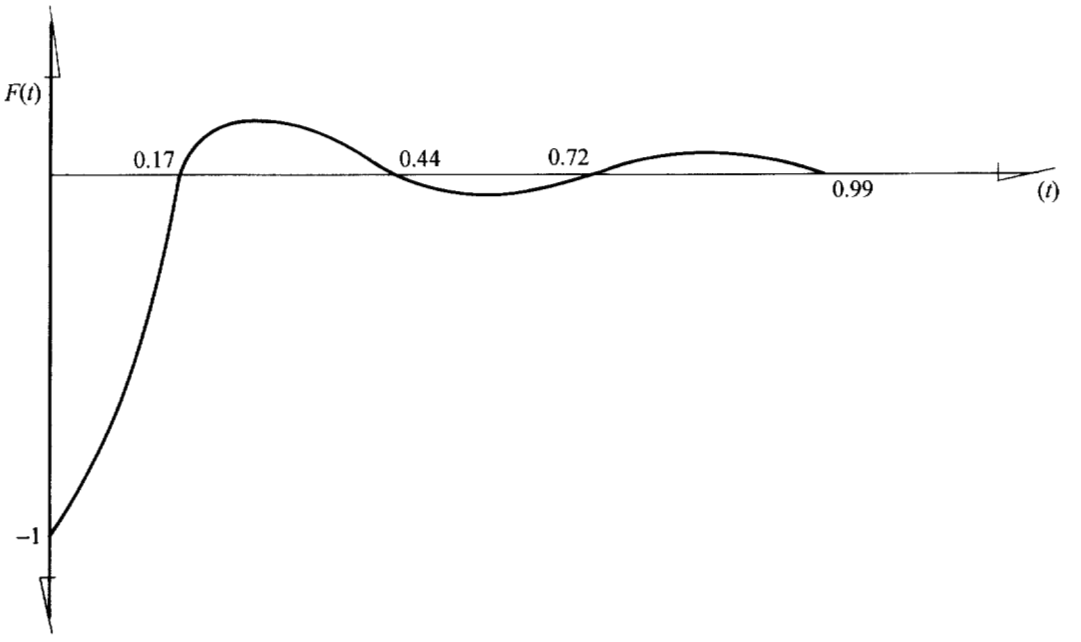


Figura 2.12 Gráfica aproximada de la función $F(t)$

Componente	Composición z_i	$K_i = y_i/x_i$
CO_2	0.0046	1.65
CH_4	0.8345	1.8
C_2H_6	0.0381	0.94
C_3H_8	0.0163	0.55
i–C_4H_{10}	0.0050	0.40
n–C_4H_{10}	0.0074	0.38
C_5H_{12}	0.0287	0.22
C_6H_{14}	0.0220	0.14
C_7H_{16}	0.0434	0.09

SOLUCIÓN

Con base en la figura 2.13

Un balance total de materia da $F = L + V$ (1)

Un balance de materia para cada componente da:

$$F\,z_i = L\,x_i + V\,y_i \qquad i = 1,2,\ldots,n \tag{2}$$

Las relaciones de equilibrio líquido-vapor establecen

$$K_i = \frac{y_i}{x_i} \qquad i = 1,2,\ldots,n \tag{3}$$

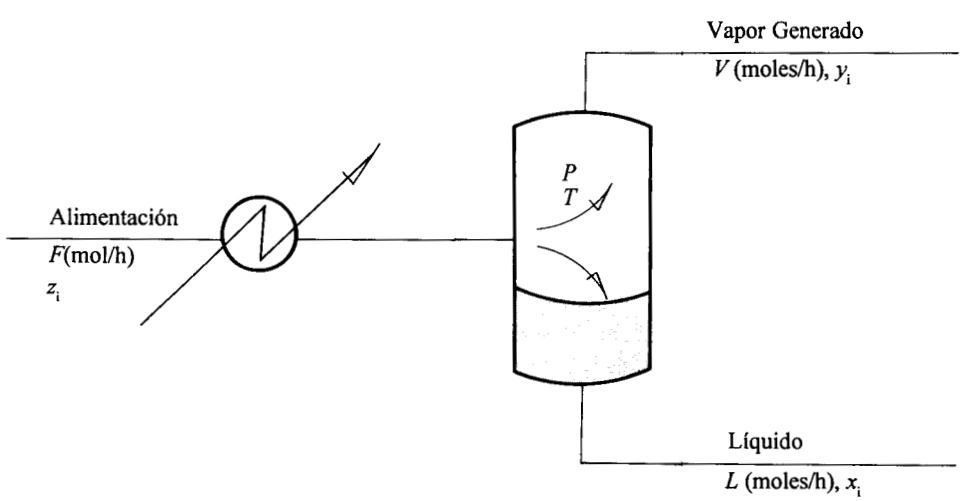


Figura 2.13 Esquema de un vaporización instantánea (*flash*) de una mezcla multicomponente.

Sustituyendo la ecuación 3 en la 2 se obtiene

$$F z_i = L x_i + V K_i x_i \qquad i = 1,2, \ldots, n \tag{4}$$

o bien

$$F z_i = x_i (L + V K_i) \qquad i = 1,2, \ldots, n$$

de donde

$$x_i = \frac{F z_i}{L + V K_i} \qquad i = 1, 2, \ldots, n$$

Sustituyendo la ecuación 1 en esta última se obtiene

$$x_i = \frac{F z_i}{F + V(K_i - 1)} \qquad i = 1,2, \ldots, n \tag{5}$$

Las restricciones de composición establecen

$$\sum_{i=1}^{n} x_i = 1 \quad ; \quad \sum_{i=1}^{n} y_i = 1$$

Por lo que puede escribirse

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i = 0$$

o bien

$$\sum_{i=1}^{n} K_i x_i - \sum_{i=1}^{n} x_i = 0$$

o simplemente

$$\sum_{i=1}^{n} x_i (K_i - 1) = 0 \tag{6}$$

sustituyendo la ecuación 5 en la 6 se obtiene

$$\sum_{i=1}^{n} \frac{F z_i (K_i - 1)}{F + V(K_i - 1)} = 0 \tag{7}$$

Valores iniciales

El valor de V que satisface la ecuación 7 está comprendido en el intervalo $0 \leq V \leq F$, por lo que la estimación de un valor inicial de V es difícil, ya que el valor de F puede ser muy grande. Esta dificultad se reduce normalizando el valor de V; esto es, dividiendo numerador y denominador de la ecuación 7 entre F, para obtener

$$\sum_{i=1}^{n} \frac{z_i (K_i - 1)}{1 + \psi (K_i - 1)} = 0 \tag{8}$$

donde $\psi = V/F$

La ecuación 8 equivale a la 7, pero expresada en la nueva variable ψ cuyos límites son

$$0 \leq \psi \leq 1$$

La ecuación 8 es no lineal en una sola variable (ψ), que se resolverá con el método de Newton-Raphson. Debe notarse que esta ecuación es monotónica decreciente, por lo que el valor inicial puede ser cualquier número dentro del intervalo [0, 1], por ejemplo $\psi_0 = 0$.

El programa 2.4 del disco, usa $\psi_0 = 0$ como estimado inicial y

$$f'(\psi) = \sum_{i=1}^{n} \frac{-z_i (K_i - 1)^2}{[1 + \psi (K_i - 1)]^2}$$

A continuación se muestran los valores que adquiere ψ y $f(\psi)$ a lo largo de las iteraciones realizadas.

Iteración	ψ	$f(\psi)$
1	0.9328799	-8.79×10^{-2}
2	0.8968149	-1.29×10^{-2}
3	0.8895657	-3.5×10^{-4}
4	0.8893582	-2.68×10^{-7}
5	0.8893580	-1.5×10^{-13}

Resultados

Para $F = 1$ moles/h
Vapor generado, $V = 0.889358$ moles/h
Líquido generado, $L = 0.110642$ moles/h
Composiciones del líquido y del vapor generados

Componente	Líquido (x_i)	Vapor (y_i)
CO_2	0.00291	0.00481
CH_4	0.48759	0.87766
C_2H_6	0.04025	0.03783
C_3H_8	0.02718	0.01495
i–C_4H_{10}	0.01072	0.00429
n–C_4H_{10}	0.01650	0.00627
C_5H_{12}	0.09370	0.02061
C_6H_{14}	0.09356	0.01310
C_7H_{16}	0.22760	0.02048

2.7 Considere un líquido en equilibrio con su vapor. Si el líquido está formado por los componentes 1, 2, 3, y 4, con los datos dados a continuación calcule la temperatura y la composición del vapor en el equilibrio a la presión total de 75 psia.

Componente	Composición del líquido % mol	Presión de vapor de componente puro (psia) a 150°F	a 200°F
1	10.0	25.0	200.0
2	54.0	14.7	60.0
3	30.0	4.0	14.7
4	6.0	0.5	5.0

Utilice la siguiente ecuación para la presión de vapor

$$ln\ (p_i^0) = A_i + B_i/T; \qquad i = 1, 2, 3, 4; \qquad T \text{ en } °R$$

SOLUCIÓN

La presión total del sistema será: $P_T = \sum_{i=1}^{n} P_i$ (1)

Si se considera que la mezcla de estos cuatro componentes, a las condiciones de presión y temperatura de este sistema, obedece las leyes de Raoult y de Dalton

$$P_T = \sum_{i=1}^{4} p_i^0\, x_i \tag{2}$$

donde

p_i^0 = presión de vapor de cada componente
P_T = presión total del sistema
P_i = presión parcial de cada componente
x_i = fracción mol de cada componente en el líquido

De la ecuación de presión de vapor se tiene que

$$p_i^0 = \exp\ (A_i + B_i/T) \qquad i = 1, 2, 3, 4 \tag{3}$$

de las ecuaciones 1 y 2 resulta

$$P_T = \sum_{i=1}^{4} x_i \exp\ (A_i + B_i/T)$$

de donde puede establecerse

$$f(T) = P_T - \sum_{i=1}^{4} x_i \exp\ (A_i + B_i/T) = 0 \tag{5}$$

A_i y B_i pueden obtenerse como sigue

Si se hace $p^0_{1,i}$ = presión de vapor del componente i a T_1 = 150 °F = 609.56 °R

$p^0_{2,i}$ = presión de vapor del componente i a T_2 = 200 °F = 659.56 °R

entonces

$$\ln (p^0_{1,i}) = A_i + B_i / T_1 \qquad i = 1, 2, 3, 4 \tag{6}$$

y

$$\ln (p^0_{2,i}) = A_i + B_i / T_2 \qquad i = 1, 2, 3, 4 \tag{7}$$

restando la ecuación 7 de la 6 se tiene

$$\ln \left(\frac{p^0_{1,i}}{p^0_{2,i}} \right) = B_i (1 / T_1 - 1 / T_2)$$

de donde

$$B_i = \frac{\ln \left(\frac{P^0_{1,i}}{P^0_{2,i}} \right)}{\frac{1}{T_1} - \frac{1}{T_2}} \tag{8}$$

Conociendo B_i se puede obtener A_i de la ecuación 6

$$A_i = \ln (p^0_{1,i}) - B_i / T_1 \qquad i = 1, 2, 3, 4 \tag{9}$$

Valores iniciales

Para estimar un valor inicial de T para resolver la ecucación 5, se considera el componente dominante de la mezcla, en este caso el componente 2, y se usa P_T en lugar de p^0_2 en la ecuación de presión de vapor

$$\ln (P_T) = A_2 + B_2 / T$$

de donde

$$T = \frac{B_2}{\ln (P_T) - A_2} \tag{10}$$

Con este estimado inicial y las consideraciones ya anotadas, el programa 2.5 del disco, utiliza el método de Newton-Raphson con

$$f'(T) = -\sum_{i=1}^{4} x_i \exp (A_i + B_i / T) (-B_i / T^2) \tag{11}$$

y reporta los siguientes resultados después de cuatro iteraciones

Temperatura del sistema = 209.07 °F = 668.63 °R
(temperatura de burbuja)

Composición del vapor en equilibrio

Componente (i)	y_i
1	0.3761
2	0.5451
3	0.0729
4	0.0059

2.8 Se emplea un intercambiador de calor (Fig. 2.14) para enfriar aceite. Encuentre la temperatura de salida del aceite y del agua enfriadora (TH_2 y TC_2, respectivamente), para gastos de aceite de 105,000; 80,000; 50,000; 30,000 y 14,000 lbm/h.

SOLUCIÓN

Un balance de calor para el aceite da

$$Q_1 = w_1 Cp_1 \ (TH_1 - TH_2) \tag{1}$$

Un balance de calor para el agua da

$$Q_2 = w_2 Cp_2 \ (TC_2 - TC_1) \tag{2}$$

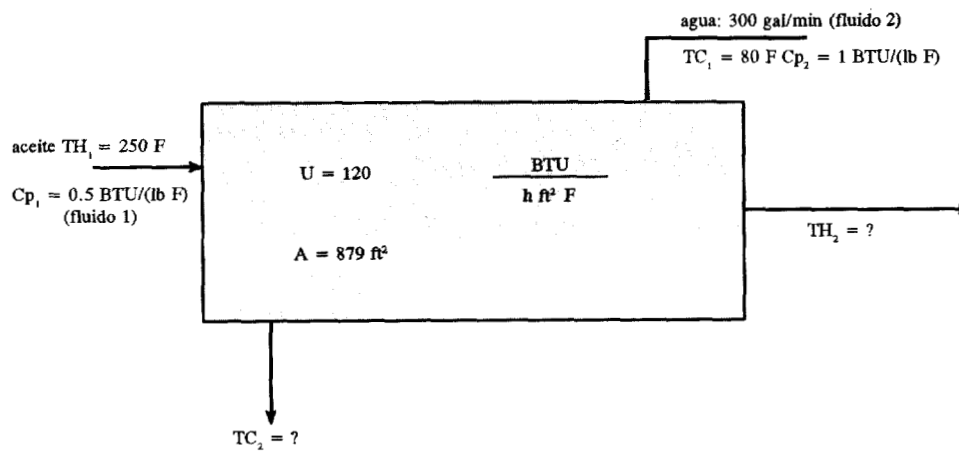


Figura 2.14 Esquema de un intercambiador de calor con flujo a contracorriente.

La ecuación que rige la transferencia de calor a través de este equipo es

$$Q = U\,A\,\Delta Tm \tag{3}$$

donde

U = coeficiente global de transferencia de calor
A = área total de transferencia de calor

$$\Delta Tm = \frac{(TH_1 - TC_2) - (TH_2 - TC_1)}{ln\left(\frac{TH_1 - TC_2}{TH_2 - TC_1}\right)} \tag{4}$$

Para encontrar TH_2 y TC_2 debe cumplirse que $Q_1 = Q_2 = Q$, o bien

$$\frac{Q}{Q_1} - 1 = 0 \tag{5}$$

Pero Q sólo podrá calcularse cuando se conozcan todas la temperaturas. Para resolver este problema se propone el siguiente procedimiento

Establecer que TH_2 sea la única variable; entonces, Q_2 puede escribirse en función de TH_2 como sigue

$$Q_2 = Q_1 = w_1\,Cp_1\,(TH_1 - TH_2) = w_2\,Cp_2\,(TC_2 - TC_1) \tag{6}$$

de donde puede despejarse TC_2

$$TC_2 = \frac{w_1\,Cp_1}{w_2\,Cp_2}\,(TH_1 - TH_2) + TC_1 \tag{7}$$

Con todo esto ya puede establecerse Q en función de TH_2 y así escribir la ecuación 5 también en función de dicha variable única

$$f(TH_2) = \frac{UA\,\dfrac{\left(TH_1 - \frac{w_1\,Cp_1}{w_2\,Cp_2}\,(TH_1 - TH_2) - TC_1\right) - (TH_2 - TC_1)}{\ln\dfrac{\left((TH_1 - \frac{w_1\,Cp_1}{w_2\,Cp_2}\,(TH_1 - TH_2) - TC_1\right)}{TH_2 - TC_1}}}{w_1\,Cp_1\,(TH_1 - TH_2)} - 1 = 0 \tag{8}$$

Valores iniciales

Para estimar un valor inicial de TH_2 cabe apoyarse en la figura 2.15, la cual muestra una gráfica de temperaturas en este tipo de intercambiadores de calor

De acuerdo con esta gráfica, se tienen las siguientes restricciones

$$TC_1 < TH_2 < TH_1$$

y

$$TC_1 < TC_2 < TH_1$$

Como en este caso no se dispone de mayor información, el programa 2.6 del apéndice usa el método de la bisección con $TH_{2I} = TC_1 + 0.5$ y $TH_{2D} = TH_1 - 0.5$ para resolver la ecuación 8.

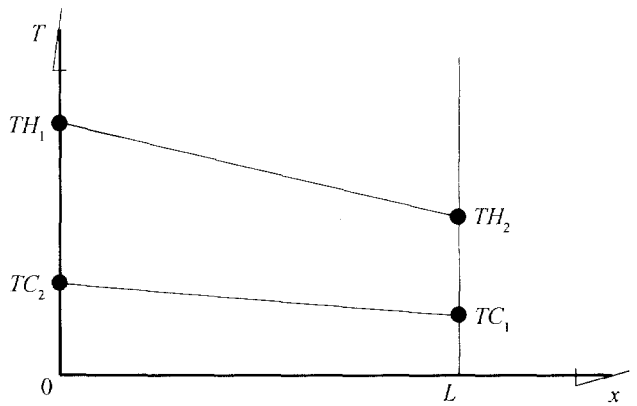


Figura 2.15 Gráfica de temperaturas contra longitud en un intercambiador de calor con flujo a contracorriente.

Para un gasto de aceite de 105,000 *lbm*/h

RESULTADOS

$$TH_2 = 113.0809 \qquad TC_2 = 127.9582$$

2.9 El siguiente circuito representa en forma muy simplificada un generador de impulsos para probar el aislamiento de un transformador en circuito abierto.

Considérese el *gap* como un interruptor.

Las condiciones iniciales en el transformador y la inductancia son cero. Use los siguientes datos para encontrar $v_2\,(t)$:

$C_1 = 12.5 \times 10^{-9}\ fd, \qquad C_2 = 0.3 \times 10^{-9} fd, \qquad L_1 = 0.25 \times 10^{-3}\ Hy$

$R_1 = 2\ Kohms, \qquad R_2 = 3\ Kohms, \qquad V_1 = 300\ Kv$

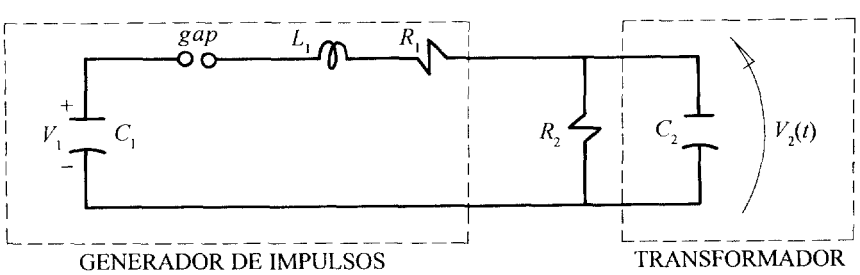


Figura 2.16 Circuito representativo de un generador de impulsos.

SOLUCIÓN

Estableciendo las ecuaciones para el circuito

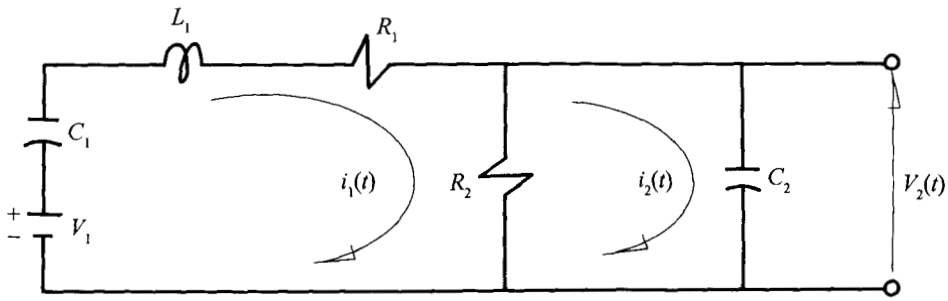


Figura 2.17 Circuito

$$V_1 = (R_1 + R_2)i_1(t) + L_1\frac{di_1(t)}{dt} + \frac{1}{C_1}\int i_1(t)\,dt - R_2 i_2(t) \qquad (1)$$

$$0 = -R_2 i_1(t) + R_2 i_2(t) + \frac{1}{C_2}\int i_2(t)\,dt \qquad (2)$$

$$v^2(t) = \frac{1}{C_2}\int i_2(t)\,dt \qquad (3)$$

Aplicando la transformada de Laplace y considerando que las condiciones iniciales son cero, se tiene

$$\frac{V_1}{s} = (R_1 + R_2)I_1(s) + L_1 sI_1(s) + \frac{1}{C_1}\frac{I_1(s)}{s} - R_2 I_2(s) \qquad (4)$$

$$0 = -R_2 I_1(s) + R_2 I_2(s) + \frac{1}{C_2}\frac{I_2(s)}{(s)} \qquad (5)$$

Despejando $I_2(s)$ de la ecuación 5 y sustituyendo en la ecuación cuatro se tiene

$$I_2(s) = \frac{V_1}{(R_1 + L_1 s + 1/C_1 s)(s + 1/R_2 C_2) - 1/C_2} \qquad (6)$$

al aplicar la transformada inversa de Laplace a la ecuación 3 y recordando que las condiciones iniciales son cero

$$V_2(s) = \frac{1}{C_2}\frac{I_2(s)}{s} \qquad (7)$$

Se sustituye la ecuación 6 en la 7

$$V_2(s) = \frac{\frac{V_1}{C_2}}{s\left((R_1 + L_1 s + \frac{1}{C_1 s})(s + \frac{1}{R_2 C_2}) + \frac{1}{C_2}\right)}$$

y simplificando se llega a

$$V_2(s) = \frac{V}{s^3 + P_1 s^2 + P_2 s + P_3} \tag{8}$$

con

$$P_1 = \frac{R_1 R_2 C_2 + L_1}{R_2 C_2 L_1} = 9.1111 \times 10^6$$

$$P_2 = \frac{R_1 C_1 + R_2 C_2 + R_2 C_1}{R_2 C_1 C_2 L_1} = 22.5422 \times 10^{12}$$

$$P_3 = \frac{1}{R_2 C_1 C_2 L_1} = 355.556 \times 10^{15}$$

$$V = \frac{V_1}{C_2 L_1} = \frac{V_1}{75 \times 10^{-15}}$$

La ecuación 8 puede escribirse

$$V_2(s) = \frac{V}{(s + a)(s + b)(s + c)}$$

cuya transformada inversa de Laplace es

$$v_2(t) = V\left(\frac{e^{-at}}{(b-a)(c-a)} + \frac{e^{-bt}}{(c-b)(a-b)} + \frac{e^{-ct}}{(a-c)(b-c)}\right) \tag{9}$$

donde a, b y c son las raíces de la ecuación

$$s^3 + P_1 s^2 + P_2 s + P_3 = 0$$

La primera raíz, obtenida con el programa 2.3 del apéndice, es

$$a = 1.5874581 \times 10^4$$

Se reduce el grado del polinomio y aplicando la fórmula cuadrática, se tiene

$$b = 4.547613 \times 10^6 + 1.310359 \times 10^6\, i$$

$$c = 4.547613 \times 10^6 - 1.310359 \times 10^6\, i$$

Estos valores se sustituyen en la ecuación 9 y se tiene

$$v_2(t) = 300 \left(0.6\, e^{-1.587458 \times 10^4 t} - e^{-4.54761 \times 10^6 t} \left[0.6 \cos(1.310359 \times 10^6 t) + 2.072102 \operatorname{sen}(1.310359 \times 10^6 t) \right] \right)$$

donde t está en *segundos* y $v_2(t)$ en *Kvolts*.

Problemas

2.1 Dadas las siguientes expresiones para $x = g(x)$, obténga $g'(x)$ y dos valores iniciales que satisfagan la condición $|g'(x)| < 1$

a) $x = \dfrac{1}{(x+1)^2}$ *b)* $x = 4 + \left(\dfrac{x-1}{x+1}\right)^{1/3}$ *c)* $x = \operatorname{sen} x$

d) $\tan x = \ln x$ *e)* $x = \left(\dfrac{6 - x - x^3}{4}\right)^{1/2}$ *f)* $x = \dfrac{\sec x}{2}$

2.2 Determine una $g(x)$ y un valor inicial x_0 tales que $|g'(x)| < 1$ en las siguientes ecuaciones

a) $2x = 4x$ *b)* $x^3 - 10x - 5 = 0$ *c)* $\operatorname{sen} x + \ln x = 0$

d) $e^x - \tan x = 0$ *e)* $\sqrt{x^3 \operatorname{sen} x} - \ln(\cos x) = 3$

2.3 Resuelva por el método de punto fijo las ecuaciones de los problemas anteriores.

2.4 Generalmente hay muchas maneras de pasar de $f(x) = 0$ a $x = g(x)$ e incluso se pueden obtener distintas formas de $g(x)$ al "despejar" x de un mismo término de $f(x)$.

Por ejemplo, en la ecuación polinomial

$$x^3 - 2x - 2 = 0$$

al despejar x del primer término se puede llegar a

a) $x = \sqrt[3]{2x + 2}$ *b)* $x = \sqrt{2 + 2/x}$ *c)* $x = \dfrac{2}{x} + \dfrac{2}{x^2}$

¿Cuál $g(x)$ sería mas ventajosa para encontrar la raíz que está en el intervalo $(1, 2)$? Calcule con un mismo valor inicial dicha raíz empleando las tres $g(x)$ y compare resultados.

2.5 Utilice la fórmula de Francis (véase ejercicio 2.3)

a) Encuentre una expresión $H = g(H)$ tal que usando como valor inicial $H = B/2$, el método de punto fijo prometa convergencia (quizá sea necesaria una $g(H)$ distinta para cada pareja B, Q dada).

b) Con los valores de B y Q dados en el ejercicio 2.3 y los resultados del inciso (a), calcule los respectivos valores de H que satisfacen la ecuación de Francis.

2.6 Sea el polinomio de grado n en su forma más general

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + \ldots + a_1 x + a_0$$

a) Calcule el número de multiplicaciones y sumas algebraicas necesarias para evaluar $f(x)$ en un punto dado mediante el método de Horner.

b) Calcule el número de multiplicaciones y sumas algebraicas requeridas para evaluar $f(x)$ en un punto dado usando la forma tradicional. Al comparar las cantidades de los incisos (a) y (b), encontrará que el número de multiplicaciones y sumas algebraicas en el arreglo de Horner se reduce prácticamente a la mitad. Como cada multiplicación involucra errores de redondeo, este método de evaluación es más exacto y rápido.

2.7 Elabore un programa que evalúe polinomios según la regla de Horner.

2.8 Resuelva las siguientes ecuaciones con el método de Newton-Raphson

a) $\ln x - x + 2 = 0$ *b*) $xe^x - 2 = 0$

c) $x - 2 \cos x = 0$ *d*) $x^3 - 5x = -1$

2.9 Resuelva los siguientes sistemas de ecuaciones con el método de Newton-Raphson

a) $2x^3 - y = 0$
$x^3 - 2 - y^3 = 0$

b) $2x^2 - y = 0$
$x = 2 - y^2$

c) $x^2 + 5xy^2 - 3z + 1 = 0$
$x - sen\ y = 1$
$y - e^{-z} = 0$

d) $(x - 1)^{1/2} + yx - 5 = 0$
$y - sen\ x^2 = 0$

2.10 La manera más simple de evitar el cálculo de $f'(x)$ en el método de Newton Raphson es remplazar $f'(x)$ en la ecuación 2.12 con un valor constante m. La fórmula resultante

$$x_{i+1} = x_i - \frac{f(x_i)}{m}$$

define un método de convergencia lineal para m en cierto intervalo de valores.

a) Utilice este algoritmo, conocido como el método de Wittaker, para encontrar una raíz real de la ecuación

$$f(x) = x^3 + 2x^2 + 10x - 20 = 0$$

b) Con este algoritmo encuentre una raíz en el intervalo (1.5, 2.5) de la ecuación

$$f(x) = x^3 - 12x^2 + 36x - 32 = 0$$

2.11 Demuestre que en el método de Newton-Raphson $g'(\bar{x}) = 0$ y $g''(\bar{x}) \neq 0$ para raíces reales no repetidas.

2.12 Dado un polinomio de grado n

$$p_n(x) = a_n x^n + a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + \ldots + a_0 \qquad (1)$$

elaborare un programa para encontrar todas las raíces reales y complejas de $p_n(x)$, mediante el método de Newton-Raphson.
El programa deberá tener incorporada la división sintética para

a) Evaluar polinomios
b) Degradar polinomios cada vez que se encuentre una raíz (véase sec. 2.10)

2.13 Demuestre que en el método de Newton-Raphson

$$\in_{i+1} \approx \frac{f''(\bar{x})}{2!f'(\bar{x})} \in_i^2$$

Sugerencia: Utilice la ecuación

$$\in_{i+1} = g'(\bar{x}) \in_i + g''(\bar{x}) \frac{\in_i^2}{2!} + g'''(\bar{x}) \frac{\in_i^3}{3!} + \ldots$$

y los resultados del problema 2.11

2.14 El siguiente algoritmo se conoce como método de Richmond y es de tercer orden

$$x_{i+1} = x_i - \frac{2f(x_i)f'(x_i)}{2[f'(x_i)]^2 - f(x_i)f''(x_i)}$$

Resuelva las ecuaciones de los problemas 2.8 y 2.9 con este algoritmo y compare los resultados con los obtenidos con el método de Newton Raphson; por ejemplo, la velocidad de convergencia y el número de cálculos por iteración.

2.15 Obtenga la expresión 2.14 del algoritmo de posición falsa, utilizando la semejanza de los triángulos rectángulos cuyos vértices son A x_I x_M y B x_D x_M en la figura 2.7.

2.16 La expresión 2.13, puede escribirse también

$$x_{i+1} = \frac{x_{i-1}f(x_i) - x_i f(x_{i-1})}{f(x_i) - f(x_{i-1})}$$

Explique por qué, en general, es más eficiente la ecuación 2.13 que la ecuación anterior en la aplicación del método de la secante.

2.17 Resuelva por el método de la secante, posición falsa o bisección las siguientes ecuaciones

$$a)\ x \log x - 10 = 0$$

$$b)\ \operatorname{sen} x - \csc x + 1 = 0$$

$$c)\ e^x + 2^{-x} + 2\cos x - 6 = 0$$

$$d)\ e^x + x^3 + 2x^2 + 10x - 20 = 0$$

Sugerencia: Utilice un análisis preliminar de estas funciones para obtener valores iniciales apropiados.

2.18 Elabore un programa para encontrar una raíz de $f(x) = 0$, por el método de posición falsa, dada $f(x)$ como una tabla de valores.

2.19 Encuentre una aproximación a $\sqrt[3]{2}$ y a $\sqrt{3}$ mediante el método de la bisección.

El cálculo deberá ser correcto en cuatro dígitos significativos.

Sugerencia:Considere $f(x) = x^3 - 2 = 0$ y $f(x) = x^2 - 3 = 0$, respectivamente.

2.20 Utilice la expresión 2.15 para hallar el número aproximado de iteraciones n a fin de encontar una raíz de

$$x^2 + 10 \cos x = 0$$

con una aproximación de 10^{-3}. Encuentre además dicha raíz.

2.21 Aplique el método de bisección y el de posición falsa a la ecuación

$$\frac{7x-3}{(x-0.45)^2} = 0$$

Use los intervalos (0.4,0.5) y (0.39,0.53)

Explique gráficamente los resultados.

2.22 Demuestre que en el caso de convergencia de una sucesión de valores $x_0, x_1, x_2, \ldots$ a una raíz $\bar{x}$ en el método de punto fijo se cumple que

$$\lim_{i \to \infty} \frac{\in_{i+1}}{\in_i} = g'(\bar{x})$$

2.23 Las siguientes sucesiones convergen y los límites de convergencia de cada una se dan al lado derecho

a) $x_k = \dfrac{(-1)^k}{k}$ $\qquad \lim_{k \to \infty} \{ x_k \} = 0$

b) $x_n = n \ln (1+1/n)$ $\qquad \lim_{n \to \infty} \{ x_n \} = 1$

c) $x_k = \dfrac{2^{k+1} + (-1)^k}{2^k}$ $\qquad \lim_{k \to \infty} \{ x_k \} = 2$

d) $x_k = 1+e^{-k}$ $\qquad \lim_{k \to \infty} \{x_k\} = 1$

Genere en cada inciso la sucesión finita: $x_1, x_2, x_3, \ldots, x_{10}$

Aplique después el algoritmo de Aitken a estas sucesiones para generar las nuevas sucesiones $x_1', x_2', x_3', \ldots$ observe qué ocurre y dé sus conclusiones.

2.24 Modifique el algoritmo 2.5 de Steffensen, incorporando una prevención para el caso en que el denominador de la ecuación 2.22 sea muy cercano a cero.

2.25 Encuentre una aproximación $\sqrt[3]{2}$ y a $\sqrt{3}$ con el método de Steffensen. El cálculo deberá ser correcto en cuatro dígitos significativos. Compare los resultados con los obtenidos en el problema 2.19.

2.26 Aproxime una solución para cada una de las siguientes ecuaciones con una aproximación de 10^{-5}, usando el método de Steffensen con $x_0 = 0$.

a) $3x-x^2+e^x-2=0$ b) $4.1x^2 - 1.3e^x = 0$ c) $x^2 + 2xe^x + e^{2x} = 0$

2.27 Encuentre la gráfica aproximada de las siguientes funciones en los intervalos indicados

a) $f(x) = e^{x^2} + x - 1000$; $\qquad (1, 10)$

$$b)\ f(x) = x^4 - 2x + 10; \quad (-\infty, \infty)$$

$$c)\ f(x) = 4(x-2)^{1/3} + \text{sen}(3x); \quad [0, \infty)$$

$$d)\ f(x) = \frac{1}{\sqrt{2}} e^{-x^2/2}; \quad -\infty < x < \infty$$

$$e)\ f(x) = \frac{1}{(\frac{\nu}{2} - 1)!\, 2^{\nu/2}} x^{\nu/2-1} e^{-\nu/2}; \quad x > 0$$

$$f)\ f(x) = x^2 - 4 + \ln 3x + 5 \operatorname{sen} x$$

2.28 Utilizando el método de Newton-Raphson con valores iniciales complejos $(a + b_i)$, encuentre las raíces complejas del polinomio

$$f(x) = x^3 + 4x + 3x^2 + 12$$

2.29 Utilizando el método de Müller con valores iniciales reales, encuentre las raíces complejas del polinomio del problema 2.28.

2.30 Encuentre las raíces faltantes de la ecuación polinomial usada a lo largo del capítulo para ilustrar los distintos métodos

$$f(x) = x^3 + 2x^2 + 10x - 20 = 0,$$

pero usando ahora el método de Newton-Raphson con valores iniciales complejos.

2.31 La solución general de la ecuación polinomial

$$p(x) = a_0 + a_1 x + a_2 x^2$$

es

$$x_1 = \frac{-a_1 + \sqrt{a_1^2 - 4a_0a_2}}{2a_2} \qquad x_2 = \frac{-a_1 - \sqrt{a_1^2 - 4a_0a_2}}{2a_2}$$

a) Demuestre que $x_1x_2 = a_0 / a_2$

b) Utilizando a), demuestre que una forma alterna para encontrar las raíces de

$$p(x) = a_0 + a_1 x + a_2 x^2 = 0$$

es

$$x_1 = \frac{2a_0}{-a_1 + \sqrt{a_1^2 - 4a_0a_2}}; \quad x_2 = \frac{2a_0}{-a_1 - \sqrt{a_1^2 - 4a_0a_2}}$$

c) Calcule la raíz x_2 de

$$p(x) = x^2 + 81x - 0.5 = 0$$

usando aritmética de cuatro dígitos con las dos formas presentadas y sustituya ambos resultados en $p(x)$. Compare la exactitud de los resultados y explique la diferencia. Puede usar Mathematica o Fortran.

d) Calcule la raíz x_1 de

$$p(x) = x^2 + 81x - 0.5 = 0$$

2.32 Elabore un programa de propósito general para encontrar todas las raíces reales y complejas de una ecuación polinomial de la forma

$$p_n(x) = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n$$

con el método de Müller.

2.33 El siguiente algoritmo, de orden tres, es conocido como método de Laguerre

$$x_{i+1} = x_i - \frac{n\,p(x_i)}{p'(x_i) \pm \sqrt{H(x_i)}} \quad , \quad i = 0, 1, 2, \ldots$$

donde n es el grado de la ecuación polinomial $p(x) = 0$, cuyas raíces se desea encontrar.

$$H(x_i) = (n-1)\,[\,(n-1)\,(p'(x_i))^2 - np(x_i)\,p''(x_i)\,]$$

y el signo del radical queda determinado por el signo de $p'(x_i)$.

Este método, que funciona con orden 3 para polinomios cuyas raíces son todas reales y distintas, converge sólo linealmente para raíces múltiples. En el caso de raíces complejas poco se sabe del orden de convergencia; no obstante, ésta es alta para raíces complejas simples. Finalmente, se hace la observación de que un valor de x_i real puede producir una $H(x_i)$ negativa y, por tanto, generar un valor de x_{i+1} complejo y eventualmente llevar a una raíz compleja de la ecuación $p(x) = 0$.

Resuelva las siguientes ecuaciones con el método de Laguerre

$$a)\ x^4 - 8.2\,x^3 + 39.41\,x^2 - 62.26\,x + 30.25 = 0$$

$$b)\ x^4 - 15.2\,x^3 + 59.7\,x^2 - 81.6\,x + 36 = 0$$

$$c)\ x^5 - 10\,x^4 + 40\,x^3 - 80\,x^2 + 79\,x - 30 = 0$$

$$d)\ x^5 - 3.7\,x^4 + 7.4\,x^3 - 10.8\,x^2 + 10.8\,x - 6.8 = 0$$

2.34 Se ha encontrado una simplificación* al algoritmo de Müller (véase algoritmo 2.6), y es

$$x_{i+1} = x_i - \frac{2\lambda_i}{1 + \sqrt{1 - 4\lambda_i + \mu_i}} \qquad i = 2, 3, 4 \ldots \qquad (1)$$

donde

$$\lambda_i = \frac{f_i}{w_i}\ ,\ \mu_i = \frac{f[x_i, x_{i-1}, x_{i-2}]}{w_i}$$

y

$$w_i = f[x_i, x_{i-1}] + (x_i - x_{i-1})\,f[x_i, x_{i-1}, x_{i-2}]$$

$$= f[x_i, x_{i-1}] + (f_i - f_{i-1})\frac{f[x_i, x_{i-1}, x_{i-2}]}{f[x_i, x_{i-1}]}$$

Para esta modificación el orden de convergencia está dado por

$$\epsilon_{i+1} \approx -\frac{f'''(x)}{6f'(x)}\,\epsilon_i\,\epsilon_{i-1}\,\epsilon_{i-2}$$

Resuelva las ecuaciones dadas en los problemas 2.17, 2.26 y 2.33 con este algoritmo.

2.35 Con la fórmula 1 del problema 2.34 y algunas consideraciones teóricas que se omiten por ser más bien tema del análisis numérico, se llega a modificaciones del método de la secante, con lo cual se consigue en éstas un orden de convergencia mayor de 2

---

*Hildebrand, B. Introduction to Numerical Analysis. 2nd Ed McGraw-Hill (1974) p. 580-581.

a) La primera modificación está dada por la ecuación

$$x_{i+1} = x_i - \frac{2\lambda_i}{1 + \sqrt{1 - 4\,\mu_i\,\lambda_i}}\,,\quad i = 1, 2, \ldots \tag{1}$$

pero ahora

$$\lambda_i = \frac{f_i}{f_i'}\,,\quad \mu_i = \frac{f_i' - f\,[x_i, x_{i-1}]}{(x_i - x_{i-1})\,f_i'}$$

y

$$w_i = f_i'$$

La interpretación geométrica de este método consiste en remplazar la función $f(x)$ en cierto intervalo con una parábola que pasa por el punto (x_{i-1}, f_{i-1}) y es tangente a la curva de $f(x)$ en (x_i, f_i). Para la ecuación 1 se tiene que

$$\epsilon_{i+1} \approx -\frac{f'(\bar{x})}{6f'(\bar{x})}\,\epsilon_i^2\,\epsilon_{i-1}$$

y se ha encontrado que es aproximadamente de orden 2.41

b) La segunda modificación está dada por la expresión

$$x_{i+1} = x_{i-1} + \frac{2\,(f_i/f_i')}{1 + \sqrt{1 - 2\,[f_i\,f_i''/(f_i')^2]}} \tag{2}$$

en ésta el orden de convergencia es 3, y se sabe que

$$\epsilon_{i+1} \approx -\frac{f'(\bar{x})}{6f'(\bar{x})}\,\epsilon_i^3$$

Aquí, la curva que reemplaza a $f(x)$ en cierto intervalo es una parábola que coincide con la curva de $f(x)$ en x_i y tiene la misma pendiente y curvatura que $f(x)$ en x_i. Resuelva las ecuaciones dadas en los problemas 2.17, 2.26 y 2.33, usando las modificaciones de los incisos (a) y (b).

c) De estas fórmulas pueden obtenerse otras más simples mediante aproximaciones. Por ejemplo, si f_i es pequeña, puede hacerse

$$\left(1 - 2\,\frac{f_i\,f_i''}{(f_i')^2}\right)^{1/2} = 1 - \frac{f_i\,f_i''}{(f_i')^2}$$

en la ecuación 2 y obtener la fórmula simplificada

$$x_{i+1} = x_i - \frac{f_i/f_i'}{1 - f_i\,f_i''/2\,(f_i')^2}$$

para la cual

$$\epsilon_{i+1} \approx \left[\left(\frac{f''(\bar{x})}{2f'(\bar{x})}\right)^2 - \frac{f'''(\bar{x})}{6f'(\bar{x})}\right]\,\epsilon_i^3$$

obsérvese que también es de tercer orden, pero sin raíz cuadrada. Esta fórmula se atribuye a Halley. Los métodos iterativos basados en esta expresión algunas veces se denominan métodos de Bailey o métodos de Lambert.

d) Si se aproxima

$$\left[1 - \frac{f_i f_i''}{2(f_i')^2} \right]^{-1} \approx 1 + \frac{f_i f_i''}{2(f_i')^2}$$

en la fórmula de Halley, se obtiene la iteración

$$x_{i+1} = x_i - \frac{f_i}{f_i'} \left[1 + \frac{f_i f_i''}{2(f_i')^2} \right] \tag{4}$$

con

$$\epsilon_{i+1} \approx \left[2 \left(\frac{f''(\bar{x})}{2f'(\bar{x})} \right)^2 - \frac{f'''(\bar{x})}{6f'(\bar{x})} \right] \epsilon_i^3$$

cuyo orden es 3 también y se llama fórmula de Chebyshev.

Resuelva las ecuaciones dadas en los problemas 2.17, 2.26 y 2.33, usando los algoritmos de Halley y Chebyshev cuando sean aplicables y compare los resultados obtenidos con los algoritmos de los incisos (a) y (b).

2.36 La ecuación de estado de Beattie-Bridgeman en su forma virial es

$$pV = RT + \frac{\beta}{V} + \frac{\gamma}{V^2} + \frac{\delta}{V^3}$$

donde

P = presión en *atm*
T = temperatura en K
V = volumen molar en *l/gmol*
R = Constante universal de los gases en *atm-l/ (gmol K)*
$\beta = R\,T\,B_0 - A_0 - R\,c/T^2$
$\gamma = -R\,T\,B_0\,b + A_0\,a - R\,B_0\,c/T^2$
$\delta = R\,B_0\,b\,c/T^2,y$
A_0, B_0, a, b, c = constantes particulares para cada gas.

Calcule el volumen molar *V* a 50 *atm* y 100 °C para los siguientes gases

Gas	A_0	a	B_0	b	$c \times 10^{-4}$
He	0.0216	0.05984	0.01400	0.000000	0.0040
H_2	0.1975	–0.00506	0.02096	–0.43590	0.0504
O_2	1.4911	0.02562	0.04624	0.004208	4.8000

2.37 La ecuación de estado de Redlich-Kwong es

$$\left[P + \frac{a}{T^{1/2} V(V + b)} \right] (V - b) = RT$$

donde

P = presión en *atm*
T = temperatura en K
V = volumen molar en *l/gmol*
R = constante univesal de los gases en *atm-l/(gmol K)*

$$a = 0.4278 \frac{R^2\, Tc^{2.5}}{Pc}, \quad b = 0.0867 \frac{R\,Tc}{Pc}$$

Calcule el volumen molar V a 50 *atm* y 100°C para los siguientes gases

Gas	Pc (atm)	Tc (K)
He	2.26	5.26
H_2	12.80	33.30
O_2	49.70	154.40

Compare los resultados obtenidos con los del problema 2.36.

2.38 Mediante la ecuación de estado de Van der Walls (véase ejercicio 2.1), encuentre el volumen molar V del CO_2 a 80 °C y 10 *atm*, utilizando los métodos de Newton-Raphson y de Richmond (véase Probl. 2.14).

2.39 Descomponga en fracciones parciales las siguientes funciones racionales

$$\text{a) } F(s) = \frac{52.5\,s\,(s+1)(s+1.5)(s+5)}{s^4 + 20.75\,s^3 + 92.6\,s^2 + 73.69\,s}$$

$$\text{b) } F(s) = \frac{10A}{s^3 + 101.4\,s^2 + 142.7\,s + 100}$$

$$\text{c) } F(s) = \frac{0.47\,K_G\,(s^3 + 4.149\,s^2 + 6.362\,s + 4.255)}{s^4 + 7\,s^3 + 11\,s^2 + 5\,s}$$

$$\text{d) } F(s) = \frac{100\,(s^2 + 3.4\,s + 2.8)}{s^5 + 10\,s^4 + 32\,s^3 + 38\,s^2 + 15\,s}$$

2.40 Una forma alterna para resolver el problema de vaporización instantánea (véase ejercicios 2.6) es

Tomando en cuenta $\Sigma\, x_i = 1$ y que $\Sigma\, y_i = 0\ \Sigma\, K_i x_i = 1$, puede escribirse

$$\frac{\Sigma\, K_i x_i}{\Sigma\, x_i} = 1 \qquad \text{(todas las sumatorias sobre } i \text{ son de 1 a } n\text{)}$$

o también

$$\ln \frac{\Sigma\, K_i x_i}{\Sigma\, x_i} = 0$$

Siguiendo la secuencia mostrada en el ejercicio 2.6, se llega a la expresión

$$\ln \frac{\sum \dfrac{K_i\, z_i}{1 + \psi\,(K_i - 1)}}{\sum \dfrac{z_i}{1 + \psi\,(K_i - 1)}} = 0$$

Utilice el método de posición falsa y los datos del ejercicio 2.6 para resolver esta última ecuación.

2.41 Para el cálculo de la temperatura de burbuja de una mezcla multicomponente a la presión total P se utiliza la ecuación

$$f(T) = \sum_{i=1}^{n} K_i x_i - 1 = 0 \qquad (1)$$

donde x_i y K_i, $i = 1, 2, \ldots, n$ son la fracción mol en la fase líquida y la relación de equilibrio del componente i, respectivamente, y T (la raíz de la ecuación 1) es la temperatura de burbuja.

Determine la temperatura de burbuja a 10 *atm* de presión total de una mezcla cuya composición en la fase líquida es 45% *mol* de n-butano, 30% *mol* de n-pentano y 25% *mol* de n-hexano. Los valores de K_i a 10 *atm* son

Componente	K (T) con T en °C para $35 \leq T \leq 205$ °C
n-butano	$-0.17809 + 1.2479 \times 10^{-2}\ T + 3.7159 \times 10^{-5}\ T^2$
n-pentano	$0.13162 - 1.9367 \times 10^{-3}\ T + 7.1373 \times 10^{-5}\ T^2$
n-hexano	$0.13985 - 3.8690 \times 10^{-3}\ T + 5.5604 \times 10^{-5}\ T^2$

2.42 Para el cálculo de la temperatura de rocío de una mezcla multicomponente a la presión total P se utiliza la ecuación

$$f(T) = \sum_{i=1}^{n} \frac{y_i}{K_i} - 1 = 0 \qquad (1)$$

donde y_i y K_i , $i = 1, 2, \ldots, n$ son la fracción *mol* en la fase vapor y la relación de equilibrio del componente i, respectivamente, y T (la raíz de la ecuación 1) es la temperatura de rocío.

Determine la temperatura de rocío a 10 *atm* de presión total de una mezcla cuya composición en la fase líquida es 45% *mol* de n-butano, 30% *mol* de n-pentano y 25% *mol* de n-hexano. Los valores de K_i a 10 *atm* se dan en el ejercicio 2.41

2.43 Para obtener la temperatura de burbuja de una solución líquida de CCl_4 y CF_4 en equilibrio con su vapor, se llegó a la ecuación

$$760 = 0.75\ [\ 10^{(6.898 - 1221.8\ /\ (T + 227.4))}] + 0.25\ [\ 10^{(6.195 - 376.71\ /\ (T + 241.2))}]$$

Aplicando un método iterativo de dos puntos, encuentre la temperatura de burbuja T con una aproximación de 10^{-2} aplicado a $f(T)$.

2.44 En la solución de ecuaciones diferenciales ordinarias con coeficientes constantes, es necesario resolver la "ecuación auxiliar asociada", que resulta ser un polinomio cuyo grado es igual al orden de la ecuación diferencial. Así, si la ecuación diferencial está dada por

$$y^{IV} + 2y'' - 8y = 0 \qquad (1)$$

la ecuación auxiliar asociada es

$$m^4 + 2m^2 - 8 = 0$$

cuyas cuatro raíces : m_1, m_2, m_3 y m_4 se emplean de la siguiente manera

$$y = c_1 e^{m_1 x} + c_2 e^{m_2 x} + c_3 e^{m_3 x} + c_4 e^{m_4 x}$$

para dar la solución general de la ecuación 1.

Encuentre la solución general de la ecuación 1 y de las siguientes ecuaciones diferenciales

$$y^{VI} + 2\,y^{IV} + y'' = 0$$

$$y''' - 4\,y'' + 4\,y' = 0$$

2.45 La ecuación 4 del ejercicio 2.8 se aplica para calcular la ΔTm, cuando

$$TC_1 - TC_2 \neq TH_2 - TC_1$$

Cuando el gradiente $TH_1 - TC_2$ es muy cercano al gradiente $TH_2 - TC_1$ se deberá utilizar la siguiente expresión para el cálculo de ΔTm

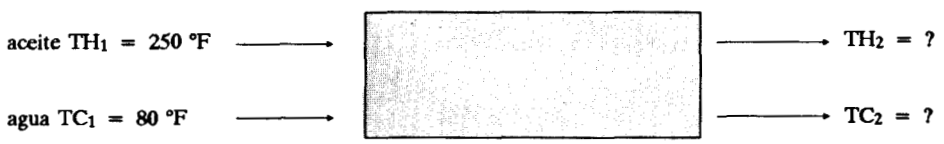
$$\Delta T_m = \frac{(TH_1 - TC_2) + (TH_2 - TC_1)}{2}$$

Modifique el programa 2.6 del ejercicio 2.8 de modo que se utilice la ΔTm dada arriba cuando

$$|\,(TH_1 - TC_2) - (TH_2 - TC_1)\,| < 10^{-2}$$

y la ecuación (4) del ejercicio 2.8 en caso contrario.

2.46 Si el cambiador de calor del ejercicio 2.8, se opera en paralelo, esto es



Encuentre TH_2 y TC_2 en estas nuevas condiciones de operación.

2.47 Suponga que el fenómeno de la transmisión de calor en un cierto material obedece en forma aproximada al modelo

$$T = T_0 + \frac{q}{k}\left(\beta\left(\frac{\alpha T}{\pi}\right)^{1/2} e^{-x^2/(4\alpha t)}\right)$$

Calcule el tiempo requerido para que la temperatura a la distancia x alcance un valor dado. Use la siguiente información

$$T_0 = 25\ °C;\ q = 300\ BTU/h\ ft^2; \qquad k = 1\ BTU/h\ ft^2\ °F$$

$$\alpha = 0.04\ ft^2/h;\ x = 1\ ft; \qquad T = 120\ °F$$

$$\beta = 2\ \frac{°F ft}{h^{1/2}}\ °C^{1/2}$$

2.48 El factor de fricción f para fluidos pseudoplásticos que siguen el modelo de Ostwald-De Waele se calcula mediante la siguiente ecuación

$$\frac{1}{f} = \frac{4}{n^{0.75}} \log\left(\mathrm{Re}\, f^{1-0.5n}\right) - \frac{0.4}{n^{1.2}}$$

Encuentre el factor de fricción f, si se tiene un número de Reynolds Re de 6000 y un valor de $n = 0.4$.

2.49 La siguiente relación entre el factor de fricción f y el número de Reynolds Re se cumple cuando hay flujo turbulento de un fluido en un tubo liso

$$\frac{1}{f} = -0.4 + 1.74 \ln\left(\mathrm{Re}\sqrt{f}\right)$$

Construya una tabla de valores de f correspondientes a números de Reynolds de 10^4 hasta 10^6 con intervalos de 10^4.

2.50 Para determinar la constante de nacimientos de una población se necesita calcular λ en la siguiente ecuación

$$1.564 \times 10^6 = 10^6 e^{\lambda} + \frac{0.435 \times 10^6}{\lambda}\left(e^{\lambda} - 1\right)$$

con una aproximación de 10^{-3}

CAPÍTULO 3

MATRICES Y SISTEMAS DE ECUACIONES LINEALES

EN ESTE CAPÍTULO se exponen las dos ideas que sustentan, en los métodos numéricos, las soluciones de los sistemas de ecuaciones lineales: Eliminación de Gauss e iteración de Jacobi con sus variantes más utilizadas.

INTRODUCCIÓN

La solución de sistemas de ecuaciones lineales es un tema clásico de las matemáticas, rico en ideas y conceptos y de gran utilidad en ramas del conocimiento tan diversas como economía, biología, física, psicología, etc. La resolución de sistemas casi de cualquier número de ecuaciones (10, 100, 1000, etc.) es una realidad hoy en día gracias a las computadoras, lo cual proporciona un atractivo especial a las ténicas de solución directas e iterativas: Su programación, la cuenta de los cálculos necesarios, la propagación de errores, etcétera.

Sin embargo, todo lo anterior requiere una revisión de los conceptos básicos sobre matrices, ortogonalización de vectores y la existencia y unicidad de las soluciones; por lo tanto, estos conceptos dan inicio al capítulo.

SECCIÓN 3.1 MATRICES

Una matriz es un conjunto de elementos ordenados en filas y columnas como

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}$$

Los elementos a_{ij} son números reales o complejos, o funciones de una o varias variables. En este libro sólo se tratarán matrices cuyos elementos son números reales.

Para denotar matrices se utilizarán las primeras letras mayúsculas del alfabeto en *cursivas A, B, C,* etc. Cuando se hace referencia a una matriz es conveniente especificar su número de filas y columnas. Así, la expresión A de $m \times n$, indica que se trata de una matriz de m filas y n columnas o de $m \times$ n elementos. A "$m \times n$" se le conoce como las dimensiones de A. Si el número de filas y de columnas es el mismo; esto es $m = n$, se tiene una matriz cuadrada de orden n o simplemente una matriz de orden n.

Para ciertas demostraciones es más conveniente la notación $[a_{ij}]$, $[b_{ij}]$, etc., en lugar de *A, B,* etcétera.

Dos matrices son iguales cuando tienen el mismo número de filas y columnas (las mismas dimensiones) y, además, los elementos correspondientes son iguales.

Por ejemplo, las matrices

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \quad \text{y} \quad B = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$$

son de orden tres y tienen los mismos elementos. Aún así son distintas, ya que los elementos correspondientes no son todos iguales. El elemento de la segunda fila y la segunda columna de A, a_{22} es 5 y el correspondiente de B, b_{22} es 5; pero el elemento de la segunda fila y la primera columna de A, a_{21} es 4 y el correspondiente a B, b_{21} es 2.

Operaciones elementales con matrices y sus propiedades

Se definirán dos operaciones en el conjunto establecido de las matrices.

Suma de matrices

Para sumar dos matrices A y B han de ser de las mismas dimensiones; si esto es cierto, la suma es una matriz C de iguales dimensiones que A y que B, y sus elementos se obtienen sumando los elementos correspondientes de A y B. Para mayor claridad

$$\begin{matrix} A & + & B & = & C \end{matrix}$$

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ & \cdot & & \cdot \\ & \cdot & & \cdot \\ & \cdot & & \cdot \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \cdots & a_{mn} + b_{mn} \end{bmatrix}$$

$$= \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{bmatrix} \tag{3.1}$$

o también

$$[a_{ij}] + [b_{ij}] = [a_{ij} + b_{ij}] = [c_{ij}] \tag{3.2}$$

Ejemplo 3.1

Sumar las matrices

$$\begin{bmatrix} 4 & 8.5 & -3 \\ 2 & -1.3 & 7 \end{bmatrix} \quad \text{y} \quad \begin{bmatrix} -1 & 2 & -4 \\ 5 & 8 & 3 \end{bmatrix}$$

SOLUCIÓN

$$\underset{2 \times 3}{\begin{bmatrix} 4 & 8.5 & -3 \\ 2 & -1.3 & 7 \end{bmatrix}} + \underset{2 \times 3}{\begin{bmatrix} -1 & 2 & -4 \\ 5 & 8 & 3 \end{bmatrix}} = \underset{2 \times 3}{\begin{bmatrix} 4-1 & 8.5+2 & -3-4 \\ 2+5 & -1.3+8 & 7+3 \end{bmatrix}} = \underset{2 \times 3}{\begin{bmatrix} 3 & 10.5 & -7 \\ 7 & 6.7 & 10 \end{bmatrix}}$$

La conmutatividad y asociatividad de la suma de matrices son propiedades heredadas de las propiedades de la suma de los números reales. Así, la conmutatividad puede verse claramente en la ecuación 3.1, ya que

$$a_{ij} + b_{ij} = b_{ij} + a_{ij} = c_{ij}$$

donde a_{ij} representa un elemento cualquiera de A y b_{ij} su correspondiente en B. Por tanto, es cierto que

$$A + B = B + A = C$$

De igual manera puede verse la asociatividad

$$(a_{ij} + b_{ij}) + d_{ij} = a_{ij} + (b_{ij} + d_{ij})$$

o bien

$$(A + B) + D = A + (B + D)$$

donde D es una matriz de las mismas dimensiones que A y que B.

Además, sí se denota con O a la matriz cuyos elementos son todos cero (matriz cero); es decir,

$$O = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

y por $-A$ la matriz cuyos elementos son los mismos que A, pero de signo contrario

$$-A = \begin{bmatrix} -a_{11} & -a_{12} & \cdots & -a_{1n} \\ -a_{21} & -a_{22} & \cdots & -a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ -a_{m1} & -a_{m2} & \cdots & -a_{mn} \end{bmatrix}$$

se tiene

$$A + O = A, \tag{3.3}$$

$$A + (-A) = O \tag{3.4}$$

A partir de la ecuación 3.4, puede definirse la resta entre A y B como

$$A + (-B)$$

o más simple

$$A - B$$

Producto de matrices por un escalar

Así como se ha definido la suma de matrices, también se puede formar el producto de un número real $\propto$ y una matriz A. El resultado, denotado por $\propto A$, es la matriz cuyos elementos son los componentes de A multiplicados por $\propto$. Así, se tiene

$$\propto A = \propto \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} \propto a_{11} & \propto a_{12} & \cdots & \propto a_{1n} \\ \propto a_{21} & \propto a_{22} & \cdots & \propto a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \propto a_{m1} & \propto a_{m2} & \cdots & \propto a_{mn} \end{bmatrix} \tag{3.5}$$

o bien

$$\propto [a_{ij}] = [\propto a_{ij}] \tag{3.6}$$

Ejemplo 3.2

Multiplique la matriz $\begin{bmatrix} 5.8 & -2.3 & 2 \\ 4 & 7.2 & 10 \\ 43 & -13 & 5 \end{bmatrix}$ por 2.

SOLUCIÓN

$$2 \begin{bmatrix} 5.8 & -2.3 & 2 \\ 4 & 7.2 & 10 \\ 43 & -13 & 5 \end{bmatrix} = \begin{bmatrix} 2(5.8) & 2(-2.3) & 2(2) \\ 2(4) & 2(7.2) & 2(10) \\ 2(43) & 2(-13) & 2(5) \end{bmatrix} = \begin{bmatrix} 11.6 & -4.6 & 4 \\ 8 & 14.4 & 20 \\ 86 & -26 & 10 \end{bmatrix}$$

Las principales propiedades algebraicas de esta multiplicación son

$$\alpha\ (A + B) = \alpha\ A + \alpha B, \quad \text{distributividad respecto de la suma de matrices} \tag{3.7}$$

$$(\alpha + \beta)\ A = \alpha A + \beta\ A, \quad \text{distributividad respecto de la suma de escalares} \tag{3.8}$$

$$(\alpha\ \beta)\ A = \alpha\ (\beta\ A), \quad \text{asociatividad} \tag{3.9}$$

$$1\ A = A, \tag{3.10}$$

donde $\propto$ y β son dos escalares cualesquiera y A y B dos matrices sumables (con igual número de filas e igual número de columnas).

Las ecuaciones 3.7 a 3.10, se comprueban con facilidad a partir de las definiciones de suma de matrices y multiplicación por un escalar. Sólo se demostrará la 3.9; las otras quedan como ejercicio para el lector.

De la definición (Ec. 3.5), aplicada al lado izquierdo de la ecuación 3.9

$$(\alpha\ \beta)\ A = \begin{bmatrix} (\propto \beta)\ a_{11} & (\propto \beta)\ a_{12} & \cdots & (\propto \beta)\ a_{1n} \\ (\propto \beta)\ a_{21} & (\propto \beta)\ a_{22} & \cdots & (\propto \beta)\ a_{2n} \\ \vdots & \vdots & & \vdots \\ (\propto \beta)\ a_{m1} & (\propto \beta)\ a_{m2} & \cdots & (\propto \beta)\ a_{mn} \end{bmatrix}$$

De la asociatividad de la multiplicación de los números reales se tiene

$$(\alpha\ \beta)\ A = \begin{bmatrix} \propto (\beta\ a_{11}) & \propto (\beta\ a_{12}) & \cdots & \propto (\beta\ a_{1n}) \\ \propto (\beta\ a_{21}) & \propto (\beta\ a_{22}) & \cdots & \propto (\beta\ a_{2n}) \\ \vdots & \vdots & & \vdots \\ \propto (\beta\ a_{m1}) & \propto (\beta\ a_{m2}) & \cdots & \propto (\beta\ a_{mn}) \end{bmatrix}$$

Al aplicar la ecuación 3.5 en sentido inverso dos veces

$$(\alpha\ \beta)\ A = \alpha \begin{bmatrix} \beta\ a_{11} & \beta\ a_{12} & \cdots & \beta\ a_{1n} \\ \beta\ a_{21} & \beta\ a_{22} & \cdots & \beta\ a_{2n} \\ \vdots & \vdots & & \vdots \\ \beta\ a_{m1} & \beta\ a_{m2} & \cdots & \beta\ a_{mn} \end{bmatrix}$$

$$(\alpha\beta)\,\mathrm{A} = \alpha \left[\beta \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \right]$$

se llega al lado derecho de la ecuación 3.9, con lo cual concluye la demostración.

Multiplicación de matrices

Dos matrices A y B son conformes en ese orden (primero A y después B), si A tiene el mismo número de columnas que B tiene de filas.

Se definirá la multiplicación sólo para matrices conformes. Dada una matriz A de $m \times n$ y una matriz B de $n \times p$, el producto es una matriz C de $m \times p$ cuyo elemento general c_{ij} se obtiene por la suma de los productos de los elementos de i - ésima fila de A y la j - ésima columna de B. Si

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{i1} & a_{i2} & \cdots & a_{in} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \quad \text{y} \quad B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1j} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2j} & \cdots & b_{2p} \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ b_{n1} & b_{n2} & \cdots & b_{nj} & \cdots & b_{np} \end{bmatrix}$$

$$A\,B = C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1j} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2j} & \cdots & c_{2p} \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ c_{i1} & c_{i2} & \cdots & c_{ij} & \cdots & c_{ip} \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ c_{m1} & c_{m2} & \cdots & c_{mj} & \cdots & c_{mp} \end{bmatrix}$$

donde

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{in}b_{nj}$$

o bien

$$c_{ij} = \sum_{k=1}^{n} a_{ik}\, b_{kj} \text{ para } i = 1, 2, \ldots, m \text{ y } j = 1, 2, \ldots, p$$

Ejemplo 3.3

Multiplicar las matrices $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & -4 & -5 \end{bmatrix}$ y $B = \begin{bmatrix} 0 & 1 & -2 \\ -1 & 2 & 3 \\ 4 & 2 & 1 \end{bmatrix}$

SOLUCIÓN

$$\underset{A}{\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & -4 & -5 \end{bmatrix}} \underset{B}{\begin{bmatrix} 0 & 1 & -2 \\ -1 & 2 & 3 \\ 4 & 2 & 1 \end{bmatrix}} = \underset{C}{\begin{bmatrix} 0-2+12 & 1+4+6 & -2+6+3 \\ 0-3+16 & 2+6+8 & -4+9+4 \\ 0+4-20 & 3-8-10 & -6-12-5 \end{bmatrix}} = \begin{bmatrix} 10 & 11 & 7 \\ 13 & 16 & 9 \\ -16 & -15 & -23 \end{bmatrix}$$

En orden inverso

$$\underset{B}{\begin{bmatrix} 0 & 1 & -2 \\ -1 & 2 & 3 \\ 4 & 2 & 1 \end{bmatrix}} \underset{A}{\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & -4 & -5 \end{bmatrix}} = \underset{C}{\begin{bmatrix} 0+2-6 & 0+3+8 & 0+4+10 \\ -1+4+9 & -2+6-12 & -3+8-15 \\ 4+4+3 & 8+6-4 & 12+8-5 \end{bmatrix}} = \begin{bmatrix} -4 & 11 & 14 \\ 12 & -8 & -10 \\ 11 & 10 & 15 \end{bmatrix}$$

Obsérvese que $A\,B \neq B\,A$; es decir, la multiplicación de matrices no es conmutativa. Este hecho deberá tenerse siempre en cuenta al multiplicar matrices.

A continuación se verán las propiedades de distributividad y asociatividad del producto de matrices.

$$\boxed{A\,(B + C) = A\,B + A\,C} \tag{3.11}$$

$$\boxed{(A\,B)\,C = A\,(B\,C)} \tag{3.12}$$

Con la notación de sumatoria se comprobará la ecuación 3.11; la 3.12 queda como ejercicio para el lector.

Demostración de la ecuación 3.11. Sea e_{ij} un elemento cualquiera de la matriz producto $A\,B$, esto es

$$e_{ij} = \sum_{k=1}^{n} a_{ik}\ \ b_{kj}$$

y d_{ij} el elemento correspondiente del producto $A\,C$

$$d_{ij} = \sum_{k=1}^{n} a_{ik}\ \ c_{kj}$$

Al sumarlos se obtiene el elemento correspondiente del lado derecho de la ecuación 3.11

$$e_{ij} + d_{ij} = \sum_{k=1}^{n} a_{ik}\, b_{kj} + \sum_{k=1}^{n} a_{ik}\, c_{kj} = \sum_{k=1}^{n} a_{ik}\, (b_{kj} + c_{kj}),$$

el cual es igual al elemento de la i-ésima fila y la j-ésima columma del lado izquierdo de la ecuación 3.11, con lo que finaliza la demostración.

A continuación se da el algoritmo para multiplicar matrices.

ALGORITMO 3.1 Multiplicación de matrices

Para multiplicar las matrices A y B, proporcionar los

DATOS: Número de filas y columnas de A y B; N, M, N1, M1, respectivamente, y sus elementos.

RESULTADOS: La matriz producto C de dimensiones N×M1 o el mensaje "LAS MATRICES A Y B NO PUEDEN MULTIPLICARSE".

PASO 1 Si M = N1 continuar, de otro modo IMPRIMIR "LAS MATRICES A Y B NO SE PUEDEN MULTIPLICAR" y TERMINAR.
PASO 2. Hacer I = 1
PASO 3 Mientras I ≤ N, repetir los pasos 4 a 12.
PASO 4 Hacer J = 1
PASO 5. Mientras J ≤ M1, repetir los pasos 6 a 11.
PASO 6. Hacer C(I, J) = 0
PASO 7. Hacer K = 1
PASO 8 Mientras K ≤ M, repetir los pasos 9 y 10.
PASO 9. Hacer C(I, J) = C(I, J) + A(I, K) * B(K, J)
PASO 10. Hacer K = K + 1
PASO 11. Hacer J = J + 1
PASO 12. Hacer I = I + 1
PASO 13. IMPRIMIR las matrices A, B y C y TERMINAR.

Ejemplo 3.4

Elaborar un programa en PASCAL para multiplicar matrices, utilizando el algoritmo 3.1

SOLUCIÓN

Ver el programa 3.1 del disco.

Sugerencia: Este material puede complementarse e incluso enriquecerse si se cuenta con un pizarrón electrónico, por ejemplo el Math-CAD, ya que permite, una vez entendida la mecánica de las operaciones matriciales, averiguar sus propiedades e incluso motivar algunas demostraciones. En adelante se hará referencia al Math-CAD, pero puede usarse un software disponible equivalente.

Matrices especiales

En una matriz cuadrada A, el conjunto de elementos en donde el primero y el segundo subíndices son iguales —es decir $i = j$— forman la diagonal principal. Por ejemplo, en la matriz de 4×4 que se da a continuación, los elementos dentro de la banda constituyen la **diagonal principal.**

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

Una matriz de orden n con todos sus elementos debajo de la diagonal principal iguales a cero se llama **matriz triangular superior.** Si todos los elementos por encima de la diagonal principal son cero en una matriz, entonces será una **matriz triangular inferior;** en caso de que una matriz tenga únicamente ceros arriba y abajo de la diagonal principal, se tiene una **matriz diagonal,** y, si en particular, todos los elementos de la diagonal son 1, entonces se obtiene la **matriz unitaria** o **matriz identidad.**

Matriz triangular superior

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22} & a_{23} & \cdots & a_{2n} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & & \cdots & a_{n-1n} \\ 0 & 0 & 0 & \cdots & a_{nn} \end{bmatrix}$$

Matriz tiangular inferior

$$\begin{bmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ a_{21} & a_{22} & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ & & & & 0 \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{n-1n} \quad a_{nn} \end{bmatrix}$$

Matriz diagonal

$$\begin{bmatrix} a_{11} & 0 & 0 & \cdots & 0 \\ 0 & a_{22} & 0 & \cdots & 0 \\ 0 & 0 & a_{33} & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & a_{nn} \end{bmatrix}$$

Matriz unitaria o identidad

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

A continuación se dan algunos casos particulares de matrices cuadradas especiales

Triangular superior	Triangular inferior	Diagonal	Unitaria
$\begin{bmatrix} 1 & 3 & -4 \\ 0 & 6 & 2 \\ 0 & 0 & -5 \end{bmatrix}$	$\begin{bmatrix} 4 & 0 & 0 \\ -2 & -1 & 0 \\ 7 & 5 & 3 \end{bmatrix}$	$\begin{bmatrix} 2 & 0 & 0 \\ 0 & -6 & 0 \\ 0 & 0 & 8 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

La matriz unitaria se denota, independientemente de su orden, como I.

Dada una matriz A de $m \times n$, la matriz de $n \times m$ que se obtiene de A intercambiando sus filas por sus columnas se denomina **matriz transpuesta** de A y se denota por A^T. Esto es

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad A^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{1n} & a_{2n} & & a_{mn} \end{bmatrix}$$

Ejemplo 3.5

Dada la matriz A, encuentre su transpuesta.

$$A = \underset{3 \times 5}{\begin{bmatrix} 1 & 0 & 3 & 4 & 1 \\ 2 & 3 & 5 & 7 & 9 \\ 8 & 6 & 2 & 5 & 0 \end{bmatrix}}$$

SOLUCIÓN

$$A^T = \underset{5 \times 3}{\begin{bmatrix} 1 & 2 & 8 \\ 0 & 3 & 6 \\ 3 & 5 & 2 \\ 4 & 7 & 5 \\ 1 & 9 & 0 \end{bmatrix}}$$

Una matriz cuadrada para la que $A^T = A$, recibe el nombre de **matriz simétrica.** Por ejemplo

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 4 \end{bmatrix} \quad \text{y} \quad A^T = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 4 \end{bmatrix}$$

son iguales, y por tanto A es simétrica.

Si A y B son dos matrices cuadradas, tales que

$$AB = I = BA ,$$

se dice que B es la **inversa** de A y se representa generalmente como A^{-1} .

Ejemplo 3.6

Demuestre que B es la inversa de A, si

$$A = \begin{bmatrix} 1 & 3 & 3 \\ 1 & 4 & 3 \\ 1 & 3 & 4 \end{bmatrix} \quad \text{y} \quad B = \begin{bmatrix} 7 & -3 & -3 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

SOLUCIÓN

$$AB = \begin{bmatrix} 1 & 3 & 3 \\ 1 & 4 & 3 \\ 1 & 3 & 4 \end{bmatrix} \begin{bmatrix} 7 & -3 & -3 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I$$

Por tanto

$$A^{-1} = B$$

En particular si A es diagonal; es decir

$$A = \begin{bmatrix} a_{11} & 0 & 0 & \dots & 0 \\ 0 & a_{22} & 0 & \dots & 0 \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ 0 & 0 & & & a_{nn} \end{bmatrix} , \text{ entonces } A^{-1} = \begin{bmatrix} 1/a_{11} & 0 & 0 & \dots & 0 \\ 0 & 1/a_{22} & 0 & \dots & 0 \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & & \cdot \\ 0 & 0 & & \dots & 1/a_{nn} \end{bmatrix}$$

La demostración se deja como ejercicio para el lector.

Es importante señalar que no todas las matrices tienen inversa. Si una matriz la tiene, se dice también que es **no singular**, y **singular** en caso contrario. Más adelante se ven métodos para encontrar la inversa de una matriz.

Matriz permutadora

Una matriz cuyos elementos son ceros y unos y donde sólo hay un uno por cada fila o columna, se conoce como **matriz permutadora** o **intercambiadora**; por ejemplo, las matrices

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

son casos particulares de matrices intercambiadoras.

El efecto de multiplicar una matriz permutadora P por una matriz A en ese orden es intercambiar las filas de A; al multiplicar en orden inverso, se intercambian las columnas de A.

Ejemplo 3.7

Multiplique la matriz A del ejemplo 3.6 por la matriz permutadora P de 3×3 dada arriba.

SOLUCIÓN

a) Cálculo de $P\ A$:

$$\underset{P}{\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}} \quad \underset{A}{\begin{bmatrix} 1 & 3 & 3 \\ 1 & 4 & 3 \\ 1 & 3 & 4 \end{bmatrix}} \quad = \quad \underset{C}{\begin{bmatrix} 1 & 4 & 3 \\ 1 & 3 & 3 \\ 1 & 3 & 4 \end{bmatrix}}$$

Obsérvese que la matriz producto C es la matriz A con la primera y segunda filas intercambiadas

b) Cálculo de $A\ P$

$$\underset{A}{\begin{bmatrix} 1 & 3 & 3 \\ 1 & 4 & 3 \\ 1 & 3 & 4 \end{bmatrix}} \quad \underset{P}{\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}} \quad = \quad \underset{D}{\begin{bmatrix} 3 & 1 & 3 \\ 4 & 1 & 3 \\ 3 & 1 & 4 \end{bmatrix}}$$

Obsérvese que la matriz producto D es la matriz A con la primera y segunda columnas intercambiadas.

La matriz identidad es un caso particular de matriz permutadora y su efecto es dejar igual la matriz por la que se multiplica (ya sea por la derecha o por la izquierda). Este hecho, junto con el ejemplo 3.7, manifiesta que cuando aparece un 1 en la diagonal principal de una matriz permutadora, la fila o columna correspondiente de la matriz por la que se multiplique no sufre cambio alguno.

Véase que hay un 1 en la posición (3,3) de la matriz P y que la fila 3 y la columna 3 de A no sufrieron intercambio en los incisos (*a*) y (*b*), respectivamente, en el ejemplo 3.7.

Ejemplo 3.8

Sin multiplicar diga qué efecto tendrá sobre una matriz cualquiera A de 4×4 la siguiente matriz

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

SOLUCIÓN

Análisis de la multiplicación PA. Los unos en las posiciones (1,1) y (3,3) indican que las filas 1 y 3 de A no sufrirán efecto alguno. Por otro lado, los unos de la segunda y cuarta filas cuyas posiciones son (2,4) y (4,2) indican que las filas 2 y 4 de A se intercambiarán (nótese que en el ejemplo 3.7, los unos fuera de la diagonal ocupan las posiciones (1,2) y (2,1) y las filas 1 y 2 se intercambian).

El lector fácilmente puede generalizar estos resultados.

SECCIÓN 3.2 VECTORES

Las matrices donde $m > 1$ y $n = 1$ (es decir, están formadas por una sola columna) son llamadas matrices columna o vectores. De igual manera, si $m = 1$ y $n > 1$, se tiene una matriz fila o vector. Los vectores se denotarán con las letras minúsculas en negritas: **a, b, x,** etc. En estos casos no será necesaria la utilización de doble subíndice para la identificación de sus elementos y un vector **x** de m elementos (en columna) queda simplemente como

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_m \end{bmatrix}$$

Un vector **y** de n elementos (en fila) queda como

$$\mathbf{y} = [\, y_1 \; y_2 \cdots y_n \,]$$

Por ejemplo, los siguientes vectores están en columna

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 3 \\ 1 \\ 0 \\ 5 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

y estos en fila

$$[0\ 1\ 0],\ [3\ 5\ 7\ 2],\ [0\ 0\ 0\ 0\ 0].$$

Obsérvese que si se tiene un vector columna, la transpuesta será un vector fila y viceversa.

$$\text{Dado } \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_m \end{bmatrix}, \qquad \mathbf{x}^T = [x_1\ x_2\ \ldots\ x_m]$$

Ejemplo 3.9

Obtener la transpuesta de los vectores columna y fila dados arriba.

SOLUCIÓN

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}^T = [1\,0\,0], \qquad \begin{bmatrix} 3 \\ 1 \\ 0 \\ 5 \end{bmatrix}^T = [3\,1\,0\,5]$$

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}^T = [0\,0\,0\,0\,0]$$

$$[0\,1\,0]^T = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad [3\,5\,7\,2]^T = \begin{bmatrix} 3 \\ 5 \\ 7 \\ 2 \end{bmatrix}$$

$$[0\,0\,0\,0\,0]^T = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Como en el texto resulta generalmente difícil expresar un vector en columna, se usará algunas veces su transpuesta.

Multiplicación de vectores

Dado que los vectores son sólo casos particulares de las matrices, siguen las mismas reglas de multiplicación que éstas. Sea por ejemplo $\mathbf{a} = [a_1\ a_2\ \ldots\ a_n]$ y $\mathbf{b}^T = [b_1\ b_2\ \ldots b_n]$, el producto **a b** es

$$\underset{1\times n}{\mathbf{a}}\ \mathbf{b} = \underset{1\times n}{[a_1\ a_2\ \ldots\ a_n]} \underset{n\times 1}{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}} = \underset{1\times 1}{a_1b_1 + a_2b_2 + \ldots + a_nb_n}$$

El producto de **a** por **b** es el número real $a_1b_1 + a_2b_2 + \ldots + a_nb_n$, que también puede verse como una matriz de 1×1.

Multiplicando en orden inverso

$$\underset{n\times 1}{\mathbf{b}}\ \mathbf{a} = \underset{n\times 1}{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}} \underset{1\times n}{[a_1\ a_2\ \ldots\ a_n]} = \underset{n\times n}{\begin{bmatrix} b_1a_1 & b_1a_2 & \ldots & b_1a_n \\ b_2a_1 & b_2a_2 & \ldots & b_2a_n \\ \vdots & \vdots & & \vdots \\ b_na_1 & b_na_2 & \ldots & b_na_n \end{bmatrix}}$$

se obtiene una matriz de $n \times n$.

Ejemplo 3.10

Dados $\mathbf{a} = [1\ 5\ 7]$ y $\mathbf{b}^T = [0\ -2\ 3]$, obtener **ab** y **ba**

SOLUCIÓN

$$\underset{1\times 3}{\mathbf{a}}\ \mathbf{b} = \underset{1\times 3}{[1\ 5\ 7]} \underset{3\times 1}{\begin{bmatrix} 0 \\ -2 \\ 3 \end{bmatrix}} = \underset{1\times 1}{1(0) + 5(-2) + 7(3)} = 11$$

y

$$\mathbf{b}\ \mathbf{a} = \begin{bmatrix} 0 \\ -2 \\ 3 \end{bmatrix} [1\ 5\ 7] = \begin{bmatrix} 0(1) & 0(5) & 0(7) \\ -2(1) & -2(5) & -2(7) \\ 3(1) & 3(5) & 3(7) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ -2 & -10 & -14 \\ 3 & 15 & 21 \end{bmatrix}$$

Puede multiplicarse también un vector por una matriz y viceversa si las dimensiones son adecuadas.

Ejemplo 3.11

Multiplique el vector $\mathbf{a} = [1 \ -2 \ 3]$ por la matriz $B = \begin{bmatrix} 0 & 4 & 3 \\ -1 & 8 & 2 \\ 3 & 1 & 5 \end{bmatrix}$

SOLUCIÓN

$$\underset{}{\mathbf{a}} \qquad \underset{}{B} \qquad = \qquad \underset{}{\mathbf{c}}$$

$$\underset{1 \times 3}{[1 \ -2 \ 3]} \quad \underset{3 \times 3}{\begin{bmatrix} 0 & 4 & 3 \\ -1 & 8 & 2 \\ 3 & 1 & 5 \end{bmatrix}} = \underset{1 \times 3}{[11 \ -9 \ 14]}$$

los elementos de **c** se calculan como

$$\begin{aligned} 1(0) + (-2)(-1) + 3(3) &= 11 \\ 1(4) + (-2)(8) + 3(1) &= -9 \\ 1(3) + (-2)(2) + 3(5) &= 14 \end{aligned}$$

Efectuar la multiplicación en orden inverso (B **a**) no es posible, por no ser conformes en ese orden. En cambio, sí puede multiplicarse B por algún vector columna **d** de tres elementos. Así

$$\underset{3 \times 3}{\begin{bmatrix} 0 & 4 & 3 \\ -1 & 8 & 2 \\ 3 & 1 & 5 \end{bmatrix}} \quad \underset{3 \times 1}{\begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}} = \begin{bmatrix} 0(1) + 4(0) + 3(2) \\ -1(1) + 8(0) + 2(2) \\ 3(1) + 1(0) + 5(2) \end{bmatrix} = \underset{3 \times 1}{\begin{bmatrix} 6 \\ 3 \\ 13 \end{bmatrix}}$$

Producto punto de vectores

Definición. Dados dos vectores **a** y **b** con igual número de elementos, por ejemplo n, su producto punto (o escalar), denotado por **a** · **b**, es un número real obtenido de la siguiente manera

$$\mathbf{a} \cdot \mathbf{b} = \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_n \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix} = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n \qquad (3.13)$$

Ejemplo 3.12

Si $\mathbf{a} = \begin{bmatrix} 2 \\ 1 \\ 6 \end{bmatrix}$ y $\mathbf{b} = \begin{bmatrix} -3 \\ 0 \\ 2.5 \end{bmatrix}$, obtenga el producto punto.

SOLUCIÓN

$$\mathbf{a} \cdot \mathbf{b} = 2(-3) + 1(0) + 6(2.5) = 9$$

Este producto punto así definido tiene las siguientes propiedades

a) $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$ conmutatividad (3.14)

b) $(\mathbf{a} + \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot \mathbf{c} + \mathbf{b} \cdot \mathbf{c}$ distributividad (3.15)

c) $(\alpha\, \mathbf{a}) \cdot \mathbf{b} = \alpha\, (\mathbf{a} \cdot \mathbf{b})$ para cualquier número real α. Asociatividad (3.16)

d) $\mathbf{a} \cdot \mathbf{a} \geq 0$ y $\mathbf{a} \cdot \mathbf{a} = 0$ si y sólo si $\mathbf{a} = \mathbf{0}$ Positividad de la definición (3.17)

Sólo se demostrará la propiedad (a) y se dejarán las restantes como ejercicio para el lector.

Demostración de (a)

$$\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n$$

y

$$\mathbf{b} \cdot \mathbf{a} = b_1 a_1 + b_2 a_2 + \ldots + b_n a_n$$

Por la conmutatividad de la multiplicación de los números reales se tiene que

$$a_i b_i = b_i a_i, \quad 1 \leq i \leq n$$

y por tanto

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$$

Enseguida se definirán conceptos tan importantes como la longitud de un vector, el ángulo entre dos vectores cualesquiera y distancia entre vectores en función del producto punto.

Cada una de estas ideas tiene un significado bien definido en los vectores de dos elementos en la geometría analítica, y es razonable pedir que cualquier definición que se adopte se reduzca a la ya conocida. Con esto en mente, se pueden obtener definiciones aceptables extendiendo las fórmulas correspondientes de la geometría analítica a vectores de n elementos.

Longitud de un vector

La noción de longitud para vectores de dos elementos está dada por la siguiente definición

Sea $\mathbf{x}$ un vector cualquiera de dos elementos, su longitud denotada por $|\ \mathbf{x}\ |$ es el número real no negativo*

$$|\ \mathbf{x}\ | = \sqrt{x_1^2 + x_2^2} \tag{3.18}$$

*Se dice que un número real es no negativo cuando sólo puede ser cero o positivo.

Gráficamente se representa así

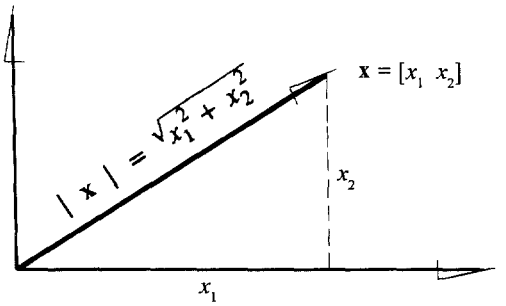


Figura 3.1 Interpretación gráfica de la longitud de un vector

La ecuación 3.18 puede escribirse en términos del producto punto como

$$| \mathbf{x} | = \sqrt{\mathbf{x} \cdot \mathbf{x}} \tag{3.19}$$

lo cual está bien definido para vectores de n elementos y puede, por tanto, tomarse como longitud de estos últimos.

Definición. La longitud (o norma) de un vector $\mathbf{x}$ de n componentes, con $n \geq 1$, está dada por el número real no negativo .*

$$\boxed{\begin{aligned} | \mathbf{x} | &= \sqrt{\mathbf{x} \cdot \mathbf{x}} \\ | \mathbf{x} | &= \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} \end{aligned}} \tag{3.20}$$

Ejemplo 3.13

Si $\mathbf{a} = \begin{bmatrix} 5 \\ 3 \\ 4 \end{bmatrix}$, encuentre su norma.

SOLUCIÓN

$$| \mathbf{a} | = \sqrt{25 + 9 + 16} = 7.0711$$

Ángulo entre vectores

Recuérdese que si se tienen dos vectores de dos componentes, ambos distintos del vector cero, la fórmula

$$\boxed{\cos \theta = \frac{\mathbf{x} \cdot \mathbf{y}}{| \mathbf{x} | \, | \mathbf{y} |}} \quad 0 \leq \theta \leq \pi \tag{3.21}$$

*Se conoce también como **norma euclideana** y algunos autores la representan por L_2.

es una consecuencia inmediata de la ley de los cosenos. Como la expresión

$$\frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}|\,|\mathbf{y}|},$$

está bien definida para vectores distintos del vector cero, de n componentes, parece conveniente usarla como definición del ángulo entre vectores de más de dos componentes. Sin embargo, sería necesario probar primero que el rango o codominio de esta expresión —usando vectores $\mathbf{x}$, $\mathbf{y}$ de n componentes— es el intervalo cerrado $[-1, 1]$, para que así se guarde consistencia con el primer miembro de la ecuación 3.21.*

La demostración está fuera de los objetivos de este libro, pero el lector interesado puede encontrarla en Kreider *et al.***

Definición. Si $\mathbf{x}$ y $\mathbf{y}$ son vectores distintos del vector 0, con n componentes, el coseno del ángulo entre ellos se define como

$$\cos\theta = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}|\,|\mathbf{y}|}$$

Si alguno de los vectores es el vector cero, se hace $\cos\theta$ igual a cero.

Ejemplo 3.14

Si $\mathbf{x}^T = [2\ -3\ 4\ 1]$ y $\mathbf{y}^T = [-1\ 2\ 4\ 2]$, calcule el ángulo entre ellos.

SOLUCIÓN

$$\cos\theta = \frac{2(-1) + (-3)(2) + 4(4) + 1(2)}{\sqrt{4 + 9 + 16 + 1}\ \sqrt{1 + 4 + 16 + 4}} = 0.3651$$

de donde $\theta = 68.58$

Distancia entre dos vectores

Uno de los tres conceptos que aún no se analiza es el de distancia entre dos vectores de n componentes. De nueva cuenta esto se hará "copiando" la definición dada en la geometría analítica, donde la distancia entre $\mathbf{x}$ y $\mathbf{y}$ es la longitud del vector $(\mathbf{x} - \mathbf{y})$ (Véase Fig. 3.2).

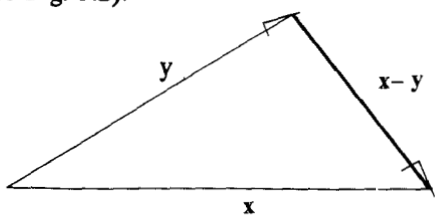


Figura 3.2. Resta de vectores en el plano.

*Recuérdese que la función cos tiene como rango el intervalo $[-1, 1]$.

**Kreider, Kuller, Ostberg, Perkins. *An Introduction to Linear Analysis.* Addison-Wesley (1966).

Definición. La distancia entre dos vectores $\mathbf{x}$ y $\mathbf{y}$ de n componentes es

$$\boxed{d(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}|} \tag{3.22}$$

definición que satisface las siguientes propiedades

a) La distancia entre dos vectores es un número real no negativo que es cero si y sólo si se trata del mismo vector; es decir

$$d(\mathbf{x}, \mathbf{y}) \geq 0 \text{ y } d(\mathbf{x}, \mathbf{y}) = 0 \text{ si y sólo si } \mathbf{x} = \mathbf{y} \tag{3.23}$$

b) Es independiente del orden en que se tomen los vectores; esto es

$$d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$$

c) Finalmente, satisface la desigualdad del triángulo, conocida en la geometría en los términos **la suma de las longitudes de los catetos de un triángulo es menor o igual a la longitud de la hipotenusa**; esto es

$$d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \geq d(\mathbf{x}, \mathbf{z})$$

para tres vectores cualesquiera $\mathbf{x}$, $\mathbf{y}$ y $\mathbf{z}$.

Ejemplo 3.15

Calcule la distancia entre $\mathbf{x}$ y $\mathbf{y}$ dadas por

$$\mathbf{x}^T = [0\ 3\ 5\ 1], \quad \mathbf{y}^T = [-2\ 1\ -3\ 1]$$

SOLUCIÓN

Primero se obtiene $\mathbf{x} - \mathbf{y}$

$$\mathbf{x} - \mathbf{y} = \begin{bmatrix} 0 \\ 3 \\ 5 \\ 1 \end{bmatrix} - \begin{bmatrix} -2 \\ 1 \\ -3 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 8 \\ 0 \end{bmatrix}$$

La norma de este vector es

$$|\mathbf{x} - \mathbf{y}| = \sqrt{2^2 + 2^2 + 8^2 + 0^2} = \sqrt{72} = 8.4853,$$

y, por tanto, la distancia entre $\mathbf{x}$ y $\mathbf{y}$ es 8.4853 unidades de longitud.

Obsérvese que ninguno de estos tres conceptos tiene representación geométrica cuando el número de componentes de los vectores es mayor de tres.

Sugerencia: Explore con el Math-CAD o algún pizarrón electrónico disponible las operaciones vistas y sus propiedades.

SECCIÓN 3.3 INDEPENDENCIA Y ORTOGONALIZACIÓN DE VECTORES

Una expresión de la forma

$$\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \ldots + \alpha_n \mathbf{x}_n , \tag{3.26}$$

donde $\alpha_1, \alpha_2, \ldots, \alpha_n$ son números reales y $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ son vectores de m elementos cada uno, se llama combinación lineal de los vectores $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$.

Ejemplo 3.16

¿La expresión

$$2.5\begin{bmatrix}1\\0\\4\\3\end{bmatrix} + 3\begin{bmatrix}-4\\2\\1.6\\5\end{bmatrix} + (-7)\begin{bmatrix}5\\-2\\0\\1\end{bmatrix}$$

es una combinación lineal?

SOLUCIÓN

Sí; es una combinación lineal de $[1\ 0\ 4\ 3]^T$, $[-4\ 2\ 1.6\ 5]^T$ y $[5\ -2\ 0\ 1]^T$, con los escalares 2.5, 3 y –7, respectivamente.

A menudo los elementos de un vector $\mathbf{x}_i$ de una combinación lineal, tendrán dos subíndices; el primero indica la fila a que pertenece y el segundo se refiere al vector a que corresponde, así

$$\mathbf{x}_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \cdot \\ \cdot \\ \cdot \\ x_{mi} \end{bmatrix}$$

Se dice que un vector $\mathbf{x} = [x_1\ x_2\ \ldots\ x_m]^T$, depende linealmente de un conjunto de vectores de m elementos $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, si se pueden encontrar escalares $\alpha_1, \alpha_2, \ldots \alpha_n$, tales que se cumpla la siguiente ecuación vectorial

$$\mathbf{x} = \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \ldots + \alpha_n \mathbf{x}_n \tag{3.27}$$

Si, por el contrario, no existen escalares que satisfagan tal ecuación, x es un vector linealmente independiente de $x_1, x_2, \ldots, x_n$. En otras palabras, x es linealmente dependiente de $x_1, x_2, \ldots, x_n$ si y sólo si x es una combinación lineal de $x_1, x_2, \ldots, x_n$.

Ejemplo 3.17

Dado el conjunto de dos vectores de dos elementos :

$$x_1 = \begin{bmatrix} 4 \\ 4 \end{bmatrix} \qquad y \qquad x_2 = \begin{bmatrix} -2 \\ 2 \end{bmatrix},$$

demuestre que el vector $x^T = [\,0 \quad 8\,]^T$ es linealmente dependiente de dicho conjunto.

SOLUCIÓN

Es suficiente encontrar dos escalares α_1 y α_2 tales que la combinación $\alpha_1 x_1 + \alpha_2 x_2$ reproduzca a x. Por observación se advierte que los números $\alpha_1 = 1$ y $\alpha_2 = 2$ cumplen este requisito.

$$\begin{bmatrix} 0 \\ 8 \end{bmatrix} = (1) \begin{bmatrix} 4 \\ 4 \end{bmatrix} + (2) \begin{bmatrix} -2 \\ 2 \end{bmatrix}$$

Generalmente, encontrar los escalares o la demostración de que no existen es un problema difícil que requiere una técnica específica, misma que se desarrolla más adelante.

Independencia de conjuntos de vectores

Un conjunto de vectores dado $y_1, y_2, \ldots, y_n$, es linealmente dependiente si por lo menos uno de ellos es combinación lineal de alguno o todos los vectores restantes. Si ninguno lo es, se dice que es un conjunto linealmente independiente.

Ejemplo 3.18

Sea el siguiente conjunto de cuatro vectores de tres elementos cada uno.

$$y_1 = \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix}, \quad y_2 = \begin{bmatrix} 0.1 \\ -3 \\ 0 \end{bmatrix}, \quad y_3 = \begin{bmatrix} 0.5 \\ -15 \\ 0 \end{bmatrix}, \quad y_4 = \begin{bmatrix} 0.03 \\ -0.9 \\ 0 \end{bmatrix}$$

Determine si es linealmente dependiente o independiente.

SOLUCIÓN

Este conjunto es linealmente dependiente, ya que $\mathbf{y}_3$ se obtiene de la combinación

$$\mathbf{y}_3 = 5 \quad \mathbf{y}_2 = 5\begin{bmatrix} 0.1 \\ -3 \\ 0 \end{bmatrix},$$

y $\mathbf{y}_4$ se obtiene de combinar $\mathbf{y}_1$ y $\mathbf{y}_2$ en la siguiente forma

$$\begin{bmatrix} 0.03 \\ -0.9 \\ 0 \end{bmatrix} = 0 \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} + 0.3 \begin{bmatrix} 0.1 \\ -3 \\ 0 \end{bmatrix}$$

Si se considera el conjunto formado sólo por $\mathbf{y}_1$ y $\mathbf{y}_2$, se tiene que es linealmente independiente, ya que ninguno se obtiene multiplicando al otro por algún escalar.

Cualquier conjunto que tenga el vector cero (vector cuyos componentes son todos cero) como uno de sus elementos, es linealmente independiente, ya que dicho vector podrá obtenerse siempre de cualquier otro vector del conjunto por la combinación

$$\mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} = 0 \begin{bmatrix} x_{1i} \\ x_{2i} \\ \cdot \\ \cdot \\ \cdot \\ x_{ni} \end{bmatrix}$$

Un conjunto formado por un sólo vector (distinto de $\mathbf{0}$) es linealmente independiente.

Interpretación geométrica de la independencia lineal

Es conveniente estudiar la independencia lineal desde el punto de vista geométrico, aunque esto sólo valga para vectores de dos y tres componentes. Considérense los tres vectores del ejemplo 3.17 en el plano x-y (Fig. 3.3). Por la geometría se sabe que dos vectores que se cortan forman un plano (por ejemplo $\mathbf{x}_1$ y $\mathbf{x}_2$ forman el plano x-y). Por tanto, es natural pensar que si se tiene un tercer vector del plano x-y, éste pueda obtenerse de alguna combinación de los que se cortan, por ejemplo $\mathbf{x}_3$ de $\mathbf{x}_1$ y $\mathbf{x}_2$, aplicando la ley del paralelogramo.

Si, por otro lado, se tienen dos vectores de dos componentes linealmente dependientes, esto se manifiesta geométricamente como paralelismo (véase los vectores $\mathbf{x}_1$ y $\mathbf{x}_2$ de la Fig. 3.4). Es evidente que estos vectores paralelos no forman un plano y un tercer vector $\mathbf{x}_3$ que no sea paralelo a ellos no podrá generarse con una combinación lineal de $\mathbf{x}_1$ y $\mathbf{x}_2$.

En conclusión, la característica geométrica de dos vectores linealmente independientes es que se cortan en un punto. En cambio, dos vectores linealmente dependientes son paralelos.

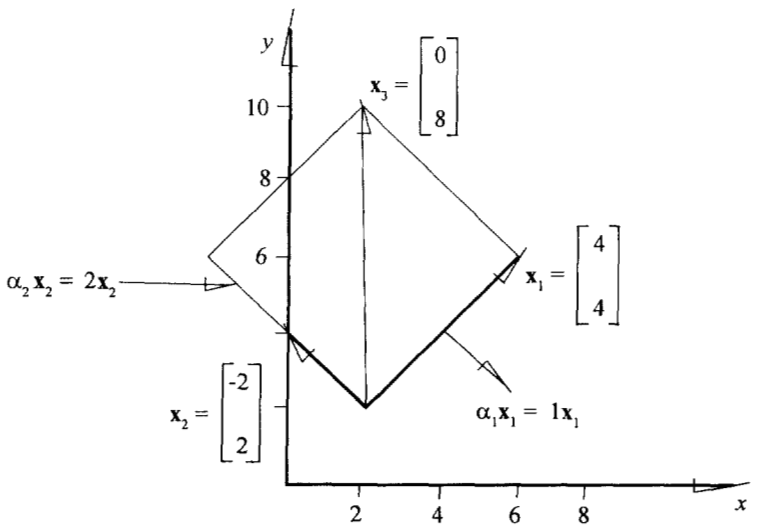


Figura 3.3. Interpretación geométrica de independencia lineal en el plano.

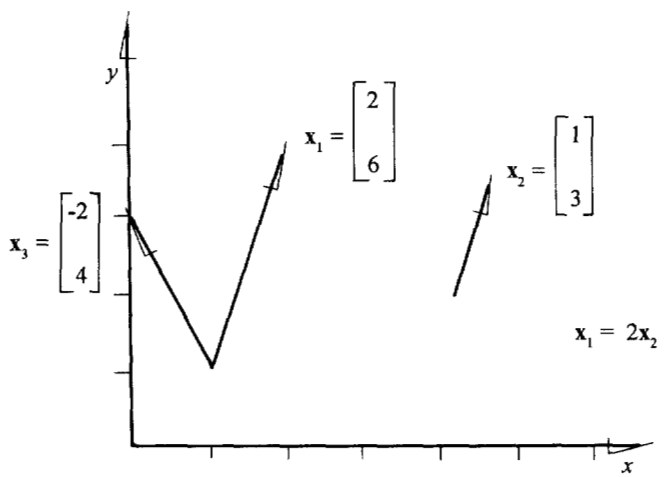


Figura 3.4. Interpretación geométrica de dependencia lineal en el plano.

Conjuntos ortogonales de vectores

Dos vectores de igual número de componentes son ortogonales o perpendiculares si el coseno del ángulo entre ellos es cero. De acuerdo con esta definición, el vector cero es ortogonal con cualquier otro vector; en general, x y y son ortogonales si y sólo si

$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n = 0,$$

derivada esta expresión del hecho que

$$\cos \theta = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}| \, |\mathbf{y}|}$$

A continuación se generaliza la definición de ortogonalidad.

Un conjunto de vectores $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ forma un conjunto ortogonal si $\mathbf{x}_i \neq \mathbf{0}$ para $1 \leq i \leq n$, y

$$\mathbf{x}_i \cdot \mathbf{y}_j = 0 \qquad 1 \leq j \leq n \tag{3.28}$$

siempre que $i \neq j$.

Ejemplo 3.19

Determine si los vectores $\mathbf{x}_1$ y $\mathbf{x}_2$ del ejemplo 3.17 son ortogonales.

SOLUCIÓN

$$\mathbf{x}_1 \cdot \mathbf{x}_2 = \begin{bmatrix} 4 \\ 4 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ 2 \end{bmatrix} = -8 + 8 = 0$$

Son perpendiculares en el sentido usual del término (véase Fig. 3.3) y esto es lo que significa la definición, dada para cualquier número de componentes.

Ejemplo 3.20

¿El conjunto siguiente es ortogonal?

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \qquad \mathbf{x}_3 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

SOLUCIÓN

Sí, ya que

$$\mathbf{x}_1 \cdot \mathbf{x}_2 = \mathbf{x}_1 \cdot \mathbf{x}_3 = \mathbf{x}_2 \cdot \mathbf{x}_3 = 0$$

En cambio, si se adiciona a este conjunto el vector

$$\mathbf{x}_4 = \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

el conjunto resultante $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ no es ortogonal, pues

$$\mathbf{x}_4 \cdot \mathbf{x}_2 = 1 \neq 0$$

Ejemplo 3.21

Corrobore si el siguiente conjunto de vectores es ortogonal

SOLUCIÓN

$$\mathbf{x}_1 = \begin{bmatrix} -3 \\ 4 \\ 1 \end{bmatrix}, \qquad \mathbf{x}_2 = \begin{bmatrix} 2 \\ 2 \\ -2.0003 \end{bmatrix}$$

$$\mathbf{x}_1 \cdot \mathbf{x}_2 = (-3)\,(2) + 4(2) + 1(-2.0003) = -0.0003$$

Obsérvese que los vectores son "casi" ortogonales. Esto ocurre con frecuencia y en los cálculos prácticos será preciso decidir con qué cercanía a cero se aceptará que un producto punto de dos vectores "es cero", y, por tanto, que los vectores son ortogonales. De nuevo ε denotará el límite de aceptación o rechazo. El valor que tome ε estará en función del instrumento con que se lleven a cabo los cálculos. Por ejemplo, para una calculadora de nueve dígitos de exactitud, ε puede ser 10^{-4}. Con $\varepsilon = 10^{-4}$ los vectores de este ejemplo no son ortogonales. Así pues, ε usado de esta manera puede llamarse **criterio de ortogonalidad.**

Ortogonalización

Se ha llegado al punto central de esta sección, donde es posible construir un conjunto de vectores ortogonales (ortogonalización) a partir de un cojunto de vectores linealmente independientes. Enseguida se considerará uno de los métodos más difundidos, la ortogonalización de Gram-Schmidt, aunque pueda representar ciertas dificultades computacionales.

Método de Gram-Schmidt

En lugar de empezar con el caso más general, se introducirá el proceso de ortogonalización con dos ejemplos; el primero se tiene cuando se toman dos vectores $\mathbf{x}_1$ y $\mathbf{x}_2$ del plano x-y, linealmente independientes y a partir de ellos se forma el conjunto ortogonal $\mathbf{e}_1$ y $\mathbf{e}_2$. La figura 3.5 muestra la manera natural de resolver este caso; simplemente se toma $\mathbf{e}_1 = \mathbf{x}_1$ y $\mathbf{e}_1$ como la "componente" de $\mathbf{x}_2$ perpendicalar a $\mathbf{x}_1$. Así, se escribe $\mathbf{e}_2$ en la forma

$$\mathbf{e}_2 = \mathbf{x}_2 - \alpha_{1,2}\mathbf{e}_1 \tag{3.29}$$

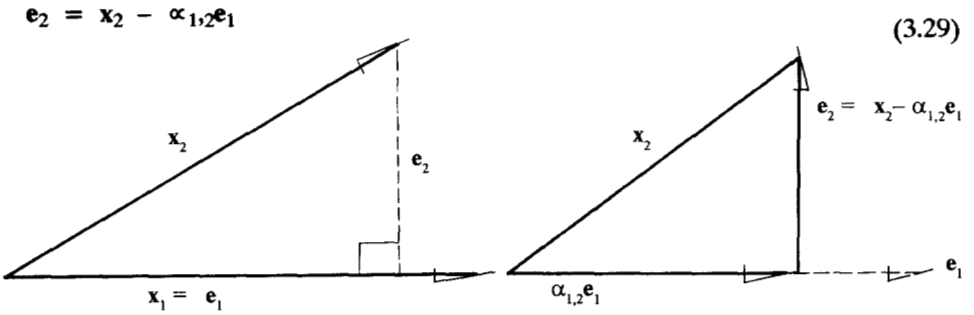


Figura 3.5. Ortogonalización en el plano x-y.

y sólo queda determinar $\alpha_{1,2}$ de manera que la condición $\mathbf{e}_1 \cdot \mathbf{e}_2 = 0$ se cumpla. Esto da la ecuación

$$\mathbf{e}_2 \cdot \mathbf{e}_1 = 0 = \mathbf{x}_2 \cdot \mathbf{e}_1 - \alpha_{1,2}\, \mathbf{e}_1 \cdot \mathbf{e}_1 \tag{3.30}$$

y finalmente

$$\alpha_{1,2} = \frac{\mathbf{x}_2 \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1} \tag{3.31}$$

De este modo $\mathbf{e}_2$ queda determinado en función de $\mathbf{x}_1$ y $\mathbf{x}_2$, y el conjunto $\mathbf{x}_1$, $\mathbf{x}_2$ se ha ortogonalizado.

Ejemplo 3.22

Ortogonalice $\mathbf{x}_1 = [2 \;\; 2]^T$ y $\mathbf{x}_2 = [3 \;\; 0]^T$

SOLUCIÓN

$$\mathbf{e}_1 = [2 \;\; 2]^T$$

y

$$\mathbf{e}_2 = \mathbf{x}_2 - \alpha_{1,2}\mathbf{e}_1$$

con

$$\alpha_{1,2} = \frac{[2 \;\; 2]^T \cdot [3 \;\; 0]^T}{[2 \;\; 2]^T \cdot [2 \;\; 2]^T} = \frac{6}{4 + 4} = \frac{3}{4}$$

Sustituyendo queda

$$\mathbf{e}_2 = [3 \;\; 0]^T - \frac{3}{4}\,[2 \;\; 2]^T = [3 \;\; 0]^T - [\tfrac{3}{2} \;\; \tfrac{3}{2}]^T = [1.5 \;\; -1.5]^T$$

Al graficar estos vectores se obtiene la siguiente figura

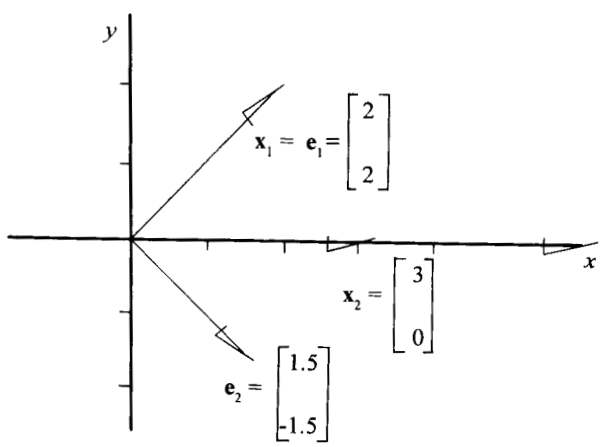


Figura 3.6. Ortogonalización de vectores

Obsérvese la perpendicularidad de $\mathbf{e}_1$ y $\mathbf{e}_2$.

Como segundo ejemplo se ortogonalizará el conjunto arbitrario $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ de vectores linealmente independientes de tres componentes. El procedimiento es esencialmente igual al que se usó antes, y se empieza escogiendo $\mathbf{e}_1 = \mathbf{x}_1$. El segundo paso es determinar $\mathbf{e}_2$ de acuerdo con el par de ecuaciones

$$\mathbf{e}_2 \cdot \mathbf{e}_1 = 0, \quad \mathbf{e}_2 = \mathbf{x}_2 - \alpha_{1,2}\,\mathbf{e}_1 \tag{3.32}$$

de las que se obtiene nuevamente que

$$\alpha_{1,2} = \frac{\mathbf{x}_2 \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1} \tag{3.33}$$

Obsérvese que $\mathbf{e}_2 \neq \mathbf{0}$; de lo contrario se cumpliría la primera de las ecuaciones 3.32 y en la segunda se tendría que $\mathbf{x}_2 = \alpha_{1,2}\mathbf{e}_1 = \alpha_{1,2}\mathbf{x}_1$. O sea que $\mathbf{x}_2$ estaría en función de $\mathbf{x}_1$, lo cual es imposible por la independencia lineal de $\mathbf{x}_1$ y $\mathbf{x}_2$.

Para el tercer vector se recurre nuevamente a una representación geométrica, en donde se verá que el proceso de ortogonalización puede completarse tomando $\mathbf{e}_3$ como la componente de $\mathbf{x}_3$ perpendicular al plano formado por los vectores $\mathbf{e}_1$ y $\mathbf{e}_2$ (Fig. 3.6)*.

De esto se tiene

$$\mathbf{e}_3 = \mathbf{x}_3 - \alpha_{1,3}\mathbf{e}_1 - \alpha_{2,3}\mathbf{e}_2 \tag{3.34}$$

y se puede encontrar $\alpha_{1,3}$ y $\alpha_{2,3}$ por medio de las condiciones de ortogonalidad

$$\mathbf{e}_1 \cdot \mathbf{e}_2 = \mathbf{e}_1 \cdot \mathbf{e}_3 = \mathbf{e}_2 \cdot \mathbf{e}_3 = 0$$

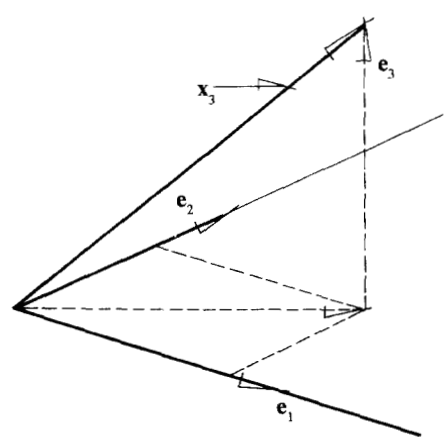


Figura 3.7. Ortogonalización en el espacio x-y-z.

* Recuérdese que dos líneas que se cortan solamente en un punto forman un plano.

Multiplicando en forma punto los dos miembros de la ecuación 3.34 por $\mathbf{e}_1$ y después por $\mathbf{e}_2$, se obtiene el par de ecuaciones

$$\begin{aligned} \mathbf{e}_3 \cdot \mathbf{e}_1 &= 0 = \mathbf{x}_3 \cdot \mathbf{e}_1 - \alpha_{1,3}\mathbf{e}_1 \cdot \mathbf{e}_1 - \alpha_{2,3}\mathbf{e}_2 \cdot \mathbf{e}_1 \\ \mathbf{e}_3 \cdot \mathbf{e}_2 &= 0 = \mathbf{x}_3 \cdot \mathbf{e}_2 - \alpha_{1,3}\mathbf{e}_1 \cdot \mathbf{e}_2 - \alpha_{2,3}\mathbf{e}_2 \cdot \mathbf{e}_2 \end{aligned} \tag{3.35}$$

o bien

$$\begin{aligned} \mathbf{x}_3 \cdot \mathbf{e}_1 &= \alpha_{1,3}\mathbf{e}_1 \cdot \mathbf{e}_1 \\ \mathbf{x}_3 \cdot \mathbf{e}_2 &= \alpha_{2,3}\mathbf{e}_2 \cdot \mathbf{e}_2 \end{aligned} \tag{3.36}$$

resolviendo para $\alpha_{1,3}$ y para $\alpha_{2,3}$, se tiene

$$\alpha_{1,3} = \frac{\mathbf{x}_3 \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1}, \qquad \alpha_{2,3} = \frac{\mathbf{x}_3 \cdot \mathbf{e}_2}{\mathbf{e}_2 \cdot \mathbf{e}_2}$$

y con esto termina la ortogonalización del conjunto $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$.

Ejemplo 3.23

Ortogonalice los vectores

$$\mathbf{x}_1 = [\,1 \quad 1 \quad 0\,]^T, \qquad \mathbf{x}_2 = [\,0 \quad 1 \quad 0\,]^T \qquad \text{y} \qquad \mathbf{x}_3 = [\,1 \quad 1 \quad 1\,]^T$$

SOLUCIÓN

$$\mathbf{e}_1 = \mathbf{x}_1, \quad \mathbf{e}_2 = \mathbf{x}_2 - \alpha_{1,2}\mathbf{e}_1, \quad \text{y}$$
$$\mathbf{e}_3 = \mathbf{x}_3 - \alpha_{1,3}\mathbf{e}_1 - \alpha_{2,3}\mathbf{e}_2,$$

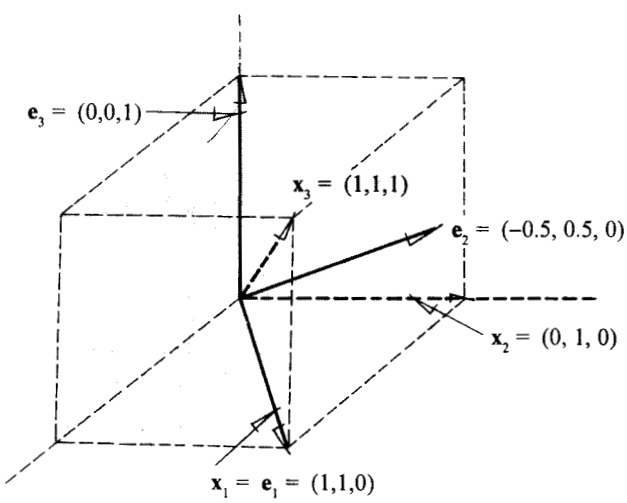


Figura 3.8. Ortogonalización en el espacio.

donde $\alpha_{1,2}$, $\alpha_{1,3}$ y $\alpha_{2,3}$

se obtienen de las ecuaciones

$$\alpha_{1,2} = \frac{\mathbf{x}_2 \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1}, \qquad \alpha_{1,3} = \frac{\mathbf{x}_3 \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1}, \qquad \alpha_{2,3} = \frac{\mathbf{x}_3 \cdot \mathbf{e}_2}{\mathbf{e}_2 \cdot \mathbf{e}_2}$$

Al verificar los cálculos se llega a

$$\alpha_{1,2} = 1/2, \quad \alpha_{1,3} = 1, \quad \alpha_{2,3} = 0$$

y sustituyendo

$$\mathbf{e}_1 = [1\ 1\ 0]^T, \ \mathbf{e}_2 = [-1/2\ \ 1/2\ \ 0]^T, \ \mathbf{e}_3 = [0\ 0\ 1]^T$$

Una vez realizado lo anterior, se puede pasar al caso general de ortogonalizar un conjunto de n vectores linealmente independientes $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ de n componentes cada uno. Primero se efectuará $\mathbf{e}_1 = \mathbf{x}_1$, después $\mathbf{e}_2 = \mathbf{x}_2 - \alpha_{1,2}\,\mathbf{e}_1$, donde $\alpha_{1,2}$ se escoge de manera que $\mathbf{e}_1 \cdot \mathbf{e}_2 = 0$.

De aquí que

$$\alpha_{1,2} = \frac{\mathbf{x}_2 \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1},$$

y la independencia lineal de $\mathbf{x}_1$ y $\mathbf{x}_2$ implica que $\mathbf{e}_2 \neq 0$.

Únicamente queda por demostrar que este proceso puede continuar hasta obtener un conjunto ortogonal $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$. Para ello, supóngase que se llegó al conjunto ortogonal $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$ con $m < n$. Para continuar un paso más efectúese

$$\mathbf{e}_{m+1} = \mathbf{x}_{m+1} - \alpha_{1,m+1}\mathbf{e}_1 - \ldots - \alpha_{m,m+1}\mathbf{e}_m$$

y determínese $\alpha_{1,m+1}, \alpha_{2,m+1}, \ldots, \alpha_{m,m+1}$, de manera que $\mathbf{e}_{m+1}$ sea ortogonal a cada elemento del conjunto $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$. Consecuentemente el conjunto de ecuaciones es

$$\begin{aligned}
\mathbf{x}_{m+1} \cdot \mathbf{e}_1 \ &- \ \alpha_{1,m+1}\ (\mathbf{e}_1 \cdot \mathbf{e}_1) \ &= 0, \\
\mathbf{x}_{m+1} \cdot \mathbf{e}_2 \ &- \ \alpha_{2,m+1}\ (\mathbf{e}_2 \cdot \mathbf{e}_2) \ &= 0, \\
&\cdots \\
\mathbf{x}_{m+1} \cdot \mathbf{e}_m \ &- \ \alpha_{m,m+1}\ (\mathbf{e}_m \cdot \mathbf{e}_m) \ &= 0,
\end{aligned}$$

y por tanto

$$\alpha_{1,m+1} = \frac{\mathbf{x}_{m+1} \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1}, \qquad \alpha_{2,m+1} = \frac{\mathbf{x}_{m+1} \cdot \mathbf{e}_2}{\mathbf{e}_2 \cdot \mathbf{e}_2}, \ldots, \qquad \alpha_{m,m+1} = \frac{\mathbf{x}_{m+1} \cdot \mathbf{e}_m}{\mathbf{e}_m \cdot \mathbf{e}_m}$$

que determinan $\mathbf{e}_{m+1}$. De nuevo, la independencia lineal de $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{m+1}$ implica que $\mathbf{e}_{m+1} \neq \mathbf{0}$. Por tanto, el proceso de ortogonalización se ha aumentado en un paso y con el mismo argumento puede continuarse hasta tener $m = n$. Lo anterior queda condensado en el siguiente teorema.

Teorema 3.1 Sean $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ un conjunto de vectores linealmente independientes de n componentes cada uno. A partir de ellos se puede construir un conjunto ortogonal $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ de la siguiente manera

$$\mathbf{e}_1 = \mathbf{x}_1$$

y

$$\mathbf{e}_{i+1} = \mathbf{x}_{i+1} - \alpha_{1,i+1}\mathbf{e}_1 - \cdots - \alpha_{i,i+1}\mathbf{e}_i \qquad 1 \leq i \leq n-1 \tag{3.37}$$

donde

$$\alpha_{1,i+1} = \frac{\mathbf{x}_{i+1} \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1}, \qquad \alpha_{2,i+1} = \frac{\mathbf{x}_{i+1} \cdot \mathbf{e}_2}{\mathbf{e}_2 \cdot \mathbf{e}_2}, \qquad \alpha_{i,i+1} = \frac{\mathbf{x}_{i+1} \cdot \mathbf{e}_i}{\mathbf{e}_i \cdot \mathbf{e}_i} \tag{3.38}$$

Ejemplo 3.24

Ortogonalice el siguiente conjunto de vectores linealmente independientes

$$\mathbf{x}_1 = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}, \qquad \mathbf{x}_2 = \begin{bmatrix} 3 \\ 2 \\ 0 \end{bmatrix}, \qquad \mathbf{x}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

SOLUCIÓN

$$\mathbf{e}_1 = \mathbf{x}_1, \quad \mathbf{e}_2 = \mathbf{x}_2 - \alpha_{1,2}\,\mathbf{e}_1,$$

donde

$$\alpha_{1,2} = \frac{\mathbf{x}_2 \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1} = \frac{\begin{bmatrix} 3 \\ 2 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}}{\begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}} = \frac{6}{5}$$

Sustituyendo

$$\mathbf{e}_2 = \begin{bmatrix} 3 \\ 2 \\ 0 \end{bmatrix} - \frac{6}{5} \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3/5 \\ 2 \\ -6/5 \end{bmatrix}$$

$$\mathbf{e}_3 = \mathbf{x}_3 - \alpha_{1,3}\mathbf{e}_1 - \alpha_{2,3}\mathbf{e}_2, \text{ donde}$$

$$\propto_{1,3} = \frac{\mathbf{x}_3 \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1} = \frac{\begin{bmatrix}1\\1\\1\end{bmatrix} \cdot \begin{bmatrix}2\\0\\1\end{bmatrix}}{\begin{bmatrix}2\\0\\1\end{bmatrix} \cdot \begin{bmatrix}2\\0\\1\end{bmatrix}} = \frac{3}{5}, \quad \propto_{2,3} = \frac{\mathbf{x}_3 \cdot \mathbf{e}_2}{\mathbf{e}_2 \cdot \mathbf{e}_2} = \frac{\begin{bmatrix}1\\1\\1\end{bmatrix} \cdot \begin{bmatrix}3/5\\2\\-6/5\end{bmatrix}}{\begin{bmatrix}3/5\\2\\-6/5\end{bmatrix} \cdot \begin{bmatrix}3/5\\2\\-6/5\end{bmatrix}} = \frac{35}{145}$$

Sustituyendo

$$\mathbf{e}_3 = \begin{bmatrix}1\\1\\1\end{bmatrix} - \frac{3}{5}\begin{bmatrix}2\\0\\1\end{bmatrix} - \frac{35}{145}\begin{bmatrix}3/5\\2\\-6/5\end{bmatrix} = \begin{bmatrix}-10/29\\15/29\\20/29\end{bmatrix}$$

A continuación se presenta un algoritmo para ortogonalizar un conjunto de n vectores de n componentes cada uno por el método visto.

ALGORITMO 3.2 Ortogonalización de Gram-Schmidt

Para ortogonalizar un conjunto de N vectores linealmente independientes de N componentes cada uno, proporcionar los

DATOS: El numero N y los vectores **x1, x2, ... , xN.**
RESULTADOS: El conjunto de vectores ortogonales **e1, e2, ..., eN.**

PASO 1. Hacer **e1** = **x1**
PASO 2. Hacer I = 1
PASO 3. Mientras I ≤ N − 1, repetir los pasos 4 a 10.
  PASO 4. Hacer **e**(I+1) = **x** (I+1)
  PASO 5. Hacer J = 1
  PASO 6. Mientras J ≤ I, repetir los pasos 7 a 9.
    PASO 7. Hacer ∝ (J,I + 1) = (**x** (I+1) · **e** (J))/(**e** (J) · **e** (J))
    PASO 8. Hacer **e** (I + 1) = **e** (I + 1) − ∝ (J,I + 1) * **e** (J)
    PASO 9. Hacer J = J + 1
  PASO 10. Hacer I = I + 1
PASO 11. IMPRIMIR los vectores **e1, e2, ... , eN** y TERMINAR.

Nota: En el paso 7, el punto indica producto escalar de dos vectores.
En el 8, ∝ (J, I + 1) es un escalar que multiplica al vector **e**(J) y la resta es vectorial.
En los pasos 1, 4, 7 y 8 se trata de asignaciones de todos los componentes de un vector a otro.

Sugerencia: Es recomendable trabajar con un programa desarrollado en un lenguaje de alto nivel (véase Probl. 3.14) basado en el algoritmo 3.2 o en un pizarrón electrónico (Math-CAD por ejemplo) para evitar cálculos y analizar la ortogonalización más finamente.

Una aplicación importante de los resultados obtenidos es determinar la independencia o dependencia lineal de un conjunto dado de vectores. Para esto se partirá de un conjunto linealmente dependiente particular; obsérvese qué ocurre en el proceso de ortogonalización.

Sean $\mathbf{x}_1 = [1\ 2]^T$ y $\mathbf{x}_2 = [-2\ -4]^T$. Obviamente $\mathbf{x}_2 = -2\ \mathbf{x}_1$

Efectuando $\mathbf{e}_1 = \mathbf{x}_1 = [1\ 2]^T$ y

$$\mathbf{e}_2 = \mathbf{x}_2 - \frac{\mathbf{x}_2 \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1}\,\mathbf{e}_1 = [-2\ -4]^T - \frac{[-2\ -4]^T \cdot [1\ 2]^T}{[1\ 2]^T \cdot [1\ 2]^T}\,[1\ 2]^T$$

$$= [-2\ \ -4]^T - (-2)\,[1\ \ 2]^T = [0\ \ 0]^T$$

y por lo tanto $\mathbf{e}_2 = \mathbf{0}$

Si $\mathbf{x}_1$ y $\mathbf{x}_2$ son vectores linealmente dependientes cualesquiera, al aplicar el proceso de ortogonalización se tiene

$$\mathbf{e}_1 = \mathbf{x}_1,$$

$$\mathbf{e}_2 = \mathbf{x}_2 - \frac{\mathbf{x}_2 \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1}\,\mathbf{e}_1$$

como $\mathbf{x}_2 = \beta\mathbf{x}_1 = \beta\mathbf{e}_1$

$$\mathbf{e}_2 = \beta\mathbf{e}_1 - \frac{\beta\mathbf{e}_1 \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1}\,\mathbf{e}_1 = \beta\mathbf{e}_1 - \beta\,\frac{\mathbf{e}_1 \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1}\,\mathbf{e}_1$$

pero $\dfrac{\mathbf{e}_1 \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1} = 1$, por lo tanto, $\mathbf{e}_2 = \mathbf{0}$ y $|\ \mathbf{e}_2\ | = 0$.

Generalmente, para determinar si un conjunto dado $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ es linealmente dependiente o independiente, se le aplica el proceso de ortogonalización de Gram-Schmidt. Supóngase que se han obtenido en dicho proceso $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_i$ a partir de $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i$. Si al querer obtener $\mathbf{e}_{i+1}$ resulta que $|\ \mathbf{e}_{i+1}\ | = 0$, o en términos prácticos su cercanía a cero satisface un criterio de ortogonalidad prestablecido $|\ \mathbf{e}_{i+1}\ | < \varepsilon$, el vector $\mathbf{x}_{i+1}$ es linealmente dependiente de los vectores $\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_i$; como consecuencia, el conjunto dado es linealmente dependiente. Si, por el contrario, se obtienen $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ tales que $|\ \mathbf{e}_j\ | > \varepsilon$ para $1 \leq j \leq n$, el conjunto en cuestión es linealmente independiente.

Ejemplo 3.25

Analice si los siguientes vectores son linealmente independientes.

$$\mathbf{x}_1 = \begin{bmatrix} 0 \\ 5 \\ 5 \\ 0 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

SOLUCIÓN

Se aplica el proceso de Gram-Schmidt

$$\mathbf{e}_1 = \mathbf{x}_1 = \begin{bmatrix} 0 \\ 5 \\ 5 \\ 0 \end{bmatrix}$$

$$\mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} - \frac{\begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 5 \\ 5 \\ 0 \end{bmatrix}}{\begin{bmatrix} 0 \\ 5 \\ 5 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 5 \\ 5 \\ 0 \end{bmatrix}} \begin{bmatrix} 0 \\ 5 \\ 5 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} - \frac{1}{5} \begin{bmatrix} 0 \\ 5 \\ 5 \\ 0 \end{bmatrix} = \mathbf{0}$$

lo cual implica que $\mathbf{x}_2$ es linealmente dependiente de $\mathbf{x}_1$. El conjunto es linealmente dependiente. Sin embargo, el proceso de ortogonalización puede continuar para ver si $\mathbf{x}_3$ es linealmente dependiente de $\mathbf{x}_1$

$$\mathbf{e}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} - \frac{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 5 \\ 5 \\ 0 \end{bmatrix}}{\begin{bmatrix} 0 \\ 5 \\ 5 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 5 \\ 5 \\ 0 \end{bmatrix}} \begin{bmatrix} 0 \\ 5 \\ 5 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} - \frac{1}{5} \begin{bmatrix} 0 \\ 5 \\ 5 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Obsérvese que en el cálculo de $\mathbf{e}_3$ se ignora a $\mathbf{e}_2$. Como $\mathbf{e}_3 \neq \mathbf{0}$, $\mathbf{x}_1$ y $\mathbf{x}_3$ son linealmente independientes.

Rango

El número de vectores linealmente independientes de un conjunto dado recibe el nombre de rango o característica del conjunto. Así, el conjunto del ejemplo 3.25 tiene un rango de 2.

Para un conjunto de m vectores, cada uno de n componentes, el rango puede ser como máximo igual al menor de m o n.

Rango de una matriz

Una matriz puede verse como un conjunto de vectores; más claramente, la matriz

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix},$$

se puede tratar como un conjunto de n vectores columna de m componentes cada uno (o bien m vectores fila de n componentes cada uno); es decir, como $A = [\mathbf{x}_1\ \mathbf{x}_2\ \dots\ \mathbf{x}_n]$

donde

$$\mathbf{x}_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \cdot \\ \cdot \\ \cdot \\ a_{m1} \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} a_{12} \\ a_{22} \\ \cdot \\ \cdot \\ \cdot \\ a_{m2} \end{bmatrix}, \dots, \quad \mathbf{x}_n = \begin{bmatrix} a_{1n} \\ a_{2n} \\ \cdot \\ \cdot \\ \cdot \\ a_{mn} \end{bmatrix}$$

o como

$$A = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{y}_m \end{bmatrix}$$

donde

$$\mathbf{y}_1 = [a_{11}\ a_{12}\ \dots\ a_{1n}],\ \mathbf{y}_2 = [a_{21}\ a_{22}\ \dots\ a_{2n}], \dots,$$
$$\mathbf{y}_m = [a_{m1}\ a_{m2}\ \dots\ a_{mn}]$$

En estas condiciones puede hablarse del rango de una matriz, en donde el rango de una matriz A está dado por el número máximo de vectores columna o vectores fila, linealmente independientes*.

Así la matriz

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 5 & 1 & 1 \\ 5 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

cuyas columnas son los elementos del conjunto dado en el ejemplo 3.25, tiene rango 2.

Cuando el rango de una matriz cuadrada de orden n es menor que n, se dice que la matriz es singular. Lo cual significa también que su determinante es cero. (Véase Prob. 3.18). Si las columnas de la matriz son "casi" linealmente dependientes, recibe el nombre de **casi singular** o **mal condicionada** (véase sistemas de ecuaciones mal condicionadas, Sec. 3.4).

En esta sección se ha considerado una serie de conceptos teóricos que, además de su interés por sí mismos, forman un marco que permitirá explicar de manera lógica ciertos algoritmos importantes de las matemáticas y también conceptos de existencia y unicidad de las soluciones de los problemas que resuelven dichos algoritmos.

---

*Puede demostrarse que el número máximo de vectores columna linealmente independientes de una matriz A, es igual al número máximo de vectores fila linealmente independientes.

SECCIÓN 3.4 SOLUCIÓN DE SISTEMAS DE ECUACIONES LINEALES

Un gran número de problemas prácticos de ingeniería se reduce al problema de resolver un sistema de ecuaciones lineales. Por ejemplo, pueden citarse la solución de sistemas de ecuaciones no lineales, la aproximación polinomial, la solución de ecuaciones diferenciales parciales, entre otros.

Un sistema de m ecuaciones lineales en n incógnitas tiene la forma general

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ \cdot \quad\quad \cdot \quad\quad\quad \cdot \quad\quad \cdot & \\ \cdot \quad\quad \cdot \quad\quad\quad \cdot \quad\quad \cdot & \\ \cdot \quad\quad \cdot \quad\quad\quad \cdot \quad\quad \cdot & \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m \end{aligned} \tag{3.39}$$

Con la notación matricial se puede escribir la ecuación anterior como

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \quad \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_m \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_m \end{bmatrix}$$

y concretamente como $A\ \mathbf{x} = \mathbf{b}$.

Donde A es la **matriz coeficiente** del sistema, $\mathbf{x}$ el **vector incógnita** y $\mathbf{b}$ el **vector de términos independientes.**

Dados A y $\mathbf{b}$, se entiende por resolver el sistema (Ec. 3.39) encontrar los vectores $\mathbf{x}$ que lo satisfagan. Antes de estudiar las técnicas que permiten encontrar $\mathbf{x}$ se expondrán algunas consideraciones teóricas.

Existencia y unicidad de soluciones

Si $\mathbf{b}$ es el vector cero, la ecuación 3.39 es un **sistema homogéneo**. Si por el contrario, $\mathbf{b} \neq \mathbf{0}$, el sistema es **no homogéneo.** A continuación se define la **matriz aumentada** B, formada con los elementos de la matriz coeficiente A y los del vector $\mathbf{b}$ de la siguiente manera

$$B = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{array}\right] = [A \mid \mathbf{b}]$$

Si el rango de la matriz coeficiente A y de la matriz aumentada B son iguales, se dice que el sistema (Ec. 3.39) es **consistente**. Si no ocurre esto, el sistema es inconsistente (por tanto, un sistema homogéneo siempre es consistente). Un sistema **inconsistente** no tiene solución, mientras que uno consistente tiene una solución única o un número infinito de soluciones, según como sea el rango de A en comparación con el número de incógnitas n. Si el rango de A es igual al número de incógnitas, la solución es única; si el rango de A es menor que dicho número, hay un número infinito de soluciones. (Véase Fig. 3.9).

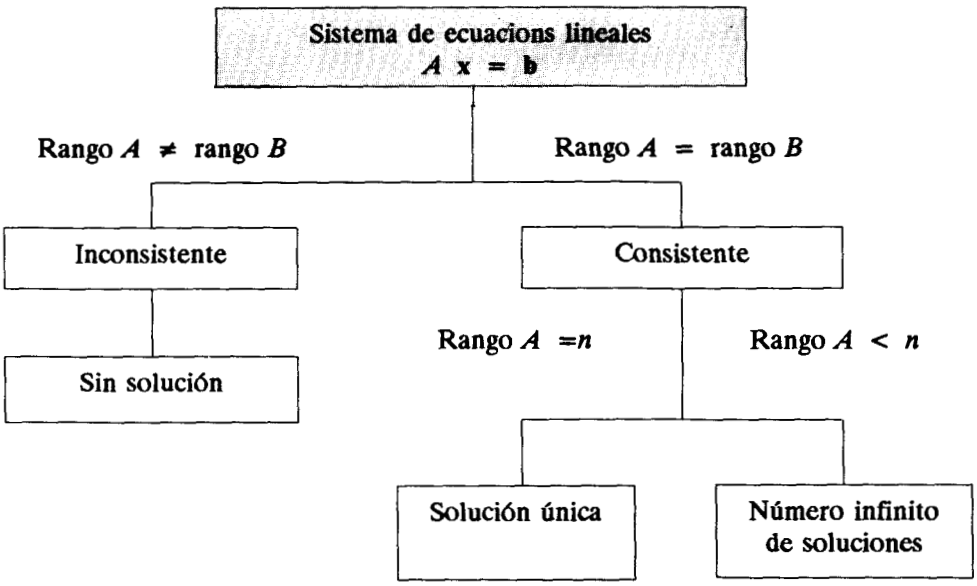


Figura 3.9. Solución de sistemas de ecuaciones lineales.

Ejemplo 3.26

Sea el sistema

$$2x_1 + 4x_2 = 6$$
$$3x_1 + 6x_2 = 5$$

La matriz aumentada es

$$\left[\begin{array}{cc|c} 2 & 4 & 6 \\ 3 & 6 & 5 \end{array}\right]$$

Puede verse fácilmente que : rango de $A = 1$, rango de $B = 2$; como rango $A \neq$ rango B, el sistema no tiene solución.

Si el sistema es homogéneo

$$2x_1 + 4x_2 = 0$$
$$3x_1 + 6x_2 = 0,$$

la matriz aumentada es

$$\left[\begin{array}{cc|c} 2 & 4 & 0 \\ 3 & 6 & 0 \end{array}\right]$$

y rango $A = 1$, rango $B = 1$, rango $A < 2 = n$; en este caso existe un número infinito de soluciones.

Ejemplo 3.27

Sea el sistema

$$\begin{aligned} 2\ x_1 + 3\ x_2 + x_3 &= 0 \\ 0\ x_1 + 2\ x_2 + x_3 &= 1 \\ x_1 + 0\ x_2 + x_3 &= 0, \end{aligned}$$

donde la matriz aumentada es

$$\left[\begin{array}{ccc|c} 2 & 3 & 1 & 0 \\ 0 & 2 & 1 & 1 \\ 1 & 0 & 1 & 0 \end{array}\right]$$

Obsérvese que la matriz coeficiente son los vectores del ejemplo 3.24, que son linealmente independientes y, por tanto, rango $A = 3$.

Al aplicar el método de Gram-Schmidt para ortogonalizar el vector de términos independientes se observa que es linealmente dependiente, y por tanto rango $B = 3$. El sistema es consistente y como rango A = número de incógnitas $=3$, puede esperarse solución única del sistema.

Esta comprobación se deja como ejercicio para el lector.

Métodos directos de solución

El prototipo de todos estos métodos se conoce como la eliminación de Gauss y se presenta a continuación.

Eliminación de Gauss

Considérese un sistema general de tres ecuaciones lineales con tres incógnitas

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \tag{3.40}$$

Como primer paso, se remplaza la segunda ecuación con lo que resulte de sumarle la primera ecuación multiplicada por $(-a_{21}/a_{11})$. Similarmente se sustituye la tercera ecuación con el resultado de sumarle la primera ecuación multiplicada por $(-a_{31}/a_{11})$.

Esto da lugar al nuevo sistema

$$\begin{aligned} a_{11}\,x_1 + a_{12}\,x_2 + a_{13}\,x_3 &= b_1 \\ a'_{22}\,x_2 + a'_{23}\,x_3 &= b'_2 \\ a'_{32}\,x_2 + a'_{33}\,x_3 &= b'_3 \end{aligned} \tag{3.41}$$

en donde las a' y las b' son los nuevos elementos que se obtienen de las operaciones ya mencionadas, y en donde x_1 se ha eliminado en la segunda y tercera ecuaciones. Ahora, multiplicando la segunda ecuación de 3.41 por $(-a'_{32}/a'_{22})$ y sumando el resultado a la tercera ecuación de 3.41, se obtiene el sistema triangular

$$\begin{aligned} a_{11}\,x_1 + a_{12}\,x_2 + a_{13}\,x_3 &= b_1 \\ a'_{22}\,x_2 + a'_{23}\,x_3 &= b'_2 \\ a''_{33}\,x_3 &= b''_3 \end{aligned} \tag{3.42}$$

donde a''_{33} y b''_3, resultaron de las operaciones realizadas y x_2 se ha eliminado de la tercera ecuación.

El proceso de llevar el sistema de ecuaciones 3.40 a la forma de la ecuación 3.42 se conoce como **triangularización.**

El sistema en la forma de la ecuación 3.42 se resuelve despejando de su última ecuación x_3, sustituyendo x_3 en la segunda ecuación y despejando x_2 de ella. Por último, con x_3 y x_2 sustituidas en la primera ecuación de 3.42 se obtiene x_1. Esta parte del proceso se llama **sustitución regresiva.**

Antes de ilustrar la eliminación de Gauss con un ejemplo particular, nótese que no es necesario conservar x_1, x_2 y x_3 en la triangularización y que ésta puede llevarse a cabo usando solamente la matriz coeficiente A y el vector **b**. Para mayor simplicidad se empleará la matriz aumentada B.

$$B = \left[\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & b_1 \\ a_{21} & a_{22} & a_{23} & b_2 \\ a_{31} & a_{32} & a_{33} & b_3 \end{array}\right] = [A \mid \mathbf{b}]$$

Con esto se incorporan la notación matricial y todas sus ventajas a la solución de sistemas de ecuaciones lineales.

Ejemplo 3.28

Resuelva por eliminación de Gauss el sistema

$$\begin{aligned} 4x_1 - 9x_2 + 2x_3 &= 5 \\ 2x_1 - 4x_2 + 6x_3 &= 3 \\ x_1 - x_2 + 3x_3 &= 4 \end{aligned} \tag{3.43}$$

SOLUCIÓN

La matriz aumentada del sistema es

$$\left[\begin{array}{ccc|c} 4 & -9 & 2 & 5 \\ 2 & -4 & 6 & 3 \\ 1 & -1 & 3 & 4 \end{array}\right] \tag{3.44}$$

Triangularización

Al sumar la primera ecuación multiplicada por (–2/4) a la segunda, y la primera ecuación multiplicada por (–1/4) a la tercera, resulta

$$\left[\begin{array}{ccc|c} 4 & -9 & 2 & 5 \\ 0 & 0.5 & 5 & 0.5 \\ 0 & 1.25 & 2.5 & 2.75 \end{array}\right] \tag{3.45}$$

Obsérvese que en este paso la primera fila se conserva sin cambio.

Sumando la segunda fila multiplicada por (–1.25/0.5) a la tercera se obtiene la matriz*

$$\left[\begin{array}{ccc|c} 4 & -9 & 2 & 5 \\ 0 & 0.5 & 5 & 0.5 \\ 0 & 0 & -10 & 1.5 \end{array}\right] \tag{3.46}$$

que en términos de sistemas de ecuaciones quedaría como

$$\begin{aligned} 4\,x_1 - 9\,x_2 + 2\,x_3 &= 5 \\ 0.5x_2 + 5\,x_3 &= 0.5 \\ -10\,x_3 &= 1.5 \end{aligned} \tag{3.47}$$

Un proceso de sustitución regresiva produce el resultado buscado. La tercera ecuación de 3.47 da el valor de $x_3 = -0.15$; de la segunda ecuación se obtiene entonces

$$0.5\,x_2 = 0.5 - 5\,x_3 = 1.25$$

y por tanto $x_2 = 2.5$

finalmente al sustituir x_2 y x_3 en la primera ecuación de la forma 3.47 resulta

$$4\,x_1 = 5 + 9\,x_2 - 2\,x_3 = 27.8,$$

de modo que $x_1 = 6.95$

Con la sustitución de estos valores en el sistema original se verifica la exactitud de los resultados**.

---

* Nótese que los vectores columna de A se han ortogonalizado en la triangularización.

** Véase matrices mal condicionadas (Sec. 3.4).

Como producto secundario de este trabajo, se puede calcular fácilmente el **determinante de la matriz** A del sistema original. La matriz coeficiente A pasa de la forma original a la matriz triangular superior

$$\begin{bmatrix} 4 & -9 & 2 \\ 0 & 0.5 & 5 \\ 0 & 0 & -10 \end{bmatrix} \tag{3.48}$$

mediante operaciones que, de acuerdo con las reglas de los determinantes, no alteran el valor de $| A |$. El determinante de la ecuación 3.48 es sólo el producto de los elementos de la diagonal principal, de modo que el resultado es

$$| A | = 4(0.5)(-10) = -20$$

Las ecuaciones para la triangularización, sustitución regresiva y cálculo del determinante de un sistema de n ecuaciones en n incógnitas $A\ \mathbf{x} = \mathbf{b}$ por el método de eliminación de Gauss son

Triangularización

$$\begin{aligned} &\text{Para } 1 \le i \le n-1 \\ &\quad \text{Para } i+1 \le k \le n \\ &\qquad b_k = b_k - (a_{k,i}/a_{i,i})\, b_i \\ &\qquad \text{Para } i+1 \le j \le n \\ &\qquad\quad a_{j,i} = 0 \\ &\qquad\quad a_{kj} = a_{kj} - \frac{a_{k,i}}{a_{i,i}}\, a_{ij} \end{aligned} \tag{3.49}$$

Sustitución regresiva

$x_n = b_n/a_{n,n}$

Para $i = n-1, n-2, \ldots, 1$

$$x_i = \frac{1}{a_{i,i}} \left[b_i - \sum_{j=i+1}^{n} a_{i,j}\, x_j \right] \tag{3.50}$$

Cálculo del determinante

$$\det A = \prod_{i=1}^{n} a_{i,i} = a_{1,1}\, a_{2,2} \cdots a_{n,n} \tag{3.51}$$

El algoritmo para resolver $A\ \mathbf{x} = \mathbf{b}$ por eliminación de Gauss queda entonces

ALGORITMO 3.3 Eliminación de Gauss

Para obtener la solución de un sistema de ecuaciones lineales $A\ \mathbf{x} = \mathbf{b}$ y el determinante de A, proporcionar los

DATOS: N número de ecuaciones, A matriz coeficiente y **b** vector de términos independientes.

RESULTADOS: El vector solución **x** y el determinante de A o mensaje de falla "*HAY UN CERO EN LA DIAGONAL PRINCIPAL*".

PASO 1. Hacer DET = 1
PASO 2. Hacer I = 1
PASO 3. Mientras I ≤ N–1, repetir los pasos 4 a 14.
  PASO 4. Hacer DET = DET * A (I,I)
  PASO 5. Si DET = 0 IMPRIMIR mensaje "*HAY UN CERO EN LA DIAGONAL PRINCIPAL*" y TERMINAR. De otro modo continuar.
  PASO 6. Hacer K = I + 1
  PASO 7. Mientras K ≤ N, repetir los pasos 8 a 13.
    PASO 8. Hacer J = I + 1
    PASO 9. Mientras J ≤ N, repetir los pasos 10 y 11.
      PASO 10. Hacer A (K, J)=A (K, J) –A (K, I)*A (I, J) / A (I, I)
      PASO 11. Hacer J = J + 1
    PASO 12. Hacer b (K) = b (K) –A (K, I) * b (I) /A (I, I)
    PASO 13. Hacer K = K + 1
  PASO 14. Hacer I = I + 1
PASO 15. Hacer DET = DET * A (N, N)
PASO 16. Si DET = 0 IMPRIMIR mensaje "*HAY UN CERO EN LA DIAGONAL PRINCIPAL*" y TERMINAR. De otro modo continuar.
PASO 17. Hacer x (N) = b (N) / A (N, N)
PASO 18. Hacer I = N – 1
PASO 19. Mientras I ≥ 1, repetir los pasos 20 a 26.
  PASO 20. Hacer x (I) = b (I)
  PASO 21. Hacer J = I + 1
  PASO 22. Mientras J ≤ N, repetir los pasos 23 y 24.
    PASO 23. Hacer x (I) = x (I) – A (I, J) * x (J)
    PASO 24. Hacer J = J + 1
  PASO 25. Hacer x (I) = x (I) / A (I, I)
  PASO 26. Hacer I = I – 1
PASO 27. IMPRIMIR **x** y DET y TERMINAR.

Eliminación de Gauss con pivoteo

En la eliminación de x_1 de la segunda y tercera ecuaciones de la forma 3.40 se tomó como base la primera, por lo cual se denomina ecuación pivote o, en términos de la notación matricial, **fila pivote**. Para eliminar x_2 de la tercera ecuación de la forma 3.41, la fila pivote utilizada fue la segunda. El coeficiente de la incógnita que se va a eliminar en la fila pivote se llama **pivote**. En la eliminación que dio como resultado el sistema de ecuaciones 3.42, los pivotes fueron a_{11} y a'_{22}. Esta elección natural de los pivotes a_{11}, a'_{22}, a''_{33} , etc., es muy conveniente tanto para trabajar con una calculadora como con una computadora; desafortunadamente falla cuando alguno de esos elementos es cero, puesto que los multiplicadores quedarían indeterminados [por ejemplo si $a_{1,1}$ fuera cero, el multiplicador $(-a_{21} \ / \ a_{11})$ no está definido]. Una manera de evitar esta posibilidad es seleccionar como pivote el coeficiente de máximo valor absoluto en la columna relevante de la matriz reducida. Como antes, se tomarán las columnas en orden natural de modo que se vayan eliminando las incógnitas también en orden natural x_1, x_2, x_3, etc. Esta técnica, llamada **pivoteo parcial**, se ilustra con la solución del siguiente sistema.

Ejemplo 3.29

Resuelva el sistema

$$\begin{aligned} 10\,x_1 + x_2 - 5\,x_3 &= 1 \\ -20\,x_1 + 3\,x_2 + 20\,x_3 &= 2 \\ 5\,x_1 + 3\,x_2 + 5\,x_3 &= 6. \end{aligned} \tag{3.52}$$

SOLUCIÓN

La matriz aumentada es

$$\left[\begin{array}{ccc|c} 10 & 1 & -5 & 1 \\ -20 & 3 & 20 & 2 \\ 5 & 3 & 5 & 6 \end{array}\right] \tag{3.53}$$

El primer pivote debe ser (–20), ya que es el elemento de máximo valor absoluto en la primera columna. Se elimina entonces x_1 de la primera y tercera filas de la ecuación 3.52. Para ello, se suma a la primera fila la segunda multiplicada por (–10 / (–20)), y a la tercera fila la segunda multiplicada por (–5 / (–20)). Con esto se obtiene la matriz reducida

$$\left[\begin{array}{ccc|c} 0 & 2.5 & 5 & 2 \\ -20 & 3 & 20 & 2 \\ 0 & 3.75 & 10 & 6.5 \end{array}\right] \tag{3.54}$$

El siguiente pivote debe seleccionarse entre la primera y tercera filas (segunda columna) y en este caso es (3.75). Sumando a la primera fila la tercera multiplicada por (–2.5 / 3.75), resulta

$$\left[\begin{array}{ccc|c} 0 & 0 & -1.666 & -2.333 \\ -20 & 3 & 20 & 2 \\ 0 & 3.75 & 10 & 6.5 \end{array}\right] \tag{3.55}$$

que puesta en forma de sistema de ecuaciones queda

$$\begin{aligned} -1.666\, x_3 &= -2.333 \\ -20\, x_1 + 3\, x_2 + 20\, x_3 &= 2 \\ 3.75\, x_2 + 10\, x_3 &= 6.5 \end{aligned} \tag{3.56}$$

De la primera ecuación de la forma 3.56

$$x_3 = \frac{-2.333}{-1.666} = 1.4 ,$$

de la tercera ecuación

$$x_2 = \frac{6.5 - 10(1.4)}{3.75} = -2 ,$$

y finalmente de la segunda ecuación

$$x_1 = \frac{2 - 3(-2) - 20(1.4)}{-20} = 1 .$$

Otra alternativa para solucionar el sistema de ecuación 3.52 es utilizar el mismo criterio de selección de los pivotes, pero llevando las filas pivote a las posiciones de modo que se obtenga la forma triangular en la eliminación. Para esto es necesario, por ejemplo en la ecuación 3.52, intercambiar la segunda fila (donde se encuentra el elemento de máximo valor absoluto) con la primera, con lo que se obtiene

$$\left[\begin{array}{ccc|c} -20 & 3 & 20 & 2 \\ 10 & 1 & -5 & 1 \\ 5 & 3 & 5 & 6 \end{array}\right] \tag{3.53'}$$

que se reduce en la primera eliminación a

$$\left[\begin{array}{ccc|c} -20 & 3 & 20 & 2 \\ 0 & 2.5 & 5 & 2 \\ 0 & 3.75 & 10 & 6.5 \end{array}\right] \tag{3.54'}$$

Como el siguiente pivote es (3.75), se intercambian la segunda y la tercera filas de la ecuación 3.54' para obtener

$$\left[\begin{array}{ccc|c} -20 & 3 & 20 & 2 \\ 0 & 3.75 & 10 & 6.5 \\ 0 & 2.5 & 5 & 2 \end{array}\right] \tag{3.54"}$$

la cual se reduce al eliminar x_2 a

$$\left[\begin{array}{ccc|c} -20 & 3 & 20 & 2 \\ 0 & 3.75 & 10 & 6.5 \\ 0 & 0 & -1.666 & -2.333 \end{array}\right] \tag{3.55'}$$

que tiene ya la forma triangular y está lista para la sustitución regresiva. **En adelante, cualquier referencia a la eliminación con pivoteo que se haga, entraña la segunda alternativa.**

La sustitución regresiva proporciona los siguientes valores

$$x_3 = 1.4, \quad x_2 = -2, \quad x_1 = 1$$

El determinante de A se calcula de nuevo, multiplicando entre sí los elementos de la diagonal principal de la matriz triangularizada (Ec. 3.55'), pero dicho producto es afectado por un cambio de signo por cada intercambio de filas que se verifique en la triangularización. En el caso en estudio

$$\det A = (-1)^2 \ (-20) \ (3.75) \ (-1.666) = 125$$

ya que hubo dos intercambios de fila para llegar a la ecuación 3.55'.

A fin de elaborar el algoritmo de este método, se utilizarán las ecuaciones 3.49 para la triangularización después de cada búsqueda del elemento de máximo valor absoluto y del intercambio de filas correspondiente. Una vez realizada la triangularización, se hará la sustitución regresiva con las ecuaciones 3.50 y el cálculo del determinante de la siguiente forma

$$\det A = (-1)^r \prod_{i=1}^{n} a_{i,i} \tag{3.57}$$

donde r es el número de intercambios de filas que hubo en el proceso de triangularización.

ALGORITMO 3.4 Eliminación de Gauss con pivoteo

Para obtener la solución de un sistema de ecuaciones lineales A **x** = **b** y el determinante de A, proporcionar los

DATOS: N número de ecuaciones, A matriz coeficiente y **b** vector de términos independientes.

RESULTADOS: El vector solución **x** y el determinante de A o mensaje "*MATRIZ SINGULAR, SISTEMA SIN SOLUCIÓN*".

PASO 1. Hacer DET = 1

PASO 2. Hacer R = 0

PASO 3. Hacer I = 1
PASO 4. Mientras I ≤ N − 1 repetir los pasos 5 a 12.
  PASO 5. Encontrar PIVOTE (elemento de mayor valor absoluto en la parte relevante de la columna I de A) y P la fila donde se encuentra PIVOTE.
  PASO 6 Si PIVOTE = 0 IMPRIMIR "*MATRIZ SINGULAR, SISTEMA SIN SOLUCION*" y TERMINAR. En caso contrario continuar.
  PASO 7. Si P = I ir al paso 10. De otro modo realizar los pasos 8 y 9.
    PASO 8. Intercambiar la fila I con la fila P.
    PASO 9. Hacer R = R + 1
  PASO 10. Hacer DET = DET * A(I, I)
  PASO 11. Realizar los pasos 6 a 13 del algoritmo 3.3
  PASO 12. Hacer I = I + 1
PASO 13. Hacer DET = DET * A(N, N) * (−1)**r
PASO 14. Realizar los pasos 17 a 26 del algortimo 3.3
PASO 15. IMPRIMIR **x** y DET y TERMINAR.

Para terminar el tema, se compararán las técnicas de eliminación de Gauss con pivoteo y sin éste. Por brevedad, la primera se denominará GP y la segunta G.

1. La búsqueda del coeficiente de mayor valor absoluto que se usará como pivote y el intercambio de filas significa mayor programación en GP.
2. Los factores ($a_{k,i}$ / $a_{i,i}$) de las ecuaciones 3.49 siempre serán menores que la unidad en valor absoluto en GP, con esto los elementos de A | **b** se conservan dentro de cierto intervalo, circunstancia valiosa en los cálculos computacionales.
3. Encontrar en GP un pivote igual a cero significaría que se trata de una matriz coeficiente A singular (det A = 0) y que el sistema A **x** = **b** no tiene solución única. Encontrar en G un pivote igual a cero, no proporciona información alguna acerca del determinante de A y sí detendría el proceso de triangularización.

A pesar de la programación adicional y el mayor tiempo de máquina que se emplea en el método de Gauss con pivoteo, sus otras ventajas borran totalmente estas desventajas en la práctica; por tanto, el pivoteo natural se usa sólo en circunstancias especiales, por ejemplo cuando se sabe por adelantado que no hay pivotes más grandes que los que van resultando en la diagonal principal.

Eliminación de Jordan

Es posible extender los métodos vistos de modo que las ecuaciones se reduzcan a una forma en que la matriz coeficiente del sistema sea diagonal y ya no se requiera la sustitución regresiva. Los pivotes se eligen como en el método de Gauss con pivoteo, y una vez intercambiadas las filas se eliminan los elementos arriba y debajo del pivote. El sistema del ejemplo 3.28 ilustra este método.

Ejemplo 3.30

Por eliminación de Jordan, resuelva el sistema

$$\begin{aligned} 4\,x_1 - 9\,x_2 + 2\,x_3 &= 5 \\ 2\,x_1 - 4\,x_2 + 6\,x_3 &= 3 \\ x_1 - x_2 + 3\,x_3 &= 4. \end{aligned}$$

SOLUCIÓN

La matriz aumentada del sistema es

$$\left[\begin{array}{ccc|c} 4 & -9 & 2 & 5 \\ 2 & -4 & 6 & 3 \\ 1 & -1 & 3 & 4 \end{array}\right]$$

Como en la primera columna el elemento de máximo valor absoluto se encuentra en la primera fila, ningún intercambio es necesario y el primer paso de eliminación produce

$$\left[\begin{array}{ccc|c} 4 & -9 & 2 & 5 \\ 0 & 0.5 & 5 & 0.5 \\ 0 & 1.25 & 2.5 & 2.75 \end{array}\right]$$

El elemento de máximo valor absoluto en la parte relevante de la segunda columna (filas 2 y 3) es 1.25; por tanto, la fila 3 debe intercambiarse con la 2.

$$\left[\begin{array}{ccc|c} 4 & -9 & 2 & 5 \\ 0 & 1.25 & 2.5 & 2.75 \\ 0 & 0.5 & 5 & 0.5 \end{array}\right]$$

Sumando la segunda multiplicada por (–(–9) / 1.25), a la primera fila y la segunda multiplicada por (–0.5 / 1.25) la tercera, se obtiene el nuevo arreglo

$$\left[\begin{array}{ccc|c} 4 & 0 & 20 & 24.8 \\ 0 & 1.25 & 2.5 & 2.75 \\ 0 & 0 & 4 & -0.6 \end{array}\right]$$

donde se han eliminado los elementos de arriba y abajo del pivote (nótese que en este paso el primer pivote no se modifica porque sólo hay ceros debajo de él).

Por último, sumando la tercera multiplicada por (–20/4) a la primera fila y la tercera multiplicada por (–2.5/4) a la segunda

$$\left[\begin{array}{ccc|c} 4 & 0 & 0 & 27.8 \\ 0 & 1.25 & 0 & 3.125 \\ 0 & 0 & 4 & -0.6 \end{array}\right]$$

que escrita de nuevo como sistema de ecuaciones da

$$\begin{aligned} 4\,x_1 &= 27.8 \\ 1.25\,x_2 &= 3.125 \\ 4\,x_3 &= -0.6 \end{aligned}$$

de donde el resultado final se obtiene fácilmente

$$x_1 = \frac{27.8}{4} = 6.95\,, \quad x_2 = \frac{3.125}{1.25} = 2.5\,, \quad x_3 = \frac{-0.6}{4} = -0.15$$

El determinante también puede calcularse

$$|\,A\,| = (-1)^1\,(4)\,(1.25)\,(4) = -20,$$

donde la potencia 1 indica que sólo hubo un intercambio de filas.

Si sólo se requiere calcular $|\,A\,|$ y no la solución del sistema, el método de Jordan requiere mayor trabajo que el método de eliminación de Gauss con pivoteo.

Sugerencia: Utilice el software del libro para estudiar y analizar este método o bien para resolver sistemas lineales.

Cálculo de inversas

Si se tienen varios sistemas por resolver que comparten la misma matriz coeficiente; es decir

$$A\,\mathbf{x}_1 = \mathbf{b}_1\,, \quad A\,\mathbf{x}_2 = \mathbf{b}_2\,, \text{ etc.}$$

pueden resolverse todos a un tiempo si se aplica al arreglo

$$[\,A \;|\; \mathbf{b}_1 \;|\; \mathbf{b}_2 \;\ldots\,]$$

el proceso de eliminación como antes y después se realiza una sustitución regresiva particular para cada columna del lado derecho de A. Como caso particular es factible encontrar A^{-1} si $\mathbf{b}_1 = \mathbf{e}_1$, $\mathbf{b}_2 = \mathbf{e}_2$, ..., $\mathbf{b}_n = \mathbf{e}_n$.* Las n soluciones obtenidas forman las n columnas de la matriz inversa A^{-1}.

Cálculo de la inversa con el método de Gauss con pivoteo

Como ejemplo se usará la matriz coeficiente del sistema (3.44) para obtener su inversa. Primero se forma el arreglo

$$\left[\begin{array}{ccc|ccc} 4 & -9 & 2 & 1 & 0 & 0 \\ 2 & -4 & 6 & 0 & 1 & 0 \\ 1 & -1 & 3 & 0 & 0 & 1 \end{array}\right], \tag{3.58}$$

*En este caso $\mathbf{e}_1$, $\mathbf{e}_2$, etc., son vectores de n elementos cuyo único elemento distinto de cero es el de la fila 1, 2, etc., y su valor es 1.

nótese que a la derecha de A se tiene la matriz identidad correspondiente. Eliminando los elementos debajo del primer pivote (4), se llega al sistema

$$\left[\begin{array}{ccc|ccc} 4 & -9 & 2 & 1 & 0 & 0 \\ 0 & 0.5 & 5 & -0.5 & 1 & 0 \\ 0 & 1.25 & 2.5 & -0.25 & 0 & 1 \end{array}\right], \tag{3.59}$$

Se intercambian la segunda y tercera filas.

$$\left[\begin{array}{ccc|ccc} 4 & -9 & 2 & 1 & 0 & 0 \\ 0 & 1.25 & 2.5 & -0.25 & 0 & 1 \\ 0 & 0.5 & 5 & -0.5 & 1 & 0 \end{array}\right]. \tag{3.60}$$

Ahora se elimina el segundo elemento de la tercera fila y el arreglo cambia a

$$\left[\begin{array}{ccc|ccc} 4 & -9 & 2 & 1 & 0 & 0 \\ 0 & 1.25 & 2.5 & -0.25 & 0 & 1 \\ 0 & 0 & 4 & -0.4 & 1 & -0.4 \end{array}\right].$$

Con la sustitución regresiva para el primer vector al lado derecho de la matriz triangular resulta

$$4\ x_3 = -\ 0.4, \text{ de donde } x_3 = -\ 0.1;$$

al sustituir x_3 en la fila 2 se tiene

$$1.25\ x_2 = -0.25 - 2.5(-0.1) \text{ y } x_2 = 0;$$

y remplazando x_3 y x_2 en la fila 1, se obtiene

$$4\ x_1 = 1 + 9(0) - 2(-0.1) = 1.2 \text{ y } x_1 = 0.3$$

Este primer vector solución representa la primera columna de A^{-1}. Del mismo modo se calculan la segunda y tercera columnas de A^{-1} con el segundo y tercer vectores del lado derecho de la matriz triangular

$$A^{-1} = \begin{bmatrix} 0.3 & -1.25 & 2.3 \\ 0 & -0.5 & 1.0 \\ -0.1 & 0.25 & -0.1 \end{bmatrix}$$

Cálculo de la inversa con el método de Jordan

Se parte del mismo arreglo (Ec. 3.58) y también se eliminan los elementos debajo del primer pivote para llegar a la ecuación 3.50. Se intercambian la segunda y tercera filas y se llega al sistema de ecuaciones 3.60. En este último arreglo se eliminan los elementos arriba y debajo del pivote (1.25) para llegar a

$$\left[\begin{array}{ccc|ccc} 4 & 0 & 20 & -0.8 & 0 & 7.2 \\ 0 & 1.25 & 2.5 & -0.25 & 0 & 1 \\ 0 & 0 & 4 & -0.4 & 1 & -0.4 \end{array}\right]$$

arreglo que todavía se reduce a

$$\left[\begin{array}{ccc|ccc} 4 & 0 & 0 & 1.2 & -5 & 9.2 \\ 0 & 1.25 & 0 & 0 & -0.625 & 1.25 \\ 0 & 0 & 4 & -0.4 & 1 & -0.4 \end{array}\right],$$

y que con la primera columna a la derecha de la matriz diagonal produce

$$x_1 = \frac{1.2}{4} = 0.3\ , \quad x_2 = \frac{0}{1.25} = 0\ , \quad x_3 = \frac{-0.4}{4} = -0.1\ ,$$

con la segunda columna

$$x_1 = \frac{-5}{4} = -1.25\ , \quad x_2 = \frac{-0.625}{1.25} = -0.5\ , \quad x_3 = 0.25$$

De igual manera con la tercera columna para llegar a

$$A^{-1} = \begin{bmatrix} 0.3 & -1.25 & 2.3 \\ 0 & -0.5 & 1.0 \\ -0.1 & 0.25 & -0.1 \end{bmatrix}$$

Los métodos de eliminación vistos proporcionan la solución del sistema $A\ \mathbf{x} = \mathbf{b}$, el det A y A^{-1}, siempre que A sea no singular.

Obsérvese por otro lado que si se tiene un conjunto de vectores $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ de n componentes cada uno y se quieren ortogonalizar, se aplica alguna de las eliminaciones vistas al conjunto dado tomado como una matriz. La técnica de Gauss con pivoteo también puede aplicarse —por ejemplo— para determinar si dicho conjunto es linealmente independiente o no (cuando un elemento pivote $a_{i,i}$ es igual a cero, la fila correspondiente es linealmente dependiente de las filas anteriores).

La sección que sigue puede omitirse sin pérdida de continuidad en los siguientes temas.

Cuenta de operaciones

Para establecer la velocidad de cálculo y el "trabajo computacional", se requiere conocer cuántos cálculos de los diferentes tipos se realizan. Considérese para ello la reducción del sistema general

$$\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \quad\cdot \qquad\quad \cdot \qquad\qquad\quad \cdot \qquad\quad \cdot \\ \quad\cdot \qquad\quad \cdot \qquad\qquad\quad \cdot \qquad\quad \cdot \\ \quad\cdot \qquad\quad \cdot \qquad\qquad\quad \cdot \qquad\quad \cdot \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_m \end{array} \tag{3.61}$$

a la forma triangular

$$\begin{aligned} t_{11}x_1 + t_{12}x_2 + \ldots + t_{1n}x_n &= c_1 \\ t_{22}x_2 + \ldots + t_{2n}x_n &= c_2 \\ \vdots \quad & \quad \vdots \\ t_{nn}x_n &= c_n \end{aligned} \tag{3.62}$$

o en notación matricial más compacta de $[A \mid \mathbf{b}]$ a $[T \mid \mathbf{c}]$, matrices ambas de $n \times (n + 1)$. Sea

M_n = número de multiplicaciones o divisiones
S_n = número de sumas y restas

necesarias para ir del sistema 3.61 al 3.62

Evidentemente $M_1 = 0$ y $S_1 = 0$, ya que cualquier matriz A de 1×1 es triangular. Si $n > 1$, se considera la eliminación en la primera columna. Si la primera columna de A es distinta del vector cero, generalmente se intercambian filas con el fin de llevar el elemento de máximo valor absoluto de la primera columna a la posición (1,1). Denomínese de nuevo $[A \mid \mathbf{b}]$ el sistema resultante de este intercambio. Ahora debe restarse un múltiplo de la nueva primera fila

$$a_{11}\ a_{12}\ a_{13}\ \ldots\ a_{1n}\ b_1$$

de cada fila

$$a_{i1}\ a_{i2}\ a_{i3}\ \ldots\ a_{in}\ b_i \qquad 2 \leq i \leq n \tag{3.63}$$

para producir filas de la forma

$$0\ a'_{i2}\ a'_{i3}\ \ldots\ a'_{in}\ b'_i \qquad 2 \leq i \leq n \tag{3.64}$$

Explícitamente, si $r_i = a_{i1} / a_{11}$,

$$\begin{aligned} a'_{ij} &= a_{ij} - r_i\, a_{1j} \\ b'_i &= b_i - r_i\, b_1 \end{aligned} \qquad 2 \leq j \leq n \tag{3.65}$$

Se efectúa una división para producir r_i. La fórmula 3.65 requiere n multiplicaciones y un número igual de restas. Como se forman $(n-1)$ filas, la eliminación en la primera columna se logra con

$(n + 1)\ (n - 1)$ divisiones o multiplicaciones y
$n\ (n - 1)$ restas. (3.66)

La primera columna ya tiene ceros debajo de la posición (1,1). Queda por reducir la matriz de $(n - 1) \times n$, matriz debajo de la primera fila y a la derecha de la primera columna. De la fórmula 3.66, se obtienen las fórmulas

$$\begin{aligned} M_n &= (n + 1)\ (n - 1) + M_{n-1} \\ S_n &= n(n - 1) + S_{n-1} \end{aligned} \qquad n \geq 2 \tag{3.67}$$

Como $M_1 = S_1 = 0$, se tiene para $n \geq 2$

$$\begin{aligned} M_n &= (2 + 1)\ 1 + (3 + 1)\ 2 + \ldots + (n + 1)\ (n - 1) \\ S_n &= 2(1) + 3(2) + \ldots\ n(n - 1) \end{aligned} \tag{3.68}$$

Fácilmente se verifica por inducción que

$$\sum_{t=1}^{n-1} t = \frac{1}{2}(n - 1)n; \qquad \sum_{t=1}^{n} t^2 = \frac{1}{6}(n - 1)\ n\ (2n - 1);$$

Por tanto, como

$$M_n = \sum_{t=1}^{n-1} (t + 1 + 1)\ t$$

y

$$S_n = \sum_{t=1}^{n-1} (t + 1)\ t$$

Entonces

$$\begin{aligned} M_n &= \frac{1}{6}(n - 1)n\ (2n - 1) + (n - 1)n \\ S_n &= \frac{1}{6}(n - 1)n\ (2n - 1) + \frac{1}{2}(n-1)n \end{aligned} \tag{3.69}$$

Se determinará el número m_n de multiplicaciones o divisiones y el número s_n de sumas o restas requeridas para resolver el sistema triangular $[\ T\ |\ \mathbf{x}\ ] = \mathbf{c}$. Sean $n \geq 2$ y todas las $t_{ii} \neq 0$. Supóngase que se han calculado $x_n, x_{n-1}, \ldots, x_2$; llámense m_{n-1} y s_{n-1} las operaciones realizadas para ello.

Sea ahora

$$x_1 = \frac{c_1 - t_{12}x_2 - \ldots - t_{1r}x_n}{t_{11}} \tag{3.70}$$

El cálculo de x_1 requiere $(n-1)$ multiplicaciones, una división y $(n-1)$ restas. Entonces, para $n \geq 2$

$$\begin{aligned} m_n &= (n - 1 + 1) + m_{n-1} \\ s_n &= (n - 1) + s_{n-1} \end{aligned} \tag{3.71}$$

Como $m_1 = 1$ y $s_1 = 0$, se tiene

$$\begin{aligned} m_n &= 1 + 2 + 3 + \ldots + n = \frac{1}{2} n(n + 1) \\ s_n &= 1 + 2 + 3 + \ldots + (n - 1) = \frac{1}{2}(n-1)\ n \end{aligned} \tag{3.72}$$

El resultado final se resume a continuación.

El sistema 3.61 con matriz coeficiente A y determinante distinto de cero, puede resolverse por el método de eliminación con pivoteo con

$$\begin{aligned} u_n &= M_n + m_n = \frac{1}{6}(n-1)\,n\,(2n-1) + (n-1)n + \frac{1}{2}n\,(n+1) \\ &= \frac{1}{3}n^3 + n^2 - \frac{1}{3}n \quad \text{multiplicaciones o divisiones y} \\ v_n &= S_n + s_n = \frac{1}{6}(n-1)\,n\,(2n-1) + \frac{1}{2}(n-1)\,n + \frac{1}{2}(n-1)\,n \\ &= \frac{1}{3}n^3 + \frac{1}{2}n^2 - \frac{5}{6}n \quad \text{sumas o restas.} \end{aligned} \tag{3.63}$$

Obviamente, el "trabajo computacional" para resolver la ecuación 3.61 es función del número de operaciones necesarias (Ec. 3.73); por tanto, puede decirse que es proporcional a n^3. Por otro lado, las necesidades de memoria de máquina serán proporcionales a n^2.

Sistemas especiales

Con frecuencia la matriz coeficiente del sistema $A\ \mathbf{x} = \mathbf{b}$ por resolver es simétrica, o bien gran número de sus componentes son cero (matrices dispersas). En estos casos algunos de los métodos conocidos pueden adaptarse, con lo cual se reduce el trabajo computacional y la memoria de máquina. Primero se tratará el caso de las matrices bandeadas (matrices dispersas particulares); las matrices simétricas serán abordadas como un caso particular de los métodos L-U.

Primero se darán algunos ejemplos particulares de matrices bandeadas

$$\begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 7 & 0 \\ 0 & 0 & 0 & 0 & 6 \end{bmatrix} \qquad \begin{bmatrix} 4 & 0 & 0 & 0 & 0 \\ 7 & 8 & 1 & 0 & 0 \\ 0 & 0 & 5 & 2 & 0 \\ 0 & 0 & 1 & 3 & 5 \\ 0 & 0 & 0 & 3 & 4 \end{bmatrix} \qquad \begin{bmatrix} 8 & 7 & 6 & 0 & 0 \\ 9 & 3 & 0 & -2 & 0 \\ 3 & -1 & 8 & 9 & 10 \\ 0 & 0 & 3 & 5 & 8 \\ 0 & 0 & 7 & 4 & 0 \end{bmatrix}$$

Diagonal Tridiagonal Pentadiagonal

Generalizando : Una matriz A de $n \times n$ es tridiagonal si

$$a_{i,j} = 0 \text{ siempre que } |\ i - j\ | > 1\ ,$$

pentadiagonal si

$$a_{i,j} = 0 \text{ siempre que } |\ i - j\ | > 2, \text{ etc.}$$

El ancho de banda es 1, 3, 5, etc., en las matrices diagonales, tridiagonales, pentadiagonales, etc., respectivamente.

Enseguida se adapta la eliminación de Gauss para la solución del sistema tridiagonal $A\ \mathbf{x} = \mathbf{b}$; es decir, A es tridiagonal.

Método de Thomas

Sea el sistema tridiagonal de tres ecuaciones en tres incógnitas

$$\begin{aligned} b_1 x_1 + c_1 x_2 &= d_1 \\ a_2 x_1 + b_2 x_2 + c_2 x_3 &= d_2 \\ a_3 x_2 + b_3 x_3 &= d_3 \end{aligned}$$

Triangularización

Si $b_1 \neq 0$, se elimina x_1 **sólo** en la segunda ecuación, con lo que se obtiene como nueva segunda ecuación

$$b'_2 x_2 + c'_2 x_3 = d'_2$$

con

$$b'_2 = b_2 - a_2 c_1/b_1;\ c'_2 = c_2;\ d'_2 = d_2 - a_2 d_1/b_1$$

Si $b'_2 \neq 0$, x_2 se elimina **sólo** en la tercera ecuación, y así se obtiene como nueva tercera ecuación

$$b'_3 x_3 = d'_3$$

con

$$b'_3 = b_3 - a_3 c_2/b'_2;\ d'_3 = d_3 - a_3 d'_2/b'_2$$

Generalizando: Para un sistema tridiagonal de n ecuaciones en n incógnitas.

Triangularización

Para $i = 1, 2, \ldots, n-1$

si $b'_i \neq 0$ se elimina x_i sólo en la $(i + 1)$-ésima ecuación, con lo que se obtiene como nueva $(i + 1)$-ésima ecuación

$$b'_{i+1} x_{i+1} + c'_{i+1} x_{i+2} = d'_{i+1}$$

con

$$b'_{i+1} = b_{i+1} - a_{i+1} c_i/b'_i\ ;\ d'_{i+1} = d_{i+1} - a_{i+1} d'_i / b'_i$$

y

$$c'_{i+1} = c_i$$

Sustitución regresiva

$$x_n = d'_n/b'_n$$

y para $i = n-1, n-2, \ldots, 1$

$$x_i = \frac{d_i' - c_i' x_{i+1}}{b_i'}$$

Esta simplificación del algoritmo de Gauss, válida para sistemas tridiagonales se conoce como método de Thomas. Con su aplicación se consiguen las siguientes ventajas:

- La memoria de máquina se reduce al no tener que almacenar los elementos de A que son cero. Obsérvese que en lugar de almacenar la matriz A, se guardan sólo los vectores $\mathbf{a} = (a_1, a_2, \ldots, a_n)$, $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ y $\mathbf{c} = (c_1, c_2, \ldots, c_n)$ con $a_1 = c_n = 0$, empleando $3n$ localidades en lugar de $n \times n$ localidades, ventaja muy importante cuando n es grande ($n \geq 50$).
- No se requiere pivotear.
- Sólo se elimina durante el i-ésimo paso de la triangularización la variable x_i en la ecuación $i + 1$, con lo que se reduce el número de operaciones.
- Por último, en la sustitución regresiva debe remplazarse sólo x_{i+1} en la i-ésima ecuación para obtener x_i.

Ejemplo 3.31

Resuelva el sistema tridiagonal

$$\begin{aligned} 3x_1 - 2x_2 &= 1.0 \\ x_1 + 5x_2 - 0.2x_3 &= 5.8 \\ 4x_2 + 7x_3 &= 11.0 , \end{aligned}$$

por el método de Thomas.

SOLUCIÓN

En este sistema

$$\mathbf{b} = \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} 0 \\ 1 \\ 4 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} -2 \\ -0.2 \\ 0 \end{bmatrix} \quad \text{y} \quad \mathbf{d} = \begin{bmatrix} 1.0 \\ 5.8 \\ 11.0 \end{bmatrix}$$

Como $b_1 \neq 0$ se calculan las componentes de la nueva segunda fila

$$b'_2 = b_2 - a_2 c_1/b_1 = 5 - 1(-2) / 3 = 5.6666$$

y

$$c'_2 = c_2 = -0.2$$

$$d'_2 = d_2 - a_2 d_1/b_1 = 5.8 - 1(1/3) = 5.4666$$

Como

$b'_2 \neq 0$, se forma la nueva tercera fila

$$b'_3 = b_3 - a_3 c_2/b'_2 = 7 - 4(-0.2) / 5.6666 = 7.141176$$

$$d'_3 = d_3 - a_3 d'_2/b'_2 = 11.0 - 4(5.4666) / 5.6666 = 7.1411760$$

El sistema equivalente resultante es

$$\begin{aligned} 3\,x_1 - 2\,x_2 &= 1.0 \\ 5.6666\,x_2 - 0.2\,x_3 &= 5.4666 \\ 7.141176\,x_3 &= 7.141176 \end{aligned}$$

y por sustitución regresiva se llega a

$$x_3 = d'_3 \,/\, b'_3 = 7.141176/7.141176 = 1$$

$$x_2 = (d'_2 - c_2\,x_3) \,/\, b'_2 = (5.4666 - 0.2\,)(1))/5.6666 = 1$$

$$x_1 = (d'_1 - c_1\,x_2)/b'_1 = (1.0 - (-2)\,(1))/3 = 1$$

Nótese que $d'_1 = d_1$ y $b'_1 = b_1$

A continuación se da el algoritmo de Thomas.

ALGORITMO 3.5 Método de Thomas

Para obtener la solución **x** del sistema tridiagonal A **x** = **b** proporcionar los

DATOS: El número de ecuaciones N, los vectores **a**, **b**, **c**, y el vector de términos independientes **d**.

RESULTADOS: El vector solución **x** o mensaje de falla "*EL SISTEMA NO TIENE SOLUCION*"

PASO 1. Hacer I = 1
PASO 2. Mientras I ≤ N–1, repetir los pasos 3 a 6.
  PASO 3. Si b(I) ≠ 0 continuar. De otro modo IMPRIMIR el mensaje "*EL SISTEMA NO TIENE SOLUCION*" y TERMINAR.
  PASO 4. Hacer b(I+1) = b(I+1) – a (I+1) *c (I)/b(I)
  PASO 5. Hacer d(I+1) = d(I+1) – a(I+1) *d (I)/b(I)
  PASO 6. Hacer I = I + 1
PASO 7. Si b(N) ≠ 0 continuar. De otro modo IMPRIMIR mensaje "*EL SISTEMA NO TIENE SOLUCION*" y TERMINAR.
PASO 8. Hacer x(N) = d(N)/b(N)
PASO 9. Hacer I = N – 1
PASO 10. Mientras I ≥ 1, repetir los pasos 11 y 12.
  PASO 11. Hacer x(I) = (d(I) – c(I)*x(I+1))/b(I)
  PASO 12. Hacer I = I – 1
PASO 13. IMPRIMIR el vector solución x y TERMINAR.

Métodos de factorización.

Factorización de matrices en matrices triangulares

La eliminación de Gauss aplicada al sistema (véase ejemplo 3.28)

$$\begin{aligned} 4x_1 - 9x_2 + 2x_3 &= 5 \\ 2x_1 - 4x_2 + 6x_3 &= 3 \\ x_1 - x_2 + 3x_3 &= 4 \end{aligned}$$

condujo en su fase de triangularización al sistema equivalente

$$\left[\begin{array}{ccc|c} 4 & -9 & 2 & 5 \\ 0 & 0.5 & 5 & 0.5 \\ 0 & 0 & -10 & 1.5 \end{array}\right],$$

donde se aprecia una matriz triangular superior de orden 3 que se denotará como U

$$U = \begin{bmatrix} 4 & -9 & 2 \\ 0 & 0.5 & 5 \\ 0 & 0 & -10 \end{bmatrix}$$

Ahora se define una matriz triangular inferior L de orden 3, con números 1 a lo largo de la diagonal principal y con $l_{i,j}$ igual al factor que permitió eliminar el elemento $a_{i,j}$ del sistema 3.43 (por ejemplo, a fin de eliminar $a_{2,1} = 2$ se utilizó el factor $l_{2,1} = 2/4$; para eliminar $a_{3,1} = 1$, el factor $l_{3,1} = 1/4$, y para hacer cero a $a_{3,2} = -1$ se empleó $l_{3,2} = 1.25/0.5$). Así, la matriz L queda entonces

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2/4 & 1 & 0 \\ 1/4 & 1.25/0.5 & 1 \end{bmatrix},$$

cuyo producto con U resulta en

$$\begin{bmatrix} 1 & 0 & 0 \\ 2/4 & 1 & 0 \\ 1/4 & 1.25/0.5 & 1 \end{bmatrix} \begin{bmatrix} 4 & -9 & 2 \\ 0 & 0.5 & 5 \\ 0 & 0 & -10 \end{bmatrix} = A,$$

la matriz coeficiente del sistema original.

Esta descomposición de A en los factores L y U es cierta en general cuando la eliminación de Gauss puede aplicarse al sistema $A\,\mathbf{x} = \mathbf{b}$ sin intercambio de filas, o equivalentemente si y sólo si los determinantes de las submatrices de A son todos distintos de cero

$$| a_{1,1} | \neq 0, \quad \begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{vmatrix} \neq 0, \ldots, \quad \begin{vmatrix} a_{1,1} & \ldots & a_{1,n} \\ . & & . \\ . & & . \\ . & & . \\ a_{n,1} & \ldots & a_{n,n} \end{vmatrix} \neq 0$$

El resultado anterior permite resolver el sistema $A\,\mathbf{x} = \mathbf{b}$, ya que sustituyendo A por $L\,U$ se tiene

$$L\,U\,\mathbf{x} = \mathbf{b}$$

SOLUCIÓN

Se hace $U\,\mathbf{x} = \mathbf{c}$, donde $\mathbf{c}$ es un vector desconocido $[c_1\ c_2\ c_3\ \ldots\ c_n]^T$, que se puede obtener fácilmente resolviendo el sistema

$$L\,\mathbf{c} = \mathbf{b},$$

con sustitución progresiva o hacia adelante, ya que L es triangular inferior (en el sistema del Ejemplo 3.28, $\mathbf{c}$ resulta $[5\ 0.5\ 1.5]^T$).

Una vez calculado $\mathbf{c}$, se resuelve

$$U\,\mathbf{x} = \mathbf{c}$$

con sustitución regresiva, ya que U es triangular superior y de esa manera se obtiene el vector solución $\mathbf{x}$ (el sistema particular que se ha trabajado da $\mathbf{x} = [6.95\ 2.5\ -0.15]^T$).

Métodos de Doolitle y Crout

Aún cuando las matrices L y U pueden obtenerse en la triangularización de la matriz aumentada $[A \mid \mathbf{b}]$, es deseable encontrar un método más directo para su determinación. Esto es factible analizando la factorización de A en las matrices generales de orden tres L y U, dadas a continuación

$$\begin{bmatrix} l_{1,1} & 0 & 0 \\ l_{2,1} & l_{2,2} & 0 \\ l_{3,1} & l_{3,2} & l_{3,3} \end{bmatrix} \begin{bmatrix} u_{1,1} & u_{1,2} & u_{1,3} \\ 0 & u_{2,2} & u_{2,3} \\ 0 & 0 & u_{3,3} \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix}$$

Análisis

Se multiplican

a) Primera fila de L por las tres columnas de U

$$l_{1,1}u_{1,1} = a_{1,1}$$
$$l_{1,1}u_{1,2} = a_{1,2}$$
$$l_{1,1}u_{1,3} = a_{1,3}.$$

b) Segunda fila de L por las tres columnas de U

$$\begin{aligned} l_{2,1}u_{1,1} &= a_{2,1} \\ l_{2,1}u_{1,2} + l_{2,2}u_{2,2} &= a_{2,2} \\ l_{2,1}u_{1,3} + l_{2,2}u_{2,3} &= a_{2,3} \end{aligned}$$

c) Tercera fila de L por las tres columnas de U

$$\begin{aligned} l_{3,1}u_{1,1} &= a_{3,1} \\ l_{3,1}u_{1,2} + l_{3,2}u_{2,2} &= a_{3,2} \\ l_{3,1}u_{1,3} + l_{3,2}u_{2,3} + l_{3,3}u_{3,3} &= a_{3,3}, \end{aligned}$$

se llega a un sistema de nueve ecuaciones en 12 incógnitas $l_{1,1}$, $l_{2,1}$, $l_{2,2}$, $l_{3,1}$, $l_{3,2}$, $l_{3,3}$, $u_{1,1}$, $u_{1,2}$, $u_{1,3}$, $u_{2,2}$, $u_{2,3}$, $u_{3,3}$, por lo que será necesario establecer tres condiciones arbitrarias sobre las incógnitas para resolver dicho sistema. La forma de seleccionar las condiciones ha dado lugar a diferentes métodos; por ejemplo, si se toman de modo que $l_{1,1} = l_{2,2} = l_{3,3} = 1$, se obtiene el **método de Doolitle**; si en cambio se selecciona $u_{1,1} = u_{2,2} = u_{3,3} = 1$, el algoritmo resultante es llamado **método de Crout.**

Se continuará el desarrollo de la factorización. Tómese

$$l_{1,1} = l_{2,2} = l_{3,3} = 1$$

Con estos valores se resuelven las ecuaciones directamente en el orden en que están dadas

De (a) $u_{1,1} = a_{1,1}, \qquad u_{1,2} = a_{1,2}, \qquad u_{1,3} = a_{1,3}$ (3.74)

De (b) y sustituyendo los resultados (Ec. 3.74)

$$l_{2,1} = a_{2,1}/u_{1,1} = a_{2,1}/a_{1,1}$$

$$u_{2,2} = a_{2,2} - l_{2,1}u_{1,2} = a_{2,2} - \frac{a_{2,1}}{a_{1,1}}\, a_{1,2} \tag{3.75}$$

$$u_{2,3} = a_{2,3} - l_{2,1}u_{1,3} = a_{2,3} - \frac{a_{2,1}}{a_{1,1}}\, a_{1,3}$$

De (c) y sustituyendo los resultados de las ecuaciones 3.74 y 3.75

$$l_{3,1} = a_{3,1}/u_{1,1} = a_{3,1}/a_{1,1}$$

$$l_{3,2} = \frac{a_{3,2} - l_{3,1}u_{1,2}}{u_{2,2}} = \frac{a_{3,2} - \dfrac{a_{3,1}}{a_{1,1}}\, a_{1,2}}{a_{2,2} - \dfrac{a_{2,1}}{a_{1,1}}\, a_{1,2}} \tag{3.76}$$

$$u_{3,3} = a_{3,3} - l_{3,1}u_{1,3} - l_{3,2}u_{2,3} =$$

$$a_{3,3} - \frac{a_{3,1}}{a_{1,1}}a_{1,3} - \left[\frac{a_{3,2} - \frac{a_{3,1}}{a_{1,1}}a_{1,2}}{a_{2,2} - \frac{a_{2,1}}{a_{1,1}}a_{1,2}}\right]\left[a_{2,3} - \frac{a_{2,1}}{a_{1,1}}a_{1,3}\right]$$

Las ecuaciones 3.74, 3.75 y 3.76, convenientemente generalizadas constituyen un método directo para la obtención de L y U, con la ventaja sobre la triangularización de que no se tiene que escribir repetidamente las ecuaciones o arreglos modificados de $A\,\mathbf{x} = \mathbf{b}$. A continuación se resuelve un ejemplo.

Ejemplo 3.32

Resuelva por el método de Doolitle el sistema

$$\begin{aligned} 4x_1 - 9x_2 + 2x_3 &= 5 \\ 2x_1 - 4x_2 + 6x_3 &= 3 \\ x_1 - x_2 + 3x_3 &= 4 \end{aligned}$$

SOLUCIÓN

Con $l_{1,1} = l_{2,2} = l_{3,3} = 1$, se procede al

cálculo de la primera fila de U

$$u_{1,1} = 4\,;\, u_{1,2} = -9\,;\, u_{1,3} = 2$$

cálculo de la primera columna de L

$$l_{1,1} = 1 \text{ (dato)};\ l_{2,1} = 2/4 = 0.5;\ l_{3,1} = 1/4 = 0.25$$

cálculo de la segunda fila de U

$$u_{2,1} = 0 \text{ (recuérdese que } U \text{ es triangular superior)}$$
$$u_{2,2} = -4 - (2/4)\,(-9) = 0.5,\ u_{2,3} = 6 - (2/4)(2) = 5$$

cálculo de la segunda columna de L

$$l_{1,2} = 0 \text{ (ya que } L \text{ es triangular inferior)}$$
$$l_{2,2} = 1 \text{ (dato)},\ l_{3,2} = (-1 - (1/4)(-9\,))/(-4 - (2/4)(-9)) = 2.5$$

cálculo de la tercera fila de U, o más bien sus elementos faltantes, ya que por ser triangular superior

$$u_{3,1} = u_{3,2} = 0$$

$$u_{3,3} = 3-(1/4)(2)-[(-1-(1/4)(-9))/(-4-(2/4)(-9))](6-(2/4)(2)) = -10$$

Con esto se finaliza la factorización*.

Las matrices L y U quedan como sigue

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.25 & 2.5 & 1 \end{bmatrix} ; \quad U = \begin{bmatrix} 4 & -9 & 2 \\ 0 & 0.5 & 5 \\ 0 & 0 & -10 \end{bmatrix}$$

cuyo producto, como ya se comprobó, da A.

Se resuelve el sistema $L\ \mathbf{c} = \mathbf{b}$, donde $\mathbf{b}$ es el vector de términos independientes del sistema original

$$\begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0.25 & 2.5 & 1 \end{bmatrix} \quad \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \\ 4 \end{bmatrix}$$

$$c_1 = 5 \; ; \; c_2 = 3 - 0.5(5) = 0.5$$

$$c_3 = 4 - 0.25(5) - 2.5(0.5) = 1.5,$$

y, finalmente, al resolver el sistema $U\ \mathbf{x} = \mathbf{c}$ se tiene la solución del sistema original

$$\begin{bmatrix} 4 & -9 & 2 \\ 0 & 0.5 & 5 \\ 0 & 0 & -10 \end{bmatrix} \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 0.5 \\ 1.5 \end{bmatrix}$$

$$x_3 = -0.15$$

$$x_2 = (0.5 - 5(-0.15))/0.5 = 2.5$$

$$x_1 = (5 + 9(2.5) - 2(-0.15))/4 = 6.95$$

$$\mathbf{x} = \begin{bmatrix} 6.95 \\ 2.5 \\ -0.15 \end{bmatrix}$$

Las ecuaciones 3.74, 3.75 y 3.76 se generalizan para factorizar la matriz coeficiente del sistema $A\ \mathbf{x} = \mathbf{b}$, que puede resolverse por eliminación de Gauss sin intercambio de filas; se tiene entonces

---

* Los cálculos se han llevado en el orden fila-columna, fila-columna, etc., por convenir a la elaboración de los algoritmos correspondientes.

$$u_{i,j} = a_{i,j} - \sum_{k=1}^{i-1} l_{i,k}\, u_{k,j} \; ; \; j = i, i+1, \ldots, n$$

$$l_{i,j} = \frac{1}{u_{j,j}} \left(a_{i,j} - \sum_{k=1}^{j-1} u_{k,j}\, l_{i,k} \right) \; ; \; i = j+1, \ldots, n \qquad (3.77)$$

$$l_{i,i} = 1 \; ; \; i = 1, 2, \ldots, n$$

con la convención en las sumatorias que $\sum_{k=1}^{0} = 0$.

Puede observarse al seguir las ecuaciones 3.74, 3.75 y 3.76 o bien las ecuaciones 3.77, que una vez empleada $a_{i,j}$ en el cálculo de $u_{i,j}$ o $l_{i,j}$ según sea el caso, esta componente de A no vuelve a emplearse como tal, por lo que las componentes de L y U generadas pueden guardarse en A y ahorrar memoria de esa manera. El siguiente algoritmo de factorización de A ilustra esto.

ALGORITMO 3.6 Factorización directa

Para factorizar una matriz A de orden N en el producto de las matrices L y U triangulares inferior y superior respectivamente, con $l_{i,i} = 1$; i=1, 2, ..., N, proporcionar los

DATOS: El orden N y las componentes de la matriz A.
RESULTADOS: Las matrices L y U en A o mensaje de falla "*LA FACTORIZACIÓN NO ES POSIBLE*".

PASO 1. Si $A(1,1) = 0$ IMPRIMIR "*LA FACTORIZACIÓN NO ES POSIBLE*" y TERMINAR. De otro modo continuar.
PASO 2. Hacer J = 1
PASO 3. Mientras J ≤ N, repetir los pasos 4 a 25.
  PASO 4. Hacer I = J
  PASO 5. Mientras I ≤ N, repetir los pasos 6 a 13.
    PASO 6. Hacer SUMAT = 0
    PASO 7. Si J = 1 ir al paso 12. De otro modo continuar.
    PASO 8. Hacer K = 1
    PASO 9. Mientras K ≤ J − 1, repetir los pasos 10 y 11.
      PASO 10. Hacer SUMAT=SUMAT+A(J,K)*A(K,I)
      PASO 11. Hacer K = K + 1
    PASO 12. Hacer A(J,I) = A(J,I) − SUMAT
    PASO 13. Hacer I = I + 1
  PASO 14. Si J = N ir al paso 26. De otro modo continuar.
  PASO 15. Hacer I = J + 1
  PASO 16. Mientras I ≤ N, repetir los pasos 17 a 24.
    PASO 17. Hacer SUMAT = 0

PASO 18. Si J = 1 ir al paso 23. De otro modo continuar.
PASO 19. Hacer K = 1
PASO 20. Mientras K ≤ J–1, repetir los pasos 21 y 22.
PASO 21. Hacer SUMAT=SUMAT+A(K,J)*A(I,K)
PASO 22. Hacer K = K + 1
PASO 23. Hacer A(I,J)=(A(I,J)–SUMAT)/A(J,J)
PASO 24. Hacer I = I + 1
PASO 25. Hacer J = J + 1
PASO 26. Si A(N,N) = 0 IMPRIMIR "LA FACTORIZACIÓN NO ES POSIBLE" y TERMINAR. De otro modo continuar.
PASO 27. IMPRIMIR *A* y TERMINAR.

Obsérvese que cualquier elemento $a_{i,i} = 0$, impediría emplear este algoritmo; por otro lado, al no pivotear no se reduce en lo posible los errores de redondeo. Para hacer eficiente este algoritmo, debe incluirse un intercambio de filas como en la eliminación de Gauss con pivoteo. A continuación se presenta el algoritmo anterior, pero ahora con estas modificaciones.

ALGORITMO 3.7 Factorización con pivoteo

Para factorizar una matriz *A* de orden N en el producto de las matrices *L* y *U* triangulares inferior y superior respectivamente, con $l_{i,i} = 1$; i=1, 2, ..., N, con pivoteo parcial, proporcionar los

DATOS: El orden N y las componentes de la matriz *A*.
RESULTADOS: Las matrices *L* y *U* en *A* o mensaje de falla "*LA FACTORIZACIÓN NO ES POSIBLE*".

PASO 1. Hacer R = 0 (R registra el número de intercambios de fila que se llevan a cabo).
PASO 2. Hacer J = 1
PASO 3. Mientras J ≤ N, repetir los pasos 4 a 11.
PASO 4. Si J = N ir al paso 10.
PASO 5. Encontrar PIVOTE y P (ver paso 5 de algoritmo 2.4)
PASO 6. Si PIVOTE = 0 IMPRIMIR "*LA FACTORIZACIÓN NO ES POSIBLE*" y TERMINAR. De otro modo continuar.
PASO 7. Si P = J ir al paso 10. De otro modo continuar.
PASO 8. Intercambiar la fila J con la fila P de *A*.
PASO 9. Hacer R = R + 1
PASO 10. Realizar los pasos 4 a 24 del algoritmo 3.6

PASO 11. Hacer J = J + 1
PASO 12. Si A(N,N) = 0 IMPRIMIR "*LA FACTORIZACIÓN NO ES POSIBLE*" y TERMINAR. De otro modo continuar.
PASO 13. IMPRIMIR A y TERMINAR.

A continuación se resuelve un sistema por el método de Doolitle usando la factorización con pivoteo.

Ejemplo 3.33

Resuelva el sistema del ejemlo 3.29

$$\begin{aligned} 10x_1 + x_2 - 5x_3 &= 1 \\ -20x_1 + 3x_2 + 20x_3 &= 2 \\ 5x_1 + 3x_2 + 5x_3 &= 6 \end{aligned}$$

por el método de Doolitle, con pivoteo parcial.

SOLUCIÓN

Al intercambiar la primera y segunda filas resulta la matriz aumentada siguiente

$$A = \left[\begin{array}{ccc|c} -20 & 3 & 20 & 2 \\ 10 & 1 & -5 & 1 \\ 5 & 3 & 5 & 6 \end{array}\right]$$

Como la nueva $a_{1,1} \neq 0$, se forma la primera fila de U y se guarda como primera fila de A

$$a_{1,1} = u_{1,1} = -20,\ a_{1,2} = u_{1,2} = 3,\ a_{1,3} = u_{1,3} = 20$$

Cálculo de la primera columna de L y su registro, excepto $l_{1,1}$, como primera columna de A

$$\begin{aligned} l_{1,1} &= 1 \text{ (dato)}, \\ a_{2,1} = l_{2,1} &= 10/(-20) = -0.5 \\ a_{3,1} = l_{3,1} &= 5/(-20) = -0.25 \end{aligned}$$

La matriz A resultante entonces es

$$A = \left[\begin{array}{ccc|c} -20 & 3 & 20 & 2 \\ -0.5 & 1 & -5 & 1 \\ -0.25 & 3 & 5 & 6 \end{array}\right]$$

Se busca el nuevo pivote en la parte relevante de la segunda columna (segunda y tercera filas) y resulta ser el elemento $a_{3,2}$.
Se intercambia la segunda fila con la tercera y entonces queda

$$A = \left[\begin{array}{ccc|c} -20 & 3 & 20 & 2 \\ -0.25 & 3 & 5 & 6 \\ -0.5 & 1 & -5 & 1 \end{array}\right]$$

Cálculo de la segunda fila de U (mejor dicho de los elementos distintos de cero de dicha fila y almacenamiento de éstos en las posiciones correspondientes de A):

$$a_{2,2} = u_{2,2} = 3 - (-0.25)(3) = 3.75$$

$$a_{2,3} = u_{2,3} = 5 - (-0.25)(20) = 10.0$$

Cálculo de la segunda columna de L; es decir, de los elementos debajo de $l_{2,2}$ y almacenamiento de éstos en las posiciones correspondientes de A

$$a_{3,2} = l_{3,2} = \frac{1 - (-0.5)(3)}{3.75} = 0.666666$$

Con estos valores la matriz A resultante es

$$A = \left[\begin{array}{ccc|c} -20 & 3 & 20 & 2 \\ -0.25 & 3.75 & 10 & 6 \\ -0.5 & 0.6666 & -5 & 1 \end{array}\right]$$

Como $a_{3,3} \neq 0$, se calcula $u_{3,3}$ que constituye la parte relevante de la tercera fila de U, y se almacena en $a_{3,3}$

$$a_{3,3} = u_{3,3} = -5 - (-0.5)(20) - (0.66666)(10) = -1.6666$$

con lo cual la matriz aumentada queda como sigue:

$$A = \left[\begin{array}{ccc|c} -20 & 3 & 20 & 2 \\ -0.25 & 3.75 & 10 & 6 \\ -0.5 & 0.6666 & -1.6666 & 1 \end{array}\right]$$

Al resolver los sistemas

$$L\,\mathbf{c} = \mathbf{b}' \text{ con } L = \begin{bmatrix} 1 & 0 & 0 \\ -0.25 & 1 & 0 \\ -0.5 & 0.6666 & 1 \end{bmatrix} \quad \text{y} \quad \mathbf{b}' = \begin{bmatrix} 2 \\ 6 \\ 1 \end{bmatrix}$$

se tiene

$$c_1 = 2$$

$$c_2 = 6 + 0.25(2) = 6.5$$

$$c_3 = 1 + 0.5(2) - 0.6666(6.5) = -2.33329$$

y

$$U\mathbf{x} = \mathbf{c} \text{ con } U = \begin{bmatrix} -20 & 3 & 20 \\ 0 & 3.75 & 10 \\ 0 & 0 & -1.6666 \end{bmatrix} \text{ y } \mathbf{c} \text{ como arriba.}$$

se tiene

$$x_3 = \frac{-2.33329}{-1.66666} = 1.3999796$$

$$x_2 = \frac{6.5 - 10(1.3999796)}{3.75} = -1.9999456$$

$$x_1 = \frac{2 - 3(-1.9999456) - 20(1.3999796)}{-20} = 0.99999$$

A continuación se da el algoritmo de Doolitle

ALGORITMO 3.8 Método de Doolitle

Para obtener la solución del sistema $A\mathbf{x} = \mathbf{b}$ y el determinante de A, proporcionar los

DATOS: N el número de ecuaciones, A la matriz aumentada del sistema.

RESULTADOS: El vector solución x y el determinante de A o mensaje "*LA FACTORIZACIÓN NO ES POSIBLE*".

PASO 1. Realizar los pasos 1 a 12 del algoritmo 3.7
PASO 2. Hacer c(1) = A(1,N+1)
PASO 3. Hacer DET = A(1,1)
PASO 4. Hacer I = 2
PASO 5. Mientras I ≤ N, repetir los pasos 6 a 12.
  PASO 6. Hacer DET = DET * A(I,I)
  PASO 7. Hacer c(I) = A(I,N+1)
  PASO 8. Hacer J = 1
  PASO 9. Mientras J ≤ I–1, repetir los pasos 10 y 11.
    PASO 10. Hacer c(I) = c(I) – A(I,J) * c(J)
    PASO 11. Hacer J = J + 1
  PASO 12. Hacer I = I + 1
PASO 13. Hacer x(N) = c(N)/A(N,N)
PASO 14. Hacer I = N – 1
PASO 15. Mientras I ≤ 1, repetir los pasos 16 a 22.
  PASO 16. Hacer x(I) = c(I)
  PASO 17. Hacer J = I + 1
  PASO 18. Mientras J ≥ N, repetir los pasos 19 y 20.

PASO 19. Hacer x(I) = x(I) - A(I,J) * x(J)
PASO 20. Hacer J = J + 1
PASO 21. Hacer x(I) = x(I)/A(I,I)
PASO 22. Hacer I = I - 1
PASO 23. Hacer DET = DET * (-1) ** R
PASO 24. IMPRIMIR x y DET y TERMINAR.

Sistemas simétricos

En el caso de que la matriz coeficiente del sistema A x = b sea simétrica, los cálculos de la factorización (si es posible) se simplifican, ya que la segunda de las ecuaciones 3.77 se reduce a

$$l_{ij} = \frac{a_{ji}}{a_{jj}} \qquad i = j+1, \ldots, n; \qquad j = 1, 2, \ldots, n-1 \tag{3.78}$$

Esto disminuye considerablemente el trabajo, en particular cuando n es grande.

Ejemplo 3.34

Resuelva el sistema simétrico siguiente

$$\begin{bmatrix} 2 & 1 & 3 \\ 1 & 0 & 4 \\ 3 & 4 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$$

SOLUCIÓN

Cálculo de la primera fila de U y su registro en A.

$$a_{1,1} = u_{1,1} = 2,$$
$$a_{1,2} = u_{1,2} = 1,$$
$$a_{1,3} = u_{1,3} = 3.$$

Cálculo de los elementos relevantes de la primera columna de L , usando la ecuación 3.78 y su registro en A

$$a_{2,1} = l_{2,1} = \frac{a_{1,2}}{a_{1,1}} = 0.5$$

$$a_{3,1} = l_{3,1} = \frac{a_{1,3}}{a_{1,1}} = 1.5$$

Cálculo de los elementos relevantes de la segunda fila de U y su registro en las posiciones correspondientes de A

$$a_{2,2} = u_{2,2} = a_{2,2} - l_{2,1}\, u_{1,2} = 0 - 0.5(1) = -0.5$$
$$a_{2,3} = u_{2,3} = a_{2,3} - l_{2,1}\, u_{1,3} = 4 - 0.5(3) = 2.5$$

Cálculo de los elementos relevantes de la segunda columna de L mediante la ecuación 3.78 y su registro en las posicones correspondientes de A

$$a_{3,2} = l_{3,2} = \frac{a_{2,3}}{a_{2,2}} = -5$$

Finalmente se calcula la componente $u_{3,3}$ (único elemento relevante de la tercera fila de U) y se verifica su registro en $a_{3,3}$

$$\begin{aligned} a_{3,3} = u_{3,3} &= a_{3,3} - l_{3,1}\, u_{1,3} - l_{3,2}\, u_{2,3} \\ &= 3 - 1.5(3) - (-5)(2.5) = 11 \end{aligned}$$

La factorización da como resultado

$$\begin{bmatrix} 2 & 1 & 3 \\ 0.5 & -0.5 & 2.5 \\ 1.5 & -5 & 11 \end{bmatrix}$$

Con la resolución del sistema $L\,\mathbf{c} = \mathbf{b}$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ 1.5 & -5 & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$$

se obtiene: $\mathbf{c} = [0\ 1\ 8]^T$

y al resolver el sistema $U\,\mathbf{x} = \mathbf{c}$

$$\begin{bmatrix} 2 & 1 & 3 \\ 0 & -0.5 & 2.5 \\ 0 & 0 & 11 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 8 \end{bmatrix}$$

se obtiene

$$\mathbf{x} = \begin{bmatrix} 1.9091 \\ 1.6364 \\ 0.7273 \end{bmatrix}$$

Es importante observar que no se emplea pivoteo parcial y que si alguno de los elementos $u_{i,i}$ resulta ser cero, este método no es aplicable; como consecuencia, habrá que recurrir al método de Doolitle con pivoteo, por ejemplo, con lo cual se pierde la ventaja de que A es simétrica.

A continuación se da el algoritmo correspondiente.

ALGORITMO 3.9 Factorización de matrices simétricas

Para factorizar una matriz A de orden n en el producto de las matrices L y U triangulares inferior y superior respectivamente, con $l_{i,i} = 1$; $i=1, 2, \ldots, n$, proporcionar los

DATOS: El orden N y las componentes de la matriz simétrica A.

RESULTADOS: Las matrices L y U en A o mensaje de falla *"LA FACTORIZACIÓN NO ES POSIBLE"*.

PASO 1. Hacer J = 1
PASO 2. Mientras J ≤ N, repetir los pasos 3 a 15.
  PASO 3. Hacer I = J
  PASO 4. Mientras I ≤ N, repetir los pasos 5 a 13.
    PASO 5. Hacer SUMAT = 0
    PASO 6. Si J = 1 ir al paso 11. De otro modo continuar.
    PASO 7. Hacer K = 1
    PASO 8. Mientras K≤J–1, repetir los pasos 9 y 10.
      PASO 9. Hacer SUMAT=SUMAT+ A(J,K)*A(K,I)
      PASO 10. Hacer K = K + 1
    PASO 11. Hacer A(J,I) = A(J,I) – SUMAT
    PASO 12. Si I > J Hacer A(I,J)=A(J,I)/A(J,J). De otro modo continuar.
    PASO 13. Hacer I = I + 1
  PASO 14. Si A(J,J) = 0 IMPRIMIR *"LA FACTORIZACIÓN NO ES POSIBLE"* y TERMINAR. De otro modo continuar.
  PASO 15. Hacer J = J + 1
PASO 16. IMPRIMIR A y TERMINAR.

Método de Cholesky.

Una matriz simétrica A cuyas componentes son números reales, es positiva definida si y solo si los determinantes de las submatrices de A son positivos

$$|a_{1,1}| > 0, \quad \begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{vmatrix} > 0, \ldots, \quad \begin{vmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{vmatrix} > 0$$

En el caso de tener un sistema $A\,\mathbf{x} = \mathbf{b}$, con A positiva definida, la factorización de A en la forma $L\,U$ es posible y muy sencilla ya que toma la forma $L\,L^{\mathrm{T}}$ donde L es triangular inferior:

$$L = \begin{bmatrix} l_{1,1} & 0 & \cdots & 0 \\ l_{2,1} & l_{2,2} & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & 0 \\ l_{n,1} & l_{n,2} & \cdots & l_{n,n} \end{bmatrix}$$

Los cálculos se reducen, ya que ahora basta estimar $n(n+1)/2$ elementos (los $l_{i,j} \neq 0$), en lugar de los n^2 elementos de una factorización nominal (los $l_{i,j}$ tales que $i < j$ y los $u_{i,j}$ tales que $i \geq j$). El número de cálculos es prácticamente la mitad.

Ejemplo 3.35

Resuelva el sistema de ecuaciones lineales

$$\begin{bmatrix} 4 & 1 & 2 \\ 1 & 2 & 0 \\ 2 & 0 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}$$

cuya matriz coeficiente es simétrica y positivamente definida.

SOLUCIÓN

Factorización de A

$$\begin{bmatrix} 4 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} l_{1,1} & 0 & 0 \\ l_{2,1} & l_{2,2} & 0 \\ l_{3,1} & l_{3,2} & l_{3,3} \end{bmatrix} \begin{bmatrix} l_{1,1} & l_{2,1} & l_{3,1} \\ 0 & l_{2,2} & l_{3,2} \\ 0 & 0 & l_{3,3} \end{bmatrix}$$

De la multiplicación de matrices se tiene

$l_{1,1}^2 = a_{1,1}$; $l_{1,1} = \pm\sqrt{a_{11}} = \pm 2$ se toma el valor positivo de todas las raíces

$$l_{1,1} = 2$$

$$l_{1,1}\, l_{2,1} = a_{1,2}; \quad l_{2,1} = a_{1,2}/l_{1,1} = 1/2 = 0.5$$

$$l_{1,1}\, l_{3,1} = a_{1,3}; \quad l_{3,1} = a_{1,3}/l_{1,1} = 2/2 = 1$$

$$l_{2,1}^2 + l_{2,2}^2 = a_{2,2}; \quad l_{2,2} = \sqrt{a_{2,2} - l_{2,1}^2}$$

$$l_{2,2} = \sqrt{2 - 0.5^2} = 1.32287$$

$$l_{2,1}l_{3,1} + l_{2,2}l_{3,2} = a_{2,3}\ ; \qquad l_{3,2} = -\frac{l_{2,1}\ l_{3,1} + a_{2,3}}{l_{2,2}}$$

$$l_{3,2} = \frac{0.5(1)}{1.32287} = -\ 0.37796$$

$$l_{3,1}^2 + l_{3,2}^2 + l_{3,3}^2 = a_{3,3}\ ; \qquad l_{3,3} = \sqrt{a_{3,3} - l_{3,1}^2 - l_{3,2}^2}$$

$$l_{3,3} = \sqrt{5 - 1 - 0.14286} = 1.96396$$

Al resolver el sistema

$L\ \mathbf{c} = \mathbf{b}$

$$\begin{bmatrix} 2 & 0 & 0 \\ 0.5 & 1.32287 & 0 \\ 1 & -0.37796 & 1.96396 \end{bmatrix} \quad \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}$$

$c_1 = 0.5$

$c_2 = (2 - 0.5(0.5))/1.32287 = 1.32287$

$c_3 = (4 - 0.5 + 0.37796(1.32287))/1.96396 = 2.0367$

Al resolver el sistema

$L^{\mathrm{T}}\ \mathbf{x} = \mathbf{c}$

$$\begin{bmatrix} 2 & 0.5 & 1 \\ 0 & 1.32287 & -0.37796 \\ 0 & 0 & 1.96396 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.32287 \\ 2.0367 \end{bmatrix}$$

$x_3 = 2.0367/1.96396 = 1.037$

$x_2 = (1.32287 + 0.37796(1.037))/1.32287 = 1.29629$

$x_1 = (0.5 - 0.5(1.29629) - 1.037)/2 = -\ 0.59259$

El vector solución es

$$\mathbf{x} = \begin{bmatrix} -0.59259 \\ 1.29629 \\ 1.037 \end{bmatrix}$$

Las fórmulas de este algoritmo para un sistema de n ecuaciones son

$$l_{1,1} = \sqrt{a_{1,1}}$$

$$l_{i,1} = a_{i,1}/l_{1,1} \qquad i = 2, 3, \ldots, n$$

$$l_{i,j} = \left(a_{i,j} - \sum_{k=1}^{i-1} l_{i,k}^2 \right)^{1/2} \qquad j = 2, 3, \ldots, n$$

$$l_{i,j} = \frac{1}{l_{i,i}} \left(a_{i,j} - \sum_{k=1}^{i-1} l_{i,k} l_{j,k} \right) \qquad j = 2, 3, \ldots, n$$
$$i = j+1, j+2, \ldots, n-1$$

$$l_{i,j} = 0 \qquad i < j$$

A continuación se da el algoritmo para este método.

ALGORITMO 3.10 Método de Cholesky

Para factorizar una matriz positiva definida en la forma $L\,L^T$, proporcionar los

DATOS: N, el orden de la matriz y sus elementos.
RESULTADOS: La matriz L.

PASO 1. Hacer $L(1,1) = A(1,1)$ ** 0.5
PASO 2. Hacer I = 2
PASO 3. Mientras I ≤ N, repetir los pasos 4 y 5.
  PASO 4. Hacer $L(I,1) = A(I,1)/L(1,1)$
  PASO 5. Hacer I = I + 1
PASO 6. Hacer I = 2
PASO 7. Mientras I ≤ N, repetir los pasos 8 a 24.
  PASO 8. Hacer S = 0
  PASO 9. Hacer K = 1
  PASO 10. Mientras K ≤ I–1, repetir los pasos 11 y 12.
    PASO 11. Hacer S = S + $L(I,K)$**2
    PASO 12. Hacer K = K + 1
  PASO 13. Hacer $L(I,I) = (A(I,I) - S)$**0.5
  PASO 14. Si I = N ir al paso 25.
  PASO 15. Hacer J = I + 1
  PASO 16. Mientras J ≤ N, repetir los pasos 17 a 23.
    PASO 17. Hacer S = 0
    PASO 18. Hacer K = 1
    PASO 19. Mientras K ≤ I–1, repetir los pasos 20 y 21.
      PASO 20. Hacer S=S+$L(I,K)$*$L(J,K)$
      PASO 21. Hacer K = K + 1
    PASO 22. Hacer $L(J,I)=(A(J,I)–S)/L(I,I)$
    PASO 23. Hacer J = J + 1
  PASO 24. Hacer I = I + 1
PASO 25. IMPRIMIR L y TERMINAR.

Sistemas de ecuaciones mal condicionados

Algunos autores caracterizan los métodos de solución directos como aquellos con los que se obtiene la solución exacta **x** del sistema $A\ \mathbf{x} = \mathbf{b}$ mediante un número finito de operaciones, siempre y cuando no existan errores de redondeo. Como estos errores son prácticamente inevitables, se obtendrán en general soluciones aproximadas y, cuya sustitución en el sistema producirá una aproximación del vector **b** : **b'**

$$A\ \mathbf{y} = b' \approx \mathbf{b}$$

En general, pequeños errores de redondeo producen sólo pequeños cambios en el vector solución; en estos casos se dice que el sistema está **bien condicionado**. Sin embargo, en algunos casos los errores de redondeo de los primeros pasos causan errores más adelante (se propagan), de modo que la solución obtenida y resulta ser un vector muy distinto del vector solución; peor aún, en estos sistemas la sustitución de y satisface prácticamente dicho sistema. Este tipo de sistemas se conocen como **mal condicionados**. *A* continuación se presentan dos ejemplos

Sea el sistema mal condicionado*

$$\begin{bmatrix} 1.00 & 0.99 \\ 0.99 & 0.98 \end{bmatrix} \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.99 \\ 1.97 \end{bmatrix} \tag{3.79}$$

cuya solución es $x_1 = x_2 = 1.00$, y sea la matriz aumentada siguiente

$$\left[\begin{array}{cc|c} 1.00 & 0.9900 & 1.9900 \\ 0.00 & 0.0001 & 0.0001 \end{array}\right],$$

el resultado de la triangularización. Si se redondea o corta a tres dígitos la última fila, quedaría como fila de ceros y el sistema original como un sistema sin solución única.

Si, por otro lado, por un pequeño error en los cálculos se obtiene como solución de la ecuación 3.79

$$y_1 = 0\ , y_2 = 2\ ,$$

que aunque muy distinta del vector solución da en la sustitución

$$\left[\begin{array}{cc|c} 1.00 & 0.99 & 0 \\ 0.99 & 0.98 & 2 \end{array}\right] = \begin{bmatrix} 1.98 \\ 1.96 \end{bmatrix}$$

prácticamente el vector **b**.

Aun una solución tan absurda como

$$y_1 = 100, y_2 = -99,$$

*Forsythe, G.E. and Moler, C.B. *Computer Solution of Linear Algebraic Systems*. Englewood Cliffs, N.J. Prentice Hall (1967).

da resultados sorprendentemente cercanos a **b**

$$\left[\begin{array}{cc|c} 1.00 & 0.99 & 100 \\ 0.99 & 0.98 & -99 \end{array}\right] = \begin{bmatrix} 1.99 \\ 1.98 \end{bmatrix}$$

Algunas veces los elementos de A y **b** son generados por cálculos (véase algoritmos 5.1 y 5.5) y los valores resultantes de ambos son ligeramente erróneos.

Sea el sistema mal condicionado

$$\begin{aligned} 1.001\, x_1 - x_2 &= 1 \\ x_1 - x_2 &= 0 \end{aligned} \tag{3.80}$$

que se desea resolver, pero por errores de redondeo o de otro tipo, se obtiene en su lugar

$$\begin{aligned} y_1 - 0.9999\, y_2 &= 1.001 \\ y_1 - 1.0001\, y_2 &= 0, \end{aligned} \tag{3.80'}$$

que difiere sólo "ligeramente" del sistema 3.80

Las soluciones exactas son, respectivamente

$$\mathbf{x} = \begin{bmatrix} 1000 \\ 1000 \end{bmatrix}; \qquad \mathbf{y} = \begin{bmatrix} 5005.5005 \\ 5005.0000 \end{bmatrix}$$

cuya diferencia es notable a pesar de que los sistemas son casi idénticos. Para entender esto se da a continuación una interpretación geométrica de los sistemas mal condicionados.

Interpretación geométrica de un sistema mal condicionado de orden 2

La solución de un sistema de dos ecuaciones en dos incógnitas

$$\begin{aligned} a_{1,1}\, x_1 + a_{1,2}\, x_2 &= b_1 \\ a_{2,1}\, x_1 + a_{2,2}\, x_2 &= b_2 \end{aligned} \tag{3.81}$$

es el punto de intersección de las rectas

$$x_1 = \frac{b_1}{a_{1,1}} - \frac{a_{1,2}}{a_{1,1}}\, x_2 \tag{3.82}$$

$$x_1 = \frac{b_2}{a_{2,1}} - \frac{a_{2,2}}{a_{2,1}}\, x_2 \tag{3.83}$$

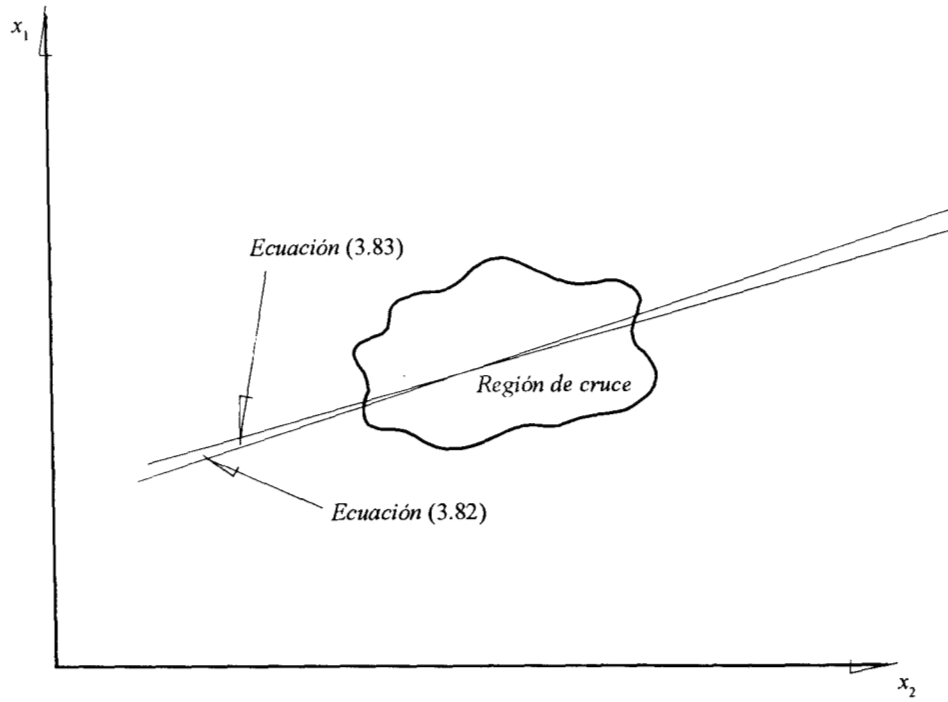


Figura 3.10. Interpretación geométrica de un sistema mal condicionado de orden 2.

en el plano x_2 –x_1. Si el sistema 3.81 es mal condicionado, las rectas 3.82 y 3.83 son casi paralelas, pero resulta difícil decir dónde se cortan exactamente * (Véase Fig. 3.10). Cualquier pequeño error de redondeo o de otro tipo puede alejar del vector solución, con lo que se produce una solución errónea **y**. No obstante esto, si **y** está en la región de cruce, el sistema 3.81 se satisface prácticamente con **y**. Obsérvese que la región de cruce es muy amplia y que algunos de sus puntos pueden estar muy alejados del vector solución.

Una vez que se ha visto el comportamiento de los sistemas mal condicionados, resulta de interés determinar si un sistema dado está mal condicionado y qué hacer en tales casos para resolverlo. Hay varias formas de detectar si un sistema está mal o bien condicionado; pero quizá la más simple de ellas es la del determinante normalizado que se describe a continuación.

Medida de condicionamiento usando el determinante normalizado

En el sistema 3.81 el determinante de la matriz coeficiente

$$\begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{vmatrix} = a_{1,1}\, a_{2,2} - a_{1,2}\, a_{2,1} ,$$

*Nótese que hay una solución única, pero resulta difícil decir dónde está.

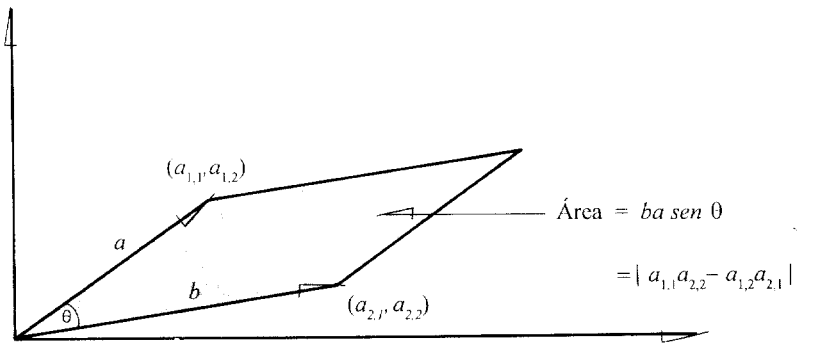


Fig. 3.11. Interpretación geométrica del determinante.

puede interpretarse en valor absoluto como el área del paralelogramo cuyos lados son los vectores fila* $[a_{1,1}\ a_{1,2}]$ y $[a_{2,1}\ a_{2,2}]$ (véase Fig. 3.11).

En el caso de un sistema general de orden 3, el determinante de la matriz coeficiente de dicho sistema es, en valor absoluto, el volumen del paralelepípedo cuyos lados son los vectores $[a_{1,1}\ a_{1,2}\ a_{1,3}]$, $[a_{2,1}\ a_{2,2}\ a_{2,3}]$ y $[a_{3,1}\ a_{3,2}\ a_{3,3}]$, (véase Fig. 3.12).

Al multiplicar cada una de las filas del sistema 3.81 por un factor, el sistema resultante es equivalente, pero la matriz coeficiente se ha modificado y, por ende, su determinante. Si por ejemplo, se divide la primera y segunda ecuaciones de 3.81, respectivamente entre

$$k_1 = \sqrt{a_{1,1}^2 + a_{1,2}^2} \qquad k_2 = \sqrt{a_{2,1}^2 + a_{2,2}^2}$$

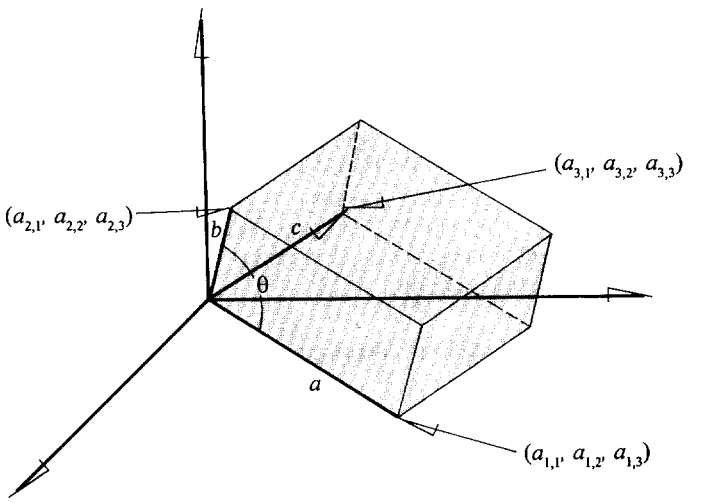


Fig. 3.12. Interpretación geométrica del determinante.

*Puede decirse lo mismo para los vectores columna.

se obtiene como nueva matriz coeficiente

$$\begin{bmatrix} \frac{a_{1,1}}{k_1} & \frac{a_{1,2}}{k_1} \\ \frac{a_{2,1}}{k_2} & \frac{a_{2,2}}{k_2} \end{bmatrix},$$

cuyo determinante en valor absoluto es menor o igual a la unidad, ya que ahora $| \mathbf{a} | = 1$ y $| \mathbf{b} | = 1$ (véase Fig. 3.11). El determinante así obtenido se conoce como **determinante normalizado** y, en general, para sistemas de orden n la matriz coeficiente resultante de dividir la i–ésima fila por los factores*

$$k_i = \sqrt{a_{i,1}^2 + a_{i,2}^2 + \ldots + a_{i,n}^2}, \qquad i = 1, 2, \ldots, n$$

tiene un determinante, en valor absoluto, menor o igual que la unidad.

Si el sistema 3.81 está mal condicionado, los vectores fila $[a_{1,1}\ a_{1,2}]$ y $[a_{2,1}\ a_{2,2}]$ son casi paralelos y el determinante normalizado estará muy cercano a cero (muy pequeño). Si por otro lado, los vectores fila son casi ortogonales (perpendiculares), el determinante estará muy cercano a la unidad, en valor absoluto.

Resumiendo y precisando: para medir el condicionamiento de un sistema de orden n, se debe obtener el determinante normalizado de la matriz coeficiente de dicho sistema y si su valor absoluto es "prominentemente menor" que 1, el sistema está mal condicionado; en caso de tener un valor absoluto prominentemente cercano a 1, el sistema está bien condicionado. Esta lejanía o cercanía de 1 queda determinada por la precisión empleada**.

Si bien la técnica es útil, no resulta práctica en sistemas grandes, ya que el cálculo del determinante toma tiempo y es casi equivalente a resolver dichos sistemas. Entonces, si se sospecha que un sistema está mal condicionado, se analiza de la manera siguiente.

a) Se resuelve el sistema original $A\ \mathbf{x} = \mathbf{b}$.
b) Se modifican los componentes de A ligeramente y se resuelve el sistema resultante $A'\ \mathbf{x} = \mathbf{b}$.
c) Si las dos soluciones son sustancialmente diferentes (estas diferencias se comparan con los cambios hechos en $a_{i,j}$), el sistema está mal condicionado.

Una vez corroborado que un sistema grande está mal condicionado, deberán emplearse los métodos de solución vistos con ciertas recomendaciones.

a) Aprovechar las características de la matríz coeficiente (matrices bandeadas, simétricas, diagonal dominantes, positivas definidas, etc.), para que el método seleccionado sea el más adecuado y se realicen, por ejemplo, menos cálculos.

---

*Llamados factores de escalamiento.

**Young, D.M. y Gregory, R.T. *A Survey Of Numerical Mathematics*, Vol. II Addison-Wesley (1973) p. 812-820.

b) Emplear pivoteo parcial o total (Véase Ejer. 3.9).
c) Emplear doble precisión en los cálculos.

Si aún después de seguir estas sugerencias persisten las dificultades, puede recurrirse a los métodos iterativos que se estudian más adelante y que son, en general, otra alternativa de solución de sistemas lineales mal y bien condicionados, con la ventaja de no ser tan sensibles a los errores de redondeo.

Matrices elementales y los métodos de eliminación.

Nótese que cualquiera de los métodos de eliminación vistos para resolver el sistema $A\ \mathbf{x} = \mathbf{b}$ involucra las siguientes operaciones sobre una matriz*:

a) Intercambio de filas.
b) Multiplicación de la fila por un escalar, y
c) Sustitución de una fila por la suma de ésta y alguna otra fila de la matriz.

Estas operaciones pueden llevarse a cabo mediante multiplicaciones de la matriz en cuestión por ciertas matrices especiales; por ejemplo, la matriz permutadora permite intercambiar filas. Multiplicando en cambio por la izquierda una matriz B cualquiera por la matriz identidad correspondiente I, pero sustituido uno de sus elementos unitarios por m (la posición (i,i) por ejemplo), se multiplica la i–ésima fila de B por m.

Ejemplo 3.36

Multiplique la matriz general B de 3×4 por la matriz identidad correspondiente I, donde se ha remplazado el 1 de la posición (2,2) con m.

SOLUCIÓN

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & m & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} & b_{1,4} \\ b_{2,1} & b_{2,2} & b_{2,3} & b_{2,4} \\ b_{3,1} & b_{3,2} & b_{3,3} & b_{3,4} \end{bmatrix} = \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} & b_{1,4} \\ mb_{2,1} & mb_{2,2} & mb_{2,3} & mb_{2,4} \\ b_{3,1} & b_{3,2} & b_{3,3} & b_{3,4} \end{bmatrix}$$

Los resultados hablan por sí solos.

Finalmente, cuando se multiplica por la izquierda una matriz general B por la matriz identidad correspondiente I, en la que se ha sustituido uno de los ceros con m (el cero de la posición (i,j) por ejemplo) se tiene el efecto de sustituir la fila i–ésima de B por la fila resultante de sumar ésta y la fila j–ésima de B multiplicada por m.

*Generalmente, se trata de la matriz aumentada $[A \mid \mathbf{b}]$.

Ejemplo 3.37

Sustituya la segunda fila de la matriz general B de 3×3 por el resultado de sumar dicha segunda fila con la primera fila de B multiplicada por m.

SOLUCIÓN

Se sustituye el cero de la posición (2,1) de la matriz I de 3×3 con m y se multiplica por la izquierda por B; es decir

$$\begin{bmatrix} 1 & 0 & 0 \\ m & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} \\ b_{2,1} & b_{2,2} & b_{2,3} \\ b_{3,1} & b_{3,2} & b_{3,3} \end{bmatrix} = \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} \\ mb_{1,1} + b_{2,1} & mb_{1,2} + b_{2,2} & mb_{1,3} + b_{2,3} \\ b_{3,1} & b_{3,2} & b_{3,3} \end{bmatrix}$$

Si se desea intercambiar columnas, multiplicarlas por un escalar o sustituir una columna por la suma de ésta y alguna otra, se procede siguiendo las mismas ideas, pero con las multiplicaciones por la derecha sobre la matriz en cuestión.

Estas matrices se conocen como elementales y se denotan como

Permutación: P

Multiplicación por un escalar: M

Sustitución: S

Para aclarar la relación que existe entre estas matrices y los métodos de eliminación, se resuelve nuevamente el ejemplo 3.30, pero ahora con matrices elementales.

Ejemplo 3.38

Resuelva por eliminación de Jordan el sistema

$$\begin{aligned} 4x_1 - 9x_2 + 2x_3 &= 5 \\ 2x_1 - 4x_2 + 6x_3 &= 3 \\ x_1 - x_2 + 3x_3 &= 4 \end{aligned}$$

con matrices P, M y S.

SOLUCIÓN

La matriz aumentada es

$$\left[\begin{array}{ccc|c} 4 & -9 & 2 & 5 \\ 2 & -4 & 6 & 3 \\ 1 & -1 & 3 & 4 \end{array}\right] = B$$

No se intercambian filas, ya que el elemento de máximo valor absoluto se encuentra en la primera. Para hacer cero el elemento (2,1), se suma la primera multiplicada por –1/2 a la segunda fila; la siguiente matriz cumple con ese fin.

$$\begin{bmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = S_1$$

Para hacer cero el elemento (3,1) se suma la primera multiplicada por –1/4 a la tercera fila; esto es

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1/4 & 0 & 1 \end{bmatrix} = S_2$$

El efecto de S_1 y S_2 sobre B resulta en

$$\left[\begin{array}{ccc|c} 4 & -9 & 2 & 5 \\ 0 & 0.5 & 5 & 0.5 \\ 0 & 1.25 & 2.5 & 2.75 \end{array}\right] = S_2\ S_1\ B$$

Como el elemento de máximo valor absoluto es 1.25, se intercambia la segunda y tercera filas, para lo cual se emplea la matriz

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = P_1$$

y queda

$$\left[\begin{array}{ccc|c} 4 & -9 & 2 & 5 \\ 0 & 1.25 & 2.5 & 2.75 \\ 0 & 0.5 & 5 & 0.5 \end{array}\right] = P_1\ S_2\ S_1\ B$$

Para hacer cero los elementos (1,2) y (3,2), se suma la segunda multiplicada por (–(–9)/1.25) a la primera fila, y la segunda multiplicada por (–0.5/1.25) a la tercera, proceso que se lleva a cabo con las matrices

$$\begin{bmatrix} 1 & -(-9)/1.25 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = S_3 \quad \text{y} \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -0.5/1.25 & 1 \end{bmatrix} = S_4$$

y queda como resultado

$$\left[\begin{array}{ccc|c} 4 & 0 & 20 & 24.8 \\ 0 & 1.25 & 2.5 & 2.75 \\ 0 & 0 & 4 & -0.6 \end{array}\right] = S_4\ S_3\ P_1\ S_2\ S_1\ B$$

Para eliminar los elementos (1,3) y (2,3), se suma la tercera multiplicada por (–20/4) a la primera fila y la tercera multiplicada por (–2.5/4) a la segunda, lo cual se logra con S_5 y S_6, respectivamente. (Se deja al lector determinar la forma que tienen S_5 y S_6). El resultado es

$$\left[\begin{array}{ccc|c} 4 & 0 & 0 & 27.8 \\ 0 & 1.25 & 0 & 3.125 \\ 0 & 0 & 4 & -0.6 \end{array}\right] = S_6\, S_5\, S_4\, S_3\, P_1\, S_2\, S_1\, B$$

Todavía se puede multiplicar la primera fila por $m_1 = 1/4$, la segunda por $m_2 = 1/1.25$ y la tercera por $m_3 = 1/4$ lo cual se consigue con

$$\begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = M_1\text{, etc.}$$

finalmente queda

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & 6.95 \\ 0 & 1 & 0 & 2.5 \\ 0 & 0 & 1 & -0.15 \end{array}\right] = M_3\, M_2\, M_1\, S_6\, S_5\, S_4\, S_3\, P_1\, S_2\, S_1\, B$$

que puesta nuevamente como un sistema de ecuaciones da

$$\begin{aligned} x_1 &= 6.95 \\ x_2 &= 2.5 \\ x_3 &= -0.15, \end{aligned}$$

directamente la solución del sistema original $A\,\mathbf{x} = \mathbf{b}$.

Si el producto de las matrices elementales se denota por E

$$E = M_3\, M_2\, M_1\, S_6\, S_5\, S_4\, S_3\, P_1\, S_2\, S_1\,,$$

se tiene

$$E\,B = E\,[A \mid \mathbf{b}\,] = [\,I \mid \mathbf{x}\,],$$

de donde

$$E\,A = I \qquad \text{y} \qquad E\,B = \mathbf{x}$$

resulta que E es la inversa de A.

$$E = A^{-1}$$

Por otro lado, se sabe que el determinante del producto de dos o más matrices es igual al producto de los determinantes de cada una de las matrices

$$\det A\,B \ldots = \det A \det B \ldots$$

de donde

$$\det E\,A = \det I$$

o bien

$$\det E \det A = 1$$

y

$$\frac{1}{\det E} = \det A$$

de modo que el determinante de A está dado como la inversa del determinante de E y sólo queda obtener det E. Esto parece complicado a simple vista; sin embargo, observando que en general (véase Probl. 3.52).

$\det P = -1$, el determinante de una matriz permutadora es -1.

$\det M = m$, el determinante de una matriz multiplicadora es el factor m, que deberá ser distinto de cero.

$\det S = 1$, el determinante de una matriz del tipo S es 1.

Se tiene

$$\det E = \det M_3 \det M_2 \det M_1 \det S_6 \det S_5 \det S_4 \det S_3 \det P_1 \det S_2 \det S_1$$

sustituyendo

$$\det E = m_3\, m_2\, m_1\, (-1) = -\frac{1}{4}\left(\frac{1}{1.25}\right)\frac{1}{4} = -0.05$$

y

$$\det A = \frac{1}{-0.05} = -20$$

Finalmente, para obtener E y por tanto A^{-1} se toma S_1 como matriz pivote y sobre ella se efectúan las operaciones de intercambio de filas, multiplicación por un escalar, etc., que vayan indicando las matrices a su izquierda. Así

$$\begin{bmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ -1/4 & 0 & 1 \end{bmatrix} = S_2\, S_1$$

ya que según se dijo, S_2 tiene como efecto multiplicar la primera fila de S_1 por $-1/4$ y sumarla a la tercera fila de S_1.

Con P_1 en cambio se tiene

$$\begin{bmatrix} 1 & 0 & 0 \\ -1/4 & 0 & 1 \\ -1/2 & 1 & 0 \end{bmatrix} = P_1\,(S_2\, S_1)$$

ya que P_1 intercambia las filas segunda y tercera de $(S_2\, S_1)$.

Continuando este proceso se llega a

$$\begin{bmatrix} 0.3 & -1.25 & 2.3 \\ 0 & -0.5 & 1.0 \\ -0.1 & 0.25 & -0.1 \end{bmatrix} = E = A^{-1}$$

SECCIÓN 3.5 MÉTODOS ITERATIVOS

Al resolver un sistema de ecuaciones lineales por eliminación, la memoria de máquina requerida es proporcional al cuadrado del orden de A, y el trabajo computacional es proporcional al cubo del orden de la matriz coeficiente A (véase Secc. 3.4). Debido a esto, la solución de sistemas lineales grandes ($n \geq 50$), con matrices coeficiente densas*, se vuelve costoso y difícil en una computadora con los métodos de eliminación, ya que se requiere amplia memoria; además, como el número de operaciones que se debe ejecutar es muy grande, se pueden producir errores de redondeo también muy grandes. Sin embargo, se han resuelto sistemas de orden 1000, y aun mayor, con los métodos que se estudiarán en esta sección.

Estos sistemas de un número muy grande de ecuaciones se presentan en la solución numérica de ecuaciones diferenciales parciales, en la solución de los modelos resultantes en la simulación de columnas de destilación, etc. En favor de estos sistemas, puede decirse que tienen matrices con pocos elementos distintos de cero y que éstas poseeen ciertas propiedades (simétricas, bandeadas, diagonal dominantes, etc.), que permiten garantizar el éxito en la aplicación de los métodos de esta sección.

Métodos de Jacobi y Gauss-Seidel

Los métodos iterativos más sencillos y conocidos son una generalización del método de punto fijo, estudiado en el capítulo 2. Se puede aplicar la misma técnica a fin de elaborar métodos para la solución de $A\,\mathbf{x} = \mathbf{b}$, de la manera siguiente.

Se parte de $A\,\mathbf{x} = \mathbf{b}$ para obtener la ecuación

$$A\,\mathbf{x} - \mathbf{b} = \mathbf{0}, \tag{3.84}$$

ecuación vectorial correspondiente a $f(x) = 0$. Se busca ahora una matriz B y un vector $\mathbf{c}$, de manera que la ecuación vectorial

$$\mathbf{x} = B\,\mathbf{x} + \mathbf{c}, \tag{3.85}$$

sea sólo un arreglo de la ecuación 3.84; es decir, de manera que la solución de una sea también la solución de la otra. La ecuación 3.85 correspondería a $x = g(x)$. A continuación se propone un vector inicial $\mathbf{x}^{(0)}$ como primera aproximación al vector solución $\mathbf{x}$. Luego, se calcula con la ecuación 3.86 la sucesión vectorial $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, ..., de la siguiente manera

$$\mathbf{x}^{(k+1)} = B\,\mathbf{x}^{(k)} + \mathbf{c},\ k = 0, 1, 2, \ldots$$

donde

$$\mathbf{x}^{(k)} = [x_1^k \quad x_2^k \quad \ldots \quad x_n^k]^T \tag{3.86}$$

*Una matriz densa tiene pocos ceros como elementos.

Para que la sucesión $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}, \ldots$, converja al vector solución $\mathbf{x}$ es necesario que eventualmente x_j^m , $1 \leq j \leq n$ (los componentes del vector $\mathbf{x}^{(m)}$), se aproximen tanto a x_j, $1 \leq j \leq n$ (los componentes correspondientes a $\mathbf{x}$) que todas las diferencias $\mid x_j^m - x_j \mid$, $1 \leq j \leq n$ sean menores que un valor pequeño previamente fijado, y que se conserven menores para todos los vectores siguientes de la iteración; es decir

$$\lim_{m \to \infty} x_j^m = x_j \qquad 1 \leq j \leq n \tag{3.87}$$

La forma como se llega a la ecuación 3.85 define el algoritmo y su convergencia. Dado el sistema $A\ \mathbf{x} = \mathbf{b}$, la manera más sencilla es despejar x_1 de la primera ecuación, x_2 de la segunda, etc. Para ello, es necesario que todos los elementos de la diagonal principal de A, por razones obvias, sean distintos de cero. Para ver esto en detalle considérese el sistema general de tres ecuaciones (naturalmente puede extenderse a cualquier número de ecuaciones).

Sea entonces

$$\begin{aligned} a_{11}\, x_1 + a_{12} x_2 + a_{13} x_3 &= b_1 \\ a_{21}\, x_1 + a_{22} x_2 + a_{23} x_3 &= b_2 \\ a_{31}\, x_1 + a_{32} x_2 + a_{33} x_3 &= b_3 \end{aligned}$$

con a_{11}, a_{22} y a_{33} distintos de cero.

Se despeja x_1 de la primera ecuación, x_2 de la segunda y x_3 de la tercera con lo que se obtiene

$$\begin{aligned} x_1 &= \qquad - \frac{a_{12}}{a_{11}} x_2 - \frac{a_{13}}{a_{11}} x_3 + \frac{b_1}{a_{11}} \\ x_2 &= - \frac{a_{21}}{a_{22}} x_1 \qquad - \frac{a_{23}}{a_{22}} x_3 + \frac{b_2}{a_{22}} \\ x_3 &= - \frac{a_{31}}{a_{33}} x_1 - \frac{a_{32}}{a_{33}} x_2 \qquad + \frac{b_3}{a_{33}} \end{aligned} \tag{3.88}$$

que en notación matricial queda

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 & - \frac{a_{12}}{a_{11}} & - \frac{a_{13}}{a_{11}} \\ - \frac{a_{21}}{a_{22}} & 0 & - \frac{a_{23}}{a_{22}} \\ - \frac{a_{31}}{a_{33}} & - \frac{a_{32}}{a_{33}} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} \frac{b_1}{a_{11}} \\ \frac{b_2}{a_{22}} \\ \frac{b_3}{a_{33}} \end{bmatrix} \tag{3.89}$$

y ésta es la ecuación 3.86 desarrollada, con

$$B = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} \\ -\frac{a_{31}}{a_{33}} & -\frac{a_{32}}{a_{33}} & 0 \end{bmatrix} \quad \text{y} \quad \mathbf{c} = \begin{bmatrix} \frac{b_1}{a_{11}} \\ \frac{b_2}{a_{22}} \\ \frac{b_3}{a_{33}} \end{bmatrix}$$

Una vez que se tiene la forma 3.89, se propone un vector inicial $\mathbf{x}^{(0)}$ que puede ser $\mathbf{x}^{(0)} = \mathbf{0}$, o algún otro que sea aproximado al vector solución $\mathbf{x}$.

Para iterar existen dos variantes

1. Iteración de Jacobi (método de desplazamientos simultáneos)

Si

$$\mathbf{x}^{(k)} = \begin{bmatrix} x_1^k \\ x_2^k \\ x_3^k \end{bmatrix} \tag{3.90}$$

es el vector aproximación a la solución $\mathbf{x}$ después de k iteraciones, entonces se tiene para la siguiente aproximación

$$\mathbf{x}^{(k+1)} = \begin{bmatrix} x_1^{k+1} \\ x_2^{k+1} \\ x_3^{k+1} \end{bmatrix} = \begin{bmatrix} \frac{1}{a_{11}} (b_1 - a_{12} x_2^k - a_{13} x_3^k) \\ \frac{1}{a_{22}} (b_2 - a_{21} x_1^k - a_{23} x_3^k) \\ \frac{1}{a_{33}} (b_3 - a_{31} x_1^k - a_{32} x_3^k) \end{bmatrix} \tag{3.91}$$

O bien, para un sistema de n ecuaciones con n incógnitas y usando notación más compacta y de mayor utilidad en programación, se tiene

$$\boxed{x_i^{k+1} = -\frac{1}{a_{ii}} \left[-b_i + \sum_{\substack{j=1 \\ j \neq i}} a_{ij} x_j^k \right], \text{ para } 1 \leq i \leq n} \tag{3.92}$$

2. Iteración de Gauss-Seidel (método de desplazamientos sucesivos)

En este método los valores que se van calculando en la $(k+1)$–ésima iteración se emplean para calcular los valores faltantes de esa misma iteración; es decir, con $\mathbf{x}^{(k)}$ se calcula $\mathbf{x}^{(k+1)}$ de acuerdo con

$$\mathbf{x}^{(k+1)} = \begin{bmatrix} x_1^{k+1} \\ x_2^{k+1} \\ x_3^{k+1} \end{bmatrix} = \begin{bmatrix} \frac{1}{a_{11}} & (b_1 - a_{12}\, x_2^k - & a_{13}x_3^k) \\ \frac{1}{a_{22}} & (b_2 - a_{21}\, x_1^{k+1} - & a_{23}x_3^k) \\ \frac{1}{a_{33}} & (b_3 - a_{31}\, x_1^{k+1} - & a_{32}x_2^{k+1}) \end{bmatrix} \tag{3.93}$$

O bien, para un sistema de n ecuaciones

$$x_i^{k+1} = -\frac{1}{a_{ii}} \left[-b_i + \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} + \sum_{j=i+1}^{n} a_{ij}\, x_j^k \right], \text{ para } 1 \leq i \leq n \tag{3.94}$$

Sugerencia: El empleo de un pizarrón electrónico para los siguientes ejemplos o el de una calculadora programable, atenuaría considerablemente el trabajo de los cálculos.

Ejemplo 3.39

Resuelva el siguiente sistema por los métodos de Jacobi y Gauss-Seidel

$$\begin{aligned} 4x_1 - x_2 &= 1 \\ -x_1 + 4x_2 - x_3 &= 1 \\ - x_2 + 4x_3 - x_4 &= 1 \\ - x_3 - 4x_4 &= 1 \end{aligned} \tag{3.95}$$

SOLUCIÓN

Despejando x_1 de la primera ecuación, x_2 de la segunda, etc., se obtiene

$$\begin{aligned} x_1 &= x_2/4 + 1/4 \\ x_2 &= x_1/4 - x_3/4 - 1/4 \\ x_3 &= x_2/4 - x_4/4 - 1/4 \\ x_4 &= - x_3/4 - 1/4 \end{aligned} \tag{3.96}$$

Vector inicial

Cuando no se tiene una aproximación al vector solución, se emplea generalmente como vector inicial el vector cero, esto es

$$\mathbf{x}^{(0)} = [0\ 0\ 0\ 0]^T$$

a) **Método de Jacobi**

El cálculo de $\mathbf{x}^{(1)}$ en el método de Jacobi se obtiene remplazando $\mathbf{x}^{(0)}$ en cada una de las ecuaciones de 3.96

$$\begin{aligned} x_1 &= \quad 0/4 \quad + \quad 1/4 = 1/4 \\ x_2 &= 0/4 \quad + \quad 0/4 \quad + \quad 1/4 = 1/4 \\ x_3 &= \quad 0/4 \quad + \quad 0/4 \quad + \quad 1/4 = 1/4 \\ x_4 &= \quad 0/4 \quad + \quad 1/4 = 1/4 \end{aligned}$$

y entonces $\mathbf{x}^{(1)} = (1/4\ 1/4\ 1/4\ 1/4)^T$

Para calcular $\mathbf{x}^{(2)}$ se sustituye $\mathbf{x}^{(1)}$ en cada una de las ecuaciones de 3.96. Para simplificar la notación se han omitido los superíndices.

$$\begin{aligned} x_1 &= \quad 1/16 \quad + \quad 1/4 = 0.3125 \\ x_2 &= 1/16 \quad + \quad 1/16 \quad + \quad 1/4 = 0.3750 \\ x_3 &= \quad 1/16 \quad + \quad 1/16 \quad + \quad 1/4 = 0.3750 \\ x_4 &= \quad 1/16 \quad + \quad 1/4 = 0.3125 \end{aligned}$$

A continuación se presentan los resultados de subsecuentes iteraciones, en forma tabular

k	x_1^k	x_2^k	x_3^k	x_4^k
0	0.0000	0.0000	0.0000	0.0000
1	0.2500	0.2500	0.2500	0.2500
2	0.3125	0.3750	0.3750	0.3125
3	0.3438	0.4219	0.4219	0.3438
4	0.3555	0.4414	0.4414	0.3555
5	0.3604	0.4492	0.4492	0.3604
6	0.3623	0.4524	0.4524	0.3623
7	0.3631	0.4537	0.4537	0.3631
8	0.3634	0.4542	0.4542	0.3634
9	0.3635	0.4544	0.4544	0.3635
10	0.3636	0.4545	0.4545	0.3636

Tabla 3.1 Solución del sistema 3.95 por el método de Jacobi.

b) **Método de Gauss-Seidel**

Para el cálculo del primer elemento del vector $\mathbf{x}^{(1)}$, se sustituye $\mathbf{x}^{(0)}$ en la primera ecuación de 3.96, para simplificar la notación se han omitido los superíndices.

$$x_1 = 0/4 + 1/4 = 1/4$$

Para el cálculo de x_2 de $\mathbf{x}^{(1)}$, se emplea el valor de x_1 ya obtenido (1/4) y los valores de x_2, x_3 y x_4 de $\mathbf{x}^{(0)}$. Así

$$x_2 = \frac{1}{4(4)} + 0/4 + 1/4 = 0.3125$$

Con los valores de x_1 y x_2 ya obtenidos y con x_3 y x_4 de $\mathbf{x}^{(0)}$ se evalúa x_3 de $\mathbf{x}^{(1)}$.

$$x_3 = 0.3125/4 + 0/4 + 1/4 = 0.3281$$

Finalmente, con los valores de x_1, x_2 y x_3 calculados previamente y con x_4 de $\mathbf{x}^{(0)}$, se obtiene la última componente de $\mathbf{x}^{(1)}$

$$x_4 = 0.3281/4 + 1/4 = 0.3320$$

Entonces $\mathbf{x}^{(1)} = [0.25\ 0.3125\ 0.3281\ 0.3320]^T$

Para la segunda iteración (cálculo de $\mathbf{x}^{(2)}$) se procede de igual manera.

$$x_1 = 0.3125/4 + 1/4 = 0.3281$$

$$x_2 = 0.3281/4 + 0.3281/4 + 1/4 = 0.4141$$

$$x_3 = 0.4141/4 + 0.3320/4 + 1/4 = 0.4365$$

$$x_4 = 0.4365/4 + 1/4 = 0.3591$$

Con lo que $\mathbf{x}^{(2)} = [\ 0.3281\ 0.4141\ 0.4365\ 0.3591]^T$.

En la tabla 3.2 se presentan los resultados de las iteraciones subsecuentes.

k	x_1^k	x_2^k	x_3^k	x_4^k
0	0.0000	0.0000	0.0000	0.0000
1	0.2500	0.3125	0.3281	0.3320
2	0.3281	0.4141	0.4365	0.3591
3	0.3535	0.4475	0.4517	0.3629
4	0.3619	0.4534	0.4541	0.3635
5	0.3633	0.4544	0.4545	0.3636
6	0.3636	0.4545	0.4545	0.3636

Tabla 3.2 Solución del sistema 3.95 por el método de Gauss-Seidel.

En la aplicación de estas dos variantes son válidas las preguntas siguientes:

1. ¿La sucesión de vectores $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, $\mathbf{x}^{(3)}$, ..., converge o se aleja del vector solución $\mathbf{x} = [x_1\, x_2 \ldots x_n]^T$?
2. ¿Cuándo se detendrá el proceso iterativo?

Las respuestas correspondientes, conocidas como criterio de convergencia, se dan a continuación

1. Si la sucesión converge a $\mathbf{x}$, cabe esperar que los elementos de $\mathbf{x}^{(k)}$ se vayan acercando a los elementos correspondientes de $\mathbf{x}$; es decir, x_1^k, a x_1, x_2^k a x_2, etc., o que se alejen en caso contrario.
2. Cuando

   a) Los valores absolutos $| x_1^{k+1} - x_1^k |$, $| x_2^{k+1} - x_2^k |$, etc., sean todos menores de un número pequeño ε cuyo valor será dado por el programador.

O bien

   b) Si el número de iteraciones ha excedido un máximo predeterminado MAXIT.

Por otro lado, es natural pensar que si la sucesión $\mathbf{x}^{(0)}$, $\mathbf{x}^{(1)}$, ..., converge a $\mathbf{x}$, la distancia (véase Sec. 3.2) de $\mathbf{x}^{(0)}$ a $\mathbf{x}$, de $\mathbf{x}^{(1)}$ a $\mathbf{x}$, etc., se va reduciendo; también es cierto que la distancia entre cada dos vectores consecutivos $\mathbf{x}^{(0)}$ y $\mathbf{x}^{(1)}$, $\mathbf{x}^{(1)}$ y $\mathbf{x}^{(2)}$, etc., se decrementa conforme el proceso iterativo avanza; esto es, la sucesión de números reales

$$\begin{matrix} | \mathbf{x}^{(1)} - \mathbf{x}^{(0)} | \\ | \mathbf{x}^{(2)} - \mathbf{x}^{(1)} | \\ \vdots \\ | \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} | \end{matrix} \tag{3.97}$$

convergirá a cero.

Si, por el contrario, esta sucesión de números diverge, entonces puede pensarse que el proceso diverge. Con esto, un criterio más es

   c) Detener el proceso una vez que $| \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} | < \varepsilon$

Al elaborar un programa de cómputo para resolver sistemas de ecuaciones lineales, generalmente se utilizan los criterios (a), (b) y (c) o la combinación de (a) y (b), o la de (b) y (c).

Si se observan las columnas de las tablas 3.1 y 3.2, se advertirá que todas son sucesiones de números convergentes, por lo que ambos métodos convergen a un vector, presumiblemente la solución del sistema 3.95.

Si se tomara el criterio (a) con $\varepsilon = 10^{-2}$ y el método de Jacobi, ε se satisface en la sexta iteración de la tabla 3.1; en cambio si $\varepsilon = 10^{-3}$, se necesitan de 10 iteraciones.

Si se toma $\varepsilon = 10^{-3}$, el método de Gauss-Seidel y el criterio (a), se requerirían sólo seis iteraciones, como puede verse en la tabla 3.2.

Aunque hay ejemplos en los que Jacobi converge y Gauss-Seidel diverge y viceversa, en general puede esperarse convergencia más rápida por Gauss-Seidel, o una manifestación más rápida de divergencia. Esto se debe al hecho de ir usando los valores más recientes de $\mathbf{x}^{(k+1)}$ que permitirán acercarse o alejarse más rápidamente de la solución.

Rearreglo de ecuaciones.

Para motivar el rearreglo de ecuaciones, se propone resolver el siguiente sistema con el método de Gauss-Seidel y con $\varepsilon = 10^{-2}$ aplicado a $| \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} |$.

$$\begin{aligned} -x_1 + 3x_2 + 5x_3 + 2x_4 &= 10 \\ x_1 + 9x_2 + 8x_3 + 4x_4 &= 15 \\ x_2 + x_4 &= 2 \\ 2x_1 + x_2 + x_3 - x_4 &= -3 \end{aligned} \tag{3.98}$$

Al resolver para x_1 de la primera ecuación, para x_2 de la segunda, x_3 de la cuarta y x_4 de la tercera se obtiene

$$\begin{aligned} x_1 &= 3x_2 + 5x_3 + 2x_4 - 10 \\ x_2 &= -x_1/9 - (8/9)x_3 - (4/9)x_4 + 15/9 \\ x_3 &= -2x_1 - x_2 + x_4 - 3 \\ x_4 &= -x_2 + 2 \end{aligned}$$

Con el vector cero como vector inicial, se tiene la siguiente sucesión de vectores. Nótese que el proceso diverge.

k	x_1^k	x_2^k	x_3^k	x_4^k	$\| \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \|$
0	0.000	0.000	0.000	0.000	
1	−10.000	2.7778	14.222	−0.7778	17.62
2	67.8889	−18.172	−121.2	20.17	159.0
3	−631.1	170.7	1108.0	−168.71	1439.05

Tabla 3.3. Aplicación del método de Gauss-Seidel al sistema 3.98.

Si el proceso iterativo diverge, como es el caso, un rearreglo de las ecuaciones puede originar convergencia; por ejemplo, en lugar de despejar x_1 de la primera

ecuación, x_2 de la segunda, etc., cabe despejar las diferentes x_i de diferentes ecuaciones, teniendo cuidado de que los coeficientes de las x_i despejadas sean distintos de cero.

Esta sugerencia presenta, para un sistema n de ecuaciones, $n!$ distintas formas de rearreglar dicho sistema. A fin de simplificar este procedimiento, se utilizará el siguiente teorema

Teorema 3.2 Los procesos de Jacobi y Gauss-Seidel convergirán si en la matriz coeficiente cada elemento de la diagonal principal es mayor (en valor absoluto) que la suma de los valores absolutos de todos los demás elementos de la misma fila o columna (matriz diagonal dominante). Es decir, se asegura la convergencia si,

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \qquad 1 \leq i \leq n$$

y

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ji}| \qquad 1 \leq i \leq n \tag{3.99}$$

Este teorema no será de mucha utilidad si se toma al pie de la letra, ya que contados sistemas de ecuaciones lineales poseen matrices coeficiente diagonalmente dominantes; sin embargo, si se arreglan las ecuaciones para tener el sistema lo más cercano posible a las condiciones del teorema, algún beneficio se puede obtener. Ésta es la pauta para reordenar las ecuaciones y obtener o mejorar la convergencia, en el mejor de los casos. A continuación se ilustra esto, rearreglando el sistema 3.98, despejando x_1 de la ecuación 4, x_2 de la ecuación 2, x_3 de la ecuación 1 y x_4 de la ecuación 3 para llegar a

$$\begin{aligned}
x_1 &= -x_2/2 - x_3/2 + x_4/2 - 3/2 \\
x_2 &= -x_1/9 - 8x_3/9 - 4x_4/9 + 15/9 \\
x_3 &= x_1/5 - 3x_2/5 - 2x_4/5 + 10/5 \\
x_4 &= -x_2 + 2
\end{aligned}$$

Los resultados para las primeras 18 iteraciones con el vector cero como vector inicial se muestran en la tabla 3.4

Antes de continuar las iteraciones, puede observarse en la tabla 3.4 que los valores de $\mathbf{x}^{(18)}$ parecen converger al vector

$$\mathbf{x} = [-1\ 0\ 1\ 2]^T$$

Con la sustitución de estos valores en el sistema 3.98, se comprueba que $x_1 = 1, x_2 = 0.0, x_3 = 1$ y $x_4 = 2$ es el vector solución y por razones obvias se detiene el proceso.

Finalmente, las ecuaciones 3.99 son equivalentes (en sistemas de ecuaciones) a la expresión 2.10 del capítulo 2 que establece el criterio de convergencia del método iterativo para resolver $f(x) = 0$.

k	x_1^k	x_2^k	x_3^k	x_4^k	$\lvert \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \rvert$
0	0.0000	0.0000	0.0000	0.0000	
1	–1.5000	1.8333	0.6000	0.1667	2.44
2	–2.6333	1.3519	0.5956	0.6481	1.32
3	–2.1496	1.0881	0.6580	0.9119	0.6140
4	–1.9171	0.8895	0.7181	1.1105	0.3695
5	–1.7486	0.7291	0.7686	1.2704	0.2867
6	–1.6134	0.5978	0.8102	1.4022	0.2337
7	–1.5030	0.4903	0.8444	1.5097	0.1907
8	–1.4125	0.4020	0.8724	1.5980	0.1567
9	–1.3382	0.3297	0.8953	1.6703	0.1285
10	–1.2774	0.2704	0.9142	1.7296	0.10529
11	–1.2275	0.2217	0.9296	1.7783	0.08643
12	–1.1865	0.1818	0.9423	1.8182	0.07089
13	–1.1530	0.1491	0.9527	1.8509	0.06162
14	–1.1254	0.1223	0.9612	1.8777	0.04764
15	–1.1029	0.1003	0.9682	1.8997	0.03903
16	–1.0844	0.0822	0.9739	1.9178	0.03209
17	–1.0692	0.0674	0.9786	1.9326	0.02629
18	–1.0567	0.0553	0.9824	1.9447	0.02152

Tabla 3.4. Aplicación del método de Gauss-Seidel al sistema 3.98, rearreglando las ecuaciones para obtener una aproximación a un sistema diagonal dominante.

Se presenta a continuación un algoritmo para resolver sistemas de ecuaciones lineales con el método iterativo, en sus dos versiones, desplazamientos simultáneos y desplazamientos sucesivos

ALGORITMO 3.11 Métodos de Jacobi y Gauss-Seidel

Para encontrar la solución aproximada del sistema de ecuaciones $A\,\mathbf{x} = \mathbf{b}$ proporcionar los

DATOS: El número de ecuaciones N, la matriz coeficiente A, el vector de términos independientes **b**, el vector inicial **x0**, el número máximo de iteraciones

MAXIT, el valor de EPS y M = 0 para usar JACOBI o M ≠ 0 para usar GAUSS-SEIDEL.

RESULTADOS: La solución aproximada **x** y el número de iteraciones K en que se alcanzó la convergencia o mensaje *"NO SE ALCANZO LA CONVERGENCIA"*, la última aproximación a **x** y MAXIT.

PASO 1. Arreglar la matriz aumentada de modo que la matriz coeficiente quede lo más cercana posible a la diagonal dominante (véase Probl. 3.55).
PASO 2. Hacer K = 1
PASO 3. Mientras K ≤ MAXIT repetir los pasos 4 a 18.
  PASO 4. Si M = 0 ir al paso 5. De otro modo Hacer* **x** = **x0**.
  PASO 5. Hacer I = 1
  PASO 6. Mientras I ≤ N, repetir los pasos 7 a 14.
    PASO 7. Hacer SUMA = 0.
    PASO 8. Hacer J = 1
    PASO 9. Mientras J ≤ N, repetir los pasos 10 a 12.
      PASO 10 Si J = I ir al paso 12.
      PASO 11. Hacer SUMA = SUMA+A(I,J)*$x0$(J)
      PASO 12. Hacer J = J + 1
    PASO 13. Si M = 0, hacer x(I)=–(b(I)–SUMA)/A(I,I) De otro modo hacer $x0$(I)=(b(I)–SUMA)/A(I,I)
    PASO 14. Hacer I = I + 1
  PASO 15. Si | **x** – **x0** | ≤ EPS ir al paso 19. De otro modo continuar.
  PASO 16. Si M = 0, hacer **x0** = **x**
  PASO 17. Hacer K = K + 1
PASO 18. IMPRIMIR mensaje "NO SE ALCANZO LA CONVERGENCIA", el vector **x**, MAXIT y el mensaje "ITERACIONES" y TERMINAR.
PASO 19. IMPRIMIR el mensaje "VECTOR SOLUCION", **x**, K y el mensaje "ITERACIONES" y TERMINAR.

*Operaciones vectoriales.

Sugerencia: Una vez más se recomienda programar el algoritmo 3.11 en un lenguaje de alto nivel (véase programa 3.3 en el disco), o bien en una calculadora o en un pizarrón electrónico, donde las operaciones vectoriales se ejecutan con sólo indicarlas.

Aceleración de convergencia

Si aún después de arreglado el sistema por resolver $A\,\mathbf{x} = \mathbf{b}$, conforme la pauta del teorema 3.2, no se obtiene convergencia por los métodos de Jacobi y Gauss-Seidel o es muy lenta (como sucedió con el sistema 3.98 de la sección anterior), puede recurrirse a los métodos de **relajación** que, como se hará notar posteriormente, son los métodos de Jacobi y Gauss-Seidel afectados por un factor de peso w que, elegido adecuadamente, puede producir convergencia o acelerarla si ya existe. Se describen a continuación estos métodos para un sistema de n ecuaciones en n incógnitas.

Llámese N la matriz coeficiente del sistema por resolver, una vez que haya sido llevada a la forma más cercana posible a diagonal dominante, y después de dividir la primera fila entre $a_{1,1}$, la segunda entre $a_{2,2}$, ..., y la n-ésima entre $a_{n,n}$. N es una matriz con unos en la diagonal principal. A continuación descompóngase N en la siguiente forma

$$N = L + I + U,$$

donde L es una matriz cuyos elementos por debajo de su diagonal principal son idénticos a los correspondientes de N y ceros en cualquier otro sitio, I es la matriz identidad y U una matriz cuyos elementos arriba de la diagonal principal son idénticos a los correspondientes de N y cero en cualquier otro sitio. Sustituyendo esta descomposición de N, el sistema que se quiere resolver quedaría:

$$(L + I + U)\,\mathbf{x} = \mathbf{b} \tag{3.100}$$

Si ahora se suma $\mathbf{x}$ a cada miembro de la ecuación 3.100 se obtiene

$$(L + I + U)\,\mathbf{x} + \mathbf{x} = \mathbf{b} + \mathbf{x}$$

"Despejando" $\mathbf{x}$ del lado izquierdo, se llega al esquema siguiente

$$\mathbf{x} = \mathbf{x} + [\mathbf{b} - L\,\mathbf{x} - \mathbf{x} - U\,\mathbf{x}], \tag{3.101}$$

que puede utilizarse para iterar a partir de un vector inicial $\mathbf{x}^{(0)}$. Nótese que la ecuación 3.101, puede reducirse a la ecuación 3.89, ya que sólo es un rearreglo de ésta.

Al aplicar la ecuación 3.101, pueden presentarse de nuevo las dos variantes que dieron lugar a los métodos de Jacobi y Gauss-Seidel, con lo que el esquema de desplazamiento simultáneos quedaría

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + [\mathbf{b} - L\,\mathbf{x}^{(k)} - \mathbf{x}^{k} - U\,\mathbf{x}^{(k)}] \tag{3.102}$$

y el de desplazamientos sucesivos así:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + [\mathbf{b} - L\,\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} - U\,\mathbf{x}^{(k)}] \tag{3.103}$$

Llegar al esquema 3.102 y 3.103 no es simplemente para tener una versión distinta de las ecuaciones 3.89, sino para someterlo a un análisis que permita proponer "nuevos métodos" o mejoras en los que ya se tienen. Por ejemplo, factorizando $\mathbf{x}^{(k)}$ dentro del paréntesis rectangular de la ecuación 3.102, se tiene

$$\mathbf{b} - (L + I + U)\,\mathbf{x}^{(k)} = \mathbf{b} - N\,\mathbf{x}^{(k)} = \mathbf{r}^{(k)} \tag{3.104}$$

vector que se denota como $\mathbf{r}^{(k)}$ y se llama **vector residuo** de la k-ésima iteración y puede tomarse como una medida de la cercanía de $\mathbf{x}^{(k)}$ al vector solución $\mathbf{x}$; si las componentes de $\mathbf{r}^{(k)}$ o $|\ \mathbf{r}^{(k)}\ |$ son pequeñas, $\mathbf{x}^{(k)}$ suele ser una buena aproximación a $\mathbf{x}$; pero si los elementos de $\mathbf{r}^{(k)}$ o $|\ \mathbf{r}^{(k)}\ |$ son grandes, puede pensarse que $\mathbf{x}^{(k)}$ no es muy cercana a $\mathbf{x}$. Aunque hay circunstancias donde esto no se cumple, por ejemplo, cuando el sistema por resolver está **mal condicionado** (véase Sec. 3.4), *es práctico tomar estos criterios como válidos.*

Al sustituir en ella la ecuación 3.104, la 3.102 queda

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{r}^{(k)} \tag{3.105}$$

que puede verse como un esquema iterativo en que el vector de la $(k+1)$-ésima iteración se obtiene a partir del vector de la k-ésima iteración y el residuo correspondiente.

Si la aplicación de la ecuación 3.105 a un sistema particular da convergencia lenta, entonces $\mathbf{x}^{(k+1)}$ y $\mathbf{x}^{(k)}$ están muy cercanas entre sí, y para que la convergencia se acelere puede intentarse afectar $\mathbf{r}^{(k)}$ con un peso $w > 1$ (**sobrerrelajar** el proceso); si, en cambio, el proceso diverge $|\ \mathbf{r}^{(k)}\ |$ es grande y convendría afectar $\mathbf{r}^{(k)}$ con un factor $w < 1$ (**subrelajar** el proceso), para provocar la convergencia. El esquema 3.105 quedaría en general así

$$\boxed{\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + w\ \mathbf{r}^{(k)}} \tag{3.106}$$

o

$$\boxed{x_i^{k+1} = x_i^k + w\ [b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}\ x_j^k] \qquad 1 \leq i \leq n\,,} \tag{3.107}$$

para desplazamientos simultáneos.

Para desplazamientos sucesivos, en cambio, quedaría

$$\boxed{\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + w\ [\ \mathbf{b} - L\ \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} - U\ \mathbf{x}^{(k)}\ ]} \tag{3.108}$$

o

$$\boxed{x_i^{k+1} = x_i^k + w\ [b_i - \sum_{j=1}^{i-1} l_{ij}\ x_j^{k+1} - x_i^k - \sum_{j=i+1}^{n} u_{ij}\ x_j^k] \qquad 1 \leq i \leq n} \tag{3.109}$$

Estos métodos se abrevian frecuentemente como SOR (del inglés *Succesive Over-Relaxation).*

En general, el cálculo de w es complicado y sólo para sistemas especiales (matriz coeficiente positivamente definida y tridiagonal) se tiene una fórmula*.

*Burden, R.L. y Faires, J.D. *Análisis Numérico*. Grupo Ed. rial Iberoamérica (1985) pp 475.

Ejemplo 3.40

Resuelva el sistema 3.98

$$\begin{aligned} -x_1 + 3x_2 + 5x_3 + 2x_4 &= 10 \\ x_1 + 9x_2 + 8x_3 + 4x_4 &= 15 \\ x_2 + x_4 &= 2 \\ 2x_1 + x_2 + x_3 - x_4 &= -3 \end{aligned}$$

con desplazamientos sucesivos, $w = 1.3$ y con $\varepsilon = 10^{-2}$ aplicado a $| \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} |$. (Puede seguir los cálculos con un pizarrón electrónico).

SOLUCIÓN

La matriz N y el vector de términos independientes correspondiente son

$$N = \begin{bmatrix} 1 & 1/2 & 1/2 & -1/2 \\ 1/9 & 1 & 8/9 & 4/9 \\ -1/5 & 3/5 & 1 & 2/5 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \quad \mathbf{b} = [-3/2 \quad 15/9 \quad 10/5 \quad 2]^T$$

Descomposición de N

$$L = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1/9 & 0 & 0 & 0 \\ -1/5 & 3/5 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad U = \begin{bmatrix} 0 & 1/2 & 1/2 & -1/2 \\ 0 & 0 & 8/9 & 4/9 \\ 0 & 0 & 0 & 2/5 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Primera iteración

Obtención de $\mathbf{x}^{(1)}$ a partir del vector inicial $\mathbf{x}^{(0)} = [0\ 0\ 0\ 0]^T$ y empleando la ecuación 3.108.

Cálculo de x_1^1, esto es, $i = 1$ y $k+1 = 1$

$$x_1^1 = x_1^0 + 1.3[b_1 - \sum_{j=1}^{0} l_{1,j}\, x_j^1 - x_1^0 - \sum_{j=2}^{4} u_{1,j}\, x_j^0]$$

Obsérvese que en la primera sumatoria el valor inicial ($j=1$) es mayor que el valor final (0); la convención en estos casos es que tal sumatoria no se realiza. Por tanto

$$x_1^1 = 0 + 1.3\ [-3/2 - 0 - 1/2(0) - 1/2(0) + 1/2(0)] = -1.95$$

Cálculo de x_2^1, esto es, $i = 2$ y $k+1 = 1$

$$x_2^1 = x_2^0 + 1.3\ [b_2 - \sum_{j=1}^{1} l_{2j}\, x_j^1 - x_2^0 - \sum_{j=3}^{4} u_{2j}\, x_j^0\ ]$$

$$= 0 + 1.3[15/9 - 1/9(-1.95) - 0 - 8/9(0) - 4/9(0)] = 2.4483$$

Cálculo de x_3^1, esto es $i = 3$ y $k+1 = 1$

$$x_3^1 = x_3^0 + 1.3[\, b_3 - \sum_{j=1}^{2} l_{3j}\, x_j^1 - x_3^0 - \sum_{j=4}^{4} - u_{3j}\, x_j^0\ ]$$

$$= 0+1.3[10/5-(-1/5)(-1.95)-(3/5)(2.4483)-0-2/5(0)] = 0.1833$$

Cálculo de x_4^1, esto es, $i = 4$ y $k+1 = 1$

$$x_4^1 = x_4^0 + 1.3[\, b_4 - \sum_{j=1}^{3} l_{4j}\, x_j^1 - x_4^0 - \sum_{j=5}^{4} u_{4j}\, x_j^0\ ]$$

$$= 0 + 1.3[2 - 0(-1.95) - 1(2.4483) - 0(0.1833) - 0] = -0.5828$$

Cálculo de $|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}| = d_1$

$$d_1 = \sqrt{(x_1^1 - x_1^0)^2 + (x_2^1 - x_2^0)^2 + (x_3^1 - x_3^0)^2 + (x_4^1 - x_4^0)^2}$$

$$= \sqrt{(-1.95)^2 + (2.4483)^2 + (0.1833)^2 + (-0.5828)^2} = 3.1891$$

Los valores mostrados en la tabla 3.5 se encuentran continuando las iteraciones.

k	x_1^k	x_2^k	x_3^k	x_4^k	$\lvert \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \rvert$
0	0.0000	0.0000	0.0000	0.0000	
1	–1.9500	2.4483	0.1833	–0.5828	3.1891
2	–3.4544	2.0561	0.3462	0.1020	1.7066
3	–2.4089	1.4388	0.6945	0.6989	1.3971
4	–2.1597	0.8406	0.8110	1.2976	0.8898
5	–1.5322	0.4489	0.9384	1.6271	0.8190
6	–1.3312	0.2055	0.9674	1.8447	0.3848
7	–1.1140	0.0821	0.9968	1.9397	0.2689
8	–1.0563	0.0220	1.0005	1.9895	0.0972
9	–1.0046	–0.0004	1.0045	2.0037	0.0583
10	–0.9988	–0.0073	1.0027	2.0074	0.0103
11	–0.9919	–0.0070	1.0024	2.0066	0.0072

Tabla 3.5 Resultados obtenidos con $w = 1.3$.

Al comparar estos resultados con los obtenidos en la tabla 3.4 (método de Gauss-Seidel aplicado al sistema que aquí se resuelve), se observa que la convergencia es acelerada y los cálculos se reducen a la mitad.

Comparación de los métodos directos e iterativos

Una parte importante del análisis numérico es conocer las características (ventajas y desventajas) de los métodos numéricos básicos que resuelven una familia de problemas (en este caso $A\ \mathbf{x} = \mathbf{b}$), para así elegir el algoritmo más adecuado para cada problema.

A continuación se presentan las circunstancias donde pudiera verse como ventajosa la elección de un método iterativo y también a qué se renuncia con esta decisión.

Ventajas	Desventajas
1. Probablemente más eficientes que los directos para sistemas de orden muy alto. 2. Más simples de programar. 3. Puede aprovecharse una aproximación a la solución, si tal aproximación existe. 4. Se obtienen facilmente aproximaciones burdas de la solución. 5. Son menos sensibles a los errores de redondeo (valioso en sistemas mal condicionados). 6. Se requiere menos memoria de máquina. Generalmente, las necesidades de memoria son proporcionales al orden de la matriz.	1. Si se tienen varios sistemas que comparten la matriz coeficiente, esto no representará ahorro de cálculos ni tiempo de máquina, ya que por cada vector a la derecha de A tendrá que aplicarse el método seleccionado. 2. Aún cuando la convergencia esté asegurada, puede ser lenta y, por lo tanto, los cálculos requeridos para obtener una solución particular no son predecibles. 3. El tiempo de máquina y la exactitud del resultado dependen del criterio de convergencia. 4. Si la convergencia es lenta, los resultados deben interpretarse con cautela. 5. No se tiene ventaja particular alguna (tiempo de máquina por iteración) si la matriz coeficiente es simétrica. 6. No se obtiene A^{-1} ni det A.

Tabla 3.6 Ventajas y desventajas de los métodos iterativos comparados con los métodos directos.

EJERCICIOS

3.1 En una columna de cinco platos, se requiere absorber benceno contenido en una corriente de gas V, con un aceite L que circula a contracorriente del gas. Con-

sidérese que el benceno transferido no altera sustancialmente el número de moles de V y L fluyendo a contracorriente, que la relación de equilibrio está dada por la ley de Henry ($y = m\,x$) y que la columna opera a régimen permanente. Calcule la composición del benceno en cada plato.

Datos: V = 100 moles/min; L= 500 moles/min

y_0 = 0.09 fracción molar de benceno en V.

x_0 = 0.0 fracción molar de benceno en L (el aceite entra por el domo sin benceno).

m = 0.12

SOLUCIÓN

Los balances de materia para el benceno en cada plato son (véase Fig. 3.13).

Plato	Balance de benceno
5	$L\,(x_0 - x_5) + V\,(y_4 - y_5) = 0$
4	$L\,(x_5 - x_4) + V\,(y_3 - y_4) = 0$
3	$L\,(x_4 - x_3) + V\,(y_2 - y_3) = 0$
2	$L\,(x_3 - x_2) + V\,(y_1 - y_2) = 0$
1	$L\,(x_2 - x_1) + V\,(y_0 - y_1) = 0$

Al sustituir la información que se tiene, las consideraciones hechas y rearreglando las ecuaciones, se llega a

$$
\begin{aligned}
512\,x_5 - 12\,x_4 &= 0\\
500\,x_5 - 512\,x_4 + 12\,x_3 &= 0\\
500\,x_4 - 512\,x_3 + 12\,x_2 &= 0\\
500\,x_3 - 512\,x_2 + 12\,x_1 &= 0\\
-500\,x_2 + 512\,x_1 &= 9
\end{aligned}
$$

Con el programa 3.2 del apéndice, se obtienen los siguientes resultados

$x_1 = 0.018$, $x_2 = 4.32 \times 10^{-4}$, $x_3 = 1.037 \times 10^{-5}$

$x_4 = 2.4869 \times 10^{-7}$, $x_5 = 5.8286 \times 10^{-9}$

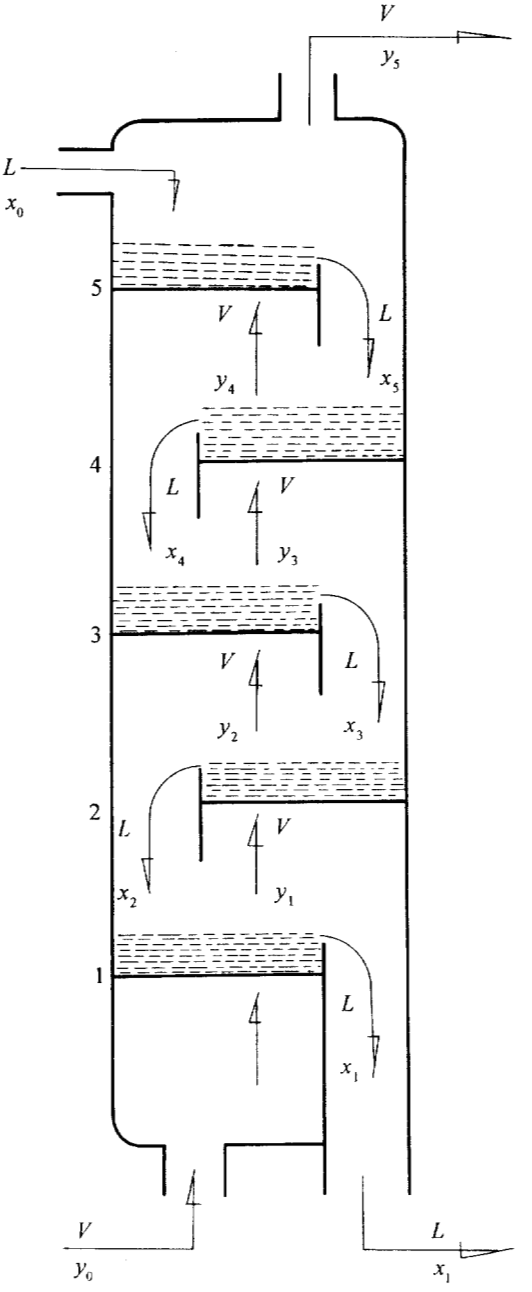


Figura 3.13. Columna de absorción de cinco platos.

3.2 Supóngase que se tiene una estructura cuadrada. Con el fin de analizarla se forma una malla imaginaria sobre dicha estructura, como se muestra en la figura siguiente

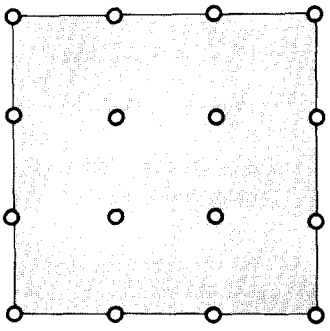


Figura 3.14. Estructura cuadrada.

y se numeran los nodos, por ejemplo, como se muestra a continuación

4	8	12	16
3	7	11	15
2	6	10	14
1	5	9	13

Cada nodo se identifica con $a_{i,j;}$ por ejemplo el 4 con $a_{1,1}$, el 6 con $a_{3,2}$, etc., y así queda formada una matriz A representativa de la estructura.

Ciertas consideraciones ingenieriles determinan que $a_{i,j} \neq 0$ siempre que los nodos i y j sean vecinos o adyacentes*. Para aclararlo, obsérvese que al nodo 5 le corresponde $a_{4,2}$ y como los nodos 4 y 2 no son vecinos, $a_{4,2} = 0$; al 11 en cambio le corresponde $a_{2,3}$ y como 2 y 3 son vecinos, $a_{2,3} \neq 0$. Por último $a_{3,3} \neq 0$, ya que el nodo 3 puede considerarse vecino consigo mismo.

Estas consideraciones generan matrices o sistemas dispersos o frecuentemente bandeados. Estos sistemas suelen ser muy grandes, ya que las mallas se construyen con un gran número de nodos.

En la aplicación del método de las rigideces**, para calcular los desplazamientos en los nodos de una estructura dada al aplicarse una carga en uno de los nodos, se obtiene el siguiente sistema de ecuaciones

$$10^5 \begin{bmatrix} 3.01687 & 0.00000 & 3.3750 & -3.0000 & 0.00000 & 0.0000 \\ 0.00000 & 3.01687 & 3.3750 & 0.0000 & -0.01687 & 3.3750 \\ 3.37500 & 3.37500 & 900.00 & 0.0000 & -3.37500 & 450.00 \\ -3.00000 & 0.00000 & 0.0000 & 3.0400 & 0.00000 & 6.0000 \\ 0.00000 & -0.01687 & -3.3750 & 0.0000 & 4.01687 & -3.3750 \\ 0.00000 & 3.37500 & 450.00 & 6.0000 & -3.37500 & 2100.0 \end{bmatrix} \begin{bmatrix} d_{xB} \\ d_{yB} \\ \theta_B \\ d_{xC} \\ d_{yC} \\ \theta_C \end{bmatrix} = \begin{bmatrix} 1600 \\ 0.00 \\ 0.00 \\ 0.00 \\ 0.00 \\ 0.00 \end{bmatrix}$$

donde la matriz coeficiente es conocida como la matriz de nodos, $d_B = [d_{xB}\ d_{yb}\ \theta_B]^T$ y $d_C = [d_{xC}\ d_{yC}\ \theta_C]^T$ son los vectores de desplazamiento de los nodos B y C, respectivamente. Resuelva dicho sistema.

*El nodo 7 por ejemplo, tiene como vecinos a los nodos 3, 6, 11 y 8.

**Carlos Magdaleno. *Análisis Matricial de Estructuras Reticulares.* Edición mimeográfica. ESIA, IPN.

SOLUCIÓN

La solución obtenida con el programa 3.2 del apéndice, se da a continuación

$$d_{xB} = 0.47185; \qquad d_{yB} = 0.00259; \qquad \theta_B = -\ 0.00125$$
$$d_{xC} = 0.46776; \qquad d_{yB} = -0.00194; \qquad \theta_C = -\ 0.00108$$

3.3 Determine las concentraciones molares de una mezcla de cinco componentes en solución a partir de los siguientes datos espectrofotométricos.

Longitud de onda i	Absorbancia molar del componente j					Absorbancia total observada
	1	2	3	4	5	
1	98	9	2	1	0.5	0.1100
2	11	118	9	4	0.88	0.2235
3	27	27	85	8	2	0.2800
4	1	3	17	142	25	0.3000
5	2	4	7	17	118	0.1400

Asúmase que la longitud de la trayectoria óptica es unitaria y que el solvente no absorbe a estas longitudes de onda.

SOLUCIÓN

Si se considera que se cumple la ley de Beer, entonces a una longitud de onda dada, i

$$A_{TOT\ i} = \sum_{j=1}^{5} \varepsilon_{i,j} C_j ,$$

donde

$A_{TOT\,i}$ es la absorbancia total observada a la longitud de onda i.

$\varepsilon_{i,j}$ es la absorbancia molar del componente j a la longitud de onda i.

C_j es la concentración molar del componente j en la mezcla.

Al sustituir los valores de la tabla se obtiene

$$\begin{aligned}
98\,C_1 + 9\,C_2 + 2\,C_2 + C_4 + 0.5\,C_5 &= 0.1100 \\
11\,C_1 + 118\,C_2 + 9\,C_3 + 4\,C_4 + 0.88\,C_5 &= 0.2235 \\
27\,C_1 + 27\,C_2 + 85\,C_3 + 8\,C_4 + 2\,C_5 &= 0.2800 \\
C_1 + 3\,C_2 + 17\,C_3 + 142\,C_4 + 25\,C_5 &= 0.3000 \\
2\,C_1 + 4\,C_2 + 7\,C_3 + 17\,C_3 + 118\,C_5 &= 0.1400
\end{aligned}$$

Un sistema de ecuaciones lineales con matriz coeficiente dominante. Esto sugiere resolver el sistema con el método de Gauss-Seidel.

El programa 3.3 del apéndice utiliza el método de Gauss-Seidel para resolver un sistema de ecuaciones lineales. Este programa se utilizó con el vector cero como vector inicial, por la relativa cercanía de cero con cada uno de los valores del lado derecho del sistema. Los resultados obtenidos son

$$C_1 = 0.000910 \; C_2 = 0.001569 \; C_3 = 0.002333$$
$$C_4 = 0.001664 \; C_5 = 0.000740$$

3.4 Determine la intensidad de corriente en cada rama del circuito que se da en la figura 3.15

SOLUCIÓN

Se asigna un sentido y una letra a cada magnitud desconocida; los sentidos supuestos son enteramente arbitrarios. Obsérvese que la intensidad de corriente en R_3, R_1 y E_1 es la misma y, por consiguiente, sólo se requiere una letra. Lo mismo ocurre para la intensidad de corriente en R_2, E_2 y R_6. Los nodos (puntos de la red en los cuales se unen tres o más conductores) se designan con las letras a, b, c, d.

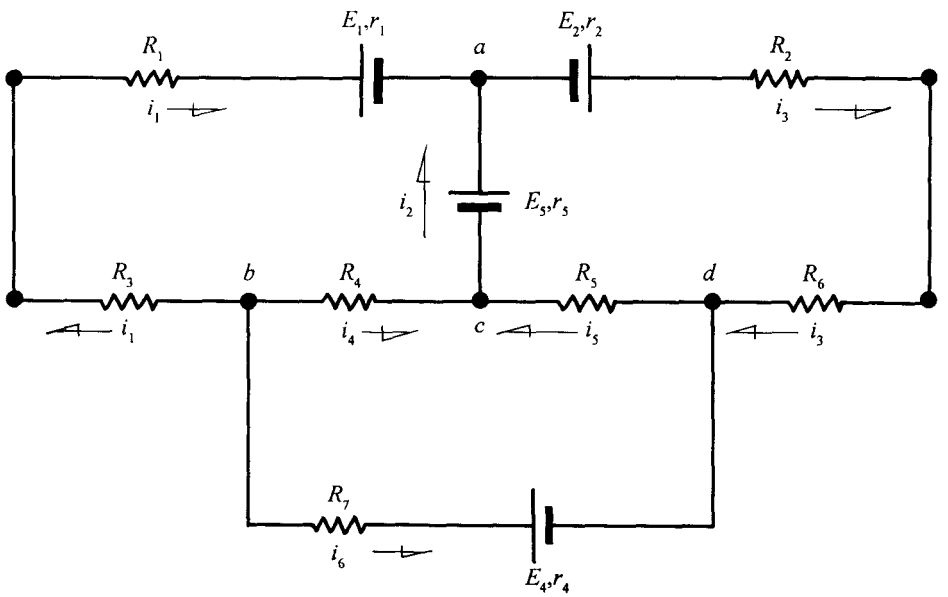


Figura 3.15. Circuito eléctrico con resitencias y fuentes de poder.

Aplicación de la **regla de los nodos de Kirchhoff** a tres nodos cualesquiera

$$\begin{array}{lrcl} \text{Nodo} & \sum i & = & 0 \\ \text{a} & i_1 + i_2 - i_3 & = & 0 \\ \text{b} & -i_1 - i_4 - i_6 & = & 0 \\ \text{c} & i_4 + i_5 - i_2 & = & 0 \end{array}$$

Si bien es cierto que hay un nodo más, el d, la aplicación de la regla daría una ecuación linealmente dependiente de las otras tres, esto es

$$\text{Nodo d} \qquad i_6 + i_3 - i_5 = 0,$$

ecuación que se obtiene sumando las tres primeras; por ello resulta redundante y en general se aplica dicha regla a n-1 nodos solamente.

En la figura 3.16 se representa el circuito cortado en mallas. Considérese en cada malla como positivo el sentido de las agujas del reloj. La regla de las mallas de Kirchhoff ($\Sigma\, E_k = \Sigma\, i_k\, R_k$) proporciona las siguientes ecuaciones

$$\text{Malla} \qquad \sum E_k = \sum i_k R_k$$

$$\text{I} \qquad -E_1 - E_5 = i_1 R_1 + i_1 r_1 - i_2 r_5 - i_4 R_4 + i_1 R_3$$

$$\text{II} \qquad E_2 + E_5 = i_3 r_2 + i_3 R_2 + i_3 R_6 + i_5 R_5 + i_2 r_5$$

$$\text{III} \qquad E_4 = i_4 R_4 - i_5 R_5 - i_6 r_4 - i_6 R_7$$

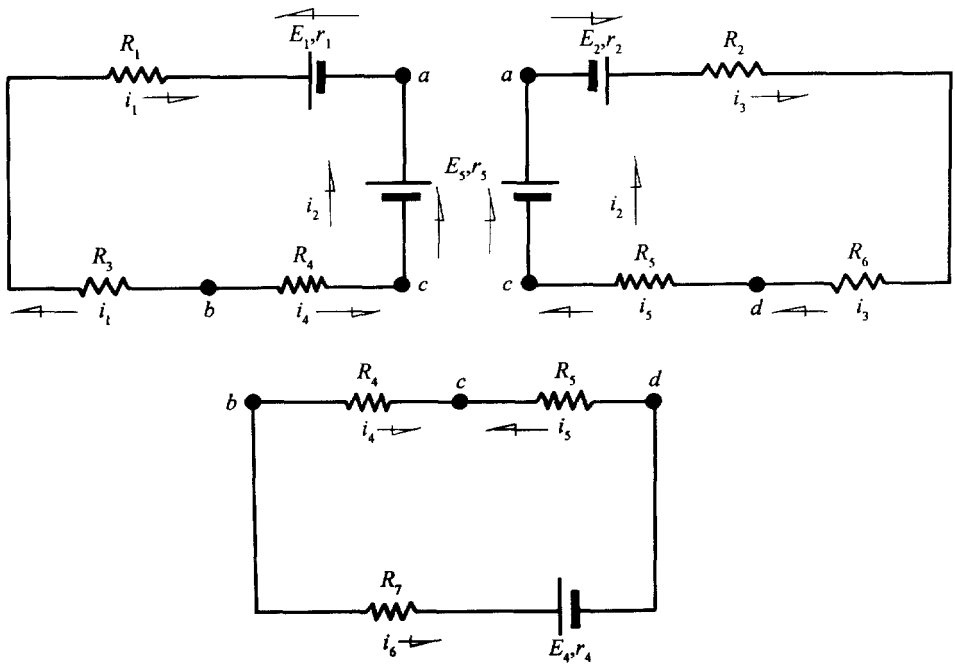


Figura 3.16. Circuito de la figura 3.15 cortado en mallas.

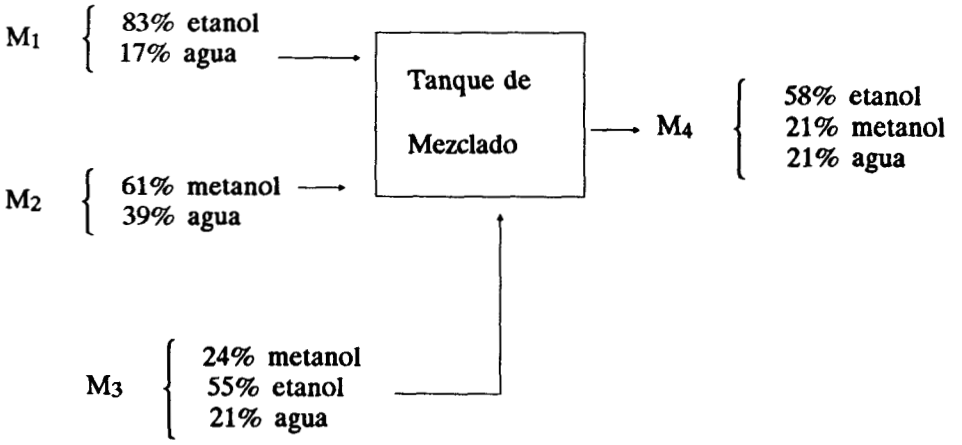
Se tienen ecuaciones independientes, donde conocidas las R_k, las E_k y las r_k, se pueden calcular las seis intensidades de corriente resolviendo el sistema. Para los siguientes datos, calcule las intensidades de corriente.

k	E_k (volts)	r_k (Ω)	R_k (Ω)
1	12	0.1	25
2	10	0.5	40
3			16
4	12	0.5	20
5	24	0.2	9
6			4
7			20

Con el programa 3.2 del disco se obtienen los siguientes valores para las intensidades de corriente

k	1	2	3	4	5	6
i_k	–0.53811	1.1934	0.6553	0.68226	0.51115	–0.14415

3.5 Con los datos del diagrama siguiente (donde los porcentajes están dados en peso), encuentre posibles valores de las corrientes M_1, M_2, M_3 y M_4.



SOLUCIÓN

Mediante balances de materia por componente y global, se tiene

Componente	Balance de materia
Etanol	$0.83\ M_1 + 0.55\ M_3 - 0.58\ M_4 = 0$
Metanol	$0.61\ M_2 + 0.24\ M_3 - 0.21\ M_4 = 0$
Agua	$0.17\ M_1 + 0.39\ M_2 + 0.21\ M_3 - 0.21\ M_4 = 0$
Global	$M_1 + M_2 + M_3 - M_4 = 0$

Obsérvese que sólo se tienen tres ecuaciones linealmente independientes, pues la ecuación del balance global de materia es la suma de las otras tres. Por ser el sistema homogéneo es consistente, y como el rango de la matriz coeficiente es menor que el número de incógnitas, el sistema tiene un número infinito de soluciones. Fijando una base de cálculo, por ejemplo $M_4 = 100$ Kg, se obtiene el sistema

$$\begin{aligned} 0.83\ M_1 + 0.55\ M_3 &= 58 \\ 0.61\ M_2 + 0.24\ M_3 &= 21 \\ 0.17\ M_1 + 0.39\ M_2 + 0.21\ M_3 &= 21 \end{aligned}$$

cuya solución se deja al lector, utilizando alguno de los programas vistos.

3.6 Un granjero desea preparar un fórmula alimenticia para engordar ganado. Dispone de maíz, desperdicios, alfalfa y cebada, cada uno con ciertas unidades de ingredientes nutritivos, de acuerdo con la tabla siguiente

Unidades de ingredientes nutritivos por kg de cada alimento disponible

Alimento / Ingrediente nutritivo	Maíz	Desperdicio	Alfalfa	Cebada	Requerimiento diario Unidades /kg
Carbohidrato	80	15	35	60	230
Proteína	28	72	57	25	180
Vitaminas	20	20	12	20	80
Celulosa	50	10	20	60	160
Costo $	18	5	7	20	

a) Determine los kilogramos necesarios de cada material para satisfacer el requerimiento diario (presentado en la última columna).
b) Determine el costo de la mezcla.

NOTA: La fórmula alimenticia debe contener los cuatro alimentos.

SOLUCIÓN

Si se llama x_1 a los kg de maíz necesarios, x_2 los de desperdicio, ... se tiene

$$\begin{aligned} 80x_1 + 15x_2 + 35x_3 + 60x_4 &= 230 \\ 28x_1 + 72x_2 + 57x_3 + 25x_4 &= 180 \\ 20x_1 + 20x_2 + 12x_3 + 20x_4 &= 80 \\ 50x_1 + 10x_2 + 20x_3 + 60x_4 &= 160 \end{aligned}$$

Con el programa 3.2 del disco se obtiene

$$x_1 = 1.8524, x_2 = 1.0318, x_3 = 0.6178, x_4 = 0.745$$

De donde el costo de la mezcla es

$$\text{Costo} = 18*1.8524 + 5*1.03 + 7*0.61 + 20*0.745 = 57.56\$$$

3.7 En un sistema monofásico en equilibrio químico existen los siguientes compuestos: CO, H_2, CH_3OH, H_2O y C_2H_6. Calcule el número de reacciones químicas independientes.

SOLUCIÓN

Se establece la matriz atómica enlistando los compuestos como cabezas de columna y los átomos como inicio de filas, de tal modo que la intersección muestre el número de átomos del compuesto correspondiente.

Compuesto / Átomo	CO	H_2	CH_3OH	H_20	C_2H_6
C	1	0	1	0	2
H	0	2	4	2	6
O	1	0	1	1	0

Si N es el número de compuestos en equilibrio químico, R el número de reacciones independientes, se tiene la siguiente relación discutida por Jouguet, Brinkey y otros*

$$R = N - C$$

donde C es el rango de la matriz atómica.

Para encontrar el rango se utilizará el método de ortogonalización de Gram-Schmidt, aplicado a las columnas de la matriz atómica. Para esto, llámense x_1, x_2, ..., x_5 las columnas CO, H_2, ..., C_2H_6.

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix},$$

$$\mathbf{e}_2 = \mathbf{x}_2 - \alpha\, \mathbf{e}_1, \text{ donde } \alpha = \frac{\mathbf{x}_2 \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1} = \frac{\begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}}{\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}} = 0$$

Por lo tanto

$$\mathbf{e}_2 = \mathbf{x}_2 = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}$$

Nótese que como $\mathbf{x}_2$ es ortogonal a $\mathbf{x}_1$, el proceso da $\mathbf{e}_2 = \mathbf{x}_2$.

$$\mathbf{e}_3 = \mathbf{x}_3 - \alpha_1\, \mathbf{e}_1 - \alpha_2\, \mathbf{e}_2, \text{ donde } \alpha_1 = \frac{\mathbf{x}_3 \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1} = \frac{\begin{bmatrix} 1 \\ 4 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}}{\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}} = 1$$

y

$$\alpha_2 = \frac{\mathbf{x}_3 \cdot \mathbf{e}_2}{\mathbf{e}_2 \cdot \mathbf{e}_2} = \frac{\begin{bmatrix} 1 \\ 4 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}}{\begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}} = 2$$

*Jouguet, J. Ec. Polyt. Paris, 2, 62 (1921) Prigogine and Defay. J. Chem. Phys 15, 614 (1947)

Por lo tanto

$$\mathbf{e}_3 = \begin{bmatrix}1\\4\\1\end{bmatrix} - (1)\begin{bmatrix}1\\0\\1\end{bmatrix} - (2)\begin{bmatrix}0\\2\\0\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix}$$

Esto indica que $\mathbf{x}_3$ es linealmente dependiente de $\mathbf{x}_1$ y $\mathbf{x}_2$. Continuando el proceso de ortogonalización, pero sin tomar en cuenta a $\mathbf{e}_3$, se tiene

$$\mathbf{e}_4 = \mathbf{x}_4 - \alpha_1 \mathbf{e}_1 - \alpha_2 \mathbf{e}_2 \text{ donde } \alpha_1 = \frac{\mathbf{x}_4 \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1} = \frac{\begin{bmatrix}0\\2\\1\end{bmatrix} \cdot \begin{bmatrix}1\\0\\1\end{bmatrix}}{\begin{bmatrix}1\\0\\1\end{bmatrix} \cdot \begin{bmatrix}1\\0\\1\end{bmatrix}} = \frac{1}{2}$$

$$\text{y } \alpha_2 = \frac{\mathbf{x}_4 \cdot \mathbf{e}_2}{\mathbf{e}_2 \cdot \mathbf{e}_2} = \frac{\begin{bmatrix}0\\2\\1\end{bmatrix} \cdot \begin{bmatrix}0\\2\\0\end{bmatrix}}{\begin{bmatrix}0\\2\\0\end{bmatrix} \cdot \begin{bmatrix}0\\2\\0\end{bmatrix}} = 1$$

Por lo tanto

$$\mathbf{e}_4 = \begin{bmatrix}0\\2\\1\end{bmatrix} - \frac{1}{2}\begin{bmatrix}1\\0\\1\end{bmatrix} - (1)\begin{bmatrix}0\\2\\0\end{bmatrix} = \begin{bmatrix}-1/2\\0\\1/2\end{bmatrix}$$

Como el número de filas de la matriz atómica es 3, el máximo número de vectores linealmente independientes es 3 y como ya se ha encontrado que $\mathbf{x}_1$, $\mathbf{x}_2$ y $\mathbf{x}_4$ son linelamente independientes, $\mathbf{x}_5$ es necesariamente dependiente de $\mathbf{x}_1$, $\mathbf{x}_2$, y $\mathbf{x}_4$ y $\mathbf{e}_5$ debe ser el vector cero (demostración que se deja al lector como ejercicio); entonces, el rango de la matriz atómica es 3.

Al aplicar la fórmula $\qquad R = N - C = 5 - 3 = 2$

se tiene que el número de reacciones independientes para llegar al sistema en equilibrio químico mencionado es 2.

3.8 Si A es una matriz de números reales de orden n y I la matriz identidad de orden n, el polinomio definido por

$$p(\lambda) = \det(A - \lambda I) \qquad (1)$$

se llama el **polinomio característico de** A.

Es fácil ver* que p es un polinomio de n-ésimo grado en λ con coeficientes reales y que, por lo tanto, la ecuación

$$p(\lambda) = 0 \tag{2}$$

tiene n raíces, de las cuales algunas suelen ser complejas. Los ceros de esta ecuación, conocidos **como valores característicos o propios** de A, están íntimamente ligados con la solución del sistema $A\,\mathbf{x} = \mathbf{b}$. Por ejemplo, el método de Gauss-Seidel, independientemente del vector inicial que se emplee, converge a la solución de $A\,\mathbf{x} = \mathbf{b}$ si y sólo si los valores propios de B (véase Ec. 3.89) son todos menores de uno en valor absoluto.

Dada la siguiente matriz, encontre sus valores propios

$$A = \begin{bmatrix} 4 & -9 & 2 \\ 2 & -4 & 6 \\ 1 & -1 & 3 \end{bmatrix}$$

SOLUCIÓN

Se forma $A - \lambda I$

$$A - \lambda I = \begin{bmatrix} 4 & -9 & 2 \\ 2 & -4 & 6 \\ 1 & -1 & 3 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 4-\lambda & -9 & 2 \\ 2 & -4-\lambda & 6 \\ 1 & -1 & 3-\lambda \end{bmatrix}$$

Se obtiene el determinante de este último arreglo

$$\begin{aligned} \det (A - \lambda I) = {} & (4 - \lambda)(-4 - \lambda)(3 - \lambda) - 4 - 54 - \\ & (2)(-4 - \lambda)(1) - (-9)(3 - \lambda)(2) - \\ & (6)(-1)(4 - \lambda) \end{aligned}$$

Al desarrollar e igualar con cero se obtiene

$$-\lambda^3 + 3\lambda^2 - 6\lambda - 20 = 0,$$

el polinomio característico de A, cuyos ceros $\lambda_1, \lambda_2, \lambda_3$ son los valores buscados.

El hecho de ser un polinomio cúbico con coeficientes reales garantiza una raíz real por lo menos. Con el método de Newton-Raphson y un valor inicial de –2 se llega a

$$\lambda_1 = -1.53968$$

*Véase Probl. 3.59

El polinomio, se degrada por división sintética

	−1	3	−6	−20
−1.53968		1.53968	−6.98965	20
	−1	4.53968	−12.98965	0

El polinomio degradado es

$$-\lambda^2 + 4.53968\,\lambda - 12.98965 = 0,$$

de donde, por aplicación de la fórmula cuadrática se tiene

$$\lambda_2 = 2.26984 + 2.795975\,i$$

$$\lambda_3 = 2.26984 - 2.795975\,i$$

3.9 Una vez obtenidos los valores propios de una matriz A de orden n (véase Ej. 3.8), los vectores $\mathbf{x} \neq \mathbf{0}$ que resuelven el sistema

$$\begin{aligned} &A\,\mathbf{x} = \lambda_i\,\mathbf{x}, \qquad i = 1, 2, \ldots, n \\ &(A - \lambda_i I)\,\mathbf{x} = \mathbf{0} \end{aligned} \tag{1}$$

se denominan **vectores propios** de A correspondientes a λ_i. Como $\det(A - \lambda_i I) = 0$ y el sistema es homogéneo, se tiene un número infinito de soluciones para cada λ_i.

Encuentre los vectores propios de la matriz del problema anterior, correspondientes al valor propio $\lambda_1 = -1.53968$.

SOLUCIÓN

Al resolver el sistema por alguno de los métodos de eliminación

$$(A - \lambda_1 I)\,\mathbf{x} = \begin{bmatrix} 4-(-1.53968) & -9 & 2 \\ 2 & -4-(-1.53968) & 6 \\ 1 & -1 & 3-(-1.53968) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

resulta una matriz triangular superior, por lo menos con una fila de ceros*. Para asegurar que esa(s) fila(s) de ceros sea(n) la(s) última(s) y que la submatriz no singular resultante esté lo mejor condicionada posible, se usa **pivoteo total** (intercambio de filas y columnas) y escalamiento.

*Pizer, M.S. *Numerical Computing and Mathematical Analysis.* S.R.A. (1975)

Sea entonces la matriz por triangularizar

$$\begin{bmatrix} 5.53968 & -9 & 2 \\ 2 & -2.46032 & 6 \\ 1 & -1 & 4.53968 \end{bmatrix} = B$$

Nótese que el vector de términos independientes no se emplea porque todos sus componentes son cero.

En lugar de emplear la **norma euclideana** para el escalamiento, se usará ahora la siguiente norma, definida para un vector cualquiera $\mathbf{y} = [y_1, y_2 \ldots y_n]^T$, como

$$\mathbf{y} = |y_1| + |y_2| + \ldots + |y_n|$$

ya que es más sencilla de calcular que la euclideana y que para la primera, segunda y tercera filas de A es, respectivamente,

$$\begin{bmatrix} 16.53968 \\ 10.46032 \\ 6.53968 \end{bmatrix}$$

Cada fila de la matriz B se divide entre su factor de escalamiento y se obtiene

$$B' = \begin{bmatrix} 0.33493 & -0.54415 & 0.12092 \\ 0.19120 & -0.23520 & 0.57360 \\ 0.15291 & -0.15291 & 0.69417 \end{bmatrix}$$

En el pivoteo total es necesario registrar los cambios de columnas que se verifican, ya que éstos afectan el orden de las incógnitas. Para ello se utilizará un vector **q**, en donde aparecen como elementos las columnas. Al principio están en orden natural y se tiene

$$\mathbf{q} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

Se busca el elemento de máximo valor absoluto de B'. En este caso es $b'_{3,3} = 0.69417$. Se intercambian las filas 1 y 3 y las columnas 1 y 3 para llevar este elemento a la posición pivote (1,1), teniendo cuidado de registrar los intercambios de columnas en **q**. Los resultados son

$$B'' = \begin{bmatrix} 0.69417 & -0.15291 & 0.15291 \\ 0.57360 & -0.23520 & 0.19120 \\ 0.12092 & -0.54415 & 0.33493 \end{bmatrix} \qquad \mathbf{q} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

Se eliminan los elementos de la primera columna que están debajo del elemento pivote, con lo cual se produce

$$B''' = \begin{bmatrix} 0.69417 & -0.15291 & 0.15291 \\ 0.0 & -0.10885 & 0.06485 \\ 0.0 & -0.51751 & 0.30830 \end{bmatrix}$$

Se busca ahora el elemento de máximo valor absoluto en las dos últimas filas; resulta ser $b'''_{3,2} = -0.51751$. Se intercambian las filas 2 y 3 y con esto se lleva este elemento a la posición pivote (2,2). Los resultados son

$$B^{IV} = \begin{bmatrix} 0.69417 & -0.15291 & 0.15291 \\ 0.0 & -0.51751 & 0.30830 \\ 0.0 & -0.10885 & 0.06485 \end{bmatrix}$$

y $\mathbf{q} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$, ya que no hubo intercambio de columnas.

Se eliminan los elementos de la segunda columna que están debajo del elemento pivote y se produce

$$B^{V} = \begin{bmatrix} 0.69417 & -0.15291 & 0.1529100000 \\ 0.00000 & -0.51751 & 0.3083000000 \\ 0.00000 & 0.00000 & -0.000000218 \end{bmatrix}$$

una matriz triangularizada con una fila de ceros, la última como se planeó. La submatriz no singular de la que se habló al principio está formada por los elementos (1,1), (1,2), (2,1) y (2,2).

Al escribir el sistema en términos de x_1, x_2 y x_3, y considerar los cambios de columnas que hubo, se tiene

$$\begin{aligned} 0.69417\, x_3 - 0.15291\, x_2 + 0.15291\, x_1 &= 0 \\ 0.0 \quad x_3 - 0.51751\, x_2 + 0.30830\, x_1 &= 0 \end{aligned}$$

Un sistema homogéneo de dos ecuaciones en tres incógnitas, cuyas infinitas soluciones pueden obtenerse en términos de alguna de las incógnitas. El sistema se resuelve en términos de x_1

$$\begin{aligned} 0.69417\, x_3 - 0.15291\, x_2 &= -0.15291\, x_1 \\ 0.0 \quad x_3 - 0.51751\, x_2 &= -0.30830\, x_1 \end{aligned}$$

de donde

$$\begin{aligned} x_2 &= 0.59573\, x_1 \\ x_3 &= -0.08905\, x_1 \end{aligned}$$

Se da un valor particular a x_1, por ejemplo $x_1 = 1$, y resulta

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.59753 \\ -0.08905 \end{bmatrix},$$

uno de los infinitos vectores propios de A correspondientes a λ_1.

Comprobación

Ya que por definición $A\,\mathbf{x} = \lambda_1\,\mathbf{x}$

$$\begin{bmatrix} 4 & -9 & 2 \\ 2 & -4 & 6 \\ 1 & -1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 0.59573 \\ -0.08905 \end{bmatrix} = -1.53968 \begin{bmatrix} 1 \\ 0.59573 \\ -0.08905 \end{bmatrix}$$

Problemas

3.1 Elabore un algoritmo general para sumar y restar matrices.

3.2 Con el algoritmo del problema anterior, elabore uno de propósito general para sumar y restar matrices.

3.3 Demuestre, partiendo de la definición del producto de una matriz por un escalar, las ecuaciones 3.7, 3.8 y 3.10.

3.4 Demuestre la ecuación 3.12, utilizando la definición de multiplicación de matrices.

3.5 Con el programa 3.1 del disco multiplique las siguientes matrices

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 3 & 4 & 5 \\ 7 & 8 & 9 \\ 11 & 12 & 13 \\ 15 & 16 & 17 \end{bmatrix},$$

$$\begin{bmatrix} 2 & 3 & 4 & 5 \\ 0 & 6 & 7 & 8 \\ 0 & 0 & 5 & 3 \\ 0 & 0 & 0 & 4 \end{bmatrix} \qquad \begin{bmatrix} 4 & 3 & 0 & 1 \\ 5 & 0.2 & -1 & 8 \\ 3 & 4 & 5 & 7 \\ 10 & 9 & 3 & -2 \end{bmatrix},$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix},$$

$$[\,0\ 1\ 2\ 3\ 4\,] \begin{bmatrix} 3 \\ 8 \\ -2 \\ 5 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} [\,3\ 8\ -2\ 5\ 1\,]$$

$$\begin{bmatrix} 3 & 4 & 5 \\ 7 & 8 & 9 \\ 11 & 12 & 13 \\ 15 & 16 & 17 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

3.6 La siguiente tabla representa las existencias en bodega de una agencia de refacciones para automóviles

Marca / Refacción	M1	M2	M3	M4	M5	M6
R1	5	13	23	8	15	98
R2	16	45	11	54	10	86
R3	34	22	77	21	65	2
R4	21	19	83	2	16	37
R5	8	97	69	27	14	3

En la siguiente tabla se dan los precios unitarios correspondientes a las refacciones de arriba.

Marca / Refacción	M1	M2	M3	M4	M5	M6
R1	65000	73450	82500	71245	62350	76450
R2	3400	3560	2560	5790	4700	5000
R3	12500	13450	16400	15600	11650	9500
R4	895	940	780	950	645	1000
R5	5350	7620	6700	3250	5890	7000

Determine la inversión en bodega de la agencia.

3.7 Responda las siguientes preguntas.

a) ¿Una matriz no cuadrada puede ser simétrica?
b) ¿Una matriz diagonal es triangular superior, triangular inferior o ambas?
c) ¿Una matriz diagonal tiene inversa con uno de sus elementos de la diagonal principal igual a cero?

3.8 Multiplique una matriz permutadora (seleccione una cualquiera) por sí misma y observe el resultado. Generalice dicho resultado.

3.9 Demuestre las ecuaciones 3.15, 3.16 y 3.17.

3.10 Obtenga la ecuación 3.21 a partir de la ley de los cosenos.

3.11 Elabore un algoritmo tal que, dados dos vectores de igual número de componentes, se determine e imprima la norma euclideana de estos vectores, su producto punto, el ángulo que guardan entre ellos y la distancia que hay entre ambos.

3.12 Codifique el algoritmo del problema 3.11 y verifique este programa con las siguientes parejas de vectores

a) $\begin{bmatrix}1\\0\\0\\0\end{bmatrix}$ $\begin{bmatrix}0\\1\\0\\0\end{bmatrix}$ *b*) $\begin{bmatrix}0\\0\\0\\0\end{bmatrix}$ $\begin{bmatrix}1\\3\\5\\-2\end{bmatrix}$ *c*) $\begin{bmatrix}5\\-2\\8\\0.1\end{bmatrix}$ $\begin{bmatrix}1.5\\-0.6\\2.4\\0.03\end{bmatrix}$ *d*) $\begin{bmatrix}0\\1\end{bmatrix}$ $\begin{bmatrix}0\\-1\end{bmatrix}$

3.13 El teorema 3.1 puede y debe emplearse también para ortogonalizar un conjunto de m vectores linealmente independientes de n componentes cada uno, con $m < n$. Por otro lado, demuestre con el teorema mencionado, que cualquier conjunto de n vectores linealmente independientes con n componentes cada uno da como resultado un conjunto linealmente dependiente al adicionársele un vector $\mathbf{x}_{n+1}$ de n componentes.

NOTA: Use como motivación algunos casos particulares sencillos: por ejemplo, a un conjunto particular de dos vectores linealmente independientes con dos componentes cada uno, añada un tercer vector y aplique la ortogonalización al conjunto resultante.

3.14 Elabore una subrutina de propósito general para ortogonalizar un conjunto de m vectores linealmente independientes de n componentes cada uno ($m < n$) con el método de Gram-Schmidt.

NOTA: Puede usar el algoritmo 3.2 como base.

3.15 Con la subrutina del problema 3.14 ortogonalice los siguientes conjuntos de vectores.

a)
$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ -2 \\ 5 \\ 7 \\ 8 \\ 0.3 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 1 \\ -2 \\ 5 \\ 7 \\ 8 \\ 0.3 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 3 \\ 0.8 \\ 4 \\ 15 \\ 3 \\ 2 \end{bmatrix}, \mathbf{x}_4 = \begin{bmatrix} 7 \\ -3 \\ 5 \\ 3.2 \\ 9 \\ 40 \end{bmatrix}, \mathbf{x}_5 = \begin{bmatrix} 5 \\ 4 \\ 3 \\ 1 \\ 7 \\ 8 \end{bmatrix},$$

b)
$$\mathbf{x}_1 = \begin{bmatrix} 4 \\ 2 \\ 1 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} -9 \\ -4 \\ -1 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 2 \\ 6 \\ 3 \end{bmatrix}$$

c)
$$\mathbf{x}_1 = \begin{bmatrix} 10 \\ -20 \\ 5 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 1 \\ 3 \\ 3 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} -5 \\ 20 \\ 5 \end{bmatrix}$$

d)
$$\mathbf{x}_1 = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 2 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 3 \\ 9 \\ 1 \\ 1 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 1 \\ 3 \\ 5 \\ -2 \end{bmatrix}, \mathbf{x}_4 = \begin{bmatrix} 2 \\ 4 \\ 1 \\ 1 \end{bmatrix}$$

3.16 Modifique el programa del problema 3.14 de modo que
a) Dado un conjunto cualquiera de m vectores de n componentes cada uno ($m<n$), se vayan ortogonalizando los linealmente independientes y se descarte los que resulten linealmente dependientes.
b) Imprima el número de vectores linealmente independientes del conjunto denotando este número como **rango** del conjunto.

Corra el programa para determinar el rango de las siguientes matrices o conjuntos de vectores columna

$$\begin{bmatrix} 4 & -9 & 2 \\ 2 & -4 & 6 \\ 1 & -1 & 3 \end{bmatrix}, \begin{bmatrix} 10 & 1 & -5 \\ -20 & 3 & 20 \\ 5 & 3 & 5 \end{bmatrix}, \begin{bmatrix} 10 & 1 & -5 & 1 \\ -20 & 3 & 20 & 2 \\ 5 & 3 & 5 & 6 \end{bmatrix}, \begin{bmatrix} 4 & -1 & 0 & 0 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & 4 \end{bmatrix}$$

3.17 Calcule el número de reacciones independientes en una reacción de pirólisis, en la cual se encuentran en equilibrio los siguientes compuestos O_2, H_2, CO, CO_2, H_2CO_3, CH_3OH, C_2H_5OH, $(CH_3)_2$ CO, CH_4, CH_3CHO y H_2O.

3.18 Dada una matriz A de orden n, los términos

a) Matriz singular (det $A = 0$)
b) Rango $A < n$
c) Los vectores columna o fila de A son linealmente dependientes

están estrechamente relacionados.
Demuestre que (a) implica tanto (b) como (c).

3.19 ¿La concidencia del número de incógnitas con el número de ecuaciones en un sistema de ecuaciones lineales implica que éste tiene solución única? Justifique su respuesta.

3.20 Dado el siguiente sistema de ecuaciones, encuentre dos valores de w que permitan tener solución única y diga qué valores de w permiten un número infinito de soluciones

$$\begin{bmatrix} w & 1 & 5 \\ 4 & 2 & -w \\ 0 & 3 & -7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 5 \\ 3 \end{bmatrix}$$

3.21 Si la matriz coeficiente del sistema $A\,\mathbf{x} = \mathbf{0}$ es tal que det $A = 0$; ¿dicho sistema tiene por ese hecho un número infinito de soluciones?

3.22 El método de eliminación de Gauss usualmente hace la transformación conocida como triangularización.

$$\left[\begin{array}{ccc|c} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} \end{array}\right] \qquad \left[\begin{array}{ccc|c} a'_{1,1} & a_{1,2,} & a'_{1,3} & a'_{1,4} \\ 0 & a'_{2,2} & a'_{2,3} & a'_{2,4} \\ 0 & 0 & a'_{3,3} & a'_{3,4} \end{array}\right]$$

En estas condiciones, una sustitución hacia atrás permite obtener la solución. Las ecuaciones 3.49 y 3.50 constituyen el algoritmo para el caso general.
Encuentre las ecuaciones correspondientes para resolver el sistema $A\,\mathbf{x} = \mathbf{b}$, pero ahora llevando a cabo la transformación

$$\left[\begin{array}{ccc|c} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} \end{array}\right] \qquad \left[\begin{array}{ccc|c} a'_{1,1} & 0 & 0 & a'_{1,4} \\ a'_{2,1} & a'_{2,2} & 0, & a'_{2,4} \\ a'_{3,1} & a_{3,2} & a_{3,3} & a'_{3,4} \end{array}\right]$$

y posteriormente una sustitución hacia adelante.

3.23 Modifique el algoritmo 3.4, de modo que una vez encontrado el elemento pivote e intercambiadas las filas (si procede), se divida la fila pivote entre el elemento pivote. En el caso de un sistema de orden 3, el resultado en la triangularización sería

$$\left[\begin{array}{ccc|c} 1 & a_{1,2} & a_{1,3} & b_1 \\ 0 & 1 & a_{2,3} & b_2 \\ 0 & 0 & 1 & b_3 \end{array}\right],$$

y, por tanto, en la sustitución regresiva no se tendría que dividir entre los coeficientes de las incógnitas.

Por otro lado, para el cálculo del determinante deben guardarse los pivotes para su empleo en la expresión

$$\det A = (-1)^{\gamma} \prod_{i=1}^{n} a_{i,i}$$

Sugerencia: Corra el programa 3.1 del apéndice.

3.24 Elabore un algoritmo para resolver un sistema de ecuaciones $A\,\mathbf{x} = \mathbf{b}$ usando la eliminación de Jordan.

3.25 Calcule el número de multiplicaciones, divisiones o ambas y la cantidad de sumas, restas o ambas que se requieren para resolver un sistema tridiagonal por el método de Thomas. Determine también las necesidades de memoria para este algoritmo.

3.26 Utilice el subprograma que se da en el ejercicio 8.2 del capítulo 8 para resolver los siguientes sistemas

a)
$$\begin{aligned} 0.5x_1 + 0.25x_2 &= 0.32 \\ 0.3x_1 + 0.8x_2 + 0.4x_3 &= 0.77 \\ + 0.2x_2 + x_3 + 0.6x_4 &= -0.6 \\ x_3 - 3x_4 &= -2 \end{aligned}$$

b)
$$\begin{aligned} x_1 - x_2 &= 1 \\ 2x_1 - x_2 + 3x_3 &= 8 \\ x_2 + x_3 &= 4 \end{aligned}$$

c)
$$\begin{aligned} 4x_1 + x_2 &= -1 \\ -8x_1 - x_2 + x_3 &= 13 \\ 3x_2 - 2x_3 + 4x_4 &= -3 \\ x_3 - x_4 + x_5 &= 2.1 \\ 2x_4 + 6x_5 &= 3.4 \end{aligned}$$

3.27 Una matriz tridiagonal por bloques (o partida) es una matriz de la forma

$$A = \begin{bmatrix} B_1 & C_1 & O & & & & O \\ A_2 & B_2 & C_2 & O & & & \\ O & A_3 & B_3 & C_3 & O & & O \\ O & O & & & & & \\ & & & & & & C_{n-1} \\ O & & & & O & A_n & B_n \end{bmatrix}$$

donde $B_1, B_2, \ldots, B_n$ son matrices de orden $n_1, n_2, \ldots, n_n$ respectivamente. $A_2, A_3, \ldots, A_n$ son matrices de orden $(n_2 \times n_1), (n_3 \times n_2), \ldots, (n_n \times n_{n-1})$ respectivamente y $C_1, C_2, \ldots, C_{n-1}$ son matrices de orden $(n_1 \times n_2), (n_2 \times n_3), \ldots, (n_{n-1} \times n_n)$, respectivamente.

Por ejemplo, las matrices

a)
$$A = \begin{bmatrix} B_1 & C_1 & O \\ A_2 & B_2 & C_2 \\ O & A_3 & B_3 \end{bmatrix} \quad \text{donde } B_i = \begin{bmatrix} 6 & -1 & 0 \\ -1 & 6 & -1 \\ 0 & -1 & 6 \end{bmatrix} \quad i = 1, 2, 3$$

y $A_{i+1} = C_i = \begin{bmatrix} -2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -2 \end{bmatrix} \quad i = 1, 2$

b)
$$\begin{bmatrix}
1 & 5 & 3 & 5 & 8 & 9 & -2 & 0 & 0 & 0 & 0 & 0 \\
2 & -1 & 0 & 1 & 4 & 0 & 7 & 0 & 0 & 0 & 0 & 0 \\
4 & 3 & 6 & 7 & 3 & 2 & 3 & 0 & 0 & 0 & 0 & 0 \\
7 & 3 & 6 & 4 & 5 & 8 & 9 & 4 & 5 & 5 & 4 & 3 \\
2 & 2 & 5 & 7 & 6 & 3 & 2 & 2 & 7 & 8 & 9 & 1 \\
3 & 7 & 3 & 4 & 1 & 0 & 1 & 0 & -3 & 5 & 7 & 2 \\
1 & 1 & 2 & 4 & 3 & 2 & 5 & 4 & 5 & 7 & 9 & 5 \\
0 & 0 & 0 & 5 & 7 & 9 & 5 & 0 & 5 & 7 & 4 & 2 \\
0 & 0 & 0 & 4 & 8 & 2 & 2 & -1 & 7 & 9 & 7 & 8 \\
0 & 0 & 0 & 3 & 2 & 1 & 1 & 4 & 8 & 4 & 3 & 2 \\
0 & 0 & 0 & 5 & 1 & 5 & 4 & 2 & 7 & 4 & 5 & -1 \\
0 & 0 & 0 & 2 & 9 & 7 & 3 & 3 & 2 & 7 & 2 & 2
\end{bmatrix}$$

son tridiagonales por bloques.

Observe que una matriz tridiagonal por bloques no es tridiagonal en el sentido de la definición original.

Elabore un algoritmo similar al algoritmo 3.5 para resolver sistema tridiagonales por bloques $A\,\mathbf{x} = \mathbf{b}$.

Sugerencia: Para el sistema

$$\begin{bmatrix} B_1 & C_1 & O \\ A_2 & B_2 & C_2 \\ O & A_3 & B_3 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \mathbf{b}_3 \end{bmatrix}$$

donde se ha segmentado a $\mathbf{x}$ y $\mathbf{b}$ de modo tal que

$\mathbf{x}_1$ y $\mathbf{b}_1$ son vectores de n_1 componentes (el orden de B_1),

$\mathbf{x}_2$ y $\mathbf{b}_2$ son vectores de n_2 componentes (el orden de B_2),

$\mathbf{x}_3$ y $\mathbf{b}_3$ son vectores de n_3 componentes (el orden de B_3),

forme la matriz aumentada

$$\begin{bmatrix} B_1 & C_1 & O & \mathbf{b}_1 \\ A_2 & B_2 & C_2 & \mathbf{b}_2 \\ O & A_3 & B_3 & \mathbf{b}_3 \end{bmatrix},$$

y elimine la matriz A_2 por medio de los elementos de la diagonal principal de B_1; posteriormente elimine la matriz A_3 con los elementos diagonales de B_2. Para iniciar la sustitución regresiva, resuelva el sistema

$$B_3\,\mathbf{x}_3 = \mathbf{b}_3,$$

con el resultado resuelva el sistema

$$B_2\,\mathbf{x}_2 = \mathbf{b}_2 - C_2\,\mathbf{x}_3$$

Finalmente, sustituyendo $\mathbf{x}_2$, resuelva

$$B_1\,\mathbf{x}_1 = \mathbf{b}_1 - C_1\,\mathbf{x}_2$$

Los sistemas pueden resolverse con alguno de los métodos vistos.

3.28 Resuelva el sistema tridiagonal por bloques

$$\begin{bmatrix} 6 & -1 & 0 & -2 & 0 & 0 & 0 & 0 & 0 \\ -1 & 6 & -1 & 0 & -2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 6 & 0 & 0 & -2 & 0 & 0 & 0 \\ -2 & 0 & 0 & 6 & -1 & 0 & -2 & 0 & 0 \\ 0 & -2 & 0 & -1 & 6 & -1 & 0 & -2 & 0 \\ 0 & 0 & -2 & 0 & -1 & 6 & 0 & 0 & -2 \\ 0 & 0 & 0 & -2 & 0 & 0 & 6 & -1 & 0 \\ 0 & 0 & 0 & 0 & -2 & 0 & 1 & 6 & -1 \\ 0 & 0 & 0 & 0 & 0 & -2 & 0 & -1 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 3 \\ 1 \\ 0 \\ 1 \\ 3 \\ 2 \\ 3 \end{bmatrix}$$

Utilice la sugerencia del problema 3.27, el algoritmo de ese ejercicio o ambos.

3.29 En la simulación de una columna de destilación de NP platos que separa una mezcla de NC componentes, el balance de materia por componente en cada plato, el balance de entalpía en cada plato y la relación de equilibrio líquido-vapor de cada componente en cada plato, resultan en un sistema de NP(2NC+1) (dos veces el número de componentes más uno multiplicado por el número de platos) ecuaciones algebraicas no lineales. En la aplicación del método de Newton-Raphson para un sistema no lineal es necesario resolver un sistema tridiagonal por bloques de orden NP(2NC+1) en cada iteración. Para una columna de cinco platos y tres componentes, el sistema tridiagonal por bloques por resolver en cada iteración es

$$\begin{bmatrix} B_1 & C_1 & & & \\ A_2 & B_2 & C_2 & & \\ & A_3 & B_3 & C_3 & \\ & & A_4 & B_4 & C_4 \\ & & & A_5 & B_5 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \mathbf{x}_4 \\ \mathbf{x}_5 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \mathbf{b}_3 \\ \mathbf{b}_4 \\ \mathbf{b}_5 \end{bmatrix}$$

donde

$$A_2 = \begin{bmatrix} 0 & 0 & 0 & 12204.1 & 9216.1 & 1262.6 & 1768.8 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} 0 & 0 & 0 & 10550.0 & 1187.3 & 1636.7 & 2291.6 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$A_4 = \begin{bmatrix} 0 & 0 & 0 & 9863.2 & 1529.5 & 2109.1 & 2951.1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$A_5 = \begin{bmatrix} 0 & 0 & 0 & 8341.5 & 2227.6 & 3073.1 & 4294.5 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B_1 = \begin{bmatrix} 5522..3 & 6518.9 & 7105.4 & 18015.3 & 916.1 & 1262.6 & 1768.8 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ -0.065 & 0.9346 & 0.9346 & 2.93858 & 0.119 & -0.4894 & -0.4894 \\ 0.0643 & -0.9357 & 0.0643 & 0.30762 & -0.033 & 0.1481 & -0.0337 \\ 0.0011 & 0.0011 & -0.9989 & 0.00643 & -0.006 & -0.0006 & 0.0524 \end{bmatrix}$$

$$B_2 = \begin{bmatrix} 5777.5 & 6941.9 & 7659.2 & 28231.1 & 1187.3 & 1636.7 & 2291.6 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ -0.137 & 0.8625 & 0.8625 & 7.2004 & 0.8898 & -1.6296 & -1.6296 \\ 0.1314 & -0.8686 & 0.1314 & 1.6619 & 0.5605 & 0.5605 & -0.2482 \\ 0.0061 & 0.0061 & -0.9989 & 0.0979 & -0.0115 & -0.0115 & 02374 \end{bmatrix}$$

$$B_3 = \begin{bmatrix} 6099.7 & 7471.9 & 8357.4 & 27837.5 & 1529.5 & 2109.1 & 2951.1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ -0.2217 & 0.7783 & 0.7783 & 5.6209 & 1.6619 & -1.6521 & -1.6296 \\ 0.1917 & -0.8029 & 0.1971 & 2.1270 & -0.4184 & 0.7336 & -0.2482 \\ 0.0246 & 0.0246 & -0.9754 & 0.3359 & -0.0522 & -0.0522 & 03355 \end{bmatrix}$$

$$B_4 = \begin{bmatrix} 6557.3 & 8540.2 & 9778.8 & 18947.5 & 2227.6 & 3073.1 & 4294.5 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ -0.4351 & 0.5649 & 0.5649 & 1.7373 & 2.2062 & -0.8346 & -0.8346 \\ 0.3456 & -0.6544 & 0.3456 & 1.5331 & -0.5106 & 0.6927 & -0.5106 \\ 0.8923 & 0.8923 & -0.9108 & 0.5003 & -0.1322 & -0.1322 & 0.3058 \end{bmatrix}$$

$$B_5 = \begin{bmatrix} 7547.5 & 9801.1 & 11480.6 & 12961.3 & 3065.4 & 4231.4 & 5904.4 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ -0.6827 & 0.3173 & 0.3173 & 0.7585 & 29.335 & -3.9259 & -3.9059 \\ 0.4311 & -0.5689 & 0.4311 & 1.4233 & -5.3074 & 9.3998 & -5.3074 \\ 0.2516 & 0.2516 & -0.7484 & 1.0573 & -3.0980 & -3.0980 & 2.8411 \end{bmatrix}$$

$$C_1 = \begin{bmatrix} -5777.5 & -6941.9 & -7659.2 & -17681.1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

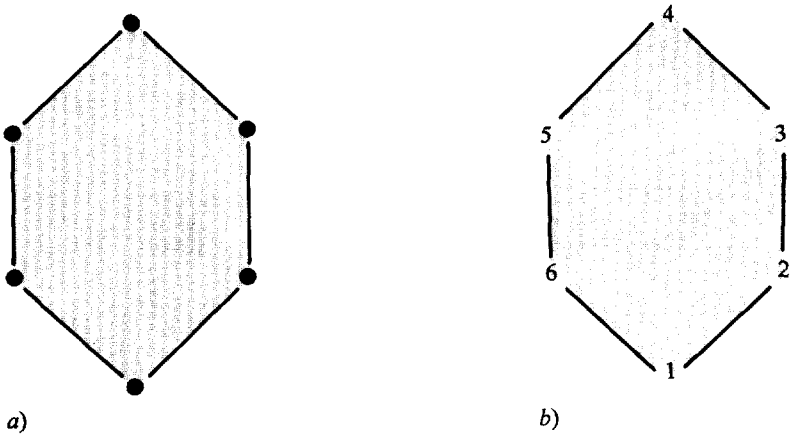
$$C_2 = \begin{bmatrix} -6099.6 & -7471.9 & -8357.4 & -17974.2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$C_3 = \begin{bmatrix} -6757.4 & -8540.2 & -9778.7 & -10606.0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$C_4 = \begin{bmatrix} -7547.5 & -9801.1 & -11480.6 & -11886.0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{b}_1 = \begin{bmatrix} -3390307.1 \\ -6.7419 \\ 13.4936 \\ 1.2278 \\ -0.009835 \\ 0.000629 \\ 0.000566 \end{bmatrix}, \mathbf{b}_2 = \begin{bmatrix} -117198.1 \\ -70.6904 \\ -16.8926 \\ 0.1614 \\ 0.0142 \\ 0.00463 \\ -0.00256 \end{bmatrix}, \mathbf{b}_3 = \begin{bmatrix} -117288.9 \\ -16.5304 \\ -18.9351 \\ -2.1684 \\ 0.02723 \\ 0.00536 \\ -0.0020 \end{bmatrix}, \mathbf{b}_4 = \begin{bmatrix} -421289.9 \\ -3.9928 \\ -52.9235 \\ 2.2335 \\ -0.0021 \\ -0.00192 \\ 0.12736 \end{bmatrix}, \mathbf{b}_5 = \begin{bmatrix} 348305.0 \\ 59.4815 \\ 35.448 \\ 3.1614 \\ -0.01917 \\ -0.01459 \\ -0.00998 \end{bmatrix}$$

3.30 Adapte la eliminación de Gauss a la solución del sistema pentadiagonal $A\ \mathbf{x} = \mathbf{b}$ (A es una matriz pentadiagonal) y obtenga las ecuaciones correspondientes a esta adaptación.



a) *b)*

Figura 3.17

3.31 Demuestre que la numeración de los nodos de la figura 3.14 con las consideraciones de que $a_{i,j} \neq 0$ siempre que los nodos sean vecinos, genera una matriz tridiagonal.

3.32 Considere la estructura hexagonal de la figura 3.17a (véase Ej. 3.2). Numere los nodos en la forma mostrada en la figura 3.17b, por ejemplo, y con consideraciones físicas que determinan que $a_{i,j} \neq 0$, cuando i y j son nodos vecinos, determine la matriz A representativa de dicha estructura.

3.33 Se tiene un sistema de tres reactores continuos tipo tanque perfectamente agitado trabajando en serie, en donde se lleva a cabo la reacción $A \longrightarrow$ *Productos* y se opera isotérmicamente (véase Fig. 3.18). Los volúmenes se mantienen constantes y son de 100, 50 y 50 litros, respectivamente.

Un balance de materia en cada reactor, de acuerdo con la ecuación de continuidad, conduce al siguiente sistema de ecuaciones

Entrada	− Salida	− lo que reacciona	= Acumulación
$FC_{A0} + FR\,C_{A3}$	$-(F+F_R)C_{A1}$	$-k_1\,V_1\,C_{A1}^n$	$= \frac{dC_{A1}}{dt}$
$(F+F_R)C_{A1}$	$-(F+F_R)C_{A2}$	$-k_1\,V_2\,C_{A2}^n$	$= \frac{dC_{A2}}{dt}$
$(F+F_R)C_{A2}$	$-(F+F_R)C_{A3}$	$-k_1\,V_3\,C_{A3}^n$	$= \frac{dC_{A3}}{dt}$

Calcule la concentración de A a régimen permanente en cada reactor si la reacción es de primer orden con repecto a A y la constante de velocidad de reacción k_1 es 0.1 min^{-1}. Las composiciones están dadas en *mol/l*.

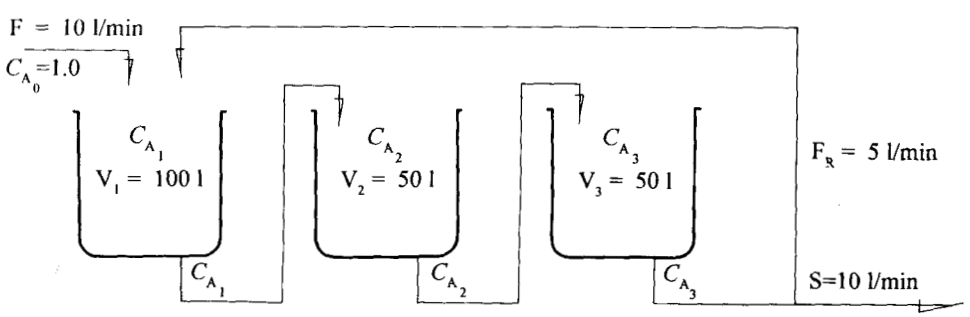


Figura 3.18. Sistema de tres reactores continuos tipo tanque agitado en donde se lleva a cabo la reacción $A \longrightarrow$ *Productos*.

3.34 Repita el problema 3.33, considerando que el reflujo es como se muestra en la figura 3.19.

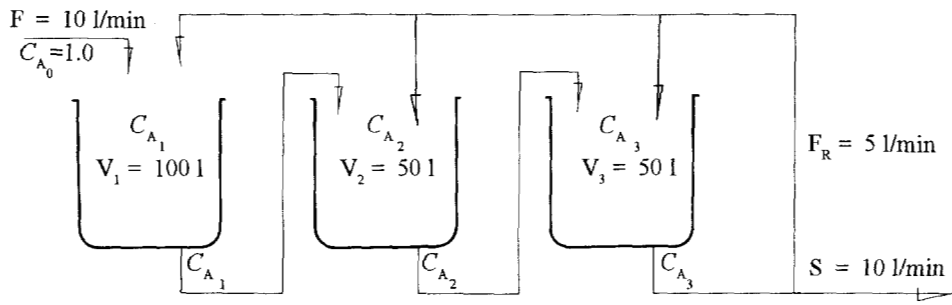


Figura 3.19. Sistema de tres reactores continuos tipo tanque agitado en donde se lleva a cabo la reacción $A \rightarrow$ *Productos*.

3.35 Calcule la composición del benceno en cada plato de la columna de absorción del ejercicio 3.1, si se modifica y_0 a 0.2 de fracción molar. Use las consideraciones del mismo ejercicio.

3.36 Las reacciones químicas pueden escribirse como

$$\sum_{i=1}^{n} x_i c_i = 0$$

donde: x_i es el coeficiente estequiométrico del compuesto i y c_i el compuesto i.

Por ejemplo, $CH_4 + 2\ O_2 \rightarrow CO_2 + 2\ H_2O$

puede escribirse como

$$CO_2 + 2\ H_2O - CH_4 - 2\ O_2 = 0$$

Dado que los átomos se conservan en una reacción química, la ecuación de conservación del elemento k es

$$\sum_{i=1}^{n} x_i\ m_{i,k} = 0; \qquad k = 1, 2, \ldots, m$$

donde $m_{i,k}$ es el número de átomos del elemento k en el compuesto i.
Esta última expresión representa un conjunto de ecuaciones lineales, donde x_i son las incógnitas. Lo anterior se conoce como el **método algebraico de balanceo de ecuaciones químicas.**

Utilice este método para balancear la ecuación química

$$Fe(CrO_2)_2 + O_2 + Na_2\ CO_3 \rightarrow Fe_2\ O_3 + Na_2\ CrO_4 + CO_2$$

3.37 Factorice las siguientes matrices en la forma $L\ U$, con el algoritmo 3.6

a) $\begin{bmatrix} 4 & 1 & -1 \\ 2 & 5 & 0 \\ 3 & 8 & 9 \end{bmatrix}$ *b)* $\begin{bmatrix} 3.444 & 16100 & -9.1 \\ 1.9999 & 17.01 & 9.6 \\ 1.6 & 5.2 & 1.7 \end{bmatrix}$

c) $\begin{bmatrix} 5.8 & 3.2 & 11.25 \\ 4.3 & 3.4 & 9.625 \\ 2.5 & 5.2 & 9.625 \end{bmatrix}$ *d)* $\begin{bmatrix} 1.002 & 4 \times 10^{-4} & 5 \times 10^{-4} & 8 \times 10^{-6} \\ 0 & 2.3 & 3 \times 10^{-3} & 4 \times 10^{-5} \\ 0 & 0 & 5 & 0.01 \\ 0 & 0 & 0 & 2 \end{bmatrix}$

e) $\begin{bmatrix} 4 & 0 & 0 & 0 \\ 7 & 10 & 0 & 0 \\ 8 & 9 & 5 & 0 \\ 10 & 3 & 3 & 1 \end{bmatrix}$ *f)* $\begin{bmatrix} 0 & 6 & 1 & 12 \\ 7 & 5 & -2 & 5 \\ -4 & 5 & 7 & 8 \\ 3 & 9 & 12 & 24 \end{bmatrix}$

3.38 Factorice las matrices del problema 3.37, con el algoritmo 3.7.

3.39 Resuelva los siguientes sistemas de ecuaciones con el algoritmo 3.8

a)
$$\begin{aligned} 4x_1 + x_2 - x_3 &= 8 \\ 2x_1 + 5x_2 &= 5 \\ 3x_1 + 8x_2 + 9x_3 &= 0 \end{aligned}$$

b)
$$\begin{aligned} 3.444x_1 + 16100x_2 - 9.1x_3 &= 0 \\ 1.9999x_1 + 17.01x_2 + 9.6x_3 &= 1 \\ 1.6x_1 + 5.2x_2 + 1.7x_3 &= 0 \end{aligned}$$

c)
$$\begin{aligned} 5.8x_1 + 3.2x_2 + 11.24x_3 &= 20.24 \\ 4.3x_1 + 3.4x_2 + 9.625x_3 &= 17.325 \\ 2.5x_1 + 5.2x_2 + 9.625x_3 &= 17.325 \end{aligned}$$

d)
$$\begin{aligned} 4x_1 + 5x_2 + 2x_3 - x_4 &= 3 \\ 5x_1 + 8x_2 + 7x_3 + 6x_4 &= 2 \\ 3x_1 + 7x_2 - 4x_3 - 2x_4 &= 0 \\ -x_1 + 6x_2 - 2x_3 + 5x_4 &= 1 \end{aligned}$$

e)
$$\begin{aligned} 2.156x_1 + 4.102x_2 - 2.3217x_3 + 6x_4 &= 18 \\ -4.102x_1 + 6x_2 + 1.2x_4 &= 6.5931 \\ -x_1 - 5.7012x_2 + 1.2222x_3 &= 3.4 \\ 6.532x_1 + 7x_2 - 4x_4 &= 0 \end{aligned}$$

3.40 Factorice las matrices simétricas siguientes, mediante el algoritmo 3.9

a) $\begin{bmatrix} -5 & 5 & 3 \\ 5 & 6 & 1 \\ 3 & 1 & 7 \end{bmatrix}$ *b)* $\begin{bmatrix} 3.33 & 4.81 & -2.22 \\ 4.81 & 10.0 & 7.45 \\ -2.22 & 7.45 & 15.0 \end{bmatrix}$

c)
$$\begin{bmatrix} 72.00 & 0.00 & 0.00 & 9.00 & 0.00 & 0.00 \\ 0.00 & 2.88 & 0.00 & 0.00 & 0.00 & -4.50 \\ 0.00 & 0.00 & 18.00 & 9.00 & 0.00 & 0.00 \\ 9.00 & 0.00 & 9.00 & 12.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 33.00 & 0.00 \\ 0.00 & -4.50 & 0.00 & 0.00 & 0.00 & 33.00 \end{bmatrix}$$

3.41 Con el algoritmo 3.9, elabore un algoritmo para resolver sistemas lineales simétricos y resuelva con él los siguientes sistemas

a)
$$\begin{aligned} -5x_1 + 5x_2 + 3x_3 &= 1 \\ 5x_1 + 6x_2 + x_3 &= 2 \\ 3x_1 + x_2 + 7x_3 &= 3 \end{aligned}$$

b)
$$\begin{aligned} 3.33x_1 + 4.81x_2 - 2.22x_3 &= 5 \\ 4.81x_1 + 10.00x_2 + 745x_3 &= 0 \\ -2.22x_1 + 7.45x_2 + 15.00x_3 &= 2 \end{aligned}$$

c)
$$\begin{aligned} 72x_1 + 9x_4 &= 2 \\ 2.88x_2 - 4.5x_6 &= 0.5 \\ 18x_3 + 9x_4 &= 1 \\ 9x_3 + 12x_4 &= 0 \\ 33x_5 &= 1.2 \\ -4.5x_2 + 33x_6 &= 5 \end{aligned}$$

3.42 Use el algoritmo 3.10 para factorizar en la forma $L\,L^{\mathrm{T}}$ las siguientes matrices positivas definidas.

a)
$$\begin{bmatrix} 4 & -2 & 0 \\ -2 & 4 & -1 \\ 0 & -1 & 4 \end{bmatrix}$$

b)
$$\begin{bmatrix} 5 & 1 & 2 & -1 \\ 1 & 7 & 0 & 3 \\ 2 & 0 & 5 & 1 \\ -1 & 3 & 1 & 8 \end{bmatrix}$$

c)
$$\begin{bmatrix} 10 & 0 & 0 & -1 & 0 \\ 0 & 5 & 0 & 0 & 2 \\ 0 & 0 & 2 & 0 & 0 \\ -1 & 0 & 0 & 8 & 3 \\ 0 & 2 & 0 & 3 & 5 \end{bmatrix}$$

3.43 Mediante el algoritmo 3.10 elabore un algoritmo para resolver sistemas lineales con matriz coeficiente positiva definida y resuelva con él los siguientes sistemas.

a)
$$\begin{aligned} 4x_1 - 2x_2 &= 0 \\ -2x_1 + 4x_2 - x_3 &= 0.5 \\ - x_2 + 4x_3 &= 1 \end{aligned}$$

b)
$$\begin{aligned} 5x_1 + x_2 + 2x_3 - x_4 &= 1 \\ x_1 + 7x_2 + 3x_4 &= 2 \\ 2x_1 + 5x_3 + x_4 &= 3 \\ -x_1 + 3x_2 + x_3 + 8x_4 &= 4 \end{aligned}$$

c)
$$\begin{aligned} 10x_1 - x_4 &= 0.2 \\ 5x_2 + 2x_4 &= 0.4 \\ 2x_3 &= 1.0 \\ - x_1 + 8x_4 + 3x_5 &= 0.6 \\ 2x_2 + 3x_4 + 5x_5 &= 0.8 \end{aligned}$$

3.44 Si la factorización de A en las matrices L y U es posible, puede imponerse que $u_{i,i} = 1$ con i = 1, 2, ..., n. Con estas condiciones obtenga las ecuaciones correspondientes a las ecuaciones 3.74, 3.75 y 3.76, para el caso del orden de A igual a 3. También obtenga las ecuaciones correspondientes a la ecuación 3.77 para el caso general, orden de A igual a n. Este método, como se recordará, es conocido como **algoritmo de Crout.**

3.45 Elabore un algoritmo para resolver un sistema de ecuaciones lineales con el método de Crout (véase algoritmo 3.8); resuelva los sistemas del problema 3.43 con el algoritmo encontrado.

3.46 Demuestre que en la solución del sistema lineal $A\,\mathbf{x} = \mathbf{b}$, donde A es positiva definida, con el método de Cholesky se requiere efectuar

n raíces cuadradas

$$\frac{n^3 + 9n^2 + 2n}{6} \text{ multiplicaciones o divisiones}$$

y

$$\frac{n^3 + 6n^2 - 7n}{6} \text{ sumas o restas}$$

cuando el orden de A es n.

3.47 Demuestre que si una matriz A es positiva definida, entonces $a_{i,i} > 0$ para $i = 1, 2, ..., n$.

3.48 Los algoritmos de factorización, cuando son aplicables, se pueden simplificar considerablemente en el caso de matrices bandeadas, debido al gran número de ceros que aparecen en estas matrices. Adapte el método de Doolitle o el de Crout para sistemas tridiagonales y una vez obtenidas las ecuaciones correspondientes, elabore un algoritmo eficiente.

3.49 En la solución de una estructura doblemente empotrada se obtuvo el siguiente sistema

$$\frac{1}{EI}\,\mathbf{c} + \frac{1}{EI}\,A\,\mathbf{p} = 0$$

donde EI es el módulo de elasticidad del elemento,

$$\mathbf{c} = \begin{bmatrix} -1.80 \\ 22.50 \\ -67.50 \\ 0.00 \\ 165.00 \end{bmatrix}; \qquad \mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix};$$

y

$$A = \begin{bmatrix} 72.00 & 0.00 & 0.00 & 9.00 & 0.00 & 0.00 \\ 0.00 & 2.88 & 0.00 & 0.00 & 0.00 & -4.50 \\ 0.00 & 0.00 & 18.00 & 9.00 & 0.00 & 0.00 \\ 9.00 & 0.00 & 9.00 & 12.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 33.00 & 0.00 \\ 0.00 & -4.50 & 0.00 & 0.00 & 0.00 & 33.00 \end{bmatrix}$$

Encuentre **p**.

3.50 Determine si el sistema que sigue está mal condicionado

$$\begin{bmatrix} 3.4440 & 16100 & -9.1 \\ 1.9999 & 17.01 & 9.6 \\ 1.6000 & 5.20 & 1.7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 16000.00 \\ 29.00 \\ 8.42 \end{bmatrix}$$

Resuélvalo usando la eliminación de Gauss y aritmética de cinco dígitos.

3.51 La matriz $H^{(n)}$ de orden n o matriz de Hilbert, definida por

$$h_{i,j} = \frac{1}{i + j - 1}; \quad 1 \leq i \leq n; \quad 1 \leq j \leq n,$$

es una matriz mal condicionada que surge, por ejemplo, al resolver las ecuaciones normales del método de aproximación por mínimos cuadrados (véase Cap. 5). Encuentre $H^{(4)}$, $H^{(5)}$ y sus inversas por alguno de los métodos vistos; además, resuelva el sistema

$$H^{(4)}\,\mathbf{x} = [1\ 0\ 1\ 0]^{\mathrm{T}}$$

3.52 Demuestre que

det $P = -1$, el determinante de una matriz permutadora es -1.

det $M = m$, el determinante de una matriz multiplicadora es el factor $m (m \neq 0)$.

det $S = 1$, el determinante de una matriz del tipo S es 1.

Sugerencia: Utilice la función determinante de una matriz de orden n.

3.53 Dados los siguientes sistemas de ecuaciones lineales

$$a)\ \begin{bmatrix} 3 & 5 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \qquad b)\ \begin{bmatrix} 3 & 1 & 4 \\ 2 & 0 & 1 \\ -1 & 5 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 3 \end{bmatrix} \qquad c)\ \begin{bmatrix} -2 & 4 & 5 \\ 4 & 8 & 3 \\ 5 & 3 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

calcule las inversas, determinantes y soluciones correspondientes, usando matrices elementales.

3.54 Resuelva los siguientes sistemas de ecuaciones lineales mediante los métodos de Gauss-Seidel y de Jacobi

$$a)\ \begin{bmatrix} 5 & -1 & -1 \\ 1 & -1 & 2 \\ 3 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \\ 4 \end{bmatrix} \qquad b)\ \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \\ 4 \end{bmatrix}$$

c) inciso (c) del problema 3.41.

$$d)\ \begin{bmatrix} 1 & 2 & 2^2 & 2^3 & 2^4 \\ 1 & 6 & 6^2 & 6^3 & 6^4 \\ 1 & 10 & 10^2 & 10^3 & 10^4 \\ 1 & 20 & 20^2 & 20^3 & 20^4 \\ 1 & 30 & 30^2 & 30^3 & 30^4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 13.4 \\ 30.4 \\ 41.8 \\ 57.9 \\ 66.5 \end{bmatrix}$$

$$e)\ \begin{bmatrix} 8 & 0 & 6 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 & 0 & 0 & 5 & 0 & 2 & 1 & 0 \\ 0 & 1 & 7 & 0 & 0 & 1 & 2 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 6 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 9 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 2 & 0 & 10 & 1 & 0 & 3 & 0 & 0 \\ 0 & 0 & 5 & 0 & 2 & 2 & 10 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 6 & 1 & 0 & 0 & 15 & 0 & 2 & 0 \\ 0 & 2 & 0 & 0 & 4 & 0 & 1 & 1 & 20 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 & 6 & 5 & 25 & 1 \\ 1 & 0 & 3 & 1 & 5 & 0 & 7 & 0 & 0 & 1 & 12 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \\ x_{11} \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \\ 8 \\ 0 \\ 8 \\ 1 \\ 0 \\ 3 \\ 0 \\ 1 \\ 2 \end{bmatrix}$$

3.55 Elabore un algoritmo para arreglar la matriz aumentada de un sistema de modo que la matriz coeficiente quede lo más cercana posible a diagonal dominante.

3.56 Elabore un algoritmo para resolver un sistema de ecuaciones lineales, usando los métodos SOR con $w > 1$ y con $w < 1$.

Sugerencia: Puede obtenerlo fácilmente modificando el algoritmo 3.11.

3.57 Aplique la segunda ley de Kirchhoff a la malla formada por el contorno del circuito de la figura 3.15; es decir, no considere E_5, R_4 y R_5. Demuestre que la ecuación resultante es linealmente dependiente de las tres obtenidas al seccionar en mallas dicho circuito.

3.58 Resuelva los sistemas de ecuaciones lineales del problema 3.54 con el algoritmo elaborado en el problema 3.56.

3.59 Demuestre que la ecuación 1 del ejercicio 3.8 es un polinomio de grado n si A es una matriz de orden n y I es la matriz identidad correspondiente.

3.60 Encuentre los valores característicos (eigenvalores) de la matriz coeficiente del siguiente sistema

$$\begin{bmatrix} 4 & 1 & 0.3 & -1 \\ 1 & 7 & 2 & 0 \\ -0.3 & 2 & 5 & -2 \\ -1 & 0 & -2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 2 \\ 3 \end{bmatrix}$$

3.61 Encuentre los vectores característicos (eigenvectores) correspondientes al valor característico dominante (el de máximo valor absoluto) del ejemplo anterior.

3.62 El método de las potencias permite calcular el valor y el vector característicos dominantes de una matriz A de orden n, cuando dicha matriz tiene n vectores característicos linealmente independientes: $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ y un valor característico λ_1 estrictamente dominante en magnitud

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \ldots |\lambda_n| .$$

Se muestra a continuación dicho método

Dada la independencia lineal de los vectores característicos, cualquier vector $\mathbf{v}$ de n componentes puede expresarse como una combinación lineal de ellos

$$\mathbf{v} = a_1 \mathbf{v}_1 + a_2 \mathbf{v}_2 + \ldots + a_n \mathbf{v}_n$$

Multiplicando la ecuación anterior por la izquierda por A se tiene

$$\begin{aligned} A\mathbf{v} &= a_1 A\mathbf{v}_1 + a_2 A\mathbf{v}_2 + \ldots + a_n A\mathbf{v}_n \\ &= a_1 \lambda_1 \mathbf{v}_1 + a_2 \lambda_2 \mathbf{v}_2 + \ldots + a_n \lambda_n \mathbf{v}_n \end{aligned}$$

Multiplicando repetidamente por A se llega a

$$A^k \mathbf{v} = a_1 \lambda_1^k \mathbf{v}_1 + a_2 \lambda_2^k \mathbf{v}_2 + \ldots + a_n \lambda_n^k \mathbf{v}_n$$

y factorizando

$$A^k \mathbf{v} = \lambda_1^k \left[a_1 \mathbf{v}_1 + a_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k \mathbf{v}_2 + \ldots + a_n \left(\frac{\lambda_n}{\lambda_1}\right)^k \mathbf{v}_n \right]$$

y como λ_1 es el mayor, todos los términos dentro del paréntesis rectangular tienden a cero cuando k tiende a ∞, excepto el primer término (si $a_1 \neq 0$).
Para k grande $A^k \mathbf{v} \approx \lambda_1^k a_1 \mathbf{v}_1$

Al tomar la relación de cualesquiera componentes correspondientes de A^k v y A^{k+1} v, se obtiene una sucesión de valores convergentes a λ_1, ya que

$$\frac{\lambda_1^{k+1}\, a_1\, \mathbf{v}_{1,i}}{\lambda_1^{k}\, a_1\, \mathbf{v}_{1,i}} \approx \lambda_1$$

Además, la sucesión $\lambda_1^{-k} A^k \mathbf{v}$ convergirá al vector característico $\mathbf{v}_1$ multiplicado por a_1.Con este método, encuentre el valor característico y el vector característico dominantes de la matriz coeficiente del siguiente sistema

$$\begin{bmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$$

Sugerencia: Como generalmente no se conocen los vectores característicos, sino que ése es el propósito, se empieza a iterar con $\mathbf{v} = \mathbf{e}_1 = [1\ 0\ 0]^T$.

3.63 Encuentre el valor característico dominante y los vectores característicos correspondientes del sistema de ecuaciones del problema 3.60.

3.64 Se tienen tres tanques cilíndricos iguales de 6 pies de diámetro, comunicados entre sí por medio de tubos de 4 pulgadas de diámetro y 2 pies de largo, como se muestra en la Fig. 3.30. El tercer tanque tiene una salida a través de un tubo de 4 pulgadas de diámetro y 8 pies de largo. Al primer tanque llega un fluido a razón de 0.1 pies cúbicos por minuto e inicialmente su nivel tiene una altura de 20 pies, mientras que el segundo y tercer tanques están vacíos. El fluido es un aceite viscoso cuya densidad es 51.45 lb_m/pie^3 y cuya viscosidad es 100 centipoises. Calcule la altura del fluido en cada tanque cuando se alcance el régimen permanente.

Sugerencia: Use la ecuación de Poiselle para el cálculo de la velocidad media del fluido a través de los tubos.

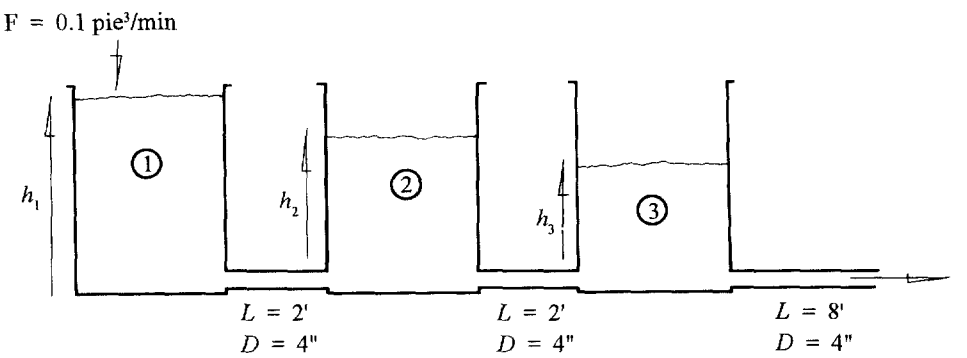


Figura 3.30. Tres tanques interconectados.

CAPÍTULO 4

SISTEMAS DE ECUACIONES NO LINEALES

ESTE CAPÍTULO tratará la generalización de los métodos de los capítulos 2 y 3 para dar lugar a algoritmos que permiten resolver sistemas no lineales.

INTRODUCCIÓN

En el capítulo 2 se vio cómo encontrar las raíces de una ecuación de la forma

$$f(x) = 0.$$

Por otro lado, en el capítulo 3 se estudiaron las técnicas iterativas de solución de un sistema de ecuaciones lineales $A\,\mathbf{x} = \mathbf{b}$.

Estos dos son casos particulares de la situación más general, donde se tiene un sistema de varias ecuaciones con varias incógnitas, cuya representación es:

$$\begin{aligned} f_1(x_1, x_2, x_3, \ldots, x_n) &= 0 \\ f_2(x_1, x_2, x_3, \ldots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, x_2, x_3, \ldots, x_n) &= 0, \end{aligned} \tag{4.1}$$

donde $f_i(x_1, x_2, x_3, \ldots, x_n)$ para $1 \leq i \leq n$ es una función (lineal o no) de las variables independientes $x_1, x_2, x_3, \ldots, x_n$.

Si por ejemplo la ecuación 4.1 consiste sólo en una ecuación de una incógnita ($n = 1$), se tiene la ecuación 2.1. En cambio la ecuación 4.1 se reducirá al caso (3.39) si $n > 1$ y $f_1, f_2, \ldots f_n$ son todas funciones lineales de $x_1, x_2, x_3, \ldots, x_n$.

Por todo esto, es fácil entender que los métodos iterativos de solución de la ecuación 4.1 son extensiones de los métodos para ecuaciones no lineales en una incógnita y emplean las ideas que se aplicaron al desarrollar los algoritmos iterativos para resolver $A\,\mathbf{x} = \mathbf{b}$.

A continuación se dan algunos ejemplos.

$$a)\ f_1(x_1, x_2) = x_1^2 + x_2^2 - 4 = 0$$

$$f_2(x_1, x_2) = x_2 - x_1^2 = 0$$

$$b)\ f_1(x_1, x_2) = 10\,(x_2 - x_1^2) = 0$$

$$f_2(x_1, x_2) = 1 - x_1 = 0$$

$$c)\ f_1(x_1, x_2, x_3) = x_1 x_2 x_3 - 10x_1^3 + x_2 = 0$$

$$f_2(x_1, x_2, x_3) = x_1 + 2x_2 x_3 + \operatorname{sen} x_2 - 15 = 0$$

$$f_3(x_1, x_2, x_3) = x_2^2 - 5x_1 x_3 - 3x_3^3 + 3 = 0$$

SECCIÓN 4.1 DIFICULTADES EN LA SOLUCIÓN DE SISTEMAS DE ECUACIONES NO LINEALES

Antes de desarrollar los métodos iterativos para resolver sistemas de ecuaciones no lineales con varias incógnitas, se destacarán algunas de las dificultades que se presentan al aplicar estos métodos.

— Es imposible graficar las superficies multidimensionales definidas por las ecuaciones de los sistemas para $n > 2$.
— No es fácil encontrar "buenos" valores iniciales.

Para atenuar estas dificultades se darán algunas sugerencias aplicables antes de un intento formal de solución de la ecuación 4.1.

Reducción de ecuaciones

Resulta muy útil tratar de reducir analíticamente el número de ecuaciones y de incógnitas antes de intentar una solución numérica. En particular, trátese de resolver alguna de las ecuaciones para alguna de las incógnitas. Después, sustitúyase la ecuación resultante para esa incógnita en todas las demás ecuaciones; con esto el sistema se reduce en una ecuación y una incógnita. Continúese de esta manera hasta donde sea posible.

Por ejemplo, en el sistema

$$f_1(x_1, x_2) = 10\,(x_2 - x_1^2\,) = 0$$

$$f_2(x_1, x_2) = 1 - x_1 = 0$$

se despeja x_1 en la segunda ecuación

$$x_1 = 1$$

y se sustituiye en la primera

$$10(x_2 - 1^2) = 0$$

cuya solución, $x_2 = 1$, conjuntamente con $x_1 = 1$ proporciona una solución del sistema dado, sin necesidad de resolver dos ecuaciones con dos incógnitas.

Partición de ecuaciones

A veces resulta más sencillo dividir las ecuaciones en subsistemas menores y resolverlos por separado. Considérese por ejemplo el siguiente sistema de cinco ecuaciones con cinco incógnitas

$$\begin{aligned} f_1(x_1, x_2, x_3, x_4, x_5) &= 0 \\ f_2(x_1, x_2, x_4) &= 0 \\ f_3(x_1, x_3, x_4, x_5) &= 0 \\ f_4(x_2, x_4) &= 0 \\ f_5(x_1, x_4) &= 0 \end{aligned}$$

En vez de atacar las cinco ecuaciones al mismo tiempo, se resuelve el subsistema formado por f_2, f_4 y f_5. Las soluciones de este subsistema se utilizan después para resolver el subsistema compuesto por las ecuaciones f_1 y f_3.

En general, una partición de ecuaciones es la división de un sistema de ecuaciones en subsistemas llamados bloques. Cada bloque de la partición es el sistema de ecuaciones más pequeño que incluye todas las variables que es preciso resolver.

Tanteo de ecuaciones

Supóngase que se quiere resolver el siguiente sistema de cuatro ecuaciones con cuatro incógnitas

$$\begin{aligned} f_1(x_2, x_3) &= 0 \\ f_2(x_2, x_3, x_4) &= 0 \\ f_3(x_1, x_2, x_3, x_4) &= 0 \\ f_4(x_1, x_2, x_3) &= 0 \end{aligned}$$

No se pueden dividir en subsistemas, sino que es preciso resolverlas simultáneamente. Sin embargo, es posible abordar el problema por otro camino. Supóngase que se estima un valor de x_3. Se podría obtener así x_2 a partir de f_1, x_4 de f_2 y x_1 de f_3. Finalmente, se comprobaría con f_4 la estimación hecha de x_3. Si f_4 fuese cero o menor en magnitud que un valor predeterminado o criterio de exactitud ε, la estimación x_3 y los valores de x_2, x_4 y x_1 obtenidos con ella, serían una aproximación a la solución del sistema dado. En caso contrario, habría que proponer un nuevo valor de x_3 y repetir el proceso.

Nótese la íntima relación que guarda este método con el método de punto fijo (Cap. 2), ya que un problema multidimensional se reduce a uno unidimensional en x_3

$$h(x_3) = 0.$$

Valores iniciales

a) De consideraciones físicas

Si el sistema de ecuaciones 4.1 tiene un significado físico, con frecuencia es posible acotar los valores de las incógnitas a partir de consideraciones físicas. Por

ejemplo, si alguna de las variables x_i representa la velocidad de flujo de un fluido, ésta no podrá ser negativa. Por tanto $x_i \geq 0$. En el caso de que x_i represente una concentración expresada como fracción peso o fracción molar de una corriente de alimentación, se tiene que $0 \leq x_i \leq 1$. (Para mayores detalles ver los problemas resueltos al final del capítulo).

b) De consideraciones geométricas

En el caso particular de tener un sistema de dos ecuaciones con dos incógnitas

$$f_1(x_1, x_2) = 0$$
$$f_2(x_1, x_2) = 0,$$

cada una define, en general, una curva en el plano x_1–x_2, y el problema de resolver el sistema puede verse como el problema de encontrar el punto o los puntos de intersección de estas dos curvas. Graficando (puede usarse el software GC, el Math-CAD, o un programa que grafique) pueden obtenerse buenos valores iniciales.

Sea por ejemplo el sistema

$$f_1(x_1, x_2) = x_1^2 + x_2^2 - a^2 = 0$$
$$f_2(x_1, x_2) = x_2 - x_1^2 = 0$$

Al graficar f_1 y f_2 se obtiene la figura 4.1, en donde se podrán apreciar valores iniciales muy cercanos a la solución.

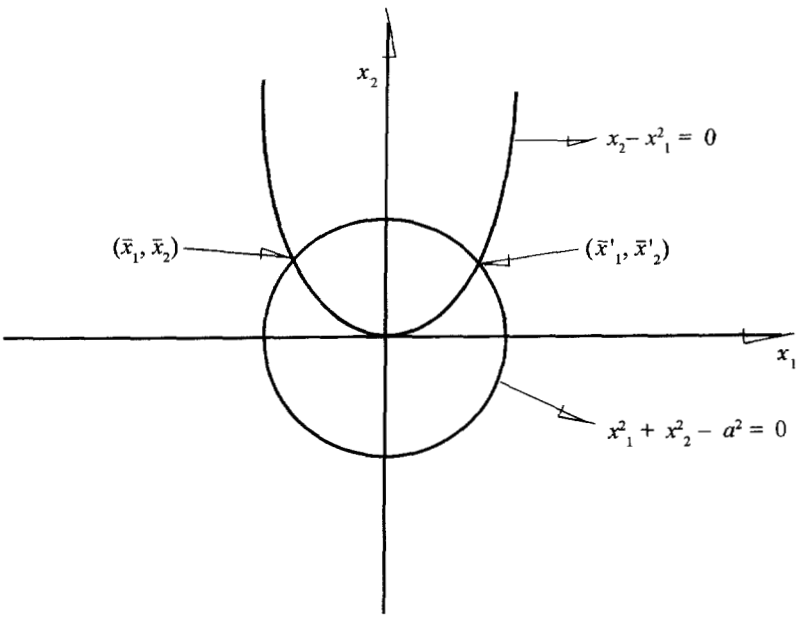


Figura 4.1 Solución gráfica de un sistema de dos ecuaciones.

Por último, resulta muy conveniente conocer bien las características de cada método de solución del sistema 4.1 para efectuar la elección más adecuada del mismo.

Se iniciará el estudio de dichos métodos con la extensión del método de punto fijo a sistemas de ecuaciones no lineales.

SECCIÓN 4.2 MÉTODO DE PUNTO FIJO MULTIVARIABLE

Los algoritmos discutidos en este capítulo son, en principio, aplicables a sistemas de cualquier número de ecuaciones. Sin embargo, para ser más concisos y evitar notación complicada, se considerará sólo el caso de dos ecuaciones con dos incógnitas. Éstas generalmente se escribirán como

$$\begin{aligned} f_1(x, y) &= 0 \\ f_2(x, y) &= 0 \end{aligned} \tag{4.2}$$

y se tratará de encontrar pares de valores (x,y) que satisfagan ambas ecuaciones.

Como en el método de punto fijo (Sec. 2.1) y en los métodos de Jacobi y Gauss-Seidel (Sec. 3.5), se resolverá la primera ecuación para alguna de las variables, x por ejemplo, y la segunda para y.

$$\begin{aligned} x &= g_1(x, y) \\ y &= g_2(x, y) \end{aligned} \tag{4.3}$$

Al igual que en los métodos mencionados, se tratará de obtener la estimación $(k + 1)$-ésima a partir de la estimación k-ésima con la expresión

$$\begin{aligned} x^{k+1} &= g_1(x^k, y^k) \\ y^{k+1} &= g_2(x^k, y^k) \end{aligned} \tag{4.4}$$

Se comienza con valores iniciales x^0, y^0, se calculan nuevos valores x^1, y^1 y se repite el proceso, esperando que después de cada iteración los valores de x^k, y^k se aproximen a la raíz buscada $\bar{x}$, $\bar{y}$, la cual cumple con

$$\begin{aligned} \bar{x} &= g_1(\bar{x}, \bar{y}) \\ \bar{y} &= g_2(\bar{x}, \bar{y}) \end{aligned}$$

Por analogía con los casos discutidos, puede predecirse el comportamiento y las características de este método de punto fijo multivariable.

Como se sabe, en el caso de una variable la manera particular de pasar de $f(x) = 0$ a $x = g(x)$, afecta la convergencia del proceso iterativo. Entonces debe esperarse que la forma en que se resuelve para $x = g_1(x,y)$ y $y = g_2(x,y)$ afecte la convergencia de las iteraciones (4.4).

Por otro lado, se sabe que el reordenamiento de las ecuaciones en el caso lineal afecta la convergencia, por lo que puede esperarse que la convergencia del método en estudio dependa de si se despeja x de f_2 o de f_1.

Finalmente, como en el método iterativo univariable y en el de Jacobi y de Gauss-Seidel, la convergencia —en caso de existir— es de primer orden, cabe esperar que el método iterativo multivariable tenga esta propiedad.

Ejemplo 4.1

Encuentre una solución del sistema de ecuaciones no lineales

$$f_1(x, y) = x^2 - 10x + y^2 + 8 = 0$$
$$f_2(x, y) = xy^2 + x - 10y + 8 = 0.$$

SOLUCIÓN

Con el despeje de x del término $(-10x)$ en la primera ecuación y de y del término $(-10y)$ en la segunda ecuación, resulta

$$x = \frac{x^2 + y^2 + 8}{10}$$

$$y = \frac{xy^2 + x + 8}{10}$$

o con la notación de la ecuación 4.4

$$x^{k+1} = \frac{(x^k)^2 + (y^k)^2 + 8}{10}$$

$$y^{k+1} = \frac{x^k(y^k)^2 + x^k + 8}{10}$$

Con los valores iniciales $x^0=0$, $y^0=0$, se inicia el proceso iterativo

Primera iteración

$$x^1 = \frac{0^2 + 0^2 + 8}{10} = 0.8$$

$$y^1 = \frac{0(0)^2 + 0 + 8}{10} = 0.8$$

Segunda iteración

$$x^2 = \frac{(0.8)^2 + (0.8)^2 + 8}{10} = 0.928$$

$$y^2 = \frac{0.8(0.8)^2 + 0.8 + 8}{10} = 0.9312$$

Al continuar el proceso iterativo, se encuentra la siguiente sucesión de vectores

k	x^k	y^k
0	0.00000	0.00000
1	0.80000	0.80000
2	0.92800	0.93120
3	0.97283	0.97327
4	0.98937	0.98944
5	0.99578	0.99579
6	0.99832	0.99832
7	0.99933	0.99933
8	0.99973	0.99973
9	0.99989	0.99989
10	0.99996	0.99996
11	0.99998	0.99998
12	0.99999	0.99999
13	1.00000	1.00000

Para observar la convergencia del proceso iterativo, se pudieron usar los criterios del capítulo anterior, como distancia entre dos vectores consecutivos, o bien las distancias componente a componente de dos vectores consecutivos. También existe un criterio de convergencia equivalente al de las ecuaciones 2.10 y 3.99 que puede aplicarse antes de iniciar el proceso iterativo mencionado, y que dice

Una condición suficiente aunque no necesaria, para asegurar la convergencia es que

$$\left|\frac{\partial g_1}{\partial x}\right| + \left|\frac{\partial g_2}{\partial x}\right| \leq M < 1 \; ; \quad \left|\frac{\partial g_1}{\partial y}\right| + \left|\frac{\partial g_2}{\partial y}\right| \leq M < 1 \qquad (4.5)$$

para todos los puntos (x, y) de la región del plano que contiene todos los valores (x^k, y^k) y la raíz buscada $(\bar{x}, \bar{y})$.

Por otro lado; si M es muy pequeña en una región de interés, la iteración converge rápidamente; si M es cercana a 1 en magnitud, entonces la iteración puede converger lentamente. Este comportamiento es similar al del caso de una función univariable discutido en el capítulo 2.

Por lo general es muy difícil encontrar el sistema 4.3 a partir de la ecuación 4.2, de modo que satisfaga la condición 4.5.

De todas maneras, cualquiera que sea el sistema (4.4) a que se haya llegado y que se vaya a resolver con este método, puede aumentarse la velocidad de convergencia usando desplazamientos sucesivos en lugar de los desplazamientos simultáneos del esquema 4.4. Es decir, se iteraría mediante

$$\boxed{\begin{aligned} x^{k+1} &= g_1(x^k, y^k) \\ y^{k+1} &= g_2(x^{k+1}, y^k) \end{aligned}} \tag{4.6}$$

Como en el caso lineal (Jacobi y Gauss-Seidel), si la iteración por desplazamientos simultáneos diverge, generalmente el método por desplazamientos sucesivos divergería más rápido; es decir, se detecta más rápido la divergencia, por lo que se recomienda en general el uso de desplazamientos sucesivos en lugar de desplazamientos simultáneos.

Ejemplo 4.2

Resuelva el sistema del ejemplo 4.1 utilizando el método de punto fijo multivariable con desplazamientos sucesivos

$$f_1(x, y) = x^2 - 10x + y^2 + 8 = 0$$
$$f_2(x, y) = xy + x - 10y + 8 = 0$$

Sugerencia: Se pueden seguir los cálculos con un pizarrón electrónico o se programa una calculadora.

SOLUCIÓN

Al despejar x del término $(-10x)$ y y del término $(-10y)$ de la primera y segunda ecuaciones, respectivamente, resulta

$$x^{k+1} = g_1(x^k, y^k) = \frac{(x^k)^2 + (y^k)^2 + 8}{10}$$

$$y^{k+1} = g_2(x^k, y^k) = \frac{x^{k+1}(y^k)^2 + x^{k+1} + 8}{10}$$

Al derivar parcialmente, se obtiene

$$\frac{\partial g_1}{\partial x} = \frac{2x^k}{10} \qquad \frac{\partial g_1}{\partial y} = \frac{2y^k}{10}$$

$$\frac{\partial g_2}{\partial x} = \frac{(y^k)^2 + 1}{10} \qquad \frac{\partial g_2}{\partial y} = \frac{2x^{k+1}y^k}{10}$$

y evaluadas en $x^0 = 0$ y en $y^0 = 0$

$$\left.\frac{\partial g_1}{\partial x}\right|_{\substack{x^0\\y^0}} = 0 \qquad \left.\frac{\partial g_2}{\partial x}\right|_{\substack{x^0\\y^0}} = 1/10$$

$$\left.\frac{\partial g_1}{\partial y}\right|_{\substack{x^0\\y^0}} = 0 \qquad \left.\frac{\partial g_2}{\partial y}\right|_{\substack{x^0\\y^0}} = 0$$

con lo que se puede aplicar la condición 4.5

$$\frac{\partial g_1}{\partial x} + \frac{\partial g_2}{\partial x} = 0 + 1/10 = 1/10 < 1$$

$$\frac{\partial g_1}{\partial y} + \frac{\partial g_2}{\partial y} = 0 + 0 = 0 < 1$$

la cual se satisface; si los valores sucesivos de la iteración: $x^1, y^1; x^2, y^2; x^3, y^3;...$ la satisfacen también, se llega entonces a $\bar{x}, \bar{y}$.

Primera iteración

$$x^1 = \frac{0^2 + 0^2 + 8}{10} = 0.8$$

$$y^1 = \frac{0.8(0)^2 + 0.8 + 8}{10} = 0.88$$

Cálculo de la distancia entre el vector inicial y el vector $[x^1, y^1]^T$

$$\left|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right| = \sqrt{(0.8 - 0.0)^2 + (0.88 - 0.0)^2} = 1.18929$$

Segunda iteración

$$x^2 = \frac{(0.8)^2 + (0.88)^2 + 8}{10} = 0.94144$$

$$y^2 = \frac{0.94144\,(0.88)^2 + 0.94144 + 8}{10} = 0.96704$$

Cálculo de la distancia entre $[x^2, y^2]^T$ y $[x^1, y^1]^T$:

$$\| \mathbf{x}^{(2)} - \mathbf{x}^{(1)} \| = \sqrt{(0.94144 - 0.8)^2 + (0.96704 - 0.88)^2} = 0.16608$$

A continuación se muestran los resultados de las iteraciones

k	x^k	y^k	$\| \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \|$
0	0.00000	0.00000	
1	0.80000	0.88000	1.18929
2	0.94144	0.96705	0.16608
3	0.98215	0.99006	0.04677
4	0.99448	0.99693	0.01411
5	0.99829	0.99905	0.00436
6	0.99947	0.99970	0.00135
7	0.99983	0.99991	0.00042
8	0.99995	0.99997	0.00013
9	0.99998	0.99999	0.00004
10	0.99999	1.00000	0.00001
11	1.00000	1.00000	0.00001

Nótese que se requirieron once iteraciones para llegar al vector solución (1, 1) contra 13 del ejemplo 4.1, donde se usaron desplazamientos simultáneos.

A continuación se presenta un algoritmo para el método de punto fijo multivariable en sus versiones de desplazamientos simultáneos y desplazamientos sucesivos.

ALGORITMO 4.1 Método de punto fijo multivariable

Para encontrar una solución aproximada de un sistema de ecuaciones no lineales $\mathbf{g}(\mathbf{x}) = \mathbf{x}$, proporcionar las funciones G(I, x), I=1,2,..., N y los

DATOS: El número de ecuaciones N, el vector de valores iniciales **x**, el criterio de convergencia EPS, el número máximo de iteraciones MAXIT y M = 0 para desplazamientos sucesivos o M = 1 para desplazamientos simultáneos.

RESULTADOS: Una solución aproximada **x** o mensaje "NO HUBO CONVERGENCIA".

PASO 1. Hacer K = 1
PASO 2. Mientras K ≤ MAXIT, repetir los pasos 3 a 14.
PASO 3. Si M = 0, hacer **xaux** = x. De otro modo continuar.
PASO 4. Hacer I = 1
PASO 5. Mientras I ≤ N, repetir los pasos 6 y 7.
PASO 6. Si M = 0, hacer X(I) = G(I,x). De otro modo hacer XAUX(I) = G(I,x)
PASO 7. Hacer I = I + 1
PASO 8. Hacer I = 1
PASO 9. Mientras I ≤ N, repetir los pasos 10 y 11.
PASO 10. Si ABS(XAUX(I) - X(I)) > EPS ir al paso13. De otro modo continuar.
PASO 11. Hacer I = I + 1
PASO 12. IMPRIMIR x Y TERMINAR.
PASO 13. Si M = 1 hacer x = **xaux**. De otro modo continuar.
PASO 14. Hacer K = K + 1
PASO 15. IMPRIMIR mensaje "NO HUBO CONVERGENCIA" y TERMINAR.

Sugerencia: Desarrolle este algoritmo con Math-CAD o un software equivalente.

SECCIÓN 4.3 MÉTODO DE NEWTON-RAPHSON

El método iterativo para sistemas de ecuaciones converge linealmente. Como en el método de una incógnita, puede crearse un método de convergencia cuadrática; es decir, el método de Newton-Raphson multivariable. A continuación se obtendrá este procedimiento para dos variables; la extensión a tres o más variables es viable generalizando los resultados.

Supóngase que se está resolviendo el sistema

$$f_1(x, y) = 0$$
$$f_2(x, y) = 0,$$

donde ambas funciones son continuas y diferenciables, de modo que puedan expanderse en serie de Taylor. Esto es

$$f(x, y) = f(a, b) + \frac{\partial f}{\partial x}(x-a) + \frac{\partial f}{\partial y}(y-b) + \frac{1}{2!}\left[\frac{\partial^2 f}{\partial x \partial x}(x-a)^2 + 2\frac{\partial^2 f}{\partial x\, \partial y}(x-a)(y-b) + \frac{\partial^2 f}{\partial y\, \partial y}(y-b)^2\right] + \ldots$$

donde $f(x,y)$ se ha expandido alrededor del punto (a,b) y todas las derivadas parciales están evaluadas en (a,b).

Expandiendo f_1 alrededor de (x^k, y^k),

$$f_1(x^{k+1},y^{k+1}) = f_1(x^k,y^k) + \frac{\partial f_1}{\partial x}(x^{k+1} - x^k) + \frac{\partial f_1}{\partial y}(y^{k+1} - y^k) + \frac{1}{2!}\left[\frac{\partial^2 f_1}{\partial x\,\partial x}(x^{k+1} - x^k)^2 + 2\frac{\partial^2 f_1}{\partial x \partial y}(x^{k+1} - x^k)(y^{k+1} - y^k) + \frac{\partial^2 f_1}{\partial y\,\partial y}(y^{k+1} - y^k)^2\right] + \ldots \tag{4.7}$$

donde todas las derivadas parciales están evaluadas en (x^k, y^k). De la misma forma puede expanderse f_2 como sigue

$$f_2(x^{k+1},y^{k+1}) = f_2(x^k,y^k) + \frac{\partial f_2}{\partial x}(x^{k+1} - x^k) + \frac{\partial f_2}{\partial y}(y^{k+1} - y^k) + \frac{1}{2!}\left[\frac{\partial^2 f_2}{\partial x\,\partial x}(x^{k+1} - x^k)^2 + 2\frac{\partial^2 f_2}{\partial x \partial y}(x^{k+1} - x^k)(y^{k+1} - y^k) + \frac{\partial^2 f_2}{\partial y\,\partial y}(y^{k+1} - y^k)^2\right] + \ldots \tag{4.8}$$

De igual manera que en la ecuación 4.7, todas las derivadas parciales de (4.8) están evaluadas en (x^k,y^k).

Ahora supóngase que x^{k+1} y y^{k+1} están tan cerca de la raíz buscada $(\bar{x}, \bar{y})$ que los lados izquierdos de las dos últimas ecuaciones son casi cero; además, asúmase que x^k y y^k están tan próximos de x^{k+1} y y^{k+1} que pueden omitirse los términos a partir de los que se encuentran agrupados en paréntesis rectangulares. Con esto las ecuaciones 4.7 y 4.8 se simplifican a

$$\begin{aligned} 0 &\approx f_1(x^k,y^k) + \frac{\partial f_1}{\partial x}(x^{k+1} - x^k) + \frac{\partial f_1}{\partial y}(y^{k+1} - y^k) \\ 0 &\approx f_2(x^k,y^k) + \frac{\partial f_2}{\partial x}(x^{k+1} - x^k) + \frac{\partial f_2}{\partial y}(y^{k+1} - y^k) \end{aligned} \tag{4.9}$$

Para simplificar aún más, se cambia la notación con

$$\begin{aligned} x^{k+1} - x^k &= h \\ y^{k+1} - y^k &= j, \end{aligned} \tag{4.10}$$

y así queda la $(k+1)$-ésima iteración en términos de la k-ésima, como se ve a continuación

$$\begin{aligned} x^{k+1} &= x^k + h \\ y^{k+1} &= y^k + j \end{aligned} \tag{4.11}$$

La sustitución de la ecuación 4.10 en la 4.9 y el rearreglo dan como resultado

$$\boxed{\begin{aligned} \frac{\partial f_1}{\partial x} h + \frac{\partial f_1}{\partial y} j &= -f_1(x^k, y^k) \\ \frac{\partial f_2}{\partial x} h + \frac{\partial f_2}{\partial y} j &= -f_2(x^k, y^k), \end{aligned}} \tag{4.12}$$

el cual es un sistema de ecuaciones lineales en las incógnitas h y j (recuérdese que las derivadas parciales de la ecuación 4.12, así como f_1 y f_2 están evaluadas en $(x^k y^k)$ y, por tanto, son números reales).

Este sistema de ecuaciones lineales resultante tiene solución única, siempre que el determinante de la matriz de coeficientes o matriz jacobiana J no sea cero; es decir si

$$\left| J \right| = \begin{vmatrix} \dfrac{\partial f_1}{\partial x} & \dfrac{\partial f_1}{\partial y} \\ \dfrac{\partial f_2}{\partial x} & \dfrac{\partial f_2}{\partial y} \end{vmatrix} \neq 0$$

Precisando: El método de Newton-Raphson consiste fundamentalmente en formar y resolver el sistema 4.12, esto último por alguno de los métodos vistos en el capítulo 3. Con la solución y la ecuación 4.11 se obtiene la siguiente aproximación.

Este procedimiento se repite hasta satisfacer algún criterio de convergencia establecido.

Es interesante notar que como en el caso unidimensional, este método puede obtenerse encontrando un plano tangente a cada f de la ecuación 4.2 en (x^k, y^k), y luego encontrar el cero común de estos planos; es decir, hallar un plano tangente en $(x^k y^k)$ tanto a la superficie f_1 como a la superficie f_2, y luego la intersección de cada plano tangente con el plano x-y, con lo cual se obtienen dos líneas rectas en el plano x-y y, por último, la intersección de estas dos líneas rectas, que da el cero común de los planos tangentes.

Cuando converge este método, lo hace con orden dos, y requiere que el vector inicial $(x^0 y^0)$ esté muy cerca de la raíz buscada $(\bar{x}, \bar{y})$.

Ejemplo 4.3

Use el método de Newton-Raphson para encontrar una solución aproximada del sistema

$$\begin{aligned} f_1(x, y) &= x^2 - 10x + y^2 + 8 = 0 \\ f_2(x, y) &= xy^2 + x - 10y + 8 = 0 \end{aligned}$$

con el vector inicial: $[x^0, y^0]^T = [0, 0]^T$.

SOLUCIÓN

Primero se forma la matriz coeficiente del sistema 4.12, también conocida como matriz de derivadas parciales

$$\begin{vmatrix} \frac{\partial f_1}{\partial x} = 2x - 10 & \frac{\partial f_1}{\partial y} = 2y \\ \frac{\partial f_2}{\partial x} = y^2 + 1 & \frac{\partial f_2}{\partial y} = 2xy - 10 \end{vmatrix},$$

que aumentada en el vector de funciones resulta en

$$\left[\begin{array}{cc|c} 2x-10 & 2y & -x^2 + 10x - y^2 - 8 \\ y^2 + 1 & 2xy-10 & -xy^2 - x + 10y - 8 \end{array}\right]$$

Primera iteración

Al evaluar la matriz en $[x^0, y^0]^T$ se obtiene

$$\left[\begin{array}{cc|c} -10 & 0 & -8 \\ 1 & -10 & -8 \end{array}\right]$$

que al resolverse por eliminación de Gauss da

$$h = 0.8, j = 0.88$$

al sustituir en la ecuación 4.11 se obtiene

$$x^1 = x^0 + h = 0 + 0.8 = 0.8$$
$$y^1 = y^0 + j = 0 + 0.88 = 0.88$$

Cálculo de la distancia entre $\mathbf{x}^{(0)}$ y $\mathbf{x}^{(1)}$

$$| \mathbf{x}^{(1)} - \mathbf{x}^{(0)} | = \sqrt{(0.8 - 0)^2 + (0.88 - 0)^2} = 1.18929$$

Segunda iteración

Al evaluar la matriz en $[x^1, y^1]^T$ resulta

$$\left[\begin{array}{cc|c} -8.4 & 1.76 & -1.41440 \\ 1.7744 & -8.592 & -0.61952 \end{array}\right]$$

que por eliminación gaussiana da como nuevos resultados de h y j

$$h = 0.19179, j = 0.11171$$

de donde

$$x^2 = x^1 + h = 0.8 + 0.19179 = 0.99179$$
$$y^2 = y^1 + j = 0.88 + 0.11171 = 0.99171$$

Cálculo* de la distancia entre $\mathbf{x}^{(1)}$ y $\mathbf{x}^{(2)}$:

$$| \mathbf{x}^{(2)} - \mathbf{x}^{(1)} | = \sqrt{(0.99179 - 0.8)^2 + (0.99171 - 0.88)^2} = 0.22190$$

Con la continuación de este proceso iterativo se obtienen los resultados siguientes

k	x^k	y^k	$\| \mathbf{x}^{k+1} - \mathbf{x}^k \|$
0	0.00000	0.00000	
1	0.80000	0.88000	1.18929
2	0.99179	0.99171	0.22190
3	0.99998	0.99997	0.00830
4	1.00000	1.00000	0.00004

Obsérvese que se requirieron cuatro iteraciones para llegar al vector solución (1,1) contra once del ejemplo 4.2, donde se usó el método de punto fijo con desplazamientos sucesivos. Sin embargo, esta convergencia cuadrática implica mayor número de cálculos, ya que —como se puede observar— en cada iteración se requiere

a) La evaluación de 2 × 2 derivadas parciales
b) La evaluación de 2 funciones
c) La solución de un sistema de ecuaciones lineales de orden 2.

Sugerencia: Los cálculos, incluidas las derivadas parciales y la inversa de la matriz, se pueden ejecutar en Math-CAD o con el software que acompaña al libro.

Generalización

Para un sistema de n ecuaciones no lineales con n incógnitas (véase Ec. 4.1) y retomando la notación vectorial y matricial, las ecuaciones 4.12 quedan

$$\begin{aligned}
\frac{\partial f_1}{\partial x_1} h_1 + \frac{\partial f_1}{\partial x_2} h_2 + \ldots + \frac{\partial f_1}{\partial x_n} h_n &= -f_1 \\
\frac{\partial f_2}{\partial x_1} h_1 + \frac{\partial f_2}{\partial x_2} h_2 + \ldots + \frac{\partial f_2}{\partial x_n} h_n &= -f_2 \\
\vdots \qquad\qquad \vdots \qquad\qquad & \vdots \\
\frac{\partial f_n}{\partial x_1} h_1 + \frac{\partial f_n}{\partial x_2} h_2 + \ldots + \frac{\partial f_n}{\partial x_n} h_n &= -f_n
\end{aligned} \tag{4.12'}$$

o

$$J\,\mathbf{h} = -\mathbf{f}$$

*Nótese que $| \mathbf{x}^{(1)} - \mathbf{x}^{(0)} | = \sqrt{h^2 + j^2}$

donde las funciones f_i y las derivadas parciales $\partial f_i / \partial x_j$, $i = 1,2, ..., n$; $j = 1,2, ..., n$ están evaluadas en el vector $\mathbf{x}^{(k)}$ y

$$h_i = x_i^{k+1} - x_i^k \qquad 1 \leq i \leq n \tag{4.10'}$$

De donde

$$x_i^{k+1} = x_i^k + h_i \qquad 1 \leq i \leq n \tag{4.11'}$$

o

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

y la matriz de derivadas parciales (matriz jacobiana), ampliada en el vector de funciones queda

$$\left[\begin{array}{cccc|c} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} & -f_1 \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} & -f_2 \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} & -f_n \end{array}\right] \tag{4.12''}$$

o bien

$$[\, J \mid \mathbf{f} \,]$$

Se presenta a continuación un algoritmo para este método.

ALGORITMO 4.2 Método de Newton-Raphson Multivariable

Para encontrar una solución aproximada de un sistema de ecuaciones no lineales $\mathbf{f(x)} = \mathbf{0}$, proporcionar la matriz jacobiana ampliada con el vector de funciones (véase Ec. 4.12") y los

DATOS: El número de ecuaciones N, el vector de valores iniciales **x**, el número máximo de iteraciones MAXIT y el criterio de convergencia EPS.

RESULTADOS: El vector solución **xn** o mensaje "NO CONVERGE".

PASO 1. Hacer K = 1
PASO 2. Mientras K ≤ MAXIT, repetir los pasos 3 a 9.

PASO 3. Evaluar la matriz jacobiana aumentada (4.12").
PASO 4. Resolver el sistema lineal (4.12').
PASO 5. Hacer* **xn = x + h**
PASO 6. Si |**xn - x**| >EPS ir al paso 8. De otro modo continuar.
PASO 7. IMPRIMIR **xn** y TERMINAR.
PASO 8. Hacer **x = xn**
PASO 9. Hacer K = K + 1
PASO 10. IMPRIMIR "NO CONVERGE" Y TERMINAR.

Ejemplo 4.4

Con el algoritmo 4.2, elabore un programa de propósito general para resolver sistemas de ecuaciones no lineales. Luego úselo para resolver el sistema

$$f_1(x_1, x_2, x_3) = 3x_1 - \cos(x_2 x_3) - 0.5 = 0$$

$$f_2(x_1, x_2, x_3) = x_1^2 - 625x_2^2 = 0$$

$$f_3(x_1, x_2, x_3) = e^{-x_1 x_2} + 20x_3 + \frac{(10\pi - 3)}{3} = 0$$

SOLUCIÓN

En el apéndice se presenta el programa 4.1, que consta de los subprogramas GAUSS JORDAN y PIVOTEO, de propósito general; es decir, no dependen del sistema de ecuaciones por resolver.

El usuario deberá escribir el programa principal que llama al subprograma FUNCIONES, donde proporcionará la matriz jacobiana ampliada (Ec. 4.12").

La matriz jacobiana ampliada para el sistema es

$$\left[\begin{array}{ccc|c} 3 & x_3 \operatorname{sen}(x_2 x_3) & x_2 \operatorname{sen}(x_2 x_3) & -3x_1 + \cos(x_2 x_3) + 0.5 \\ 2x_1 & -1250x_2 & 0 & -x_1^2 + 625x_2^2 \\ -x_2 e^{-x_1 x_2} & -x_1 e^{-x_1 x_2} & 20 & -e^{-x_1 x_2} - 20x_3 - \dfrac{10\pi - 3}{3} \end{array}\right]$$

El programa queda finalmente como se muestra en el apéndice (programa 4.1) su ejecución con el vector inicial $[1\ 1\ 1]^T$ produce los siguientes resultados

*Operaciones vectoriales.

k	x_1	x_2	x_3	Distancia
0	1.00000	1.00000	1.00000	
1	0.90837	0.50065	–0.50286	1.5863
2	0.49927	0.25046	–0.51904	0.47982
3	0.49996	0.12603	–0.52045	0.12444
4	0.49998	0.06460	–0.52199	0.61446E-01
5	0.49998	0.03540	–0.52272	0.29214E-01
6	0.49998	0.02335	–0.52302	0.12052E-01
7	0.49998	0.02024	–0.52309	0.31095E-02
8	0.49998	0.02000	–0.52310	0.23879E-03
9	0.49998	0.02000	–0.52310	0.14280E-05

LA SOLUCIÓN DEL SISTEMA ES

$$x_1 = 0.49998176$$
$$x_2 = 0.19999269E\text{-}01$$
$$x_3 = -0.52310085$$

Nótese que en cada iteración se requiere

a) La evaluación de n^2 derivadas parciales
b) La evaluación de n funciones
c) La solución de un sistema de ecuaciones lineales de orden n,

lo que representa una inmensa cantidad de cálculos. Debido a esto, se han elaborado métodos donde los cálculos no son tan numerosos y cuya convergencia es, en general, superior a la del método de punto fijo (superlineal). A continuación se presentan dos de estos métodos, el de Newton-Raphson modificado y el método de Broyden, siendo este último también una modificación del método de Newton-Raphson.

SECCIÓN 3.4 METODO DE NEWTON-RAPHSON MODIFICADO

El método de Newton-Raphson modificado que se describe a continuación consiste en aplicar el método de Newton-Raphson univariable dos veces (para el caso de un sistema de n ecuaciones no lineales en n incógnitas, se aplicará n veces), una para cada variable. Cada que se hace esto, se consideran las otras variables fijas.

Considérese de nuevo el sistema

$$f_1(x, y) = 0$$
$$f_2(x, y) = 0$$

Tomando los valores iniciales x^0, y^0, se calcula a partir del método de Newton-Raphson univariable un nuevo valor x^1 así

$$\boxed{x^1 = x^0 - \frac{f_1(x^0, y^0)}{\partial f_1/\partial x},}$$

$\partial f_1/\partial x$ evaluada en x^0, y^0.

Nótese que se ha obtenido x^1 a partir de f_1 y los valores más recientes de x y y: x^0, y^0.

Ahora úsese f_2 y los valores más recientes de x y y: x^1, y^0 para calcular y^1

$$\boxed{y^1 = y^0 - \frac{f_2(x^1, y^0)}{\partial f_2/\partial y},}$$

donde $\partial f_2/\partial y$ se evalúa en x^1, y^0. Se tiene ahora x^1 y y^1. Con estos valores se calcula x^2, después y^2, y así sucesivamente.

Este método converge a menudo si x^0, y^0 está muy cerca de $\bar{x}, \bar{y}$, y requiere la evaluación de sólo $2n$ funciones por paso (cuatro para el caso de dos ecuaciones que se está manejando).

Nótese que se han empleado desplazamientos sucesivos, pero los desplazamientos simultáneos también son aplicables.

Ejemplo 4.5

Resuelva el sistema

$$f_1(x, y) = x^2 - 10x + y^2 + 8 = 0$$
$$f_2(x, y) = xy^2 + x - 10y + 8 = 0,$$

con el método de Newton-Raphson modificado, usando los valores iniciales $x^0 = 0, y^0 = 0$. Puede seguir los cálculos con un pizarrón electrónico.

SOLUCIÓN

Primero se obtiene

$$\frac{\partial f_1}{\partial x} = 2x - 10 \qquad \text{y} \qquad \frac{\partial f_2}{\partial y} = 2xy - 10$$

Primera iteración

Se evalúan f_1 y $\partial f_1/\partial x$ en $[0,0]^T$:

$$f_1(0,0) = 8$$

y

$$\left.\frac{\partial f_1}{\partial x}\right|_{\substack{x^0\\y^0}} = -10$$

se sustituye

$$x^1 = 0 - \frac{8}{-10} = 0.8$$

Para el cálculo de y^1 se necesita evaluar f_2 y $\partial f_2/\partial y$ en x^1, y^0

$$f_2(0.8,0) = 0.8(0) + 0.8 - 10(0) + 8 = 8.8$$

$$\left.\frac{\partial f_2}{\partial y}\right|_{\substack{x^1\\y^0}} = 2(0.8)(0) - 10 = -10$$

se sustituye

$$y^1 = 0 - \frac{8.8}{-10} = 0.88$$

Segunda iteración

$$f_1(0.8, 0.88) = 1.4144 \qquad \text{y} \qquad \left.\frac{\partial f_1}{\partial x}\right|_{\substack{x^1\\y^1}} = -8.4$$

$$x^2 = 0.8 - \frac{1.4144}{-8.4} = 0.96838$$

Ahora se evalúan f_2 y $\partial f_2/\partial y$ en (x^2,y^1):

$$f_2(0.96838, 0.88) = 0.91829 \qquad \text{y} \qquad \left.\frac{\partial f_2}{\partial y}\right|_{\substack{x^2\\y^1}} = -8.29565$$

de donde

$$y^2 = 0.88 - \frac{0.91829}{-8.29565} = 0.99070$$

Se deja como ejercicio al lector continuar las iteraciones y calcular las distancias entre cada dos vectores consecutivos. Continúe hasta que $x^k \approx 1$ y $y^k \approx 1$. Compare además la velocidad de convergencia de este método con la velocidad de convergencia del método de Newton-Raphson y el de punto fijo para este sistema particular.

En la aplicación de este método se pudo tomar f_2 para evaluar x^1 y f_1 a fin de evaluar y^1, así

$$x^1 = x^0 - \frac{f_2(x^0, y^0)}{\partial f_2/\partial x}$$

$$y^1 = y^0 - \frac{f_1(x^1, y^0)}{\partial f_1/\partial y}$$

Esto puede producir convergencia en alguno de los arreglos y divergencia en el otro. Es posible saber de antemano si la primera o la segunda forma convergirán* para el caso de sistemas de dos ecuaciones, pero cuando $3 \leq n$ las posibilidades son varias ($n!$) y es imposible conocer cuál de estos arreglos tiene viabilidad de convergencia, por lo cual la elección se convierte en un proceso aleatorio. Esta aleatoriedad es la mayor desventaja de este método.

En general, para un sistema de n ecuaciones en n incógnitas: $x_1, x_2, \ldots, x_n$, el algoritmo toma la forma

$$x_i^{k+1} = x_i^k - \frac{f_i(x_1^{k+1}, x_2^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^k, \ldots, x_n^k)}{\frac{\partial f_i}{\partial x_i}\Big|(x_1^{k+1}, x_2^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^k, \ldots, x_n^k)} \quad 1 \leq i \leq n \tag{4.13}$$

ALGORITMO 4.3 Método de Newton-Raphson modificado

Para encontrar una solución aproximada de un sistema de ecuaciónes no lineales **f** (**x**) = **0**, proporcionar las funciones F (I,x) y las derivadas parciales D (I,x) y los

DATOS: El número de ecuaciones N, el vector de valores iniciales **x**, el número máximo de iteraciones MAXIT, el criterio de convergencia EPS y M=0 para desplazamientos sucesivos o M=1 para desplazamientos simultáneos.

RESULTADOS: El vector solución **xn** o mensaje "NO CONVERGE".

*Peter A Stark. Introduction to Numerical Methods. Ed. McMillan.

PASO 1. Hacer K = 1
PASO 2. Mientras K ≤ MAXIT, repetir los pasos 3 a 11.
PASO 3. Si M = 0 hacer* **xaux** = x
PASO 4. Hacer I = 1
PASO 5. Mientras I ≤ N, repetir los pasos 6 y 7.
PASO 6. Si M = 0 hacer
X(I) = X(I)-F(I, x)/D(I, x)
De otro modo hacer
XAUX(I) = X(I) - F(I, x)/D(I, x)
PASO 7. Hacer I = I + 1
PASO 8. Si | **xaux** - x | >EPS ir al paso 10.
De otro modo continuar.
PASO 9. IMPRIMIR x y TERMINAR.
PASO 10. Si M=1 hacer x = **xaux**
PASO 11. Hacer K = K + 1
PASO 12. IMPRIMIR "NO CONVERGE" Y TERMINAR.

* Operaciones vectoriales.

El método siguiente puede saltarse sin pérdida de continuidad.

SECCIÓN 4.5 METODO DE BROYDEN

Considérese ahora la generalización del método de la secante a sistemas multivariables, conocido como el método de Broyden. Según se vio en el capítulo 2, el método de la secante consiste en remplazar $f'(x_k)$ del método de Newton-Raphson

$$x_{k+1} = x_k - [f'(x_k)]^{-1} f(x_k) \tag{4.14}$$

por el cociente:

$$\frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}} \approx f'(x_k),$$

obtenido con los resultados de dos iteraciones previas: x_k y x_{k-1}.

Para ver la modificación o aproximación correspondiente del método de Newton-Raphson multivariable, conviene expresarlo primero en forma congruente con la ecuación 4.14, lo que se logra sustituyendo en la ecuación vectorial (véase Ec. 4.11)

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{h}^{(k)} \tag{4.15}$$

el vector $\mathbf{h}^{(k)}$ que, como se sabe, es la solución del sistema

$$J^{(k)} \mathbf{h}^{(k)} = -\mathbf{f}^{(k)}$$

Al multiplicar esta última ecuación por $(J^{(k)})^{-1}$ se obtiene

$$\mathbf{h}^{(k)} = -(J^{(k)})^{-1}\ \mathbf{f}^{(k)} \tag{4.17}$$

y al remplazar la ecuación 4.17 en la 4.15 se llega a

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - (J^{(k)})^{-1}\ \mathbf{f}^{(k)}, \tag{4.18}$$

la ecuación correspondiente a la 4.14 para $n > 1$.

El método de la secante para sistemas de ecuaciones no lineales consiste en sustituir $J^{(k)}$ en la ecuación 4.18 con una matriz $A^{(k)}$, cuyos componentes se obtienen con los resultados de dos iteraciones previas $\mathbf{x}^{(k)}$ y $\mathbf{x}^{(k-1)}$, de la siguiente manera*

$$A^{(k)} = A^{(k-1)} + \frac{[\mathbf{f}(\mathbf{x}^{(k)}) - \mathbf{f}(\mathbf{x}^{(k-1)}) - A^{(k-1)}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})](\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})^T}{|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}|^2} \tag{4.19}$$

o bien

$$A^{(k)} = A^{(k-1)} + \frac{[\Delta\mathbf{f}^{(k)} - A^{(k-1)}\Delta\mathbf{x}^{(k)}](\Delta\mathbf{x}^{(k)})^T}{|\Delta\mathbf{x}^{(k)}|^2} \tag{4.20}$$

con la notación

$$\Delta\mathbf{f}^{(k)} = \mathbf{f}(\mathbf{x}^{(k)}) - \mathbf{f}(\mathbf{x}^{(k-1)})$$

$$\Delta\mathbf{x}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}$$

Para la primera aplicación de la ecuación 4.20 se requieren dos vectores iniciales: $\mathbf{x}^{(0)}$ y $\mathbf{x}^{(1)}$. Este último puede obtenerse de una aplicación del método de Newton-Raphson multivariable

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - (J^{(0)})^{-1}\ \mathbf{f}^{(0)},$$

cuya $J^{(0)}$ a su vez puede emplearse en 4.20, con lo cual ésta queda

$$A^{(1)} = J^{(0)} + \frac{(\Delta\mathbf{f}^{(1)} - J^{(0)}\Delta\mathbf{x}^{(1)})(\Delta\mathbf{x}^{(1)})^T}{|\Delta\mathbf{x}^{(1)}|^2} \tag{4.21}$$

La inversión de $A^{(k)}$ en cada iteración significa un esfuerzo computacional grande (del orden de n^3) que, sin embargo, puede reducirse empleando una fórmula de inversión matricial de Sherman y Morrison**. Esta fórmula establece que si A es

---

*Dennis, J.E. Jr and J.J More (1977), "Quasi-Newton Methods, Motivation and Theory". *SIAM Review*, 19, No. 1 (46-89).

**ibid.

una matriz no singular y $\mathbf{x}$ y $\mathbf{y}$ son vectores, entonces $A + \mathbf{x}\,\mathbf{y}^T$ es no singular, siempre que $\mathbf{y}^T A^{-1}\,\mathbf{x} \neq 1$. Además, en este caso,

$$(A + \mathbf{x}\mathbf{y}^T)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{x}\mathbf{y}^T A^{-1}}{1 + \mathbf{y}^T A^{-1}\mathbf{x}} \tag{4.22}$$

Esta fórmula permite calcular $(A^{(k)})^{-1}$ a partir de $(A^{(k-1)})^{-1}$, eliminando la necesidad de invertir una matriz en cada iteración. Para esto, primero se obtiene la inversa de la ecuación 4.20

$$(A^{(k)})^{-1} = \left(A^{(k-1)} + \frac{\Delta\mathbf{f}^{(k)} - A^{(k-1)}\,\Delta\mathbf{x}^{(k)}}{|\,\Delta\mathbf{x}^{(k)}\,|^2}\,(\Delta\mathbf{x}^{(k)})^T \right)^{-1}$$

Después se hace

$$A = A^{(k-1)}$$

$$\mathbf{x} = \frac{(\Delta\mathbf{f}^{(k)} - A^{(k-1)}\,\Delta\mathbf{x}^{(k)})}{|\,\Delta\mathbf{x}^{(k)}\,|^2}$$

y

$$\mathbf{y} = \Delta\mathbf{x}^{(k)},$$

con lo que la última ecuación queda

$$(A^{(k)})^{-1} = (A + \mathbf{x}\,\mathbf{y}^T)^{-1}$$

y sustituyendo la ecuación 4.22

$$(A^{(k)})^{-1} = (A^{(k-1)})^{-1} - \frac{(A^{(k-1)})^{-1}\left[\dfrac{\Delta\mathbf{f}^{(k)} - A^{(k-1)}\,\Delta\mathbf{x}^{(k)}}{|\,\Delta\mathbf{x}^{(k)}\,|^2}\,(\Delta\mathbf{x}^{(k)})^T\right](A^{(k-1)})^{-1}}{1 + (\Delta\mathbf{x}^{(k)})^T (A^{(k-1)})^{-1}\,\dfrac{\Delta\mathbf{f}^{(k)} - A^{(k-1)}\,\Delta\mathbf{x}^{(k)}}{|\,\Delta\mathbf{x}^{(k)}\,|^2}}$$

$$= (A^{(k-1)})^{-1} - \frac{[(A^{(k-1)})^{-1}\,\Delta\mathbf{f}^{(k)} - \Delta\mathbf{x}^{(k)}]\,(\Delta\mathbf{x}^{(k)})^T (A^{(k-1)})^{-1}}{|\,\Delta\mathbf{x}^{(k)}\,|^2 + (\Delta\mathbf{x}^{(k)})^T (A^{(k-1)})^{-1}\,\Delta\mathbf{f}^{(k)} - |\,\Delta\mathbf{x}^{(k)}\,|^2}$$

$$\boxed{(A^{(k)})^{-1} = (A^{(k-1)})^{-1} + \frac{[\Delta\mathbf{x}^{(k)} - (A^{(k-1)})^{-1}\,\Delta\mathbf{f}^{(k)}]\,(\Delta\mathbf{x}^{(k)})^T (A^{(k-1)})^{-1}}{(\Delta\mathbf{x}^{(k)})^T (A^{(k-1)})^{-1}\,\Delta\mathbf{f}^{(k)}}} \tag{4.23}$$

Esta fórmula permite calcular la inversa de una matriz con sumas y multiplicaciones de matrices solamente, con lo que se reduce el esfuerzo computacional al orden n^2.

Ejemplo 4.6

Use el método de Broyden para encontrar una solución aproximada del sistema

$$f_1(x, y) = x^2 - 10x + y^2 + 8 = 0$$
$$f_2(x, y) = xy^2 + x - 10y + 8 = 0,$$

tome como vector inicial: $[x^0, y^0]^T = [0, 0]^T$. Se recomienda especialmente emplear un pizarrón electrónico para llevar los cálculos y así poner la atención en el algoritmo y en el análisis de los resultados.

SOLUCIÓN

En el ejemplo 4.3 se encontró una solución aproximada de este sistema, empleando el método de Newton-Raphson y el vector cero como vector inicial. Con los resultados de la primera iteración del ejemplo 4.3

$$J^{(0)} = \begin{bmatrix} -10 & 0 \\ 1 & -10 \end{bmatrix}, \quad (J^{(0)})^{-1} = \begin{bmatrix} -0.1 & 0 \\ -0.01 & -0.1 \end{bmatrix} \quad \mathbf{x}^{(1)} = \begin{bmatrix} 0.8 \\ 0.88 \end{bmatrix}$$

se calcula $(A^{(1)})^{-1}$ con la ecuación 4.23

$$(A^{(1)})^{-1} = (J^{(0)})^{-1} + \frac{(\Delta \mathbf{x}^{(1)} - (J^{(0)})^{-1} \Delta \mathbf{f}^{(1)})(\Delta \mathbf{x}^{(1)})^T (J^{(0)})^{-1}}{(\Delta \mathbf{x}^{(1)})^T (J^{(0)})^{-1} \Delta \mathbf{f}^{(1)}}$$

$$(A^{(1)})^{-1} = \begin{bmatrix} -0.1 & 0 \\ -0.01 & -0.1 \end{bmatrix} + \frac{\begin{bmatrix} .8 \\ .88 \end{bmatrix} - \begin{bmatrix} -.1 & 0 \\ -.01 & -.1 \end{bmatrix} \begin{bmatrix} -6.5856 \\ -7.38048 \end{bmatrix} \begin{bmatrix} .8 \\ .88 \end{bmatrix}^T \begin{bmatrix} -.1 & 0 \\ -.01 & -.1 \end{bmatrix}}{\begin{bmatrix} .8 \\ .88 \end{bmatrix}^T \begin{bmatrix} -.1 & 0 \\ -.01 & -.1 \end{bmatrix} \begin{bmatrix} -6.5856 \\ -7.38048 \end{bmatrix}}$$

$$= \begin{bmatrix} -0.11015 & -0.010079 \\ -0.01546 & -0.105404 \end{bmatrix}$$

se calcula ahora $\mathbf{x}^{(2)}$ empleando la ecuación:

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} - (A^{(1)})^{-1} \mathbf{f}^{(1)}$$

$$= \begin{bmatrix} .8 \\ .88 \end{bmatrix} - \begin{bmatrix} -0.11015 & -0.010079 \\ -0.01546 & -0.105404 \end{bmatrix} \begin{bmatrix} 1.4144 \\ 0.61952 \end{bmatrix}$$

$$= \begin{bmatrix} 0.96208 \\ 0.9672 \end{bmatrix}$$

Para la segunda iteración se utilizarán las ecuaciones

$$(A^{(2)})^{-1} = (A^{(1)})^{-1} + \frac{[\Delta \mathbf{x}^{(2)} - (A^{(1)})^{-1}\Delta \mathbf{f}^{(2)}]\,(\Delta \mathbf{x}^{(2)})^T (A^{(1)})^{-1}}{(\Delta \mathbf{x}^{(2)})^T (A^{(1)})^{-1}\Delta \mathbf{f}^{(2)}}$$

y

$$\mathbf{x}^{(3)} = \mathbf{x}^{(2)} - (A^{(2)})^{-1}\mathbf{f}^{(2)}$$

Al sustituir valores se obtiene

$$\mathbf{x}^{(3)} = \begin{bmatrix} 0.997433 \\ 0.996786 \end{bmatrix}$$

La continuación de las iteraciones da

$$\mathbf{x}^{(4)} = \begin{bmatrix} 0.9999037 \\ 0.9998448 \end{bmatrix}, \qquad \mathbf{x}^{(5)} = \begin{bmatrix} 0.999998157 \\ 0.999996667 \end{bmatrix}$$

$$\mathbf{x}^{(6)} = \begin{bmatrix} 0.9999999849 \\ 0.9999999722 \end{bmatrix}, \qquad \mathbf{x}^{(7)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

que es la solución del sistema, tal como se obtuvo en los ejemplos 4.2 y 4.3.

A continuación se presenta el algoritmo para este método.

ALGORITMO 4.4 Método de Broyden

Para encontrar una solución aproximada de un sistema de ecuaciones no lineales **f** (x) = 0, proporcionar la matriz Jacobiana ampliada con el vector de funciones (véase Ec. 4.12") y los

DATOS: Número de ecuaciones N, dos vectores de valores iniciales: **x0** y **x1**, el número máximo de iteraciones MAXIT, y el criterio de convergencia EPS.

RESULTADOS: Una aproximación a una solución: **xn** o el mensaje "NO CONVERGE".

PASO 1. Calcular *AK*, la matriz inversa de la matriz Jacobiana evaluada en **x0**.

PASO 2. Hacer K = 1.

PASO 3. Mientras K ≤ MAXIT, repetir los pasos 4 a 10.

PASO 4. Calcular **f0** y **f1**, el vector de funciones evaluado en **x0** y **x1**, respectivamente.
PASO 5. Calcular (*) **dx = x1 − x0; df = f1 − f0.**
PASO 6. Calcular *AK1*, la matriz que aproxima a la inversa de la matriz Jacobiana (4.18), con la ecuación (4.23), usando como $(A^{(k+1)})^{-1}$ a *AK*
PASO 7. Calcular (*) **xn = x1 − *AK1* * f1**
PASO 8. (*) Si | **xn − x1** | ≤ EPS ir al paso 11. De otro modo continuar.
PASO 9. Hacer (*) **x0 = x1; x1 = xn;** ***AK = AK1*** (actualización de **x0**, **x1** y *AK*).
PASO 10. Hacer K = K + 1
PASO 11. Si K ≤ MAXIT, IMPRIMIR el vector **xn** y TERMINAR. De otro modo IMPRIMIR "NO CONVERGE" y TERMINAR.

(*) Operaciones matriciales.

SECCIÓN 4.6 ACELERACIÓN DE CONVERGENCIA

Al igual que en los capítulos anteriores, una vez que se tienen métodos de solución funcionales, se mejorarán o crearán nuevos algoritmos usando dicho conocimiento. También, como ya se ha visto, esto se logra con un proceso de generalización o abstracción. Se procederá en esa dirección enseguida.

En cada iteración de los algoritmos vistos se parte de un vector $\mathbf{x}^{(k)}$, que ahora se llamará punto base; desde ese punto se camina en una dirección, dada por un vector, que se denominará **dirección de exploración.** Considérese la figura 4.2 y el punto base $(x^0, y^0)^*$ = (2,2). Si desde el punto base se camina en la dirección del vector $\mathbf{d}^0 = [4,1]^T$, se terminará pasando por el punto P(6,3).

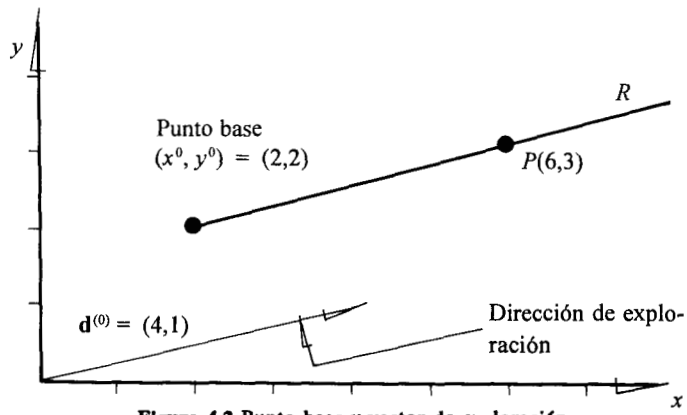


Figura 4.2 Punto base y vector de exploración

*De aquí en adelante se usará indistintamente *n*-adas ordenadas $(x_1, x_2, ..., x_n)$ para representar un vector de *n* elementos y un punto en el espacio *n*-dimensional.

Al avanzar en cierta dirección de exploración a partir de un punto base, se llega a un nuevo punto que va a ser base para la siguiente iteración, pudiera ser el punto P(6,3) o cualquier otro punto de R que dará la ecuación vectorial

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + t\,\mathbf{d}^{(0)}$$

o en forma mas general

$$\boxed{\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t\,\mathbf{d}^{(k)},} \tag{4.24}$$

donde t es el factor de tamaño de la etapa y determina la distancia del desplazamiento en la dirección especificada. Esta ecuación se obtiene fácilmente por la suma de vectores en el plano, como se muestra en la figura 4.3.

Para aclarar esta generalización, se identifica el algoritmo de Newton-Raphson para sistemas de dos ecuaciones no lineales con la ecuación 4.24.

Primero se reescribe la ecuación 4.9

$$\frac{\partial f_1}{\partial x}(x^{k+1} - x^k) + \frac{\partial f_1}{\partial y}(y^{k+1} - y^k) = -f_1(x^k, y^k)$$

$$\frac{\partial f_2}{\partial x}(x^{k+1} - x^k) + \frac{\partial f_2}{\partial y}(y^{k+1} - y^k) = -f_2(x^k, y^k)$$

para pasarla a notación matricial como sigue

$$\begin{bmatrix} \dfrac{\partial f_1}{\partial x} & \dfrac{\partial f_1}{\partial y} \\ \dfrac{\partial f_2}{\partial x} & \dfrac{\partial f_2}{\partial y} \end{bmatrix} \begin{bmatrix} x^{k+1} - x^k \\ y^{k+1} - y^k \end{bmatrix} = - \begin{bmatrix} f_1(x^k, y^k) \\ f_2(x^k, y^k) \end{bmatrix}$$

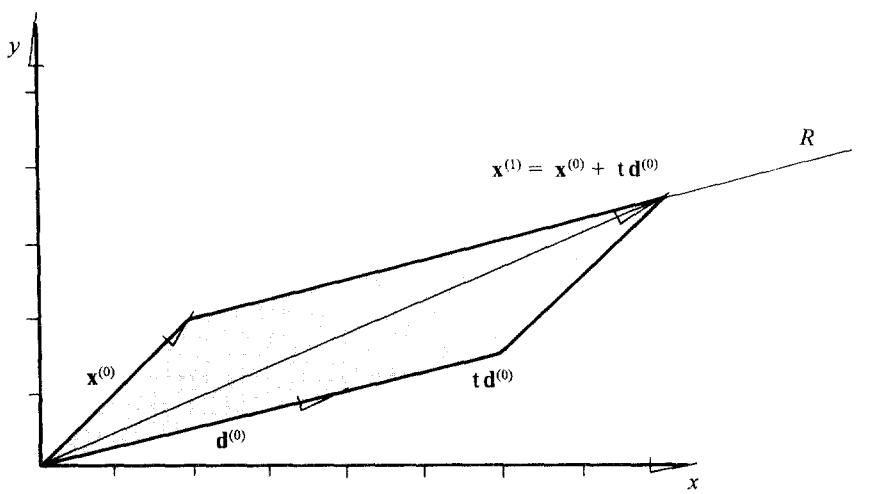


Figura 4.3. Suma de vectores en el plano.

que ahora, multiplicada por la inversa de la matriz jacobiana, llega a la forma

$$-\begin{bmatrix} \dfrac{\partial f_1}{\partial x} & \dfrac{\partial f_1}{\partial y} \\ \dfrac{\partial f_2}{\partial x} & \dfrac{\partial f_2}{\partial y} \end{bmatrix}^{-1} \begin{bmatrix} f_1(x^k, y^k) \\ f_2(x^k, y^k) \end{bmatrix} = \begin{bmatrix} x^{k+1} - x^k \\ y^{k+1} - y^k \end{bmatrix}$$

o también

$$\begin{bmatrix} x^{k+1} \\ y^{k+1} \end{bmatrix} = \begin{bmatrix} x^k \\ y^k \end{bmatrix} - \begin{bmatrix} \dfrac{\partial f_1}{\partial x} & \dfrac{\partial f_1}{\partial y} \\ \dfrac{\partial f_1}{\partial x} & \dfrac{\partial f_2}{\partial y} \end{bmatrix}^{-1} \begin{bmatrix} f_1(x^k\, y^k) \\ f_2(x^k\, y^k) \end{bmatrix}$$

y en esta última forma, ya como ecuación vectorial, se tiene la identificación total con la ecuación 4.24, con

$$\mathbf{x}^{(k+1)} = \begin{bmatrix} x^{k+1} \\ y^{k+1} \end{bmatrix}; \quad \mathbf{x}^{(k)} = \begin{bmatrix} x^k \\ y^k \end{bmatrix}; \quad t = -1 \quad \mathbf{d}^{(k)} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x} & \dfrac{\partial f_1}{\partial y} \\ \dfrac{\partial f_2}{\partial x} & \dfrac{\partial f_2}{\partial y} \end{bmatrix}^{-1} \begin{bmatrix} f_1(x^k, y^k) \\ f_2(x^k, y^k) \end{bmatrix}$$

Nótese que en el método de Newton-Raphson, el factor de tamaño de la etapa es constante en todos los pasos iterativos del proceso y que $\mathbf{d}^{(k)}$ el vector de exploración, es el resultado de multiplicar la inversa de la matriz jacobiana por el vector de funciones.

Método de Newton Raphson con optimización de *t*

Con la ecuación 4.24 puede estudiarse cómo mejorar los métodos disponibles; por ejemplo, se puede ver que en el algoritmo de Newton-Raphson, tomar distintos valores de t llevaría a distintos vectores $\mathbf{x}^{(k+1)}$, alguno más cercano a la raíz $\mathbf{x}$ que los demás (véase Fig. 4.4). La mejora es optimizar el valor de t en el método de Newton-Raphson.

Para ejemplificar, tómense los valores de la primera iteracción del ejemplo 4.3

$$k = 0; \qquad x^k = 0; \qquad y^k = 0; \qquad h = 0.8 \qquad j = 0.88$$

de aquí: $\mathbf{d}^{(k)} = [-0.8 \; -0.88]^T$

y la ecuación 4.25 queda

$$\boxed{\begin{aligned} x^{k+1} &= x^k + t\ d_1^k \\ y^{k+1} &= y^k + t\ d_2^k \end{aligned}}$$

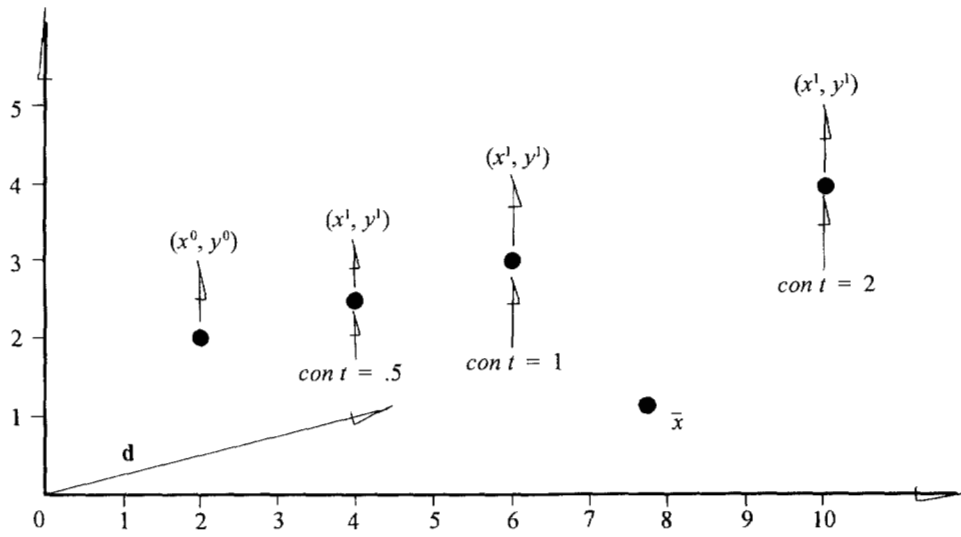


Figura 4.4 Influencia de t en el vector $\mathbf{x}^{(k+1)}$

Se enlista ahora una serie de valores de t y los correspondientes valores de $\mathbf{x}^{(k+1)}$:

t	x^{k+1}	y^{k+1}
–0.50	0.4	0.44
–0.75	0.6	0.66
–1.00	0.8	0.88
–1.25	1.0	1.10
–1.50	1.2	1.32

Para determinar cuál de los $\mathbf{x}^{(k+1)}$ está más cerca de la raíz $\mathbf{x}$, se desarrolla un nuevo criterio de convergencia o avance sustentado en la definición de **residuo** de una función $f(x, y)$, dada esta última así:

El residuo de una función $f(x,y)$ en un punto (x^k, y^k) es el valor de f en (x^k, y^k).

Así, en el sistema

$$f_1(x, y) = x^2 + y^2 - 4 = 0$$
$$f_2(x, y) = y - x^2 = 0,$$

en el punto (1,1) los residuos son

$$f_1(1,1) = 1^2 + 1^2 - 4 = -2$$

y

$$f_2(1,1) = 1 - 1^2 = 0$$

En general, el valor de la **función suma de residuos al cuadrado**

$$z_k = f_1^2\,(x^k, y^k) + f_2^2\,(x^k, y^k) \tag{4.26}$$

será indicativa de la cercanía de $\mathbf{x}^{(k)}$ con la raíz **x**.

Con la aplicación de este concepto a los distintos vectores $\mathbf{x}^{(k+1)}$ obtenidos arriba, se tiene

Para $t = -0.5$

$$z_{k+1} = [0.4^2 - 10(0.4) + 0.44^2 + 8]^2 + [0.4(0.44)^2 + 0.4 - 10(0.44) + 8]^2 = 35.57$$

Para $t = -0.75$: $z_{k+1} = 12.93$

Para $t = -1.0$: $z_{k+1} = 2.38$

Para $t = -1.25$: $z_{k+1} = 0.67$

Para $t = -1.5$: $z_{k+1} = 4.31$

De donde $\mathbf{x}^{k+1}$ correspondiente a $t = -1.25$ resulta ser el más cercano a la raíz $\mathbf{x} = [1,1]^T$.

Los valores propuestos de t anteriormente, se eligieron de manera arbitraria alrededor de -1 y aunque el valor de -1.25 es el mejor de ellos, no es el óptimo de todos los valores posibles para la primera iteración.

A continuación se da una forma de seleccionar los valores de t .

Se selecciona un intervalo de búsqueda $[a, b]$, se calculan valores de t dentro de ese intervalo de la siguiente manera:

$$t = a + \frac{b - a}{F} \qquad \text{y} \qquad t = b - \frac{b - a}{F}$$

donde F son los términos de la serie de Fibonacci

$$F = 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, \ldots$$

Para cada valor de t se calcula su correspondiente z_{k+1} y el valor mínimo de z_{k+1} proporcionará el valor óptimo de t.

Así, seleccionando el intervalo $[-1,-1.2]$, el valor mínimo de z_{k+1} $(= 0.4578)$ corresponde al valor óptimo de t $(= -1.184)$ en la primera iteración de la solución del ejemplo 4.3. Una vez encontrado el valor óptimo de t se toma el vector $\mathbf{x}^{(1)}$ correspondiente y se calcula $\mathbf{d}^{(2)}$ para proceder a optimizar el valor de t en la segunda iteración

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + t\,\mathbf{d}^{(1)}$$

Ejemplo 4.7

Modifique el programa del ejemplo 4.4 para incluir la optimización de t. Utilizando el programa resultante, resuelva el sistema del ejemplo 4.4.

SOLUCIÓN

Las modificaciones consisten en

— Elaborar un subprograma para encontrar el valor de t que minimice la función z_k, utilizando la búsqueda de Fibonacci.
— Modificar el subprograma NEWTON del ejemplo 4.4 para utilizar ahora como criterio de convergencia o avance la función z_k y la llamada a el subprograma de búsqueda de Fibonacci.

En el disco (programa 4.2) se muestran los subprogramas NEWOPT y BUSCA resultantes. El programa principal y los subprogramas SIMULT y PIVOTEO no sufren cambio alguno.

Con el programa resultante y con los valores iniciales

$$\mathbf{x}^{(0)} = [1\ 1\ 1]^T$$

se obtuvieron los siguientes resultados

VARI	1	1.00000	1.00000	1.00000
FUNC	1	1.95970	−624.00000	29.83985
SUMA		.39027E+06	TOPT =	1.833
VARI	2	.83201	.08453	−1.75525
FUNC	2	1.00701	−3.77371	−24.70092
SUMA		.62539E+03	TOPT =	.9000
VARI	3	.53770	.04775	−.64629
FUNC	3	.11359	−1.13613	−2.47923
SUMA		.74503E+01	TOPT =	.9000
VARI	4	.50380	.03001	−.53527
FUNC	4	.01153	− .30917	−.24846
SUMA		.15745E+00	TOPT =	1.167
VARI	5	.49935	.02028	−.52103
FUNC	5	−.00190	−.00767	.04138
SUMA		.17748E-02	TOPT =	.9000
VARI	6	.49992	.02003	− .52289
FUNC	6	−.00019	−.00081	.00414
SUMA		.17817E-04	TOPT =	.9000
VARI	7	.49998	.02000	−.52308
FUNC	7	−.00002	−.00008	.00041
SUMA		.17825E-06	TOPT =	.9000

LA SOLUCION DEL SISTEMA ES:

X(1) = .49998116
X(2) = .19999571E-01
X(3) = -.52309883

Obsérvense los valores de TOPT en las diferentes iteraciones.

Método de Newton-Raphson modificado con optimización de t (método SOR)

En el método de Newton-Raphson modificado, la expresión general 4.13 puede identificarse con la ecuación 4.24 directamente con $t = -1$

$$\text{y} \quad d_i^k = \frac{f_i(x_1^{k+1}, x_2^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^k, \ldots, x_n^k)}{\frac{\partial f_i}{\partial x_i} \Big| (x_1^{k+1}, x_2^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^k, \ldots, x_n^k)} \quad 1 \leq i \leq n$$

Con la optimización del valor de t en cada iteración puede acelerarse la convergencia. El método así obtenido

$$x_i^{k+1} = x_i^k - t \frac{f_i(x_1^{k+1}, x_2^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^k, \ldots, x_n^k)}{\frac{\partial f_i}{\partial x_i} \Big| (x_1^{k+1}, x_2^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^k, \ldots, x_n^k)} \quad 1 \leq i \leq n \tag{4.27}$$

se conoce como **método SOR** para sistemas no lineales. A continuación se resuelve un ejemplo con optimización de t.

Ejemplo 4.8

Resuelva el siguiente sistema de ecuaciones empleando el método SOR para sistemas no lineales.

$$f_1(x, y) = x^2 - 10x + y^2 + 8 = 0$$
$$f_2(x, y) = xy^2 + x - 10y + 8 = 0$$

Sean los valores iniciales $x^0 = 0$ y $y^0 = 0$

Sugerencia: Puede seguir los cálculos usando Math-CAD o un pizarrón elecrónico disponible.

SOLUCIÓN

Primero se obtiene

$$\frac{\partial f_1}{\partial x} = 2x - 10 \qquad \text{y} \qquad \frac{\partial f_2}{\partial y} = 2xy - 10$$

Primera iteración

Se evalúa f_1 y $\frac{\partial f_1}{\partial x}$ en $[\,0,\,0\,]^T$

$$f_1(0,0) = 8, \qquad \left.\frac{\partial f_1}{\partial x}\right|_{\substack{x^0\\y^0}} = -10$$

Se elige el intervalo de búsqueda $[-1.5,-0.5]$ y $t = b-(b-a)/F$ y el primer valor a prueba es

$t = -0.5$

$$x^1 = x^0 + t\frac{f_1(0,0)}{\left.\frac{\partial f_1}{\partial x}\right|_{(0,0)}} = 0 - .5\left(\frac{8}{-10}\right) = 0.4$$

$$f_2(0.4,0) = 8.4, \qquad \left.\frac{\partial f_2}{\partial y}\right|_{\substack{x^1\\y^0}} = -10$$

$$y^1 = y^0 + t\frac{f_2(0.4,0)}{\left.\frac{\partial f_2}{\partial y}\right|_{(0.4,0)}} = 0 - .5\left(\frac{8.4}{-10}\right) = 0.42$$

A partir del criterio de la suma de los residuos elevados al cuadrado, se tiene

$$z_1 = f_1^2\,(0.4,0.42) + f_2^2\,(0.4,0.42)$$
$$= [0.4^2 - 10(0.4) + 0.42^2 + 8]^2 + [0.4(0.42)^2 + 0.4 - 10(0.42) + 8] = 37.042$$

El segundo valor a prueba es $t = -1.0$, con lo que se obtiene

$x^1 = 0.8$ $\qquad$ $y^1 = 0.88$ $\qquad$ y $\qquad$ $z_1 = 2.3843$

Al continuar el proceso de búsqueda se tiene

t	x^1	y^1	z_1
-0.5	0.4	0.42	37.04204
-1.0	0.8	0.88	2.38433
-1.1666	0.9333	1.0422	0.61508
-1.3	1.04	1.1752	1.63124

Ahora se usa $t = a + (b\text{-}a)/F$ y se obtiene

t	x^1	y^1	z_1
–1.5	1.2	1.38	5.78774
–1.0	0.8	0.88	2.38433
–0.8333	0.6666	0.7222	8.49907

Por lo tanto, el valor óptimo de t es –1.1666 y los valores correspondientes de $[x^1, y^1]$ = [0.9333,1.0422] se toman como resultados finales de la primera iteración.

Segunda iteración

Con el mismo intervalo de búsqueda [–.5,–1.5] se tiene, con $t = b - (b\text{-}a)/F$

t	x^2	y^2	z_2
–0.5	0.971674	1.017439	0.107707
–1.0	1.010048	1.0023179	0.005725
–1.1666	1.022840	0.9947013	0.036074

y con $t = a + (b\text{-}a)/F$

t	x^2	y^2	z_2
–1.5	1.048423	0.997131	0.166988
–1.0	1.010048	1.0023179	0.005725
–0.8333	0.997257	1.006269	0.004287
–0.7	0.987024	1.010217	0.027137

El valor óptimo de t es –0.8333 y los valores correspondientes de $[x^2, y^2]$ = [0.997257,1.006269] se toman como resultados finales de la segunda iteración.

Al continuar el proceso iterativo se obtienen los siguientes valores

k	x^k	y^k	z_k	t_{opt}
0	0.00000	0.0000		
1	0.93333	1.0422	0.61508	–1.1666
2	0.997257	1.006269	0.004287	–0.8333
3	1.0008512	1.0012183	0.00008376	–0.8333
4	1.000304735	1.000076027	0.000005224	–1.0
5	1.000018997	1.000004749	0.000000047	–1.0

Método del descenso de máxima pendiente

Se ha visto cómo seguir un camino que permita ir disminuyendo z al optimizar el tamaño del paso t de un método conocido. Sin embargo, puede elaborarse un método de solución de (4.2), construyendo primero una dirección de exploración **d** que permita disminuir el valor de z en una cantidad localmente máxima y, una vez encontrada, buscar la t óptima en esa dirección. Para el desarrollo de este algoritmo, son necesarias las siguientes consideraciones.

La figura 4.5 representa una función arbitraria $z(x,y)$ por medio de sus curvas de nivel (conjunto de parejas (x,y), para las cuales z tiene el mismo valor). En el punto A, siguiendo perpendicularmente a la curva (dirección **d**), se obtiene la forma más rápida de disminuir el valor de z. Obsérvese que en cualquier dirección comprendida entre la de **d** y la recta tangente a la curva en el punto A (por ejemplo la dirección q) también disminuiría z, aunque no tan rápidamente.

Para calcular la dirección **d** del descenso de máxima pendiente en un punto, se recurre a la definición de gradiente y sus propiedades, estudiadas en el cálculo diferencial de varias variables. Como primer paso se recordará la definición de gradiente de una función escalar.

Sea $f = f(x,y)$ una función escalar* dada, con todas sus derivadas parciales de primer orden continuas, entonces el gradiente de f, denotado por grad f, existe y está dado por el vector

$$\mathbf{grad}\, f = \frac{\partial f}{\partial x}\,\mathbf{i} + \frac{\partial f}{\partial y}\,\mathbf{j},$$

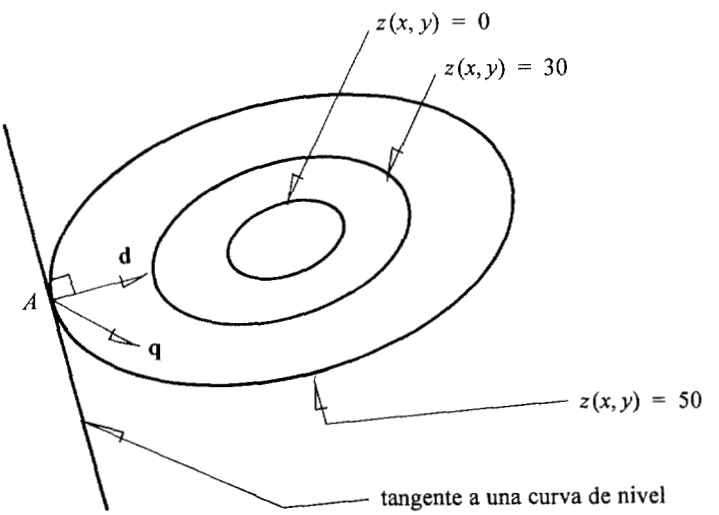


Figura 4.5. Curvas de nivel de una función arbitraria $z(x,y)$.

*$f(x,y)$ es una función escalar si a cada pareja ordenada de números reales (x,y), se asocia uno y sólo un número real: $f(x, y)$.

donde los vectores **i** y **j** son de la forma

$$\mathbf{i} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \mathbf{j} = \begin{bmatrix} 0 \\ 1 \end{bmatrix};$$

al sustituir se tiene

$$\mathbf{grad}\, f = \frac{\partial f}{\partial x}\begin{bmatrix} 1 \\ 0 \end{bmatrix} + \frac{\partial f}{\partial y}\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

y por último

$$\mathbf{grad}\, f = \begin{bmatrix} \dfrac{\partial f}{\partial x} \\ \dfrac{\partial f}{\partial y} \end{bmatrix}$$

Este vector **grad** f tiene, entre otras, las siguientes propiedades.

Sean D el dominio de definición de f y P cualquier punto de D sobre una curva de nivel N de f. Si el **grad** f en P existe y es distinto de cero, entonces

a) El vector **grad** f en P tiene la dirección en que f crece con máxima rapidez, y
b) Es perpendicular a N en P; esto es, tiene la dirección de la normal a N en P.

Por ejemplo, las curvas de nivel f = constante = c de

$$f(x, y) = \ln(x^2 + y^2)$$

son círculos con centro en el origen. El gradiente

$$\mathbf{grad}\, f = \begin{bmatrix} \dfrac{\partial f}{\partial x} \\ \dfrac{\partial f}{\partial y} \end{bmatrix} = \begin{bmatrix} \dfrac{2x}{x^2 + y^2} \\ \dfrac{2y}{x^2 + y^2} \end{bmatrix},$$

tiene la dirección de las normales a los círculos, y su dirección corresponde a la de máximo aumento de f. Así, en el punto P(2,1) sobre la curva del nivel con $c = \ln(2^2 + 1^2) = 1.6$ (círculo con centro en el origen de radio 2.236), se tiene

$$\mathbf{grad}\, f = \begin{bmatrix} 0.8 \\ 0.4 \end{bmatrix}$$

Con esta definición de gradiente y sus propiedades, se retoma el asunto del cálculo de la dirección que asegura la disminución de $z = f_1^2(\mathbf{x}) + f_2^2(\mathbf{x})$ (función escalar de x y y), en una cantidad localmente máxima en un punto. Se determina el vector gradiente de z con signo negativo en dicho punto (el signo negativo se debe a que se quiere que z disminuya). El vector gradiente de z se representa por ∇z. Por tanto, la dirección de descenso más brusco es

$$\mathbf{d} = -(\nabla z)$$

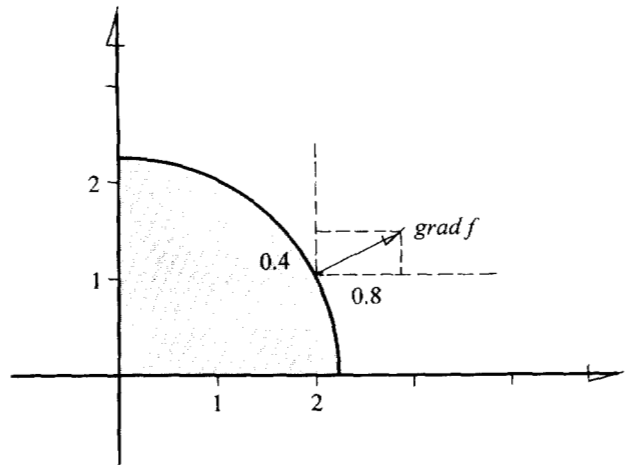


Figura 4.6. Perpendicularidad del vector gradiente a las curvas de nivel.

con cada uno de los componentes de **d** calculados como

$$d_1 = \frac{\partial z}{\partial x}, \qquad d_2 = \frac{\partial z}{\partial y}$$

Ejemplo 4.9

Obtenga la dirección del descenso de máxima pendiente del sistema

$$f_1(x_1, x_2, x_3) = 3x_1 - \cos(x_2 x_3) - 0.5 = 0$$
$$f_2(x_1, x_2, x_3) = x_1^2 - 625x_2^2 = 0$$
$$f_3(x_1, x_2, x_3) = e^{-x_1 x_2} + 20x_3 + (10\pi - 3)/3 = 0$$

use como vector inicial a

$$\mathbf{x}^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

SOLUCIÓN

$$z = [3x_1 - \cos(x_2 x_3) - 0.5]^2 + [x_1^2 - 625x_2^2]^2 + [e^{-x_1 x_2} + 20x_3 + (10\pi - 3)/3]^2$$

$$d_1 = \frac{\partial z}{\partial x_1} = 6(3x_1 - \cos(x_2 x_3) - 0.5) + 4x_1(x_1^2 - 625x_2^2)$$
$$-2x_2 e^{-x_1 x_2}[e^{-x_1 x_2} + 20x_3 + (10\pi - 3)/3]$$

$$d_2 = \frac{\partial z}{\partial x_2} = 2x_3 \operatorname{sen}(x_2 x_3)[3x_1 - \cos(x_2 x_3) - 0.5]$$
$$-2500x_2(x_1^2 - 625x_2^2) - 2x_1 e^{-x_1 x_2}[e^{-x_1 x_2} + 20x_3 + (10\pi - 3)/3]$$

$$d_3 = \frac{\partial z}{\partial x_3} = 2x_2 \operatorname{sen}(x_2 x_3)\,[\,3x_1 - \cos(x_2 x_3) - 0.5\,] +$$

$$40[\,e^{-x_1 x_2} + 20x_3 + (10\pi - 3)/3]$$

Al evaluar d_1, d_2 y d_3 en $\mathbf{x}^{(0)}$ se obtiene

$$d_1 = -9.0$$

$$d_2 = 0.0$$

$$d_3 = 418.87872$$

y entonces el vector dirección es

$$\mathbf{d} = - \begin{bmatrix} -9 \\ 0.00 \\ 418.87872 \end{bmatrix}$$

Una vez calculada la dirección, se utiliza una exploración unidimensional para localizar el mínimo en esta dirección (por ejemplo una búsqueda de Fibonacci). Ya localizado el mínimo, se calcula una nueva dirección de descenso de máxima pendiente y se repite el procedimiento. Generalmente, el método se caracteriza por cortos movimientos en zig-zag que convergen muy lentamente a la solución; sin embargo, se utiliza para acercarse a la solución y después aplicar un método de alto orden de convergencia como el de Newton-Raphson; es decir, se emplea como un método para conseguir "buenos" valores iniciales. Este método puede ejemplificarse paso a paso con el Math-CAD o un software equivalente y explorar con varios valores de t para encontrar el óptimo; cabe ensayarlo con diferentes sistemas e incluso proponer vectores de exploración; en fin, llevar la matemática a nivel experimental.

ALGORITMO 4.5 Método del descenso de máxima pendiente

Para encontrar una solución aproximada de un sistema de ecuaciones no lineales $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, proporcionar las funciones F(I,x) y las derivadas parciales de la función (ecuación (4.24)) D(I,x) y los

DATOS: Número de ecuaciones N, vector de valores iniciales x, número máximo de iteraciones MAXIT, criterio de convergencia EPS, intervalo de búsqueda [A,B] y el número de puntos de [A,B] por ensayar M.

RESULTADOS: El vector solución x o mensaje "*NO CONVERGE*".

PASO 1. Hacer $K = 1$

PASO 2. Mientras K ≤ MAXIT, repetir los pasos 3 a 27.

PASO 3. Hacer Z = 0

PASO 4. Hacer I = 1

PASO 5. Mientras I ≤ N, repetir los pasos 6 y 7.

PASO 6. Hacer Z = Z + F(I,x)^2
PASO 7. Hacer I = I + 1
PASO 8. Si Z ≤ EPS ir al paso 29. De otro modo continuar.
PASO 9. Hacer NP = 0, NU = 1, MENOR = 1E20
PASO 10. Hacer J = 1
PASO 11. Mientras J ≤ M, repetir los pasos 12 a 25.
PASO 12. Hacer S = NU + NP,
T = A + (B-A)/S, L = 1
PASO 13. Hacer **xa = x - T * dz**
PASO 14. Hacer Z = 0
PASO 15. Hacer I = 1
PASO 16. Mientras I ≤ M, repetir los pasos 17 y 18
PASO 17. Hacer
Z = Z + F(I, xa)^2
PASO 18. Hacer I = I + 1
PASO 19. Si MENOR < Z, ir al paso 21. De otro modo continuar.
PASO 20. Hacer MENOR = Z, TOPT = T
PASO 21. Si L = 0 ir al paso 24. De otro modo continuar.
PASO 22. Hacer T = B - (B-A)/S, L = 0
PASO 23. Ir al paso 13.
PASO 24. Hacer NP = NU, NU = S
PASO 25. Hacer J = J + 1
PASO 26. Hacer **x = x - TOPT * dz**
PASO 27. Hacer K = K + 1
PASO 28. IMPRIMIR "*NO CONVERGE*" y TERMINAR.
PASO 29. IMPRIMIR x y TERMINAR.

Ejemplo 4.10

Con el algoritmo 4.5, elabore un programa para resolver el sistema

$$f_1(x_1, x_2, x_3) = 3x_1 - \cos(x_2x_3) - 0.5 = 0$$
$$f_2(x_1, x_2, x_3) = x_1^2 - 625x_2^2 = 0$$
$$f_3(x_1, x_2, x_3) = e^{-x_1x_2} + 20x_3 + (10\pi\text{-}3)/3 = 0$$

empleando como vector inicial:

$$\mathbf{x}^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

SOLUCIÓN

En el disco se presenta el programa 4.3, basado en el método del descenso de máxima pendiente y con búsqueda de Fibonacci.

Para su empleo, el usuario proporcionará el procedimiento GRADTE, donde se forma la función z por minimizar y el gradiente de esta función ∇z.

Enseguida se anotan los resultados que se obtienen

1	0.00000	0.00000	0.00000	1.119E+02	−9.00000	0.00000	418.87867	0.00000
2	0.01127	0.00000	−0.52458	2.150E+00	−8.79711	0.00045	−0.78687	0.00125
3	0.33117	−0.00002	−0.49587	5.739E−01	−2.89371	−0.42689	22.10419	0.03636
4	0.33479	0.000.52	−0.52365	2.582E−01	−2.82390	−0.14421	−0.04868	0.00125
5	0.50090	0.00900	−0.52079	4.279E−02	0.41662	−4.57387	2.06911	0.05882
6	0.50006	0.01827	−0.52498	3.066E−03	0.08554	−1.84517	−1.46890	0.00203
7	0.49995	0.02058	−0.52314	2.194E−04	−0.03004	0.76145	−0.04293	0.00125
8	0.49997	0.01999	−0.52311	5.252E−08	0.00025	−0.00961	−0.00470	0.00077
9	0.49997	0.02000	−0.52310	3.010E−09	−0.00012	−0.00040	−0.00194	0.00077
10	0.49997	0.02000	−0.52310	5.884E−10	−0.00015	−0.00016	−0.00001	0.00125

LA SOLUCIÓN DEL SISTEMA ES:
X (1) = 0.49997
X (2) = 0.02000
X (3) = −0.52310

Ejercicios

4.1 Uno de los problemas de ingeniería química, que mejor ilustra la reducción de ecuaciones es el cálculo de la fracción de vapor V/F en una vaporización instantánea (véase Ejercicio 2.6), donde se tienen las ecuaciones

$$F = L + V \tag{1}$$

$$Fz_i = Lx_i + Vy_i \qquad i = 1,2,...,n \tag{2}$$

provenientes del balance de materiales; y las relaciones de equilibrio líquido-vapor

$$K_i = \frac{y_i}{x_i} \qquad i = 1,2,...,n \tag{3}$$

donde

$$K_i = \frac{P_i^0}{P} \qquad i = 1,2,...,n \tag{4}$$

y

$$P_i^0 = 10^{A_i - B_i/(C_i + T - 273.15)} \qquad i = 1,2,...,n \tag{5}$$

con las constantes A_i, B_i y C_i dadas para cada componente i.

Además se tiene

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0 \tag{6}$$

y

$$\sum_{i=1}^{n} z_i = 1 \tag{7}$$

Por otro lado, se tiene en estos problemas generalmente especificadas: z_i, $i = 1,2,...,n-1$, P, T y F.

Para un número de componentes $n = 9$ por ejemplo, se tiene entonces un sistema de 39 ecuaciones en las 39 incógnitas: L, V, x_i, y_i, K_i, P_i, $i = 1,2,...,9$ y z_9, que puede reducirse, en general, como sigue

Al combinar las ecuaciones (2) y (3) se eliminan las y_i, y se obtiene

$$x_i = \frac{z_i F}{(K_i V + F)} \tag{8}$$

Al combinar las ecuaciones (6) y (3)

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} K_i x_i = 0$$

o bien:

$$\sum_{i=1}^{n} x_i (1 - K_i) = 0 \tag{9}$$

con la sustitución de (8) en (9) se tiene

$$\sum_{i=1}^{n} \frac{z_i F (1 - K_i)}{K_i V + L} = 0 \tag{10}$$

Pero de (1) $L = F - V$, con lo que queda finalmente:

$$\sum_{i=1}^{n} \frac{z_i (1 - K_i)}{V (K_i - 1) + F} = 0 \tag{11}$$

Nótese que si se conocen los valores de z_i, $i=1,2,...,n-1$, (usando la ecuación (7) se obtiene z_n), los valores de A_i, B_i, C_i, $i=1,2,...,n$ y los valores de P y T (usando (5) y (4) se obtiene K_i, $i=1,2,...,n$) y F, la ecuación (11) es ya sólo función de V, con lo que se ha reducido el sistema de 39 ecuaciones en 39 incógnitas a una sola ecuación con una incógnita (V), cuya solución puede obtenerse con alguno de los métodos del capítulo 2.

4.2 La presión requerida para sumergir un objeto pesado grande en un terreno suave y homogéneo, que se encuentra sobre un terreno de base dura, puede predecirse a partir de la presión requerida para sumergir objetos más pequeños en el mismo suelo*. En particular, la presión p requerida para sumergir una lámina circular de radio r una distancia d en el terreno suave, donde el terreno de base dura se encuentra a una distancia $D>d$ debajo de la superficie, puede aproximarse mediante una ecuación de la forma

$$p = k_1 \exp(k_2 r) + k_3 r, \tag{1}$$

*Richard L. Burden y J. Douglas Faires. Análisis numérico. Grupo Editorial Iberoamericana (1985).

donde k_1, k_2 y k_3 son constantes que, con $k_2>0$, dependen de d y la consistencia del terreno pero no del radio de la lámina.

a) Encuentre los valores de k_1, k_2 y k_3 si se supone que una lámina de radio 1 pulgada requiere una presión de 10 lb/pulg^2 para sumergirse 1 pie en el terreno lodoso, una lámina de radio 2 pulg. requiere una presión de 12 lb/pulg^2 para sumergirse 1 pie y una lámina de radio 3 pulgadas requiere una presión de 15 lb/pulg^2 (suponiendo que el lodo tiene una profundidad mayor que 1 pie).

b) Use los cálculos de (*a*) para predecir cuál es la lámina circular de radio mínimo que se necesitaría para sostener un peso de 500 lb en este terreno, con un hundimiento de menos de 1 pie.

SOLUCIÓN

Inciso a)

Al sustituir los valores de r y p en (1) para los tres casos, se tiene

$$\begin{aligned} 10 &= k_1 \exp(k_2) + k_3 \\ 12 &= k_1 \exp(2k_2) + 2k_3 \\ 15 &= k_1 \exp(3k_2) + 3k_3 \end{aligned}$$

un sistema de tres ecuaciones no lineales en las incógnitas k_1, k_2 y k_3. Se despeja k_3 de la primera ecuación

$$k_3 = 10 - k_1 \exp(k_2)$$

Se sustituye k_3 en las dos restantes y se tiene

$$\begin{aligned} 12 &= k_1\exp(2k_2) + 2[10 - k_1 \exp(k_2)] \\ 15 &= k_1\exp(3k_2) + 3[10 - k_1\exp(k_2)] \end{aligned}$$

o bien

$$\begin{aligned} f_1(k_1, k_2) &= k_1 [\exp(2k_2) - 2\exp(k_2)] + 8 = 0 \\ f_2(k_1, k_2) &= k_1[\exp(3k_2) - 3\exp(k_2)] + 15 = 0 \end{aligned} \qquad (2)$$

un sistema de dos ecuaciones no lineales en las incógnitas k_1 y k_2.

Al dividir miembro a miembro estas dos ecuaciones

$$\frac{k_1 [\exp(2k_2) - 2\exp(k_2)]}{k_1 [\exp(3k_2) - 3\exp(k_2)]} = \frac{-8}{-15}$$

se obtiene

$$\exp(k_2) - \frac{8}{15}\exp(2k_2) - \frac{6}{15} = 0$$

o bien

$$f(k_2) = 15\exp(k_2) - 8\exp(2k_2) - 6 = 0 \qquad (3)$$

una ecuación no lineal en la incógnita k_2, cuya solución con el método de Newton-Raphson visto en el capítulo 2 es

$$k_2 = 0.259695;$$

al sustituir k_2 en cualquiera de las ecuaciones (2) y despejar se tiene:

$$k_1 = \frac{-8}{\exp(2k_2) - 2\exp(k_2)} = 8.771286,$$

por último:

$$k_3 = 10 - k_1\exp(k_2) = -1.372281$$

Inciso (b)

Un peso de 500 lb sobre un disco de radio r producirá una presión de $500/(\pi r^2)$ lb/pulg^2. Entonces

$$p = \frac{500}{\pi r^2} = k_1\exp(k_2 r) + k_3 r$$

o bien

$$f(r) = k_1\exp(k_2 r) + k_3 r - \frac{500}{\pi r^2} = 0$$

Para obtener el valor mínimo de r, se iguala $f'(r)$ con cero

$$f'(r) = k_1 k_2\exp(k_2 r) + k_3 + \frac{1000r}{[\pi r^2]^2} = 0$$

lo que origina una ecuación no lineal en la incógnita r, cuya solución con alguno de los métodos del capítulo 2 da

$$r = 3.18516 \text{ pulg}$$

que corresponde a un mínimo de $f(r)$. El lector puede verificar esto usando alguno de los criterios del cálculo diferencial.

4.3 Resuelva el siguiente sistema verificando primero su partición.

$$e_1: \quad x_1 + x_4 - 10 = 0$$

$$e_2: \quad x_2^2 x_4 x_3 - x_5 - 6 = 0$$

$$e_3: \quad x_1 x_2^{1.7}(x_4 - 5) - 8 = 0$$

$$e_4: \quad x_4 - 3x_1 + 6 = 0$$

$$e_5: \quad x_1 x_3 - x_5 + 6 = 0$$

SOLUCIÓN

Si bien la descomposición de un sistema en subsistemas es conocida como partición, la secuencia para resolver los subsistemas resultantes se denomina **orden de precedencia** del sistema. Existen algoritmos para partir un conjunto de ecuaciones y determinar el orden de precedencia. A continuación se seguirán las ideas de estos algoritmos a fin de partir el sistema dado.

a) Se forma una matriz de incidencia

$$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \left[\begin{array}{ccccc} 1 & & & 1 & \\ & 1 & 1 & 1 & 1 \\ 1 & 1 & & 1 & \\ 1 & & & 1 & \\ 1 & & 1 & & 1 \end{array}\right] \end{array}$$

donde cada fila corresponde a una ecuación y cada columna a una variable. Un 1 aparece en la fila i y la columna j si la variable x_j aparece en la ecuación e_i.

b) Se rearreglan las filas y columnas para ver mejor las particiones y el orden de precedencia. Así, después de un rearreglo se llega a

$$\begin{array}{c} \\ e_1 \\ e_4 \\ e_3 \\ e_5 \\ e_2 \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_4 & x_2 & x_3 & x_5 \end{array} \\ \left[\begin{array}{ccccc} \left[\begin{array}{cc} 1 & 1 \\ 1 & 1 \end{array}\right] & & & & \\ & & & & \\ 1 & 1 & \left[1\right] & & \\ 1 & & & \left[\begin{array}{cc} 1 & 1 \\ 1 & 1 \end{array}\right] & \end{array}\right] \end{array}$$

donde se nota de inmediato que en las ecuaciones e_1 y e_4 aparecen solamente las variables x_1 y x_4 y constituyen entonces un subsistema que puede resolverse primero

$$\begin{aligned} e_1: &\quad x_1 + x_4 = 10 \\ e_2: &\quad -3x_1 + x_4 = -6 \end{aligned}$$

resulta $x_1 = 4$ y $x_4 = 6$.

Estos valores se sustituyen en la ecuación e_3 y ésta queda en función de x_2 solamente; por tanto, como una ecuación en una incógnita

$$e_3: \quad 4x_2^{1.7} - 8 = 0$$

resulta $x_2 = 1.5034$

Finalmente, las ecuaciones e_2 y e_5 pueden resolverse para x_3 y x_5, lo que da

$$\begin{aligned} x_3 &= 1.255 \\ x_5 &= 11.0202 \end{aligned}$$

4.4 En un reactor se efectúan las siguientes reacciones en fase gaseosa:

$$A + B \rightleftharpoons C + D \qquad (1)$$

$$A + C \rightleftharpoons 2E \qquad (2)$$

A la temperatura de la reacción, las constantes de equilibrio son $kp_1=2.6$ y $kp_2 = 3.1$. Las composiciones iniciales son 2 mol/l de A y 1 mol/l de B.

Calcule la composición a la salida del reactor, asumiendo que se alcanza el equilibrio.

SOLUCIÓN

Si x_1 representa los moles de A convertidos en la reacción (1) y x_2 representa los moles de A convertidos en la reacción (2), entonces, en el equilibrio tenemos

$$\begin{aligned}
\text{moles de A} &= 2 - x_1 - x_2 \\
\text{moles de B} &= 1 - x_1 \\
\text{moles de C} &= x_1 - x_2 \\
\text{moles de D} &= x_1 \\
\text{moles de E} &= 2x_2 \\
\text{moles totales} &= 3
\end{aligned}$$

Con la aplicación de la ley de acción de masas se obtiene

Para la reacción (1)

$$2.6 = \frac{(x_1 - x_2)(x_1)}{(2-x_1-x_2)(1-x_1)}$$

Para la reacción (2)

$$3.1 = \frac{(2x_2)^2}{(2-x_1-x_2)(x_1-x_2)},$$

que es un sistema de dos ecuaciones no lineales en dos incógnitas, cuya solución por el método de Newton-Raphson, por ejemplo, exige:

- Un vector inicial cercano a la solución, obtenible a partir de consideraciones físicas del problema.
- La matriz jacobiana, ampliada con el vector de funciones, que es relativamente fácil, puesto que las derivadas parciales son directas.

Vector inicial. En virtud de las funciones y la existencia inicial de 2 moles de A y 1 mol de B, se propone $x_1 = 0.8$ y $x_2 = 0.4$.

Las derivadas parciales para la matriz jacobiana se dan a continuación

$$f_1(x_1,x_2) = \frac{(x_1 - x_2)(x_1)}{(2 - x_1 - x_2)(1 - x_1)} - 2.6 = 0$$

$$f_2(x_1,x_2) = \frac{(2x_2)^2}{(2 - x_1 - x_2)(x_1 - x_2)} - 3.1 = 0$$

$$\frac{\partial f_1(x_1,x_2)}{\partial x_1} = \frac{(2-x_1-x_2)(1-x_1)(2x_1-x_2) - (x_1 - x_2)(x_1)(-3+2x_1+x_2)}{((2-x_1-x_2)(1-x_1))^2}$$

$$\frac{\partial f_1(x_1,x_2)}{\partial x_2} = \frac{(2-x_1-x_2)(1-x_1)(x_1) + (x_1-x_2)(x_1)(x_1-1)}{((2-x_1-x_2)(1-x_1))^2}$$

$$\frac{\partial f_2(x_1,x_2)}{\partial x_1} = \frac{-(2x_2)^2(2-2x_1)}{((2-x_1-x_2)(x_1-x_2))^2}$$

$$\frac{\partial f_2(x_1,x_2)}{\partial x_2} = \frac{8(2-x_1-x_2)(x_1-x_2)x_2 - 8x_2^2(-1+x_2)}{((2-x_1-x_2)(x_1-x_2))^2}$$

Con el programa 4.1 del disco se obtienen los siguientes resultados

K	X (1)	X (2)	DISTANCIA
0	.80000	.40000	
1	.83855	.46836	.78481E-01
2	.83178	.45623	.13888E-01
3	.83144	.45566	.67206E-03
4	.83144	.45565	.15044E-05
5	.83144	.45565	.11921E-06

LA SOLUCIÓN DEL SISTEMA ES

X(1) = .83143783

X(2) = .45565480

4.5 El mezclado imperfecto en un reactor continuo de tanque agitado, se puede modelar como dos o más reactores con recirculación entre ellos, como se muestra en la figura siguiente.

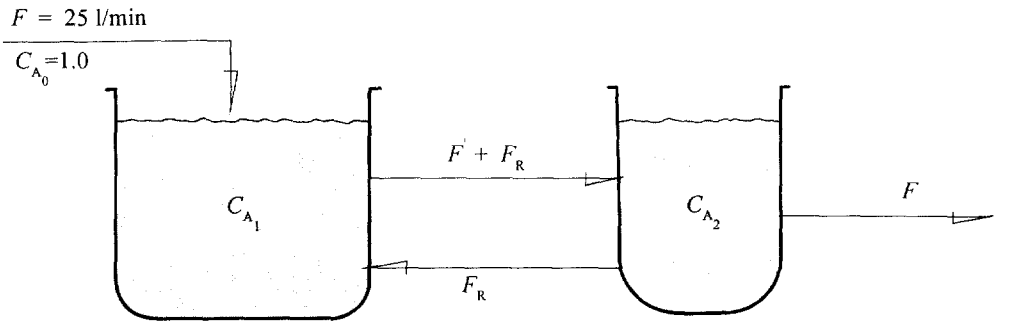


Figura 4.7. Reactores químicos con recirculación.

En este sistema se lleva a cabo una reacción isotérmica irreversible del tipo $A \xrightarrow{k} B$ de orden 1.8 con respecto al reactante A. Con los datos que se dan abajo, calcule la concentración del reactante A en los reactores 1 y 2 (C_{A1} y C_{A2} respectivamente), una vez alcanzado el régimen permanente.

Datos

$F = 25$ l/min $\qquad V_1 = 80$ l

$C_{A0} = 1$ mol/l $\qquad V_2 = 20$ l

$F_R = 100$ l/min $\qquad k = 0.2\ (\text{l/mol})^{0.8}\ (\text{min}^{-1})$

SOLUCIÓN

Con el balance del componente A en cada uno de los reactores se tiene

Entra - Sale - Reacciona = Acumulación

Reactor 1

$$F C_{A0} + F_R C_{A2} - (F + F_R) C_{A1} - V_1 k C_{A1}^n = 0 \qquad (1)$$

Reactor 2

$$(F + F_R) C_{A1} - (F_R + F) C_{A2} - V_2 k C_{A2}^n = 0 \qquad (2)$$

un sistema de dos ecuaciones no lineales en las incógnitas C_{A1} y CA_{A2}.

No obstante, se observa que despejando a C_{A2} de la ecuación (1)

$$C_{A2} = \frac{(F + F_R) C_{A1} + V_1 k C_{A1}^n - F C_{A0}}{F_R}$$

y sustituyéndola en la ecuación (2)

$$125C_{A1} - 125\,\frac{(F + F_R)\,C_{A1} + V_1 k C_{A1}^n - FC_{A0}}{F_R}$$

$$-kV_2\left[\frac{(F + F_R)\,C_{A1} + V_1\,kC_{A1}^n - FC_{A0}}{F_R}\right]^n = 0$$

el problema se reduce a una ecuación no lineal en la incógnita C_{A1} cuya solución se encuentra empleando alguno de los métodos del capítulo 2 y se deja al lector como ejercicio.

Resultados: $C_{A1} = 0.6493$
$C_{A2} = 0.6352$

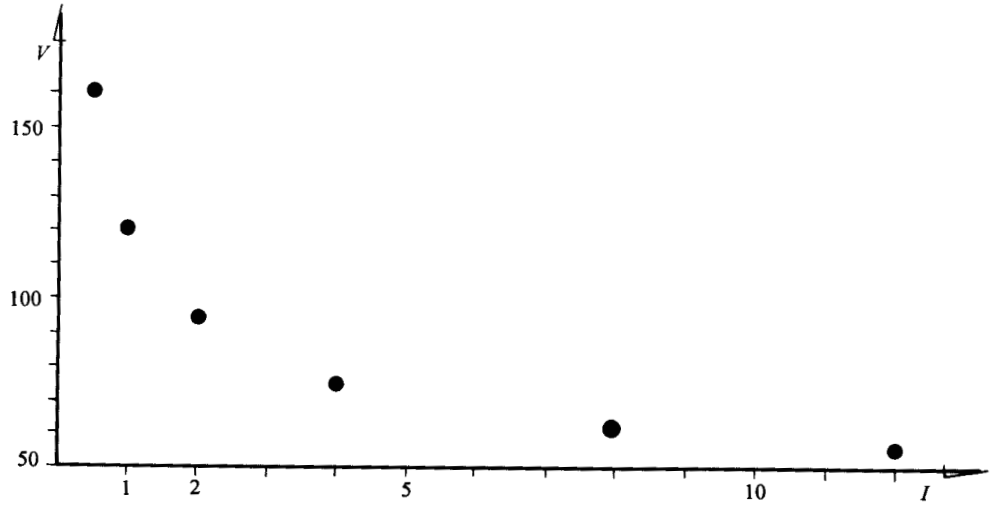
4.6 En una lámpara de arco*, de longitud de arco constante, se observa el voltaje V empleado por el arco para diversos valores de la corriente *I*

I	0.5	1	2	4	8	12
V	160	120	94	75	62	56

Encuentre la ecuación que mejor represente estos valores, empleando el criterio de mínimos cuadrados.

SOLUCIÓN

Se traza el diagrama de dispersión



*J. Lipka. *Computaciones gráficas y mecánicas*. Lipka J. CECSA.

y se observa que la curva suave que pasa por entre los puntos es hiperbólica y asintótica a alguna recta horizontal $V = c$. Con esto, se supone que los datos pueden quedar relacionados por la ecuación

$$V = a I^b + c \tag{1}$$

donde* $b < 0$.

Los parámetros a, b y c se determinan minimizando la función

$$f(a,b,c) = \sum_{i=1}^{6} (V_i - aI_i^b - c)^2 \tag{2}$$

La ecuación (2) se deriva parcialmente con respecto a a, b y c, y se igualan a cero dichas derivadas parciales para obtener

$$\sum_{i=1}^{6} V_i I_i^b - a \sum_{i=1}^{6} I_i^{2b} - c \sum_{i=1}^{6} I_i^b = 0$$

$$\sum_{i=1}^{6} V_i I_i^b \ln I_i - a \sum_{i=1}^{6} I_i^{2b} \ln I_i - c \sum_{i=1}^{6} I_i^b \ln I_i = 0 \tag{3}$$

$$\sum_{i=1}^{6} V_i - a \sum_{i=1}^{6} I_i^b - 6c = 0$$

un sistema de tres ecuaciones no lineales en las incógnitas a, b y c.

Al despejar c de la tercera ecuación

$$c = \frac{1}{6} \sum_{i=1}^{6} V_i - \frac{a}{6} \sum_{i=1}^{6} I_i^b \tag{4}$$

y sustituir en las dos primeras, se tiene (escribiendo sólo el símbolo de las sumatorias y no sus límites)

$$f_1(a,b) = \sum V_i I_i^b - a \sum I_i^{2b} - \frac{1}{6} [\sum V_i][\sum I_i^b] + [\frac{a}{6} \sum I_i^b]^2 = 0 \tag{5}$$

$$f_2(a,b) = \sum V_i I_i^b \ln I_i - a \sum I_i^{2b} \ln I_i - \frac{1}{6} [\sum V_i][\sum I_i^b \ln I_i] + \frac{a}{6} [\sum I_i^b][\sum I_i^b \ln I_i] = 0$$

un sistema de dos ecuaciones no lineales en las incógnitas a y b cuya solución requiere valores iniciales.

---

*$b > 0$ en el caso de una parábola, con ordenada al origen c.

Para estimar valores iniciales, en la ecuación (1) se sustituyen tres de los puntos dados

$$160 = a\,0.5^b + c$$

$$75 = a\,4^b + c$$

$$56 = a\,12^b + c$$

al despejar c de la tercera y sustituir en las dos primeras se tiene

$$160 = a\,0.5^b + 56 - a\,12^b$$

$$75 = a\,4^b + 56 - a\,12^b$$

o bien

$$a(0.5^b - 12^b) = 104$$

$$a(4^b - 12^b) = 19$$

Estas dos últimas ecuaciones se dividen miembro a miembro

$$\frac{0.5^b - 12^b}{4^b - 12^b} = \frac{104}{19}$$

se rearregla

$$19\,(0.5)^b + 85\,(12)^b - 104\,(4)^b = 0$$

y se resuelve esta ecuación no lineal con alguno de los métodos del capítulo 2 para obtener

$$b = -0.51952$$

de donde

$$a = 89.77$$

El sistema (5) se resuelve utilizando éstos como valores iniciales y el método de Newton-Raphson multivariable, con lo que resulta

$$a = 87.78$$

$$b = -0.532$$

y al sustituir en (4) se obtiene

$$c = 32.86$$

De tal manera que la ecuación que mejor ajusta los datos queda

$$V = 87.78\,I^{-0.532} + 32.86$$

4.7 Para la obtención de butadieno a partir de etanol en fase vapor, se propone el siguiente mecanismo de reacción

$$CH_3\text{–}CH_2\text{–}OH \rightleftharpoons CH_2 = CH_2 + H_2O \qquad (1)$$

$$CH_3\text{–}CH_2\text{–}OH \rightleftharpoons CH_2 = \underset{H}{\overset{OH}{C}} + H_2 \qquad (2)$$

$$CH_2=\underset{\displaystyle H}{\overset{\displaystyle OH}{C}} + CH_2=CH_2 \rightleftharpoons CH_2=CH-HC=CH_2 + H_2O \qquad (3)$$

Calcule las composiciones en el equilibrio a 400 °C y 1 atm, si las constantes de equilibrio son 5.97, 0.27 y 2.8 para las reacciones (1), (2) y (3) respectivamente.

SOLUCIÓN

Base de cálculo: 1 mol de etanol.

Si x_1 = moles de etileno producidas en la reacción(1)

x_2 = moles de hidrógeno producidas en la reacción (2)

x_3 = moles de agua producidas en la reacción (3)

entonces en el equilibrio se tendrá

$$\begin{aligned}
\text{moles de etanol} &= 1 - x_1 - x_2 \\
\text{moles de etileno} &= x_1 - x_3 \\
\text{moles de agua} &= x_1 + x_3 \\
\text{moles de hidrógeno} &= x_2 \\
\text{moles de acetaldehído} &= x_2 - x_3 \\
\text{moles de butadieno} &= x_3 \\
\text{moles totales} &= 1 + x_1 + x_2
\end{aligned}$$

De acuerdo con la ley de acción de masas, se tiene

$$5.97 = \frac{(x_1 + x_3)\,(x_1 - x_3)}{(1 - x_1 - x_2)} \left[\frac{P}{1 + x_1 + x_2}\right]^{\Delta n_1}$$

$$0.27 = \frac{(x_2 - x_3)\,x_2}{(1 - x_1 - x_2)} \left[\frac{P}{1 + x_1 + x_2}\right]^{\Delta n_2}$$

$$2.8 = \frac{(x_1 + x_3)\,x_3}{(x_1 - x_3)\,(x_2 - x_3)} \left[\frac{P}{1 + x_1 + x_2}\right]^{\Delta n_3}$$

donde Δn_i = número de moles de los productos-número de moles de los reactantes (en la reacción i).

Por lo tanto

$$\Delta n_1 = 2 - 1 = 1$$

$$\Delta n_2 = 2 - 1 = 1$$

$$\Delta n_3 = 2 - 2 = 0$$

Por otro lado

$$P = 1 \text{ atm.}$$

$$T = 673.2 \text{ K}$$

$$R = 0.082 \text{ atm-l/(mol-K)}$$

Vector inicial. Luego de observar las funciones y el hecho de que la base de cálculo es 1 mol de etanol, se propone

$$x_1 = 0.7, \qquad x_2 = 0.2, \qquad x_3 = 0.1$$

Nótese que x_1+x_2 no debe ser 1, para evitar la división entre cero en las dos primeras funciones.

Luego de sustituir valores y resolver el sistema de ecuaciones no lineales resultante con el programa 4.1 del disco, se llega a los siguientes resultados

$$X(1) = 0.71230$$

$$X(2) = 0.24645$$

$$X(3) = 0.15792$$

4.8 En una columna de cinco platos, se quiere absorber tolueno contenido en una corriente de gas V_0 (moles de gas sin tolueno/min), con un aceite L_0 (moles de aceite sin tolueno/min). Considérese que la relación de equilibrio está dada por la ley de Henry ($y = m\,x$), y que la columna opera a régimen permanente. Calcule la composición del tolueno en cada plato.

Datos: $V_0 = 39.6$ moles/min
$L_0 = 6.0$ moles/min
Las moles de tolueno/min que entran a la columna con el gas y el aceite son, respectivamente
$TV_0 = 5.4$ moles/min
$TL_0 = 0.0$ moles/min
$m = 0.155$

De aquí

$$y_0 = \frac{5.4}{5.4 + 39.6} = 0.12$$ fracción mol de tolueno en el gas que entra.

SOLUCIÓN

Los balances de masa para el tolueno en cada plato son (véase Fig. 4.8).

Plato	Balance de tolueno
1	$(V_0+TV_0)y_0-(V_0+TV_1)y_1+(L_0+TL_2)x_2-(L_0+TL_1)x_1 = 0$
2	$(V_0+TV_1)y_1-(V_0+TV_2)y_2+(L_0+TL_3)x_3-(L_0+TL_2)x_2 = 0$
3	$(V_0+TV_2)y_2-(V_0+TV_3)y_3+(L_0+TL_4)x_4-(L_0+TL_3)x_3 = 0$
4	$(V_0+TV_3)y_3-(V_0+TV_4)y_4+(L_0+TL_5)x_5-(L_0+TL_4)x_4 = 0$
5	$(V_0+TV_4)y_4-(V_0+TV_5)y_5+(L_0+TL_0)x_0-(L_0+TL_5)x_5 = 0$

donde TV_i, TL_i, $0 \le i \le 5$, son los moles de tolueno/min que salen del plato i con el gas y el aceite, respectivamente.

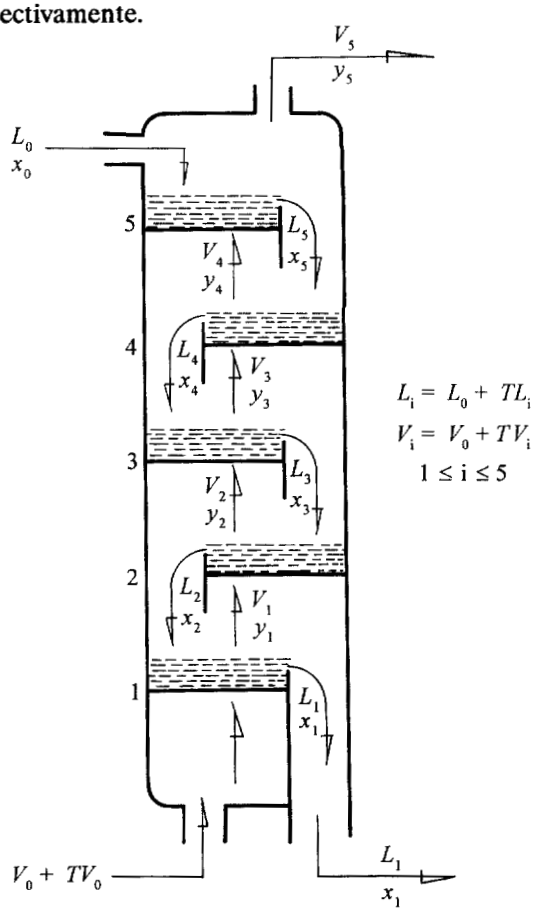


Figura 4.8. Columna de absorción de cinco platos.

Como $$y_i = \frac{TV_i}{TV_i + V_0} \quad \text{y además} \quad y_i = mx_i,$$

se obtiene $$TV_i = \frac{V_0 m x_i}{1 - m x_i}$$

Por otro lado

$$TL_i = \frac{L_0 x_i}{1 - x_i} \qquad \text{para } 0 \leq i \leq 5$$

Con la sustitución de y_i, TV_i y TL_i en los balances de masa anteriores, resulta el sistema no lineal siguiente

$$V_0 y_0 + \frac{V_0 y_0^2}{1-y_0} - V_0 m x_1 - \frac{V_0 m^2 x_1^2}{1-mx_1} + L_0 x_2 + \frac{L_0 x_2^2}{1-x_2} - L_0 x_1 - \frac{L_0 x_1^2}{1-x_1} = 0$$

$$V_0 m x_1 + \frac{V_0 m^2 x_1^2}{1-mx_1} - V_0 m x_2 - \frac{V_0 m^2 x_2^2}{1-mx_2} + L_0 x_3 + \frac{L_0 x_3^2}{1-x_3} - L_0 x_2 - \frac{L_0 x_2^2}{1-x_2} = 0$$

$$V_0 m x_2 + \frac{V_0 m^2 x_2^2}{1-x_2} - V_0 m x_3 - \frac{V_0 m^2 x_3^2}{1-mx_3} + L_0 x_4 + \frac{L_0 x_4^2}{1-x_4} - L_0 x_3 - \frac{L_0 x_3^2}{1-x_3} = 0$$

$$V_0 m x_3 + \frac{V_0 m^2 x_3^2}{1-mx_3} - V_0 x_4 - \frac{V_0 m^2 x_4^2}{1-mx_4} + L_0 x_5 + \frac{L_0 x_5^2}{1-x_5} - L_0 x_4 - \frac{L_0 x_4^2}{1-x_4} = 0$$

$$V_0 m x_4 + \frac{V_0 m^2 x_4^2}{1-mx_4} - V_0 m x_5 - \frac{V_0 m^2 x_5^2}{1-mx_5} + L_0 x_0 + \frac{L_0 x_0^2}{1-x_0} - L_0 x_5 - \frac{L_0 x_5^2}{1-x_5} = 0$$

donde $x_1, x_2, \ldots, x_5$ son las incógnitas.

Este sistema se resuelve con el programa 4.2 con los siguientes valores iniciales

$$x_1 = 0.4, x_2 = 0.3, x_3 = 0.2, x_4 = 0.1, x_5 = 0.05,$$

los cuales se obtuvieron usando un perfil lineal de concentraciones a lo largo de la columna. Los resultados obtenidos son

k	X (1)	X (2)	X (3)	X (4)	X (5)	DISTANCIA
0	.40000	.30000	.20000	.10000	.05000	
1	.45756	.30057	.19940	.12100	.06020	.62120E-01
2	.45398	.30115	.20289	.12717	.06318	.85044E-02
3	.45432	.30195	.20424	.12919	.06416	.27569E-02
4	.45444	.30222	.20468	.12986	.06449	.91471E-03
5	.45448	30231	.20483	.13008	.06460	.30494E-03
6	.45450.	.30234	.20488	.13016	.06463	.10179E-03
7	.45450	.30235	.20489	.13018	.06465	.34040E-04

LA SOLUCIÓN DEL SISTEMA ES

X (1) = .45450091
X (2) = .30234605
X (3) = .20489225
X (4) = .13018015
X (5) = .64646289E-01

Problemas

4.1 Resuelva el sistema

$$\begin{aligned} x_1 x_2 + x_6 x_4 &= 18 \\ x_2 + x_5 + x_6 &= 12 \\ x_1 + ln(x_2/x_4) &= 3 \\ x_3^2 + x_3 &= 2 \\ x_2 + x_4 &= 4 \\ x_3(x_3 + 6) &= 7, \end{aligned}$$

utilizando las sugerencias dadas al principio de este capítulo (reducción, partición, entre otros).

4.2 Resuelva el sistema

$$\begin{aligned} e_1: \quad x_1 x_3 - x_4 &= 1 \\ e_2: \quad x_2^2 x_3^2 + x_4 &= 17 \\ e_3: \quad x_1 + x_2 &= 6 \\ e_4: \quad \ln x_3 x_4^2 + x_3 x_4^2 &= 1 \end{aligned}$$

mediante tanteo de ecuaciones.

4.3 A partir de consideraciones geométricas demuestre que el sistema no lineal

$$\begin{aligned} x^2 + y^2 - x &= 0 \\ x^2 - y^2 - y &= 0, \end{aligned}$$

tiene una solución no trivial única. Ademas obtenga una estimación inicial x^0, y^0 y aproxime dicha solución, empleando el método de punto fijo.

4.4 Dado el sistema de ecuaciones no lineales

$$\begin{aligned} x^2 + y &= 37 \\ x - y^2 &= 5, \end{aligned}$$

determine un arreglo de la forma

$$\begin{aligned} g_1(x, y) &= x \\ g_2(x, y) &= y \end{aligned}$$

y un vector inicial $\mathbf{x}^{(0)}$ que prometa convergencia a una solución; es decir, que se satisfaga el sistema de desigualdades (Ec. 4.5).

4.5 Encuentre una solución del sistema de ecuaciones del problema anterior, por medio del método de Newton-Raphson y tomando como valor inicial

a) $(x, y) = (5,0)$

b) $(x, y) = (5,-1)$

¿Qué criterios se pueden aplicar para saber si el proceso converge y, en tal caso, cómo se puede verificar que efectivamente se trata de una solución?
Sugerencia: Emplee el software del libro.

4.6 Utilice el método de punto fijo multivariable para encontrar una solución de cada uno de los siguientes sistemas

a)
$$x_1(4 - 0.0003x_1 - 0.0004x_2) = 0$$
$$x_2(2 - 0.0002x_1 - 0.0001x_2) = 0$$

b)
$$x_1^2 + 2x_2^2 - x_2 - 2x_3 = 0$$
$$x_1^2 - 8x_2^2 + 10x_3 = 0.0001$$
$$x_1^2/(7x_2 x_3) - 1 = 0$$

c)
$$2x_1 + x_2 + x_3 - 4\log(10x_1) = 0$$
$$x_1 + 2x_2 + x_3 - 4\log(10x_2) = 0$$
$$x_1 x_2 x_3 - \log(10x_3) = 0$$

d)
$$3x_1 \operatorname{sen} x_2 - \cos(x_2 x_3) \operatorname{sen} x_2 - \operatorname{sen}^{-1}(-0.52356) \operatorname{sen} x_2 = 0$$
$$x_1^2 - 625x_2^2 = 0$$
$$\exp(-x_1 x_2) + 20x_3 = 9.471975$$

Sugerencia: Utilice el Math CAD o Sofware equivalente.

4.7 Elabore un programa para resolver sistemas de ecuaciones no lineales. Utilice para ello el algoritmo 4.1.

4.8 Emplee el programa del problema 4.7 para resolver los sistemas del problema 4.6.

4.9 Mediante el programa 4.1 del apéndice (véase Ej. 4.4), resuelva los siguientes sistemas de ecuaciones no lineales

a)
$$(x_1 + \cos x_1 x_2 x_3 - 1)^{1/2} = 0$$
$$(1 - x_1)^{1/4} + x_2 + x_3(0.05x_3 - 0.15) = 1$$
$$1 + x_1^2 + 0.1x_2^2 - 0.01x_2 - x_3 = 0$$

b)
$$0.5 \operatorname{sen}(x_1 x_2) - x_2/(4\pi) - 0.5x_1 = 0$$
$$0.920423[\exp(2x_1) - \exp(1)] + 8.65256x_2 - 2\exp(x_1) = 0$$

Emplee $EPS = 10^{-4}$.

4.10 Si en la aplicación del método de Newton-Raphson, en algún punto del proceso iterativo, por ejemplo $\mathbf{x}^{(i)}$, el determinante de la matriz jacobiana evaluado en ese punto es cero, o muy cercano a cero, dicho proceso no puede continuarse. ¿Qué hacer en tales casos? (véase Probl. 2.10).

4.11 Los métodos estudiados en este capítulo son aplicables también a sistemas de ecuaciones lineales y a ecuaciones no lineales en una variable, ya que estos dos son sólo casos particulares del caso general de sistemas de ecuaciones no lineales. Por ejemplo, si se aplicara el método de Newton-Raphson para resolver el sistema lineal

$$4x_1 - 9x_2 + 2x_3 = 5$$
$$2x_1 - 4x_2 + 6x_3 = 3$$
$$x_1 - x_2 + 3x_3 = 4$$

la matriz de derivadas parciales sería

$$J = \begin{bmatrix} 4 & -9 & 2 \\ 2 & -4 & 6 \\ 1 & -1 & 3 \end{bmatrix}$$

Encuentre la solución utilizando el algoritmo 4.2 con un vector inicial adecuado.

4.12 Resuelva el problema 3.33 (considerando ahora que la reacción es de orden 0.5 con respecto a A y la constante de velocidad de reacción k_1 es 0.05 $l^{-0.5}mol^{0.5}min^{-1}$. Emplee el programa del problema 4.7, o bien el programa 4.1 del apéndice.

4.13 Repita el problema 3.34, considerando que la reacción es de orden 0.5 y que la constante de velocidad de reacción es 0.05 $l^{-0.5}mol^{0.5}min^{-1}$. ¿La conversión de A mejora recirculando los tres tanques en lugar de recircular solamente el primero?

4.14 Utilice el método iterativo de punto fijo para resolver el sistema de ecuaciones no lineales del ejemplo 4.4, con el vector inicial

$$\mathbf{x}^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

a) con desplazamientos sucesivos

b) con desplazamientos simultáneos

Compare la convergencia en los dos casos.

Sugerencia: Emplee el Math-CAD o un software equivalente.

4.15 Resuelva el ejercicio 4.8, usando $TV_0 = 9.9$

4.16 Resuelva los sistemas de los problemas 4.6 y 4.9 por el método de Newton-Raphson modificado.

4.17 Elabore un programa para resolver sistemas de ecuaciones no lineales por el método de Newton-Raphson modificado, utilizando para ello el algoritmo 4.3. Resuelva con dicho programa el sistema

$$x_1^2 + 2x_2^2 + \exp(x_1 + x_2) = 6.1718 - x_1 x_3$$
$$10x_2 = -x_2 x_3$$
$$sen(x_1 x_3) + x_2^2 = 1.141 - x_1$$

utilizando como vector inicial a $\mathbf{x}^{(0)} = [1,1,1]^T$.

4.18 La siguiente tabla representa las temperaturas observadas T(°C) a diferentes tiempos t (*min*) del agua en un tanque de enfriamiento

t	0	1	2	3	5	7	10	15	20
T	92.0	85.3	79.5	74.5	67.0	60.5	53.5	45.0	39.5

Encuentre la ecuación de enfriamiento que mejor represente estos valores, empleando el criterio de mínimos cuadrados. Véase ejercicio 4.6.

4.19 La relación entre el rendimiento de un cultivo y la cantidad de fertilizante x, aplicado a ese cultivo, se ha formulado así

$$y = a - b\,d^{x}$$

donde $0 < d < 1$

Dados los siguientes datos

x	0	1	2	3	4
y	44.4	54.6	63.8	65.7	68.9

obtenga estimaciones de a, b y d empleando el método de los mínimos cuadrados. (Véase ejercicio 4.6).

4.20 Resuelva los sistemas de ecuaciones no lineales del problema 4.9 con el método de Broyden. Compare el número de iteraciones requerido con el número requerido en los métodos de punto fijo y de Newton-Raphson multivariable. Emplee en la comparación $EPS = \| \mathbf{x}^{(i)} - \mathbf{x}^{(i-1)} \| < 10^{-4}$.

4.21 Elabore un programa de cómputo para resolver sistemas de ecuaciones no lineales con el método de Broyden.
Emplee para ello el algoritmo 4.4. Resuelva con dicho programa el sistema

$$x_1^2 + 2x_2^2 + \exp(x_1 + x_2) = 6.1718 - x_1 x_3$$
$$10x_2 = -x_2 x_3$$
$$\operatorname{sen}(x_1 x_3) + x_2^2 = 1.141 - x_1,$$

utilizando como vector inicial a $\mathbf{x}^{(0)} = [1,1,1]^{T}$

4.22 El método de Broyden pertenece a una familia conocida como métodos de **Cuasi-Newton**. Otro de los miembros de dicha familia se obtiene al reemplazar a $J^{(k)}$ de la ecuación 4.18 con una matriz $A^{(k)}$, cuyos componentes son las derivadas parciales numéricas; esto es, consiste en aproximar las derivadas parciales analíticas de la matriz jacobiana J por sus correspondientes derivadas parciales numéricas. Por ejemplo, para una función de dos variables $f(x, y)$ las derivadas parciales numéricas quedan así

$$\frac{\partial f}{\partial x} = \frac{f(x+h, y) - f(x, y)}{h}$$

y

$$\frac{\partial f}{\partial y} = \frac{f(x, y+h) - f(x, y)}{h}$$

donde h es un valor pequeño.
Con las ideas dadas, encuentre una solución aproximada del sistema de ecuaciones no lineales siguiente usando como vector inicial $[x^0, y^0]^T = [0,0]^T$.

$$f_1(x, y) = x^2 - 10x + y^2 + 8 = 0$$
$$f_2(x, y) = xy^2 + x - 10y + 8 = 0$$

4.23 Resuelva los sistemas de ecuaciones no lineales de los problemas 4.6 y 4.9, mediante el método de Newton-Raphson con optimización de t.

Sugerencia: Emplee el programa 4.2 del apéndice.

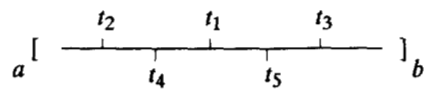
4.24 Otra forma de seleccionar los valores del tamaño de la etapa t (véase Sec. 4.6), consiste en dividir el intervalo de búsqueda $[a,b]$ en dos partes iguales sucesivamente. Esto es

$$t_1 = (a + b)/2\,,$$

$$t_2 = (a + t_1)/2\,, \qquad t_3 = (t_1 + b)/2$$

$$t_4 = (t_2 + t_1)/2, \qquad t_5 = (t_3 + t_1)/2, \text{ etc.}$$

Gráficamente:



Para cada valor de t se calcula el correspondiente z_{k+1} y el valor mínimo de z_{k+1} proporcionará el valor óptimo de t. Encuentre el valor óptimo de t en la primera iteración de la solución del ejemplo 4.3 usando este método de cálculo de t y el intervalo $[-1.2,-1]$.

4.25 Modifique el programa 4.2 del apéndice de modo que se empleen los valores de t calculados en la forma indicada en el anterior.

4.26 Resuelva el siguiente sistema de ecuaciones algebraicas no lineales, proponiendo en cada caso vectores iniciales. Emplee en cada caso los métodos que juzgue más convenientes y el software de que disponga.

a) $$y\,sen\,x + cos\,x - z = 0$$

$$\exp(x + y) - x^2\,cos\,x - \pi/1.15 = 0$$

$$y + 3xz + x^3 = 0$$

b) $$ln(x\,y) + x^2 y^2 = 8$$

$$sen\,x + y\,\exp(x) = 2$$

c) $$x_1^3 + x_3^3 - x_3^3 = 129$$

$$x_1^2 + x_2^2 - x_3^2 = 9.75$$

$$x_1 + x_2 - x_3 = 9.49$$

4.27 Se desea concentrar una solución con una concentración inicial de sólidos de 20% a una concentración final de 60% en un evaporador de doble efecto. Se dispone de vapor saturado a 0.68 atm (10 psig) y el segundo efecto que opera con una presión de vacío de 0.136 atm (2 psia). (ver figura 4.9).
Si la alimentación al sistema, 18,240.6 kg/h, entra al primer efecto a 93.3 °C, determine el área de los evaporadores, A_1 y A_2 y la cantidad de vapor requerido.

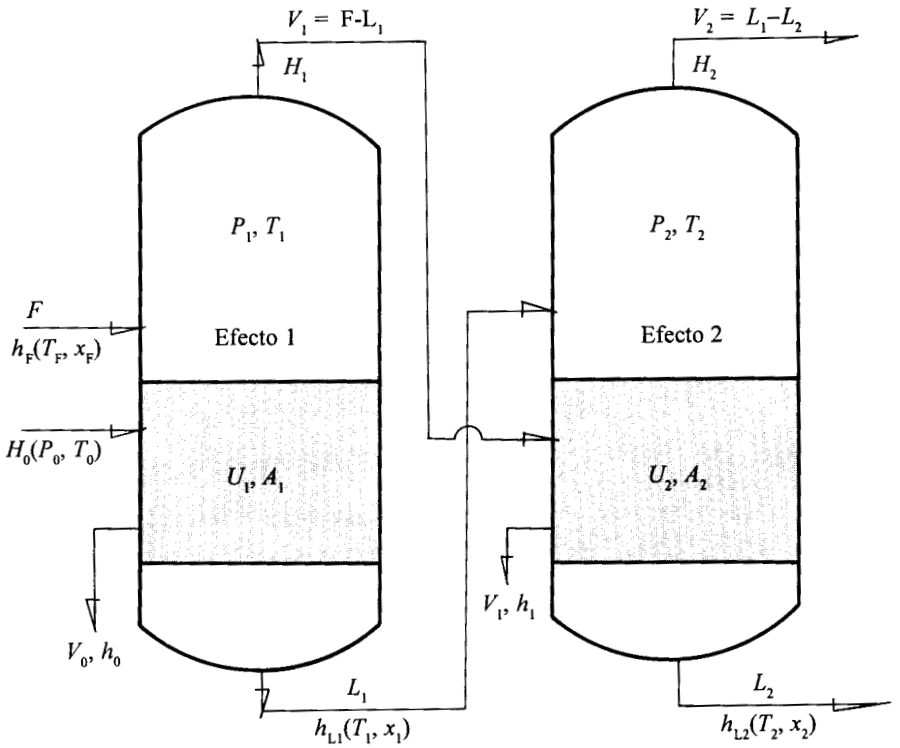


Figura 4.9 Sistema de evaporación de doble efecto.

Otros datos:

$C_{p,F} = 0.9$ kcal/(kg °C) $\qquad U_1 = 3{,}516.5$ kcal/(hm^2)

$C_{p,\,L1} = 0.8$ " $\qquad U_2 = 2{,}440.4$ "

$C_{p,\,L2} = 0.8$ "

4.28 El método del eigenvalor (valor propio) dominante* para resolver sistemas de ecuaciones no lineales, consiste en emplear el siguiente algoritmo

$$\mathbf{x}^{(k+2)} = \mathbf{x}^{(k)} + \frac{1}{1 - \lambda_1} [\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}]$$

donde λ_1 es el eigenvalor dominante de la matriz jacobiana J (véase ecuación (4.12")), evaluada en $\mathbf{x}^{(k+1)}$ y aproximado de la siguiente manera

$$\lambda_1 = \frac{|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}|}{|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}|}$$

*E. Kehat and M. Shacham. Chemical Processes Simulation Programs-3: Solution of systems of Non-Linear Equations. *Process Technology International*, Vol. 18, pag. 181 (1973).

o bien

$$\lambda_1 = \frac{\left(\sum_{i=1}^{n} (x_i^{k+1} - x_i^{k})^2 \right)^{1/2}}{\left(\sum_{i=1}^{n} (x_i^{k} - x_i^{k-1})^2 \right)^{1/2}}$$

Obsérvese que para la primera aplicación de este algoritmo se requieren tres vectores iniciales: $\mathbf{x}^{(0)}$, $\mathbf{x}^{(1)}$ y $\mathbf{x}^{(2)}$, los cuales pueden obtenerse, por ejemplo, con el método de punto fijo multivariable.

Mediante este algoritmo resuelva el sistema

$$f_1(x, y) = x^2 - 10x + y^2 + 8 = 0$$
$$f_2(x, y) = xy^2 + x - 10y + 8 = 0$$

usando como vector inicial: $[x^0, y^0] = [0,0]^T$ y los resultados de las dos primeras iteraciones del ejemplo 4.1.

4.29 La convergencia del método del eigenvalor dominante (véase Probl. 4.28), puede acelerarse usando un factor t de la siguiente manera

$$\mathbf{x}^{(k+2)} = \mathbf{x}^{(k)} + \frac{t}{1 - \lambda_1} [\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}],$$

y ensayando varios valores de t como se hizo en los métodos de Newton-Raphson con optimización de t y del descenso de máxima pendiente. El valor de t puede calcularse también calcularse en cada iteración con una fórmula dada por Broyden**, o se usa un valor constante.

Obtenga una aproximación a una solución del sistema dado en el problema 4.28 utilizando un valor de $t = 0.7$

4.30 Resuelva los sistemas de ecuaciones no lineales de los problemas 4.6 y 4.9, empleando el método del descenso de máxima pendiente para obtener valores iniciales; luego, con esos valores aplique el método de Newton-Raphson o el método de Broyden.

4.31 Obsérvese que en el método del descenso de máxima pendiente se encuentra el mínimo local de la función $z_k = f_1^2 + f_2^2 + \ldots + f_n^2$. Este método puede emplearse para aproximar el mínimo local de una función dada analíticamente, tomando dicha función como z. Modifique el algoritmo 4.5 para aproximar los mínimos de las funciones siguientes, usando EPS $= 10^{-5}$.

a) $z(x, y) = sen\,(x + y) + sen\,x - cos\,y$

b) $z(x_1, x_2, x_3) = x_1^2 + x_2^2 - 3x_3^2$

c) $z(x_1, x_2, x_3) = x_1^2 + 2x_2^4 + 3x_3^3 - 1$

Sugerencia: Grafique la superficie el inciso (a) usando el Math-CAD o el Graphics Calculus (GC).

**C.G. Broyden. *A Class of Methods for Solving Nonlinear Simultaneous Equations*. Math Comp. 19 pág. 577 (1965).

CAPÍTULO 5

APROXIMACIÓN FUNCIONAL E INTERPOLACIÓN

EN ESTE CAPÍTULO se estudiará la aproximación de funciones disponibles en forma discreta (puntos tabulados), con funciones analíticas sencillas, o bien de aproximación de funciones cuya complicada naturaleza exija su remplazo por funciones más simples.

INTRODUCCIÓN

La enorme ventaja de aproximar información discreta o funciones "complejas", con funciones analíticas sencillas, radica en su mayor facilidad de evaluación y manipulación, situación necesaria en el campo de la ingeniería.

Las funciones de aproximación se obtienen por combinaciones lineales de elementos de familias de funciones denominadas elementales. En general tendrán la forma

$$a_0 g_0(x) + a_1 g_1(x) + \ldots + a_n g_n(x), \tag{5.1}$$

donde a_i, $0 \le i \le n$, son constantes por determinar y $g_i(x)$, $0 \le i \le n$ funciones de una familia particular. Los monomios en x (x^0, x, x^2, ...) constituyen la familia o grupo más empleado; sus combinaciones generan aproximaciones del tipo polimonial

$$a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \tag{5.2}$$

El grupo conocido como funciones de Fourier

$$1, \operatorname{sen} x, \cos x, \operatorname{sen} 2x, \cos 2x, \ldots ,$$

al combinarse linealmente, genera aproximaciones del tipo

$$a_0 + \sum_{i=1}^{n} a_i \cos ix + \sum_{i=1}^{n} b_i \operatorname{sen} ix \tag{5.3}$$

El grupo de las funciones exponenciales

$$1, e^{x}, e^{2x}, \ldots$$

también puede usarse del modo siguiente

$$\sum_{i=0}^{n} a_i e^{ix} \tag{5.4}$$

De estos tres tipos de aproximaciones funcionales, las más comunes por su facilidad de manejo en evaluaciones, integraciones, derivaciones, etc., son las aproximaciones polinomiales (5.2) y son las que se estudiarán a continuación.

Sea una función $f(x)$ dada en forma tabular

Puntos	0	1	2	...	n
x	x_0	x_1	x_2	...	x_n
$f(x)$	$f(x_0)$	$f(x_1)$	$f(x_2)$	...	$f(x_n)$

Para aproximar a $f(x)$ por medio de un polinomio del tipo 5.2, se aplica alguno de los criterios siguientes: el de **ajuste exacto** o el de **mínimos cuadrados.**

La técnica del ajuste exacto consiste en encontrar una función polinomial **que pase por** los puntos dados en la tabla (véase Fig. 5.1). El método de mínimos cuadrados consiste en hallar un polinomio **que pase entre** los puntos y que satisfaga la condición de minimizar la suma de las desviaciones (d_i) elevadas al cuadrado; es decir, que se cumpla

$$\sum_{i=0}^{n} (d_i)^2 = \text{mínimo.}$$

Cuando la información tabular de que se dispone es aproximada hasta cierto número de cifras significativas, por ejemplo la de tablas de logaritmos o de funciones de Bessel, se recomienda usar ajuste exacto. En cambio si la información tiene errores considerables, como en el caso de datos experimentales, no tiene sentido encontrar un polinomio que pase por esos puntos sino más bien que pase entre ellos; entonces, el método de mínimos cuadrados es aplicable.

Una vez que se obtiene el polinomio de aproximación, éste puede usarse para obtener puntos adicionales a los existentes en la tabla, mediante su evaluación, lo que se conoce como **interpolación.** También puede derivarse o integrarse a fin de obtener información adicional de la función tabular.

A continuación se describen distintas formas de aproximar con polinomios obtenidos por ajuste exacto y su uso en la interpolación. En la sección 5.8 se describe la aproximación polinomial por mínimos cuadrados y en el capítulo 6 la derivación y la integración.

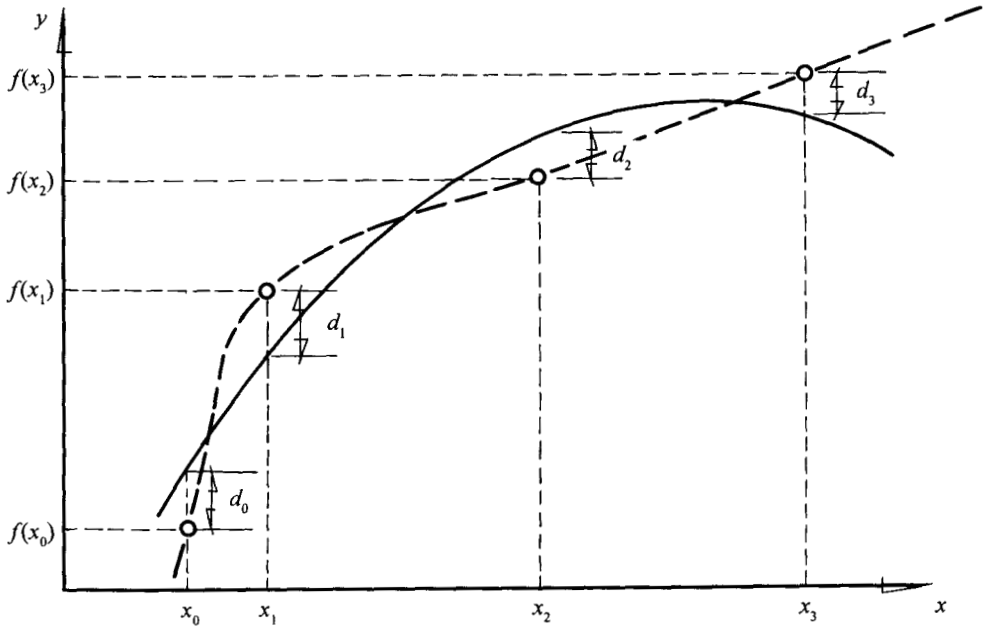


Figura 5.1. Aproximación polinomial con criterio de ajuste exacto (curva discontinua) y con mínimos cuadrados (curva llena).

SECCIÓN 5.1 APROXIMACIÓN POLINOMIAL SIMPLE E INTERPOLACIÓN

La interpolación es de gran importancia en el campo de la ingeniería, ya que al consultar fuentes de información presentadas en forma tabular, es frecuente no encontrar el valor buscado como un punto en la tabla. Por ejemplo las tablas 5.1 y 5.2 presentan la temperatura de ebullición de la acetona (C_3H_6O) a diferentes presiones.

Puntos	0	1	2	3	4	5	6
T (°C)	56.5	78.6	113.0	144.5	181.0	205.0	214.5
P (atm)	1	2	5	10	20	30	40

Tabla 5.1 Temperatura de ebullición de la acetona a diferentes presiones.

Puntos	0	1	2	3
T (°C)	56.5	113.0	181.0	214.5
P (atm)	1	5	20	40

Tabla 5.2 Temperatura de ebullición de la acetona a diferentes presiones.

Supóngase que sólo se dispusiera de la segunda y se desease calcular la temperatura de ebullición de la acetona a 2 atm de presión.

Una forma muy común de resolver este problema es sustituir los puntos (0) y (1) en la ecuación de la línea recta: $p\ (x) = a_0 + a_1 x$, de tal modo que resultan dos ecuaciones con dos incógnitas que son a_0 y a_1. Con la solución del sistema se consigue una aproximación polinomial de primer grado, lo que permite efectuar interpolaciones lineales; es decir, se sustituye el punto (0) en la ecuación de la línea recta y se obtiene

$$56.5 = a_0 + 1\ a_1$$

y al sustituir el punto (1)

$$113 = a_0 + 5\ a_1,$$

sistema que al resolverse da $a_0 = 42.375$ y $a_1 = 14.125$

Por lo tanto, estos valores generan la ecuación

$$p(x) = 42.375 + 14.125\ x \tag{5.5}$$

La ecuación resultante puede emplearse para aproximar la temperatura cuando la presión es conocida. Al sustituir la presión $x = 2$ atm se obtiene una temperatura de 70.6 °C. A este proceso se le conoce como interpolación.

Gráficamente la tabla 5.2 puede verse como una serie de puntos (0), (1), (2) y (3) en un plano P *vs* T (Fig. 5.2), en donde si se unen con una línea los puntos (0) y (1), por búsqueda gráfica se obtiene T $\approx$ 70.6 °C, para P = 2 atm.

En realidad, esta interpolación sólo ha consistido en aproximar una función analítica desconocida [T = f(P)] dada en forma tabular por medio de una línea recta que pasa por los puntos (0) y (1).

Para aproximar el valor de la temperatura correspondiente a P = 2 atm se pudieron tomar otros dos puntos distintos, por ejemplo (2) y (3), pero es de suponer que el resultado tendría un margen de error mayor, ya que el valor que se busca está entre los puntos (0) y (1).

Si se quisiera una aproximación mejor al valor "verdadero" de la temperatura buscada, podrían unirse más puntos de la tabla con una curva suave (sin picos), por ejemplo tres (0), (1) y (2) (véase Fig. 5.3) y gráficamente obtener T correspondiente a P = 2 atm.

Analíticamente el problema se resuelve al aproximar la función desconocida [T = f (P)] con un polinomio que pase por los tres puntos (0), (1) y (2). Este polinomio es una parábola y tiene la forma general

$$p_2(x) = a_0 + a_1 x + a_2 x^2, \tag{5.6}$$

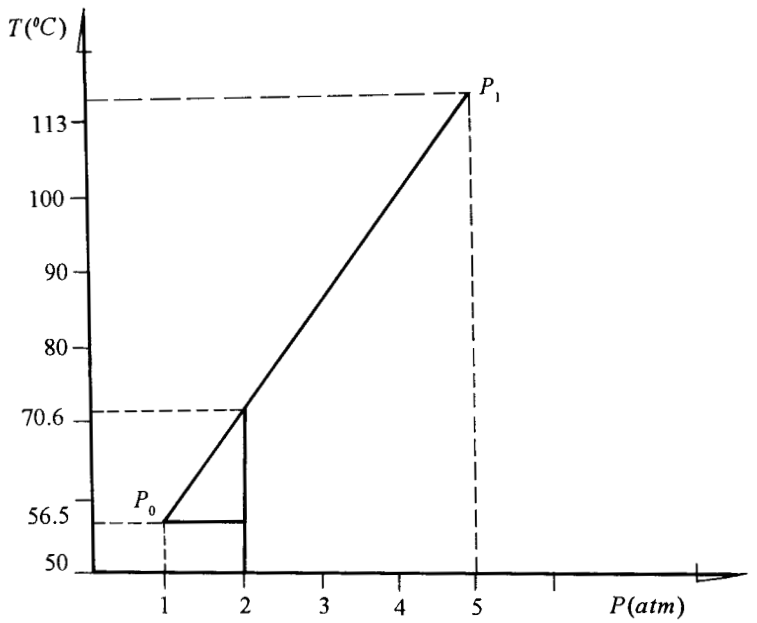


Fig. 5.2 Interpolación gráfica de la temperatura de ebullición de la acetona a 2 atm.

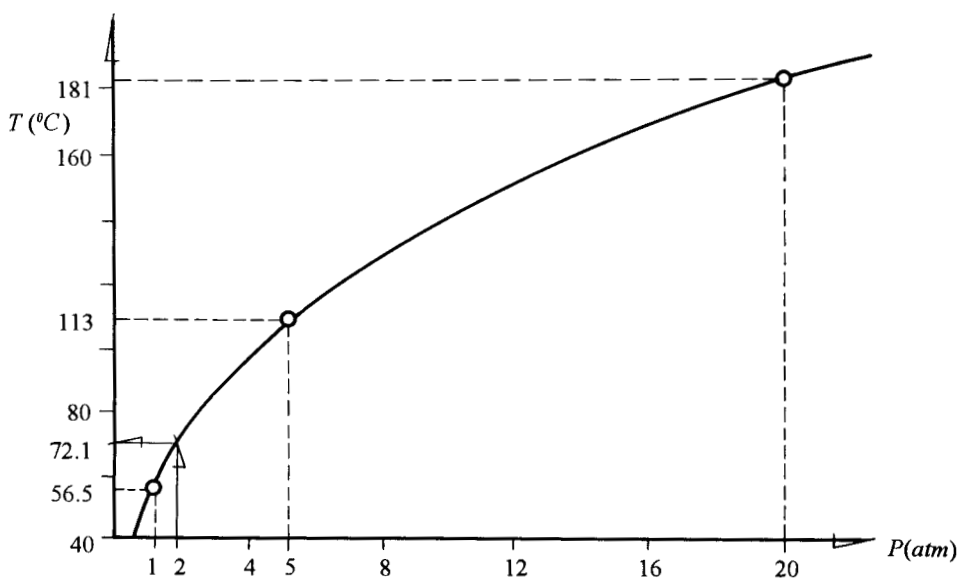


Fig. 5.3. Interpolación gráfica con tres puntos.

donde los parámetros a_0, a_1 y a_2 se determinan sustituyendo cada uno de los tres puntos conocidos en la ecuación 5.6; es decir

$$\begin{aligned} 56.5 &= a_0 + a_1 1 + a_2 1^2 \\ 113 &= a_0 + a_1 5 + a_2 5^2 \\ 181 &= a_0 + a_1 20 + a_2 20^2 \end{aligned} \tag{5.7}$$

Al resolver el sistema se obtiene

$$a_0 = 39.85, \qquad a_1 = 17.15, \qquad a_2 = -0.50482$$

De tal modo que la ecuación polinomial queda

$$p_2(x) = 39.85 + 17.15x - 0.50482x^2 \tag{5.8}$$

y puede emplearse para aproximar algún valor de la temperatura correspondiente a un valor de presión. Por ejemplo si $x = 2$ *atm*, entonces

$$T \approx p_2(2) = 39.85 + 17.15\,(2) - 0.50482(2)^2 \approx 72.1\ ^\circ C$$

La aproximación a la temperatura "correcta" es obviamente mejor en este caso.

Obsérvese que ahora se ha aproximado la función desconocida [$T = f(P)$] con un polinomio de segundo grado (parábola) que pasa por los tres puntos más cercanos al valor buscado. En general, si se desea aproximar una función con un polinomio de grado n, se necesitan $n+1$ puntos, que sustituidos en la ecuación polinomial de grado n:

$$p_n(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \tag{5.9}$$

generan un sistema de $n+1$ ecuaciones lineales en las incógnitas a_i, $i = 0,1,2,\ldots, n$.

Una vez resuelto el sistema se sustituyen los valores de a_i en la ecuación (5.9), con lo cual se obtiene el polinomio de aproximación. A este método se le conoce como **aproximación polinomial simple**.

Por otro lado, como se dijo al principio de este capítulo, puede tenerse una función conocida pero muy complicada, por ejemplo

$$f(x) = k\,x\,\ln x + \frac{1}{x}\sum_{m=0}^{\infty} C_m x^m \tag{5.10}$$

o

$$f(x) = (2/x)^{1/2}\,\text{sen}\,x \tag{5.11}$$

la cual conviene, para propósitos prácticos, aproximar con otra función más sencilla, como un polinomio. El procedimiento es generar una tabla de valores mediante la función original y a partir de dicha tabla aplicar el método descrito arriba.

ALGORITMO 5.1 Aproximación polinomial simple

Para obtener los $(n+1)$ coeficientes del polinomio de grado n $(n > 0)$ que pasa por $(n+1)$ puntos, proporcionar los

DATOS: El grado del polinomio N y las N+1 parejas de valores (X(I), FX(I), I=0,1,...,N).

RESULTADOS: Los coeficientes A(0), A(1), ..., A(N) del polinomio de aproximación.

PASO 1. Hacer I = 0
PASO 2. Mientras I≤ N, repetir los pasos 3 a 9.
  PASO 3. Hacer B(I,0) = 1
  PASO 4. Hacer J = 1
  PASO 5. Mientras J≤N, repetir los pasos 6 y 7.
    PASO 6. Hacer B(I,J) = B(I,J-1) * X(I)
    PASO 7. Hacer J = J+1
  PASO 8. Hacer B(I,N+1) = FX(I)
  PASO 9. Hacer I = I+1
PASO 10. Resolver el sistema de ecuaciones lineales B **a** = **fx** de orden N+1 con alguno de los algoritmos del capítulo 3.
PASO 11. IMPRIMIR A(0), A(1), ..., A(N) y TERMINAR.

SECCIÓN 5.2 POLINOMIOS DE LAGRANGE

El método de aproximación polinomial dado en la sección anterior, requiere la solución de un sistema de ecuaciones algebraicas lineales que, cuando el grado del polinomio es alto, puede presentar inconvenientes. Existen otros métodos de aproximación polinomial en que no se requiere resolver un sistema de ecuaciones lineales y los cálculos se realizan directamente; entre éstos se encuentra el de aproximación polinomial de Lagrange.

Se parte nuevamente de una función desconocida $f(x)$ dada en forma tabular y se asume que un polinomio de primer grado (ecuación de una línea recta) puede escribirse:

$$p(x) = a_0(x - x_1) + a_1(x - x_0) \tag{5.12}$$

donde x_1 y x_0 son los argumentos de los puntos conocidos $[x_0, f(x_0)]$, $[x_1, f(x_1)]$, y a_0 y a_1 son dos coeficientes por determinar. Para encontrar el valor de a_0, se hace $x = x_0$ en la ecuación 5.12, que al despejar da

$$a_0 = \frac{p(x_0)}{x_0 - x_1} = \frac{f(x_0)}{x_0 - x_1} \tag{5.13}$$

y para hallar el valor de a_1, se sustituye el valor de x con el de x_1, con lo que resulta

$$a_1 = \frac{p(x_1)}{x_1 - x_0} = \frac{f(x_1)}{x_1 - x_0} \tag{5.14}$$

de tal modo que al sustituir las ecuaciones 5.13 y 5.14 en la 5.12 queda

$$p(x) = \frac{f(x_0)}{x_0 - x_1}(x - x_1) + \frac{f(x_1)}{x_1 - x_0}(x - x_0) \tag{5.15}$$

o en forma más compacta

$$p(x) = L_0(x) f(x_0) + L_1(x) f(x_1) \tag{5.16}$$

donde

$$L_0(x) = \frac{x - x_1}{x_0 - x_1} \quad y \quad L_1(x) = \frac{x - x_0}{x_1 - x_0} \tag{5.17}$$

De igual manera, un polinomio de segundo grado (ecuación de una parábola) puede escribirse

$$p_2(x) = a_0(x - x_1)(x - x_2) + a_1(x - x_0)(x - x_2) + a_2(x - x_0)(x - x_1) \tag{5.18}$$

donde x_0, x_1 y x_2 son los argumentos correspondientes a los tres puntos conocidos $[x_0, f(x_0)]$, $[x_1, f(x_1)]$, $[x_2, f(x_2)]$; los valores de a_0, a_1 y a_2 se encuentran sustituyendo $x = x_0$, $x = x_1$ y $x = x_2$, respectivamente, en la ecuación 5.18 para obtener

$$a_0 = \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)}, \quad a_1 = \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} \text{ y}$$
$$a_2 = \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)} \tag{5.19}$$

cuyo remplazo en dicha ecuación genera el siguiente polinomio

$$p_2(x) = L_0(x) f(x_0) + L_1(x) f(x_1) + L_2(x) f(x_2) \tag{5.20}$$

donde

$$L_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}, \quad L_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}$$
y
$$L_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} \tag{5.21}$$

Por inducción el lector puede obtener polinomios de tercer, cuarto o *n-ésimo* grado; éste último queda como se indica a continuación

$$p_n(x) = L_0(x) f(x_0) + L_1(x) f(x_1) + \ldots + L_n(x) f(x_n)$$

donde

$$L_0(x) = \frac{(x - x_1)(x - x_2)\ldots(x - x_n)}{(x_0 - x_1)(x_0 - x_2)\ldots(x_0 - x_n)}$$

$$L_1(x) = \frac{(x - x_0)(x - x_2)\dots(x - x_n)}{(x_1 - x_0)(x_1 - x_2)\dots(x_1 - x_n)}$$

$$\vdots$$

$$L_n(x) = \frac{(x - x_0)(x - x_1)\dots(x - x_{n-1})}{(x_n - x_0)(x_n - x_1)\dots(x_n - x_{n-1})}$$

que en forma más compacta y útil para programarse en un lenguaje de computadora quedaría

$$p_n(x) = \sum_{i=0}^{n} L_i(x) f(x_i) \tag{5.22}$$

donde*

$$L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)} \tag{5.23}$$

Al combinarse linealmente con $f(x_i)$, los polinomios $L_i(x)$, denominados polinomios de Lagrange, generan la aproximación polinomial de Lagrange a la información dada en forma tabular.

Ejemplo 5.1

Para la tabla que se presenta a continuación

a) Obtenga la aproximación polinomial de Lagrange con todos los puntos
b) Interpole el valor de la función $f(x)$ para $x = 1.8$

i	0	1	2	3
$f(x_i)$	–3	0	5	7
x_i	0	1	3	6

SOLUCIÓN

a) Obsérvese que hay cuatro puntos en la tabla, por lo que el polinomio será de tercer grado. Al sustituir los cuatro puntos en las ecuaciones generales 5.22 y 5.23 se obtiene

$$p_3(x) = (x-1)(x-3)(x-6)\frac{-3}{(0-1)(0-3)(0-6)} +$$

$$(x-0)(x-3)(x-6)\frac{0}{(1-0)(1-3)(1-6)}$$

* $\prod_{i=1}^{n} (x - x_i) = (x - x_1)(x - x_2)\dots(x - x_n)$

$$+ (x-0)(x-1)(x-6)\frac{5}{(3-0)(3-1)(3-6)}$$

$$+ (x-0)(x-1)(x-3)\frac{7}{(6-0)(6-1)(6-3)}$$

al efectuar las operaciones queda

$$p_3(x)=(x^3-10x^2+27x-18)(1/6)+(x^3-7x^2+6x)(-5/18)+(x^3-4x^2+3x)(7/90)$$

y finalmente resulta

$$p_3(x) = -\frac{3}{90}x^3 - \frac{3}{90}x^2 + \frac{276}{90}x - 3$$

b) El valor de $x=1.8$ se sustituye en la aproximación polinomial de Lagrange de tercer grado obtenida arriba y se tiene $f(1.8) \approx 2$.

Obsérvese que si se remplaza x con cualquiera de los valores dados en la tabla, en la aproximación polinomial, se obtiene el valor de la función dado por la misma tabla.

Ejemplo 5.2

Encuentre tanto la aproximación polinomial de Lagrange a la tabla 5.2 como el valor de la temperatura para una presión de 2 *atm* utilizando esta aproximación.

SOLUCIÓN

a) Aproximación polinomial de Lagrange mediante dos puntos ($n = 1$)

$$p(x) = \frac{x - x_1}{x_0 - x_1} f(x_0) + \frac{x - x_0}{x_1 - x_0} f(x_1) \qquad (5.24)$$

al sustituir los primeros dos puntos de la tabla resulta

$$p(x) = \frac{x - 5}{1 - 5}56.5 + \frac{x - 1}{5 - 1}113$$

Observe que la ecuación 5.24 es equivalente a la 5.5 y, por lo tanto, al sustituir $x = 2$ se obtiene el mismo resultado $T \approx 70.6$ °C, como era de esperar.

b) Aproximación polinomial de Lagrange con tres puntos (n=2)

$$p_2(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} f(x_0) + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} f(x_1)$$

$$+ \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} f(x_2)$$

al sustituir los primeros tres puntos de la tabla, se obtiene

$$p_2(x) = \frac{(x-5)(x-20)}{(1-5)(1-20)} 56.5 + \frac{(x-1)(x-20)}{(5-1)(5-20)} 113 + \frac{(x-1)(x-5)}{(20-1)(20-5)} 181 \tag{5.25}$$

polinomio que puede servir para interpolar la temperatura de ebullición de la acetona a la presión de 2 atm; así, el resultado queda T≈ 72.1. Observe que la ecuación 5.25 equivale a la 5.8.

c) La tabla 5.2 contiene cuatro puntos, por lo que la aproximación polinomial de mayor grado posible es 3. Se desarrolla la ecuación 5.22 para n=3

$$p_3(x) = \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} f(x_0) + \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} f(x_1)$$

$$+ \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} f(x_2) + \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} f(x_3) \tag{5.26}$$

Al sustituir los puntos de la tabla, se obtiene

$$p_3(x) = \frac{(x-5)(x-20)(x-40)}{(1-5)(1-20)(1-40)} 56.5 + \frac{(x-1)(x-20)(x-40)}{(5-1)(5-20)(5-40)} 113 +$$

$$\frac{(x-1)(x-5)(x-40)}{(20-1)(20-5)(20-40)} 181 + \frac{(x-1)(x-5)(x-20)}{(40-1)(40-5)(40-20)} 214.5$$

y al simplificar queda

$$p_3(x) = 0.01077\,x^3 - 0.78323\,x^2 + 18.4923\,x + 38.774$$

el cual puede emplearse para encontrar el valor de la temperatura correspondiente a la presión de 2 atm. Con la sustitución de $x = 2$ y al evaluar $p_3(x)$ queda:

$$T = f(2) \approx p_3(2) = 0.01077(2)^3 + 0.78323(2)^2 + 18.4923(2) + 38.774 = 72.7$$

ALGORITMO 5.2 Interpolación con polinomios de Lagrange

Para interpolar con polinomios de Lagrange de grado N, proporcionar los

DATOS: El grado del polinomio N, las N+1 parejas de valores (X(I), FX(I), I=0,1,...,N) y el valor para el que se desea la interpolación XINT.

RESULTADOS: La aproximación FXINT, el valor de la función en XINT.

```
PASO  1. Hacer FXINT = 0
PASO  2. Hacer I = 0
PASO  3. Mientras I ≤ N, repetir los pasos 4 a 10
           PASO  4. Hacer L = 1
           PASO  5. Hacer J = 0
           PASO  6. Mientras J ≤ N, repetir los pasos 7 y 8
                      PASO  7. Si I ≠ J
                                 Hacer L=L*(XINT-X(J))/(X(I)-X(J))
                      PASO  8. Hacer J = J + 1
           PASO 9.  Hacer FXINT = FXINT + L*FX(I)
           PASO 10. Hacer I = I + 1
PASO 11. IMPRIMIR FXINT y TERMINAR.
```

Ejemplo 5.3

Elabore un programa para aproximar la función $f(x) = \cos x$ en el intervalo $[0, 8\pi]$, con polinomios de Lagrange de grado 1, 2, 3, ..., 10. Use los puntos que se requieran, distribuidos regularmente en el intervalo.

Determine en forma práctica el error máximo que se comete al aproximar con los polinomios de los diferentes grados y compare los resultados.

SOLUCIÓN

El programa se encuentra en el disco (programa 5.1).

Para calcular el error máximo se dividió el intervalo $[0, 8\pi]$ en 20 subintervalos y se calculó el valor con el polinomio interpolante y el valor verdadero con la función cos x, determinando el error absoluto. Se obtuvieron los siguientes resultados

Grado	Error máximo
1	2.23627
2	2.23622
3	3.17025
4	2.23627
5	4.04277
6	4.1879
7	5.68560
8	33.74134
9	12.82475
10	35.95174

Se observa que al aumentar el grado del polinomio, el error absoluto máximo va aumentando.

Antes de pasar al estudio de otra forma de aproximación polinomial (de Newton), se requiere el conocimiento de las **diferencias divididas**, las cuales se presentan a continuación.

SECCIÓN 5.3 DIFERENCIAS DIVIDIDAS

Por definición de derivada en el punto x_0 de una función analítica $f(x)$ se tiene

$$f'(x) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

Sin embargo, cuando la función está en forma tabular

Puntos	0	1	2	...	n
x	x_0	x_1	x_2	...	x_n
$f(x)$	$f(x_0)$	$f(x_1)$	$f(x_2)$	...	$f(x_n)$

la derivada sólo puede obtenerse aproximadamente; por ejemplo, si se desea la derivada en el punto x, $(x_0 < x < x_1)$, puede estimarse como sigue

$$f'(x) \approx \frac{f(x_1) - f(x_0)}{x_1 - x_0}, \qquad x_0 < x < x_1$$

El lado derecho de la expresión anterior se conoce como la primera* diferencia dividida de $f(x)$ respecto a los argumentos x_0 y x_1 y se denota generalmente como $f[x_0, x_1]$; así

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

La relación entre la primera diferencia dividida y la primera derivada queda establecida por el teorema del valor medio

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0} = f'(\xi), \xi \in (x_1, x_0)$$

siempre y cuando $f(x)$ satisfaga las condiciones de aplicabilidad de dicho teorema.

Para obtener aproximaciones de derivadas de orden más alto, se extiende el concepto de diferencias divididas a órdenes más altos como se ve en la tabla 5.3, en donde para uniformar la notación se han escrito los valores funcionales en los argumentos x_i, $0 \leq i \leq n$, como $f[x_i]$ y se les llama diferencias divididas de orden cero.

Por otro lado, de acuerdo con la tabla 5.3, la diferencia de orden i es

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

En esta expresión puede observarse que

a) Para formarla se requieren $i + 1$ puntos y
b) El numerador es la resta de dos diferencias de orden $i - 1$ y el denominador la resta de los argumentos no comunes en el numerador.

Ejemplo 5.4

La información de la tabla siguiente se obtuvo del polinomio

$$y = x^3 - 2x^2 - 2$$

Puntos	0	1	2	3	4	5
x	–2	–1	0	2	3	6
$f(x)$	–18	–5	–2	–2	7	142

A partir de ella, elabore una tabla de diferencias divididas.

*Se llama también **diferencia dividida de primer orden.**

Información		Diferencias divididas			
x	$f(x)$	Primeras	Segundas	Terceras	...
x_0	$f[x_0]$	$f[x_0,x_1]=\dfrac{f[x_1]-f[x_0]}{x_1-x_0}$	$f[x_0,x_1,x_2]=\dfrac{f[x_1,x_2]-f[x_0,x_1]}{x_2-x_0}$	$f[x_0,x_1,x_2,x_3]=\dfrac{f[x_1,x_2,x_3]-f[x_0,x_1,x_2]}{x_3-x_0}$	...
x_1	$f[x_1]$	$f[x_1,x_2]=\dfrac{f[x_2]-f[x_1]}{x_2-x_1}$	$f[x_1,x_2,x_3]=\dfrac{f[x_2,x_3]-f[x_1,x_2]}{x_3-x_1}$	$f[x_1,x_2,x_3,x_4]=\dfrac{f[x_2,x_3,x_4]-f[x_1,x_2,x_3]}{x_4-x_1}$	...
x_2	$f[x_2]$	$f[x_2,x_3]=\dfrac{f[x_3]-f[x_2]}{x_3-x_2}$	$f[x_2,x_3,x_4]=\dfrac{f[x_3,x_4]-f[x_2,x_3]}{x_4-x_2}$	$f[x_2,x_3,x_4,x_5]=\dfrac{f[x_3,x_4,x_5]-f[x_2,x_3,x_4]}{x_5-x_2}$	...
x_3	$f[x_3]$	$f[x_3,x_4]=\dfrac{f[x_4]-f[x_3]}{x_4-x_3}$	$f[x_3,x_4,x_5]=\dfrac{f[x_4,x_5]-f[x_3,x_4]}{x_5-x_3}$		
x_4	$f[x_4]$	$f[x_4,x_5]=\dfrac{f[x_5]-f[x_4]}{x_5-x_4}$			
x_5	$f[x_5]$				

Tabla 5.3 Tabulación general de diferencis divididas.

SOLUCIÓN

Las primeras diferencias divididas mediante los puntos (0), (1) y (1), (2), respectivamente, son

$$f[x_0,x_1] = \frac{-5-(-18)}{-1-(-2)} = 13; \qquad f[x_1,x_2] = \frac{-2-(-5)}{0-(-1)} = 3$$

La segunda diferencia dividida mediante los puntos (0), (1) y (2) es

$$f[x_0,x_1,x_2] = \frac{3-13}{0-(-2)} = -5$$

de igual manera se calculan las demás diferencias divididas, que se resumen en la siguiente tabla

Puntos	x	$f(x)$	1er orden	2do orden	3er orden	4o orden
0	−2	−18				
			13			
1	−1	−5		−5		
			3		1	
2	0	−2		−1		0
			0		1	
3	2	−2		3		0
			9		1	
4	3	7		9		
			45			
5	6	142				

Debe notarse que todas las diferencias divididas de tercer orden tienen el mismo valor, independientemente de los argumentos que se usen para su cálculo. Obsérvese también que las diferencias divididas de cuarto orden son todas cero, lo cual concuerda con que la tercera y cuarta derivada de un polinomio de tercer grado son —respectivamente— una constante y cero, sea cual sea el valor del argumento x. El razonamiento inverso también es válido: si al construir una tabla de diferencias divididas en alguna de las columnas el valor es constante (y en la siguiente columna es cero), la información proviene de un polinomio de grado igual al orden de las diferencias que tengan valores constantes.

ALGORITMO 5.3 Tabla de diferencias divididas

Para obtener la tabla de diferencias divididas de una función dada en forma tabular, proporcionar los

DATOS: El número de parejas M de la función tabular y las parejas de valores (X(I),FX(I), I=0, 1, 2,..., M-1).

RESULTADOS: La tabla de diferencias divididas T.

```
PASO 1.  Hacer N = M-1
PASO 2.  Hacer I = 0
PASO 3.  Mientras I ≤ N-1, repetir los pasos 4 y 5.
           PASO 4.  Hacer T(I,0)=(FX(I+1)-FX(I))/(X(I+1)-X(I))
           PASO 5.  Hacer I = I+1
PASO 6.  Hacer J = 1
PASO 7.  Mientras J ≤ N-1, repetir los pasos 8 a 12.
           PASO 8.  Hacer I = J
           PASO 9.  Mientras I ≤ N-1, repetir los pasos 10 y 11.
                      PASO 10. Hacer
                                 T(I,J)=(T(I,J-1)-
                                 T(I-1,J-1))/(X(I+1)-X(I-J))
                      PASO 11. Hacer I = I+1
           PASO 12. Hacer J = J+1
PASO 13. IMPRIMIR T y TERMINAR.
```

SECCIÓN 5.4 APROXIMACIÓN POLINOMIAL DE NEWTON

Supóngase que se tiene una función dada en forma tabular como se presenta a continuación

Puntos	0	1	2	3	...	n
x	x_0	x_1	x_2	x_3	...	x_n
$f(x)$	$f[x_0]$	$f[x_1]$	$f[x_2]$	$f[x_3]$	...	$f[x_n]$

y que se desea aproximarla preliminarmente con un polinomio de primer grado que pasa por ejemplo por los puntos (0) y (1). Sea además dicho polinomio de la forma

$$p(x) = a_0 + a_1(x - x_0), \tag{5.27}$$

donde x_0 es la abscisa del punto (0) y a_0, a_1 son constantes por determinar. Para encontrar el valor de a_0 se hace $x = x_0$ de donde $a_0 = p(x_0) = f[x_0]$ y a fin de encontrar el valor de a_1 se hace $x = x_1$, de donde $a_1 = (f[x_1] - f[x_0])/(x_1 - x_0)$, o sea la primera diferencia dividida $f[x_1, x_0]$.

Al sustituir los valores de estas constantes en la ecuación 5.27 ésta queda

$$p(x) = f[x_0] + (x - x_0) f[x_0, x_1]$$

o sea un polinomio de primer grado en términos de diferencias divididas.

Y si ahora se desea aproximar la función tabular con un polinomio de segundo grado que pase por los puntos (0), (1) y (2) y que tenga la forma

$$p_2(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1), \tag{5.28}$$

donde x_0 y x_1 vuelven a ser las abscisas de los puntos (0) y (1) y a_0, a_1 y a_2 son constantes por determinar, se procede como en la forma anterior para encontrar estas constantes; o sea

$$\text{si } x = x_0,\ a_0 = p_2(x_0) = f[x_0]$$

$$\text{si } x = x_1,\ a_1 = \frac{f[x_1] - f[x_0]}{x_1 - x_0} = f[x_0, x_1]$$

$$\text{si } x = x_2,\ a_2 = \frac{f[x_2] - f[x_0] - (x_2 - x_0)\dfrac{f[x_1] - f[x_0]}{x_1 - x_0}}{(x_2 - x_0)(x_2 - x_1)}$$

al desarrollar algebraicamente el numerador y el denominador de a_2 se llega a*

$$a_2 = \frac{\dfrac{f[x_2] - f[x_1]}{x_2 - x_1} - \dfrac{f[x_1] - f[x_0]}{x_1 - x_0}}{x_2 - x_0} = f[x_0, x_1, x_2]$$

que es la segunda diferencia dividida respecto a x_0, x_1 y x_2.

Con la sustitución de estos coeficientes en la ecuación 5.28 se obtiene

$$p_2(x) = f[x_0] + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2]$$

que es un polinomio de segundo grado en términos de diferencias divididas.

Por inducción se puede establecer que, en general, para un polinomio de grado n escrito en la forma

$$p_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \ldots + a_n(x - x_0)(x - x_1)\ldots(x - x_{n-1}) \tag{5.29}$$

*Véase el problema 5.11

y que pasa por los puntos (0), (1), (2),..., (n); los coeficientes $a_0, a_1, ..., a_n$ están dados por

$$\begin{aligned} a_0 &= f[x_0] \\ a_1 &= f[x_0, x_1] \\ a_2 &= f[x_0, x_1, x_2] \\ &\cdot \\ &\cdot \\ &\cdot \\ a_n &= f[x_0, x_1, x_2, ..., x_n] \end{aligned}$$

Esta aproximación polinomial se conoce como apoximación polinomial de Newton, la cual se puede expresar sintéticamente como

$$p_n(x) = \sum_{k=0}^{n} a_k \prod_{i=0}^{k-1} (x - x_i) \tag{5.30}$$

Ejemplo 5.5

Elabore una aproximación polinomial de Newton para la información tabular de las presiones de vapor de la acetona (tabla 5.2) e interpole la temperatura para una presión de 2 atm.

SOLUCIÓN

Para el cálculo de los coeficientes del polinomio de Newton, se construye la tabla de diferencias divididas

			Diferencia divididas		
Puntos	P	T	Primera	Segunda	Tercera
0	1	56.5			
			14.125		
1	5	113		–0.50482	
			4.533		0.01085
2	20	181		–0.08167	
			1.675		
3	40	214.5			

a) Para $n = 1$

$$p(x) = a_0+a_1(x-x_0) = f[x_0] + f[x_0, x_1](x-x_0)$$

de la tabla se tiene $f[x_0] = 56.5$ y $f[x_0, x_1] = 14.125$, de donde

$$p(x) = 56.5 + 14.125(x - 1)$$

ecuación que equivale a las obtenidas anteriormente (5.5 y 5.24).

$$\text{Si } x = 2, f(2) \approx p(2) = 56.5 + 14.125(2-1) = 70.6\ ^\circ\text{C}$$

b) Para $n = 2$

$$p_2(x) = a_0+a_1(x-x_0)+a_2(x-x_0)(x-x_1)$$
$$= f[x_0]+f[x_0, x_1](x-x_0)+f[x_0, x_1, x_2](x-x_0)(x-x_1)$$

de la tabla se obtienen $a_0 = f[x_0] = 56.5$, $a_1 = f[x_0, x_1] = 14.125$, $a_2 = f[x_0, x_1, x_2] = -0.50482$, que al sustituirse en la ecuación de arriba dan

$$p_2(x) = 56.5 + 14.125(x - 1) - 0.50482(x - 1)(x - 5)$$

ecuación que equivale a 5.8 y 5.25

$$\text{Si } x = 2, f(2) \approx p_2(2) = 56.5+14.125(2-1)-0.50482(2-1)(2-5) = 72.1\ ^\circ\text{C}$$

c) Para $n = 3$

$$p_3(x) = a_0+a_1(x-x_0)+a_2(x-x_0)(x-x_1)+a_3(x-x_0)(x-x_1)(x-x_2)$$
$$= f[x_0]+f[x_0, x_1](x-x_0)+f[x_0, x_1, x_2](x-x_0)(x-x_1)+$$
$$f[x_0, x_1, x_2, x_3](x-x_0)(x-x_1)(x-x_2)$$

de la tabla se obtienen $a_0 = f[x_0] = 56.5$, $a_1 = f[x_0, x_1] = 14.125$,

$$a_2 = f[x_0, x_1, x_2] = -0.50842,\ a_3 = f[x_0, x_1, x_2, x_3] = 0.01085.$$

que sustituidas generan el polinomio de aproximación

$$p_3(x) = 56.5+14.125(x-1)-0.50482(x-1)(x-5)+0.01085(x-1)(x-5)(x-20)$$

y que es esencialmente el polinomio obtenido con el método de Lagrange (ecuación 5.26).

$$\text{Si } x = 2, f(2) \approx p_3(2) = 56.5+14.125(2-1)-0.50482(2-1)(2-5)+$$
$$0.01085(2-1)(2-5)(2-20) = 71.6\ ^\circ\text{C}$$

Ejemplo 5.6

Aproxime la temperatura de ebullición de la acetona a una presión de 30 atm usando aproximación polinomial de Newton de grado dos (véase Ej. 5.5).

SOLUCIÓN

Se hace pasar el polinomio de Newton por los puntos (1), (2) y (3), con lo que toma la forma

$$p_2(x) = a_0 + a_1(x - x_1) + a_2(x - x_1)(x - x_2),$$

con los coeficientes dados ahora de la siguiente manera

$$a_0 = f[x_1]$$
$$a_1 = f[x_1, x_2]$$
$$a_2 = f[x_1, x_2, x_3].$$

Al sustituir

$$p_2(x) = f[x_1] + f[x_1,x_2](x - x_1) + f[x_1,x_2,x_3](x-x_1)(x-x_2)$$
$$= 113 + 4.533(x - x_1) - 0.08167(x - x_1)(x - x_2)$$

y al evaluar dicho polinomio en $x = 30$, se obtiene la aproximación buscada

$$T = p_2(30) = 113 + 4.533(30 - 5) - 0.08167(30 - 5)(30 - 20)$$
$$= 205.9$$

El valor reportado en la tabla 5.1 es 205, por lo que la aproximación es buena.

ALGORITMO 5.4 Interpolación polinomial de Newton

Para interpolar con polinomios de Newton en diferencias divididas de grado N, proporcionar los

DATOS:	El grado del polinomio N, las N+1 parejas de valores (X(I), FX(I), I=0, 1,2,...,N) y el valor para el que se desea interpolar XINT.
RESULTADOS:	La aproximación FXINT al valor de la función en XINT.

```
PASO  1. Realizar los pasos 2 a 12 del algoritmo 5.3.
PASO  2. Hacer FXINT = FX(0)
PASO  3. Hacer I = 0
PASO  4. Mientras I ≤ N-1, repetir los pasos 5 a 11.
             PASO  5. Hacer P = 1
             PASO  6. Hacer J = 0
             PASO  7. Mientras J ≤ I, repetir los pasos 8 y 9.
                          PASO  8. Hacer P = P * (XINT - X(J))
                          PASO  9. Hacer J = J + 1
             PASO 10. Hacer FXINT = FXINT + T(I,I)*P
             PASO 11. Hacer I = I + 1
PASO 12. IMPRIMIR FXINT y TERMINAR.
```

SECCIÓN 5.5 POLINOMIO DE NEWTON EN DIFERENCIAS FINITAS

Cuando la distancia h entre dos argumentos consecutivos cualesquiera, es la misma a lo largo de la tabla, el polinomio de Newton en diferencias divididas puede expresarse con más sencillez. Para este propósito se introduce un nuevo parámetro s, definido en $x = x_0+sh$, con el cual se expresa el factor productoria

$$\prod_{i=0}^{k-1} (x - x_i),$$

de la ecuación 5.30 en términos de s y h. Para esto obsérvese que $x_1 - x_0 = h$, $x_2 - x_0 = 2h$, ..., $x_i - x_0 = ih$ y que restando $x_i (0 \le i \le n)$ en ambos miembros de $x = x_0 + sh$, se obtiene

$$x - x_i = x_0 - x_i + sh = -ih + sh = h(s - i) \qquad \text{para } (0 \le i \le n)$$

Por ejemplo si $i = 1$,

$$x - x_1 = h(s - 1)$$

si $i = 2$,

$$x - x_2 = h(s - 2)$$

Al sustituir cada una de las diferencias $(x - x_i)$ con $h(s - i)$, en la ecuación 5.29, se llega a

$$\begin{aligned} p_n(x) = p_n(x_0 + sh) = {} & f[x_0] + hsf[x_0, x_1] + h^2 s(s-1) f[x_0, x_1, x_2] \\ & + h^3 s(s-1)(s-2) f[x_0, x_1, x_2, x_3] + \cdots \\ & + h^n s(s-1)(s-2)\cdots(s-(n-1)) f[x_0, x_1, \ldots, x_n] \end{aligned} \tag{5.31}$$

o en forma compacta

$$p_n(x) = \sum_{k=0}^{n} a_k h^k \prod_{i=0}^{k-1} (s - i) \tag{5.32}$$

Esta última ecuación puede simplificarse aún más si se introduce el operador lineal Δ, conocido como **operador lineal en diferencias hacia delante** y definido sobre $f(x)$ como

$$\Delta f(x) = f(x + h) - f(x)$$

La segunda diferencia hacia delante puede obtenerse como sigue

$$\begin{aligned}\Delta(\Delta f(x)) = \Delta^2 f(x) &= \Delta(f(x + h) - f(x)) \\ &= \Delta f(x + h) - \Delta f(x) \\ &= f(x + h + h) - f(x + h) - f(x + h) + f(x) \\ &= f(x + 2h) - 2f(x + h) + f(x)\end{aligned}$$

A su vez, las diferencias hacia delante de orden superior se generan como sigue

$$\Delta^i f(x) = \Delta(\Delta^{i-1} f(x))$$

Estas diferencias se conocen como diferencias finitas hacia delante. Análogamente cabe definir ∇ como **operador lineal de diferencias hacia atrás**; así, la primera diferencia hacia atrás se expresa como

$$\nabla f(x) = f(x) - f(x - h)$$

La segunda diferencia hacia atrás queda

$$\begin{aligned}\nabla^2 f(x) &= \nabla(\nabla f(x)) = \nabla(f(x) - f(x - h)) \\ \nabla^2 f(x) &= f(x) - f(x - h) - f(x - h) + f(x - 2h) \\ \nabla^2 f(x) &= f(x) - 2f(x - h) + f(x - 2h)\end{aligned}$$

de tal modo que las diferencias hacia atrás de orden superior se expresan en términos generales como

$$\nabla^i f(x) = \nabla(\nabla^{i-1} f(x)).$$

Estas diferencias se conocen como diferencias finitas hacia atrás.

Al aplicar Δ al primer valor funcional $f[x_0]$ de una tabla se tiene

$$\Delta f(x_0) = f[x_1] - f[x_0] = h f[x_0, x_1],$$

de manera que

$$f[x_0, x_1] = \frac{1}{h} \Delta f(x_0)$$

Del mismo modo

$$f[x_0, x_1, x_2] = \frac{\frac{f[x_2]-f[x_1]}{x_2-x_1} - \frac{f[x_1]-f[x_0]}{x_1-x_0}}{x_2-x_0} = \frac{f[x_2]-2f[x_1]+f[x_0]}{2h^2}$$

por lo que

$$f[x_0, x_1, x_2] = \frac{1}{2h^2}\Delta^2 f(x_0)$$

En general

$$f[x_0, x_1, \ldots, x_n] = \frac{1}{n!\,h^n}\Delta^n f(x_0) \tag{5.33}$$

De igual manera, las diferencias divididas en función de las diferencias hacia atrás quedan

$$f[x_n, x_{n-1}, \ldots, x_0] = \frac{1}{n!\,h^n}\nabla^n f(x_n) \tag{5.34}$$

Consecuentemente, al sustituir $f[x_0, x_1, ..., x_i]$, $(0 \le i \le n)$ en términos de diferencias finitas, la ecuación 5.31 queda

$$p_n(x) = p_n(x_0+sh) = f[x_0] + s\Delta f[x_0] + \frac{s(s-1)}{2!}\Delta^2 f[x_0] +$$

$$+ \frac{s(s-1)(s-2)}{3!}\Delta^3 f[x_0] + \ldots \tag{5.35}$$

$$+ \frac{s(s-1)(s-2)\ldots(s-(n-1))}{n!}\Delta^n f[x_0]$$

conocido como el **polinomio de Newton en diferencias finitas hacia delante.**

Existe una expresión equivalente a la 5.35 para diferencias hacia atrás (**polinomio de Newton en diferencias finitas hacia atrás**), cuya obtención se motiva al final del ejemplo siguiente.

Ejemplo 5.7

La siguiente tabla proporciona las presiones de vapor en lb/plg^2 a diferentes temperaturas para el 1-3 butadieno.

Puntos	0	1	2	3	4	5
T °F	50	60	70	80	90	100
P lb/plg^2	24.94	30.11	36.05	42.84	50.57	59.30

Aproxime la función tabulada por el polinomio de Newton en diferencias hacia delante e interpole la presión a la temperatura de 64 °F.

SOLUCIÓN

Primero se construye la tabla de diferencias hacia delante como sigue

Punto	x_i	$f[x_i]$	$\Delta f[x_i]$	$\Delta^2 f[x_i]$	$\Delta^3 f[x_i]$	$\Delta^4 f[x_i]$
0	50	24.94				
			$\Delta f[x_0]=5.17$			
1	60	30.11		$\Delta^2 f[x_0]=0.77$		
			$\Delta f[x_1]=5.94$		$\Delta^3 f[x_0]=0.08$	
2	70	36.05		$\Delta^2 f[x_1]=0.85$		$\Delta^4 f[x_0]=0.01$
			$\Delta f[x_2]=6.79$		$\Delta^3 f[x_1]=0.09$	
3	80	42.84		$\Delta^2 f[x_2]=0.94$		$\Delta^4 f[x_1]=-0.03$
			$\Delta f[x_3]=7.73$		$\Delta^3 f[x_2]=0.06$	
4	90	50.57		$\Delta^2 f[x_3]=1.00$		
			$\Delta f[x_4]=8.73$			
5	100	59.30				

Observe que en esta información $h=10$, el valor por interpolar es 64 y que el valor de s se obtiene de la expresión $x = x_0 + sh$; esto es

$$s = \frac{x - x_0}{h} = \frac{64 - 50}{10} = 1.4$$

Si se deseara aproximar con un polinomio de primer grado, se tomarían sólo los dos primeros términos de la ecuación 5.35; o sea,

$$p(x) = f[x_0] + s\Delta f[x_0] = 24.94 + 1.4(5.17) = 32.17$$

Nótese que realmente se está extrapolando, ya que el valor de x queda fuera del intervalo de los puntos que se usaron para formar el polinomio de aproximación.

Intuitivamente se piensa que se obtendría una aproximación mejor con los puntos (1) y (2). Sin embargo, la ecuación 5.35 se desarrolló usando x_0 como pivote y para aplicarla con el punto (1) y (2) debe modificarse a la forma siguiente

$$p_n(x) = f[x_1 + sh] = f[x_1] + s\Delta f[x_1] + \frac{s(s-1)}{2!}\Delta^2 f[x_1] + \ldots + \frac{s(s-1)\ldots(s-(n-1))}{n!}\Delta^n f[x_1] \qquad (5.36)$$

la cual usa como pivote x_1 y cuyos primeros dos términos dan la aproximación polinomial de primer grado

$$p(x) = f[x_1] + s\Delta f[x_1] \text{ donde ahora } s = \frac{x - x_1}{h} = \frac{64 - 60}{10} = 0.4;$$

al sustituir valores de la tabla se tiene

$$f(64) \approx p(64) = 30.11 + 0.4(5.94) = 32.49$$

En cambio, si se deseara aproximar con un polinomio de segundo grado, se requerirían tres puntos y sería aconsejable tomar (0), (1) y (2) en lugar de (1), (2) y (3), ya que el argumento por interpolar está más al centro de los primeros. Con esta selección y la ecuación 5.35 queda

$$p_2(x) = f[x_0] + s\Delta f[x_0] + \frac{s(s-1)}{2!}\Delta^2 f[x_0]$$

donde

$$s = \frac{x - x_0}{h} = \frac{64 - 50}{10} = 1.4;$$

este valor se sustituye arriba y queda

$$p_2(64) = 24.94 + 1.4(5.17) + \frac{1.4(1.4 - 1)}{2!}0.77 = 32.385$$

Si se quisiera interpolar el valor de la presión a una temperatura de 98 °F, tendría que desarrollarse una ecuación de Newton en diferencias hacia delante, usando como pivote el punto (4) para un polinomio de primer grado o el punto (3) para un polinomio de segundo grado, etc. Sin embargo, esto es factible usando un solo pivote (el punto 5 en este caso), independientemente del grado del polinomio por usar, si se emplean diferencias hacia atrás.

Para esto se debe desarrollar una ecuación equivalente a la 5.35 pero en diferencias hacia atrás; este desarrollo se presenta a continuación en dos pasos —el primero es un resultado necesario.

Primer paso

Obtención del polinomio de Newton en diferencias divididas hacia atrás de grado n apoyado en el punto x_n.

Para simplificar se inicia con $n = 2$ y se asume que un polinomio de segundo grado en general tiene la forma

$$p_2(x) = a_0 + a_1(x - x_n) + a_2(x - x_n)(x - x_{n-1})$$

donde a_0, a_1 y a_2 son las constantes por determinar y x_n y x_{n-1} las abscisas de los puntos (n) y $(n-1)$, respectivamente.

Si $x = x_n$, $a_0 = p(x_n) = f[x_n]$

$$\text{Si } x = x_{n-1},\ a_1 = \frac{p_2(x_{n-1}) - p_2(x_n)}{x_{n-1} - x_n} = f[x_n, x_{n-1}]$$

$$\text{Si } x = x_{n-2},\ a_2 = \frac{p_2(x_{n-2}) - p_2(x_n) - f[x_{n-1}, x_n](x_{n-2} - x_n)}{(x_{n-2} - x_n)(x_{n-2} - x_{n-1})}$$

al desarrollar algebraicamente el numerador y el denominador de a_2 se llega a

$$a_2 = f[x_n, x_{n-1}, x_{n-2}]$$

al sustituir estas constantes en el polinomio queda

$$p_2(x) = f[x_n] + (x - x_n) f[x_n, x_{n-1}] + (x - x_n)(x - x_{n-1}) f[x_n, x_{n-1}, x_{n-2}]$$

De lo anterior se puede inducir que, en general, para un polinomio de grado n escrito en la forma

$$p_n(x) = a_0 + a_1(x - x_n) + a_2(x - x_n)(x - x_{n-1}) + \ldots + a_n(x - x_n)(x - x_{n-1})\ldots(x - x_1), \qquad (5.37)$$

los coeficientes $a_0, a_1, a_2, \ldots, a_n$ están dados por

$$\begin{aligned} a_0 &= f[x_n] \\ a_1 &= f[x_n, x_{n-1}] \\ &\vdots \\ a_n &= f[x_n, x_{n-1}, x_{n-2}, \ldots, x_0] \end{aligned}$$

Segundo paso

Obtención del polinomio de Newton en diferencias finitas hacia atrás de grado n, apoyado en el punto x_n.

Las ecuaciones* siguientes se pueden construir introduciendo el parámetro s definido ahora por la expresión $x = x_n + sh$.

$$\begin{aligned} x - x_n &= sh \\ x - x_{n-1} &= x_n - x_{n-1} + sh = h(s + 1) \\ x - x_{n-2} &= x_n - x_{n-2} + sh = h(s + 2) \\ &\vdots \\ x - x_0 &= x_n - x_0 + sh = h(s + n) \end{aligned}$$

*Recuérdese que se considera aquí que la diferencia entre dos argumentos consecutivos cualesquiera es h.

Al sustituir las ecuaciones anteriores y los coeficientes $f[x_n]$, $f[x_n, x_{n-1}]$, ..., $f[x_n, x_{n-1}, ..., x_0]$ en la ecuación 5.37 en términos de diferencias finitas (Ec. 5.34), finalmente queda

$$p_n(x_n + sh) = f[x_n] + s\nabla f[x_n] + \frac{s(s+1)}{2!}\nabla^2 f[x_n] + \cdots + \frac{s(s+1)\cdots(s+(n-1))}{n!}\nabla^n f[x_n] \tag{5.38}$$

que es la ecuación de Newton en diferencias hacia atrás.

Ejemplo 5.8

Interpole el valor de la presión a una temperatura de 98 °F, utilizando la tabla de presiones de vapor del ejemplo 5.7 y el polinomio de Newton (5.38).

SOLUCIÓN

Primero se construye la tabla de diferencias hacia atrás como sigue

Punto	x_i	$f[x_i]$	$\nabla f[x_i]$	$\nabla^2 f[x_i]$	$\nabla^3 f[x_i]$	$\nabla^4 f[x_i]$
0	50	24.94				
			$\nabla f[x_1]=5.17$			
1	60	30.11		$\nabla^2 f[x_2]=0.77$		
			$\nabla f[x_2]=5.94$		$\nabla^3 f[x_3]=0.08$	
2	70	36.05		$\nabla^2 f[x_3]=0.85$		$\nabla^4 f[x_4]=0.01$
			$\nabla f[x_3]=6.79$		$\nabla^3 f[x_4]=0.09$	
3	80	42.84		$\nabla^2 f[x_4]=0.94$		$\nabla^4 f[x_5]=-0.03$
			$\nabla f[x_4]=7.73$		$\nabla^3 f[x_5]=0.06$	
4	90	50.57		$\nabla^2 f[x_5]=1.00$		
			$\nabla f[x_5]=8.73$			
5	100	59.30				

Si se usa un polinomio de primer grado, se tiene de la ecuación 5.38

$$p(98) = f[x_5] + s\,\nabla f[x_5]$$

donde

$$s = \frac{x - x_n}{h} = \frac{98 - 100}{10} = -0.2;$$

y con la tabla de diferencias finitas hacia atrás

$$p_2(98) = 59.3 - 0.2(8.73) = 57.55$$

Si en cambio se usa un polinomio de segundo grado, se emplean los tres primeros términos de la ecuación 5.38, con lo cual la aproximación queda

$$p_2(98) = f[x_5] + s\nabla f[x_5] + \frac{s(s+1)}{2!}\nabla^2 f[x_5]$$

$$= 59.3 - 0.2(8.73) + \frac{-0.2(-0.2+1)}{2!}(1) = 57.63$$

Si se deseara interpolar el valor de la presión a una temperatura de 82 °F, tendría que usarse la ecuación 5.38 pero apoyada en el punto n-1 [punto (4) en este caso]; esto es,

$$p_n(x_{n-1} + sh) = f[x_{n-1}] + s\nabla f[x_{n-1}] + \frac{s(s+1)}{2!}\nabla^2 f[x_{n-1}] + \cdots + \frac{s(s+1)\cdots(s+(n-1))}{n!}\nabla^n f[x_{n-1}] \tag{5.39}$$

Nota

Es importante hacer notar que las tablas de los ejemplos 5.7 (diferencias hacia delante) y 5.8 (diferencias hacia atrás) presentan los mismos valores numéricos aunque los operadores y subíndices de sus argumentos no sean los mismos. Por lo anterior, el polinomio de Newton en diferencias hacia adelante y su tabla correspondiente pueden usarse a fin de interpolar en puntos del final de la tabla con sólo invertir la numeración de los puntos en dicha tabla y los argumentos de cada columna de diferencias finitas.

También es útil observar que los valores de la tabla utilizados en las ecuaciones 5.35, 5.36 o alguna modificación de éstas, son los de las diagonales trazadas de arriba hacia abajo (véase tabla del ejemplo 5.7) y que los valores utilizados en 5.38, 5.39 o alguna modificación de éstas, son los de las diagonales trazadas de abajo hacia arriba (ver tabla del ejemplo 5.8).

Se resuelve un ejemplo para ilustrar esto.

Ejemplo 5.9

Con la ecuación 5.35 y la tabla de diferencias hacia delante del ejemplo 5.7, interpole la presión de vapor del 1-3 butadieno a la temperatura de 98 °F, mediante un polinomio de primer y segundo grado.

SOLUCIÓN

Invertidos la numeración de los puntos en la tabla mencionada y los argumentos de cada columna, la tabla toma el aspecto

Punto	x_i	$f[x_i]$	$\Delta f[x_i]$	$\Delta^2 f[x_i]$	$\Delta^3 f[x_i]$	$\Delta^4 f[x_i]$
5	50	24.94				
			$\Delta f[x_4]=5.17$			
4	60	30.11		$\Delta^2 f[x_3]=0.77$		
			$\Delta f[x_3]=5.94$		$\Delta^3 f[x_2]=0.08$	
3	70	36.05		$\Delta^2 f[x_2]=0.85$		$\Delta^4 f[x_1]=0.01$
			$\Delta f[x_2]=6.79$		$\Delta^3 f[x_1]=0.09$	
2	80	42.84		$\Delta^2 f[x_1]=0.94$		$\Delta^4 f[x_0]=-0.03$
			$\Delta f[x_1]=7.73$		$\Delta^3 f[x_0]=0.06$	
1	90	50.57		$\Delta^2 f[x_0]=1.00$		
			$\Delta f[x_0]=8.73$			
0	100	59.30				

Observe que todos los valores numéricos conservan su posición en la tabla.

Se emplea la ecuación 5.35 con $x = 98$, $x_0 = 100$ y $h = 10$, de donde

$$s = \frac{x - x_0}{h} = \frac{98 - 100}{10} = -0.2$$

al emplear un polinomio de primer grado se tiene

$$p(98) = 59.30 + (-0.2)(8.73) = 57.55$$

En cambio, con uno de segundo grado

$$p_2(98) = 59.30 + (-0.2)(8.73) + \frac{(-0.2)(-0.2+1)}{2!} 1 = 57.63$$

Como puede verse, son los mismos resultados que se obtuvieron en el ejemplo 5.8.

SECCIÓN 5.6 ESTIMACIÓN DE ERRORES EN LA APROXIMACIÓN

En general, al aproximar una función por un polinomio de grado n, se comete un error; por ejemplo, cuando se utiliza un polinomio de primer grado, se remplaza la función verdadera con una línea recta (Fig. 5.4). En términos matemáticos, la función se podría representar exactamente como

$$f(x) = f[x_0] + (x-x_0) f[x_0, x_1] + R_1(x) = p_1(x) + R_1(x) \qquad (5.40)$$

donde $R_1(x)$ es el error cometido al aproximar linealmente la función $f(x)$ y $p_1(x)$ es por ejemplo el polinomio de primer grado en diferencias divididas.

Al despejar $R_1(x)$ de la ecuación 5.40 y tomando como factor común $(x - x_0)$ queda

$$R_1(x) = f(x) - f[x_0] - (x - x_0) f[x_0, x_1]$$

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

$$= (x - x_0) (f[x_0, x] - f[x_0, x_1])$$

al multiplicar y dividir por $(x-x_1)$ se obtiene

$$R_1(x) = (x-x_0)(x-x_1) f[x, x_0, x_1]$$

donde $f[x, x_0, x_1]$ es la segunda diferencia dividida respecto a los argumentos x_0, x_1 y x. Resulta imposible calcular exactamente $f[x, x_0, x_1]$, ya que no se conoce la $f(x)$ necesaria para su evaluación. Sin embargo, si se tiene otro valor de $f(x)$, sea $f(x_2)$ (y si la segunda diferencia $f[x, x_0, x_1]$ no varía significativamente en el intervalo donde están los puntos x_0, x_1 y x_2), entonces $R_1(x)$ se aproxima de la siguiente manera

$$R_1(x) \approx (x-x_0)(x-x_1) f[x_0, x_1, x_2]$$

de tal modo que al sustituirlo en la ecuación original quede

$$f(x) \approx f[x_0] + (x-x_0) f[x_0, x_1] + (x-x_0)(x-x_1) f[x_0, x_1, x_2]$$

Observe que el lado derecho de esta expresión es el polinomio de segundo grado en diferencias divididas. Como se había intuido, esto confirma que —en general— se aproxima mejor la función $f(x)$ con un polinomio de grado dos que con uno de primer grado.

Por otro lado, si se aproxima la función $f(x)$ con un polinomio de segundo grado $p_2(x)$, se espera que el error $R_2(x)$ sea en general menor. La función expresada en estos términos queda

$$f(x) = p_2(x) + R_2(x) = f[x_0] + (x-x_0) f[x_0, x_1] + (x-x_0)(x-x_1) f[x_0, x_1, x_2] + R_2(x)$$

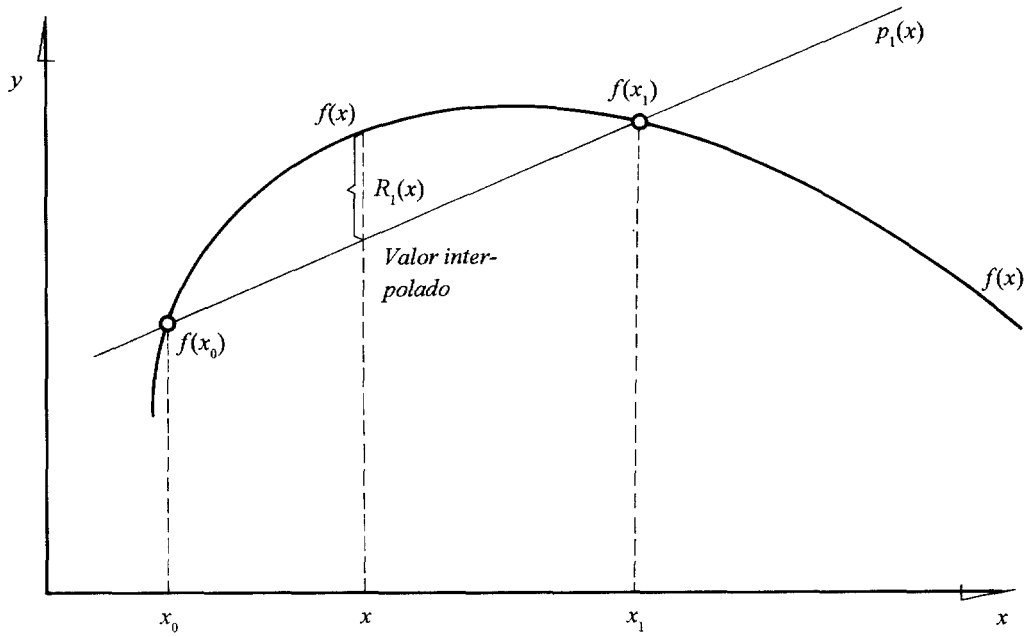


Figura 5.4.

de donde $R_2(x)$ puede despejarse

$$R_2(x) = f(x) - f[x_0] - (x-x_0) f[x_0, x_1] - (x-x_0)(x-x_1) f[x_0, x_1, x_2]$$

y como en el caso de un polinomio de primer grado, se demuestra* que el término del error para la aproximación polinomial de segundo grado es

$$R_2(x) = (x-x_0)(x-x_1)(x-x_2) f[x, x_0, x_1, x_2]$$

De igual modo que $f[x, x_0, x_1]$ en el caso lineal $f[x, x_0, x_1, x_2]$ no se puede determinar con exactitud; sin embargo, si se tiene un punto adicional $(x_3, f(x_3))$, cabe aproximar $f[x, x_0, x_1, x_2]$ con

$$f[x, x_0, x_1, x_2] \approx f[x_0, x_1, x_2, x_3],$$

que sustituida proporciona una aproximación a $R_2(x)$

$$R_2(x) \approx (x-x_0)(x-x_1)(x-x_2) f[x_0, x_1, x_2, x_3].$$

Si se continúa este proceso puede establecerse por inducción que

$$f(x) = p_n(x) + R_n(x),$$

*Veáse el problema 5.24.

donde $p_n(x)$ es el polinomio de grado n en diferencias divididas que aproxima la función tabulada, y $R_n(x)$ es el término correspondiente del error. Esto es

$$p_n(x) = f[x_0] + (x-x_0)f[x_0, x_1] + \ldots + (x-x_0)(x-x_1)\ldots(x-x_{n-1})f[x_0, \ldots, x_n]$$

y

$$R_n(x) = (x-x_0)(x-x_1)\ldots(x-x_n)\, f[x, x_0, x_1, \ldots, x_n]$$

o

$$R_n(x) = \left[\prod_{i=0}^{n}(x - x_i)\right] f[x, x_0, x_1, \ldots, x_n] \tag{5.41}$$

en donde $f[x, x_0, x_1, \ldots, x_n]$ puede aproximarse con un punto adicional $(x_{n+1}, f(x_{n+1}))$ así:

$$f[x, x_0, x_1, \ldots, x_n] \approx f[x_0, x_1, x_2, \ldots, x_n, x_{n+1}] \tag{5.42}$$

entonces $R_n(x)$ queda como

$$R_n(x) \approx \left[\prod_{i=0}^{n}(x - x_i)\right] f[x_0, x_1, x_2, \ldots, x_n, x_{n+1}]$$

La ecuación

$$f(x) = p_n(x) + R_n(x)$$

es conocida como la **fórmula fundamental de Newton en diferencias divididas.** Al analizar el factor productoria (producto acumulado)

$$\prod_{i=0}^{n}(x - x_i)$$

de $R_n(x)$, se observa que para disminuirlo (y, por ende, disminuir el error $R_n(x)$) deben usarse argumentos x_i lo más cercanos posible al valor por interpolar x (regla que se había seguido por intuición y que ahora se confirma matemáticamente). También de esta productoria se infiere que en general en una extrapolación (x fuera del intervalo de las x_i usadas) el error es mayor que en una interpolación. Puede decirse también que si bien se espera una mejor aproximación al aumentar el grado n del polinomio $p_n(x)$, es cierto que el valor del factor productoria aumenta al incrementarse n, por lo que debe existir un grado óptimo para el polinomio que se usará en el proceso de interpolación. Por último, en términos generales es imposible determinar el valor exacto de $R_n(x)$; a lo más que se puede llegar es determinar el intervalo en que reside el error.

Los ejemplos que se dan a continuación ilustran estos comentarios.

Ejemplo 5.10

Suponga que tiene la tabla siguiente de la función cos x.

Puntos	0	1	2	3
x (grados)	60	0	50	90
$f(x) = \cos x$	0.5000	1.0000	0.6400	0.0000

y desea interpolar el valor de la función en $x = 10°$.

SOLUCIÓN

Al interpolar linealmente con los puntos (1) y (2) queda

$$p(x) = f[x_1] + (x - x_1) f[x_1, x_2]$$

Al sustituir valores da $p(10) = 0.9280$

La interpolación con un polinomio de segundo grado y los puntos (0), (1) y (2) da

$$p_2(x) = f[x_0] + (x-x_0)f[x_0, x_1] + (x-x_0)(x-x_1)f[x_0, x_1, x_2]$$

Al sustituir valores resulta $p_2(10) = 0.9845$

Se interpola con un polinomio de tercer grado (usando los cuatro puntos) y queda

$$p_3(x) = f[x_0] + (x-x_0)f[x_0, x_1] + (x-x_0)(x-x_1)f[x_0, x_1, x_2] + (x-x_0)(x-x_1)(x-x_2)f[x_0, x_1, x_2, x_3]$$

Al sustituir valores da $p_3(10) = 0.9764$

El valor correcto de cos 10° hasta la cuarta cifra significativa es 0.9848, así que el error en porciento para el primer grado es 5.77, para el segundo 0.03, y para el tercero 0.85.

El grado óptimo del polinomio de aproximación para este caso particular es 2 (usando los puntos más cercanos al valor por interpolar: (0), (1) y (2)). Si se usaran los puntos (1), (2) y (3) el error sería 1.79%, como puede verificar el lector.

Ejemplo 5.11

Con la ecuación 5.41 encuentre una cota inferior del error de interpolación $R_n(x)$ para $x = 1.5$ cuando $f(x) = ln\, x$, $n=3$, $x_0=1$, $x_1=4/3$, $x_2=5/3$ y $x_3=2$.

SOLUCIÓN

La ecuación 5.41 con $n = 3$ queda

$$R_3(x) = f[x, x_0, x_1, x_2, x_3] \prod_{i=0}^{3} (x - x_i),$$

donde el factor productoria puede evaluarse directamente como sigue

$$\prod_{i=0}^{3} (x - x_i) = (1.5-1)(1.5-4/3)(1.5-5/3)(1.5-2) = 0.00694$$

En cambio, el factor $f[x, x_0, x_1, x_2, x_3]$ es —como se ha dicho antes— imposible de determinar, pues no se cuenta con el valor de $f(x)$ (necesario para su evaluación). Sin embargo, el valor de $f[x, x_0, x_1, x_2, x_3]$ está estrechamente relacionado con la cuarta derivada de $f(x)$, como lo expresa el siguiente teorema.

Teorema*

Sea $f(x)$ una función de valor real, definida en $[a,b]$ y k veces diferenciable en (a,b). Si $x_0, x_1, \ldots, x_k$ son $k+1$ puntos distintos en $[a,b]$, entonces existe $\xi \in (a,b)$ tal que

$$f[x_0, x_1, \ldots, x_k] = \frac{f^{(k)}(\xi)}{k!}$$

$$\text{con } \xi \in (\min x_i, \max x_i),\ 0 \leq i \leq n$$

Al utilizar esta información se tiene, en general,

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i) \qquad \text{con } \xi \in (\min x_i, \max x_i),\ 0 \leq i \leq n$$

y para $n = 3$

$$f[x, x_0, x_1, x_2, x_3] = \frac{f^{IV}(\xi)}{4!}$$

Se deriva sucesivamente $f(x)$ cuatro veces y se tiene

$$f'(x) = 1/x;\ f''(x) = -1/x^2;\ f'''(x) = 2/x^3;\ f^{IV}(x) = -6/x^4$$

Como $f^{IV}(x)$ es creciente en el intervalo de interés (1,2) (al aumentar x en éste se incrementa $f^{IV}(x)$), alcanza su valor mínimo en $x=1$ y, por lo tanto, la cota inferior buscada está dada por:

$$0.00694 \frac{f^{IV}(1)}{4!} = 0.00694 \frac{-6}{(1)^4 4!} = -0.00174,$$

*Para su demostración véase Conte, S.D. y De Boor C. *Análisis numérico.* Segunda edición, Mc. Graw-Hill (1967), p. 226-227.

es decir,

$$R_3\,(1.5) \geq -0.00174$$

Este valor indica que el error de interpolación cuando $x = 1.5$ es mayor o igual que —0.00174. Sin embargo, para conocer el intervalo donde reside el error, es necesario conocer la cota superior, que se calcula en el ejemplo siguiente.

Ejemplo 5.12

Calcule la cota superior del error $R_3(x)$ del ejemplo anterior y confirme que al utilizar diferencias divididas para interpolar en $x = 1.5$, el error obtenido está en el intervalo cuyos extremos son las cotas obtenidas. Use 0.40547 como valor verdadero de *ln* 1.5.

SOLUCIÓN

Como se vio, la función $-6/x^4$ es creciente en (1,2); por lo tanto alcanza su valor máximo en $x = 2$ y la cota superior está dada por

$$0.00694\,\frac{-6}{2^4\,4\,!} = -0.00011\,,$$

es decir,

$$R_3(1.5) \leq -0.00011$$

Por medio de la interpolación con diferencias divididas con un polinomio de tercer grado se obtiene

$$p_3\,(1.5) = f\,[x_0] + (1.5\text{–}x_0)f\,[x_0, x_1] + (1.5\text{–}x_0)(1.5\text{–}x_1)f\,[x_0, x_1, x_2]$$

$$+ (1.5\text{–}x_0)(1.5\text{–}x_1)(1.5\text{–}x_2)f\,[x_0, x_1, x_2, x_3]$$

$$= 0.40583$$

y el error es *ln* $1.5 - p_3\,(1.5) = -0.00036$ que, efectivamente, está en el intervalo [–0.00174,–0.00011].

SECCIÓN 5.7 APROXIMACIÓN POLINOMIAL SEGMENTARIA

En alguno de los casos previos pudo pensarse en aproximar $f(x)$ por medio de un polinomio de grado "alto", 10 o 20. Esto pudiera ser por diversas razones: porque se quiere mayor exactitud; por manejar un solo polinomio que sirva para interpolar en cualquier punto del intervalo $[a,b]$, etcétera.

Sin embargo, hay serias objeciones al empleo de la aproximación de grado "alto"; la primera es que los cálculos para obtener $p_n(x)$ son mayores, hay que verificar más cálculos para evaluar p_n (x) y, lo peor del caso, es que los resultados son poco confiables como puede verse en el ejemplo 5.10.

Si bien lo anterior es grave, lo es más que el error de interpolación aumenta en lugar de disminuir (véase Sec. 5.6 y ejemplo 5.3). Para abundar un poco más en la discusión de la sección 5.6, se retomará el factor productoria de la ecuación 5.41.

$$\prod_{i=0}^{n} (x - x_i),$$

donde, si n es muy grande, los factores $(x - x_i)$ son numerosos y, si su magnitud es mayor de 1, evidentemente su influencia será aumentar el error $R_n(x)$.

Para disminuir $R_n(x)$, atendiendo el factor productoria exclusivamente, es menester que los factores $(x - x_i)$ sean en su mayoría menores de 1 en magnitud, lo cual puede lograrse tomando intervalos pequeños alrededor de x.

Como el intervalo sobre el cual se va a aproximar $f(x)$ generalmente se da de antemano, lo anterior se logra dividiendo dicho intervalo en subintervalos suficientemente pequeños y aproximar $f(x)$ en cada subintervalo por medio de un polinomio adecuado; por ejemplo, mediante una línea recta en cada subintervalo (véase Fig. 5.5).

Esto da como aproximación de f (x) una línea quebrada o segmentos de líneas rectas —que se llamarán g_1 (x)— cuyos puntos de quiebre son $x_1, x_2, \ldots, x_{n-1}$. Las funciones f (x) y $g_1(x)$ coinciden en $x_0, x_1, x_2, \ldots, x_n$ y el error en cualquier punto x de $[x_0, x_n]$ queda acotado, de acuerdo con el teorema del ejemplo 5.11 aplicado a cada subintervalo $[x_i, x_{i+1}]$ con $i = 0,1,2,\ldots,n\text{-}1$, por

$$R_1(x) = |f(x) - g_1(x)| \leq \max_{a \leq \xi \leq b} \left| \frac{f''(\xi)}{2!} \right| \max_i |(x - x_i)(x - x_{i+1})| \tag{5.43}$$

Si f (x) fuera diferenciable dos veces en $[x_0, x_n]$, el valor máximo de $|(x-x_i)(x-x_{i+1})|$ para $x \in [x_i, x_{i+1}]$ se da en $x = (x_i + x_{i+1})/2$, el punto medio de $[x_i, x_{i+1}]$; de modo que

$$\max_i |(x - x_i)(x - x_{i+1})| = \max_i \left| \left(\frac{x_i + x_{i+1}}{2} - x_i\right)\left(\frac{x_i + x_{i+1}}{2} - x_{i+1}\right) \right|$$

$$= \max_i \left| \frac{x_{i+1} - x_i}{2} \frac{x_i - x_{i+1}}{2} \right|$$

$$= \max_i \frac{(x_{i+1} - x_i)^2}{4} = \max_i \frac{\Delta_i^2}{4}$$

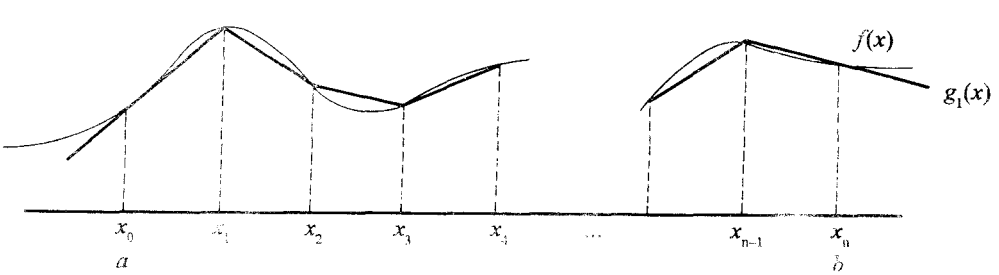


Figura 5.5 Aproximación de $f(x)$ por una línea quebrada.

Al sustituir en la ecuación 5.43

$$R_1(x) = |f(x) - g_1(x)| \leq \max_{a \leq \xi \leq b} \left| \frac{f''(\xi)}{2!} \right| \max_i \frac{\Delta x_i^2}{4} \tag{5.44}$$

Donde se aprecia que el error $R_1(x)$ puede reducirse tanto como se quiera, haciendo Δx_i pequeño para toda i; por ejemplo, tomando un número suficientemente grande de subintervalos en $[a, b]$, o bien empleando polinomios de grado dos para cada subintervalo $[x_i, x_{i+1}]$; de esta manera se consiguen segmentos polinomiales de grado dos: $g_2(x)$ cuyo término del error (5.44) correspondiente tendrá Δx_i^3 en lugar de Δx_i^2. Esto da una disminución del error respecto al empleo de líneas rectas. El empleo de polinomios de grado tres en cada subintervalo $[x_i, x_{i-1}]$ es de las técnicas más difundidas y se discute en detalle enseguida.

Aproximación cúbica segmentaria de Hermite

Se parte del hecho que se tiene una función $f(x)$ de valor real, dada en forma tabular o analítica en el intervalo $[a, b]$, con

$$a = x_0 < x_1 < x_2 < \ldots < x_n = b \tag{5.45}$$

Se quiere construir una función $g_3(x)$ con segmentos de polinomios cúbicos* $p_i(x)$ en cada $[x_i, x_{i+1}]$ con $i = 0,1,2, \ldots, n-1$, tal que

$$g_3(x_i) = f(x_i) \text{ con } i = 0, 1, 2, \ldots, n,$$

de donde

$$p_i(x_i) = f(x_i) \text{ y } p_i(x_{i+1}) = f(x_{i+1}) \quad \text{para } i = 0,1, \ldots, n-1 \tag{5.46}$$

y esta última implica que

$$p_{i-1}(x_i) = p_i(x_i) \qquad i = 1, 2, \ldots, n$$

de modo que $g_3(x)$ es continua en $[a, b]$ y tiene los puntos interiores $x_1, x_2, \ldots, x_{n-1}$ como puntos de quiebre o donde $g_3(x)$ no es diferenciable en general.

De acuerdo con el álgebra, se sabe que para que un polinomio cúbico quede determinado en forma única se requieren cuatro puntos. Hasta ahora, cada uno de los segmentos cúbicos $p_i(x)$ tiene que pasar por $(x_i, f(x_i))$ y $(x_{i+1}, f(x_{i+1}))$, de modo que quedan dos puntos o condiciones que se pueden establecer para definir en forma única $p_i(x)$.

La elección de estas dos condiciones faltantes depende, por ejemplo, de la utilización que se vaya a dar a $g_3(x)$, de $f(x)$ y del contexto donde se trabaje (ingenieril o matemático).

Por ejemplo, desde el punto de vista ingenieril sería deseable que $g_3(x)$ fuera diferenciable en los puntos interiores: $x_1, x_2, \ldots, x_{n-1}$; es decir, que $g_3(x)$ fuese suave

*En lo que sigue de esta sección, el subíndice indica el subintervalo, no el grado del polinomio como en otras ocasiones.

en $[a, b]$, en lugar de tener picos o puntos de quiebre. Esto se daría con dos condiciones como la 5.46, pero en derivadas; así

$$p_i'(x_i) = f'(x_i) \text{ y } p_i'(x_{i+1}) = f'(x_{i+1}) \qquad i = 0, 1, \ldots, n-1 \tag{5.47}$$

previsto que $f'(x)$ fuese conocida o aproximada en cada uno de los puntos $x_0, x_1, \ldots, x_n$. Con esto quedan cubiertas las dos condiciones faltantes.

De la ecuación 5.47 se infiere

$$p_{i-1}'(x_i) = p_i'(x_i) \qquad i = 1, 2, \ldots, n \tag{5.48}$$

En este punto cabe empezar a hablar del cálculo de los polinomios $p_i(x)$; por tanto, como paso siguiente se aproxima $p_i(x)$, $i=1,2, \ldots, n$ con diferencias divididas así

$$\begin{aligned} p_i(x) = f(x_i) + f[x_i, x_i](x-x_i) + f[x_i, x_i, x_{i+1}](x-x_i)^2 \\ + f[x_i, x_i, x_{i+1}, x_{i+1}](x-x_i)^2(x-x_{i+1}) \end{aligned} \tag{5.49}$$

como

$$f[x_i, x_i] = \lim_{\Delta x \to 0} \frac{f(x_i + \Delta x) - f(x_i)}{\Delta x} = f'(x_i)$$

y al sustituir $(x-x_{i+1})$ con $(x-x_i)+(x_i-x_{i+1})$ y agrupar se tiene

$$p_i(x) = f(x_i) + f'(x_i)(x-x_i) \tag{5.50}$$

$$+(f[x_i, x_i, x_{i+1}] - f[x_i, x_i, x_{i+1}, x_{i+1}]\Delta x_i)(x-x_i)^2 + f[x_i, x_i, x_{i+1}, x_{i+1}](x-x_i)^3$$

Para facilidad de manejo en su programación, la ecuación 5.50 se escribe

$$p_i(x) = c_{1,i} + c_{2,i}(x-x_i) + c_{3,i}(x-x_i)^2 + c_{4,i}(x-x_i)^3 \tag{5.51}$$

con

$$c_{1,i} = f(x_i), \; c_{2,i} = f'(x_i),$$

$$c_{3,i} = f[x_i, x_i, x_{i+1}] - f[x_i, x_i, x_{i+1} x_{i+1}]\Delta x_i$$

$$= \frac{f[x_i, x_{i+1}] - f[x_i, x_i]}{x_{i+1} - x_i} - c_{4,i}\Delta x_i = \frac{f[x_i, x_{i+1}] - c_{2,i}}{\Delta x_i} - c_{4,i}\Delta x_i$$

y

$$c_{4,i} = f[x_i, x_i, x_{i+1}, x_{i+1}] = \frac{f[x_i, x_{i+1}, x_{i+1}] - f[x_i, x_i, x_{i+1}]}{\Delta x_i} \tag{5.52}$$

$$= \frac{\dfrac{f[x_{i+1},x_{i+1}] - f[x_i,x_{i+1}]}{x_{i+1} - x_i} - \dfrac{f[x_i,x_{i+1}] - f[x_i,x_i]}{x_{i+1} - x_i}}{x_{i+1} - x_i}$$

$$= \frac{f'(x_{i+1}) - 2f[x_i,x_{i+1}] + f'(x_i)}{(x_{i+1} - x_i)^2} = \frac{f'(x_{i+1}) - 2f[x_i,x_{i+1}] + c_{2,i}}{\Delta x_i^2}$$

Ejemplo 5.13

Resuelva el problema del ejemplo 5.3 usando aproximación segmentaria, con polinomios de grado 2, 4, 6,..., 16 y estime, como antes, el error máximo en forma práctica.

SOLUCIÓN

En el disco se encuentra el programa 5.2 que realiza los cálculos solicitados

Resultados

Número de intervalos	Error máximo
2	2.23620
4	2.23622
6	0.73979
8	0.04213
10	0.09341
12	0.06417
14	0.03299
16	0.01279

En contraste con la aproximación polinomial (véase el ejemplo 5.3), el error máximo decrece conforme n crece.

Aproximación cúbica segmentaria de Bessel

La aproximación cúbica de Hermite requiere el conocimiento de $f'(x_i)$, $i=0,1,...,n$. Esta información, como se ha visto a lo largo del capítulo, no siempre existe, aún conociendo $f(x)$ analíticamente no siempre es fácil obtenerla.

La aproximación cúbica segmentaria de Bessel se distingue por emplear una aproximación de $f'(x_i)$ por

$$f'(x_i) \approx f[x_{i-1}, x_{i+1}], \qquad i = 0,1,..., n \tag{5.53}$$

y en todo lo demás se procede tal como en la aproximación de Hermite.

La expresión 5.53 requiere dos puntos adicionales a los que se tienen y son x_{-1} y x_{n+1}, ya que

$$f'(x_0) \approx f[x_{-1}, x_1] \text{ y } f'(x_n) \approx f[x_{n-1}, x_{n+1}],$$

llamadas **derivadas frontera** de $g_3(x)$.

Una forma de obtenerlos es una nueva subdivisión de $[a, b]$, como

$$a = x_{-1} < x_0 < x_1 < x_2 < \ldots < x_{n+1} = b. \tag{5.54}$$

también puede emplearse

$$f'(x_{-1}) \text{ y } f'(x_{n+1}), \tag{5.55}$$

en caso de disponer de ellas (las derivadas restantes se obtendrían de acuerdo con la ecuación 5.53).

Otra forma sería tomar $f'(x_0)$ y $f'(x_n)$ de manera que $g_3(x)$ satisfaga las condiciones de extremo libre

$$g_3''(a) = g_3''(b) = 0 \tag{5.56}$$

Independientemente de cómo se obtengan los puntos x_{-1} y x_{n+1}, las funciones $g_3(x)$ y $f(x)$ coinciden en los puntos de quiebre $x_0, x_1, x_2, \ldots x_n$. Por esto $g_3(x)$ es continua en $[a, b]$, y por la ecuación 5.48 también es continuamente diferenciable. Además es posible, y se muestra adelante, determinar $f'(x_0), f'(x_1), \ldots, f'(x_n)$ de manera que la $g_3(x)$ resultante sea dos veces continuamente diferenciable.

El método de determinar $g_3(x)$ con esta característica se conoce como **aproximación cúbica de trazador**, ya que la gráfica de $g_3(x)$ se aproxima a la forma que tomaría una varilla delgada flexible si se forzara a pasar por cada punto $(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_n, f(x_n))$.

El requisito de que $g_3(x)$ sea continuamente diferenciable dos veces puede darse como

$$p_{i-1}''(x_i) = p_i''(x_i), \qquad i = 1, 2, \ldots, n-1 \tag{5.57}$$

o

$$2c_{3,i-1} + 6c_{4,i-1}\Delta x_{i-1} = 2c_{3,i}, \qquad i = 1,2,..., n-1,$$

conforme la ecuación 5.51 (derivándola dos veces).

Al sustituir las expresiones de la ecuación 5.52 en la última ecuación, se tiene

$$\frac{2(f[x_{i-1}, x_i] - f'(x_{i-1}))}{\Delta x_{i-1}} + 4c_{4,i-1}\Delta x_{i-1} =$$

$$\frac{2\,(f\,[x_i, x_{i+1}] - f'(x_i))}{\Delta x_i} - 2\,c_{4,i}\,\Delta x_i, \qquad i = 1, 2, \ldots, n-1$$

Al continuar la sustitución y simplificar, se tiene

$$\Delta x_i f'(x_{i-1}) + 2\,(\Delta x_{i-1} + \Delta x_i) f'(x_i) + \Delta x_{i-1} f'(x_{i+1}) =$$
$$3\,(f\,[x_{i-1}, x_i]\Delta x_i + f\,[x_i, x_{i+1}]\,\Delta x_{i-1}),\ i = 1,2, \ldots, n-1 \tag{5.58}$$

Un sistema de $n-1$ ecuaciones lineales en las $(n+1)$ incógnitas $f'(x_0), f'(x_1), \ldots, f'(x_n)$. Al obtener $f'(x_0)$ y $f'(x_n)$ de alguna manera (por ejemplo mediante las ecuaciones 5.53 o 5.55) se resuelve la 5.58 para $f'(x_1), f'(x_2), \ldots, f'(x_{n-1})$ por alguno de los métodos vistos en el capítulo 3; no obstante, como el sistema 5.58 es tridiagonal, conviene utilizar el algoritmo de Thomas.

Ejemplo 5.14

La siguiente tabla muestra las viscosidades del isopentano a 59°F y a diferentes presiones

Presión (psia)	Viscosidad (micropoises)
426.690	2468
483.297	2482
497.805	2483
568.920	2498
995.610	2584
1422.300	2672
2133.450	2811
3555.750	3094
4266.900	3236
7111.500	3807

Elabore un programa para aproximar el valor de la viscosidad a las presiones de 355.575, 711.150, 2844.600, 5689.200 y 8533.801 psia, utilizando la aproximación cúbica segmentaria de Bessel.

SOLUCIÓN

En el disco se encuentra el programa 5.3, el cual proporciona los siguientes resultados

Presión (psia)	Viscosidad (micropoises)
355.575	2453.56
711.150	2531.32
2844.600	2950.92
5689.200	3520.79
8533.801	4093.21

SECCIÓN 5.8 APROXIMACIÓN POLINOMIAL CON MÍNIMOS CUADRADOS

Hasta ahora el texto se ha enfocado en encontrar un polinomio de aproximación que pase **por los puntos** dados en forma tabular. Sin embargo, a veces la información (dada en la tabla) tiene errores significativos; por ejemplo cuando proviene de medidas físicas; en estas circunstancias no tiene sentido pasar un polinomio de aproximación **por** los puntos dados, sino sólo cerca de ellos (véase Fig. 5.6).

No obstante, esto crea un problema, ya que se puede pasar un número infinito de curvas **entre los puntos.** Para determinar **la mejor** curva se establece un criterio que la fije y una metodología que la determine. El criterio más común consiste en pedir que la suma de las distancias calculadas entre el valor de la función que aproxima $p(x_i)$ y el valor de la función $f(x_i)$ dada en la tabla, sea mínima (véase Fig. 5.7); es decir, que

$$\sum_{i=1}^{m} \mid p(x_i) - f(x_i) \mid = \sum_{i=1}^{m} d_i = \text{mínimo}$$

Para evitar problemas de derivabilidad más adelante, se acostumbra utilizar las distancias d_i elevadas al cuadrado:

$$\sum_{i=1}^{m} [p(x_i) - f(x_i)]^2 = \sum_{i=1}^{m} d_i^2 = \text{mínimo}$$

En la figura 5.7 se observan los puntos tabulados, la aproximación polinomial $p(x)$ y las distancias d_i entre los puntos correspondientes, cuya suma hay que minimizar.

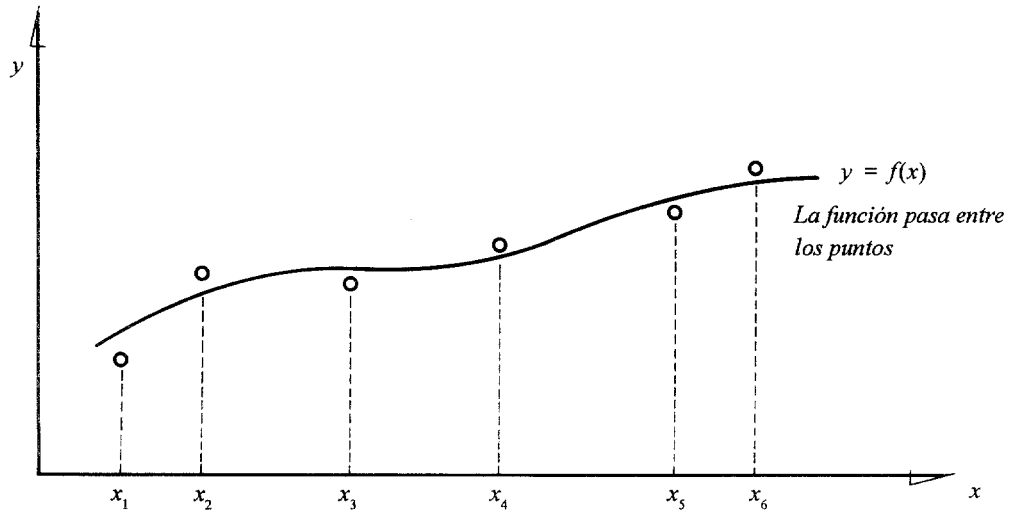


Figura 5.6. Aproximación polinomial que pasa por entre los puntos.

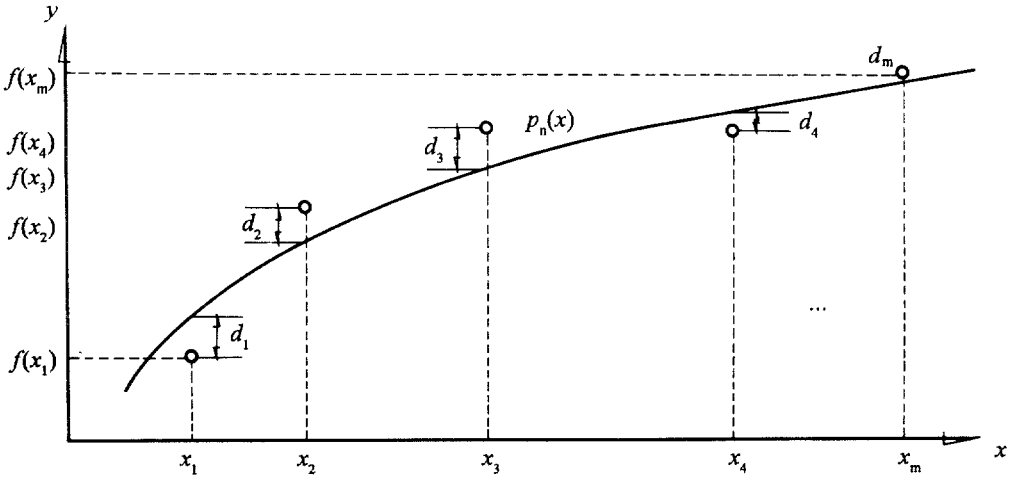


Figura 5.7. Ilustración de las distancias d_i a minimizar.

Si se utiliza

$$p(x) = a_0 + a_1 x \tag{5.59}$$

para aproximar la función dada por la tabla, el problema queda como el de minimizar

$$\sum_{i=1}^{m} [a_0 + a_1 x_i - f(x_i)]^2 \tag{5.60}$$

Nótese que del número infinito de polinomios que pasan entre los puntos, se selecciona aquel cuyos coeficientes a_0 y a_1 minimicen 5.60.

En el cálculo de funciones de una variable, el lector ha aprendido que para encontrar el mínimo o máximo de una función, se deriva y se iguala con cero esa derivada. Después se resuelve la ecuación resultante para obtener los valores de la variable que pudieran minimizar o maximizar la función. En el caso en estudio, donde se tiene una función por minimizar de dos variables (a_0 y a_1), el procedimiento es derivar parcialmente con respecto a cada una de las variables e igualar a cero cada derivada, con lo cual se obtiene un sistema de dos ecuaciones algebraicas en las incógnitas a_0 y a_1; o sea,

$$\begin{aligned} \frac{\partial}{\partial a_0} \left[\sum_{i=1}^{m} (a_0 + a_1 x_i - f(x_i))^2 \right] &= 0 \\ \frac{\partial}{\partial a_1} \left[\sum_{i=1}^{m} (a_0 + a_1 x_i - f(x_i))^2 \right] &= 0 \end{aligned} \tag{5.61}$$

Se deriva dentro del signo de sumatoria

$$\sum_{i=1}^{m} \frac{\partial}{\partial a_0} [a_0 + a_1 x_i - f(x_i)]^2 = \sum_{i=1}^{m} 2[a_0 + a_1 x_i - f(x_i)]1 = 0$$

$$\sum_{i=1}^{m} \frac{\partial}{\partial a_1} [a_0 + a_1 x_i - f(x_i)]^2 = \sum_{i=1}^{m} 2[a_0 + a_1 x_i - f(x_i)]x_i = 0$$

al desarrollar las sumatorias se tiene

$$[a_0 + a_1 x_1 - f(x_1)] + [a_0 + a_1 x_2 - f(x_2)] + \ldots + [a_0 + a_1 x_m - f(x_m)] = 0$$

$$[a_0 x_1 + a_1 x_1^2 - f(x_1)x_1] + [a_0 x_2 + a_1 x_2^2 - f(x_2)x_2] + \ldots +$$

$$+ [a_0 x_m + a_1 x_m^2 - f(x_m)x_m] = 0$$

que simplificadas quedan

$$m a_0 + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} f(x_i)$$

$$a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} f(x_i)\, x_i$$

El sistema se resuelve por la regla de Cramer y se tiene

$$a_0 = \frac{[\sum_{i=1}^{m} f(x_i)]\ [\sum_{i=1}^{m} x_i^2] - [\sum_{i=1}^{m} x_i]\ [\sum_{i=1}^{m} f(x_i) x_i]}{m \sum_{i=1}^{m} x_i^2 - [\sum_{i=1}^{m} x_i]^2}$$

$$a_1 = \frac{m \sum_{i=1}^{m} f(x_i)\, x_i - [\sum_{i=1}^{m} f(x_i)]\,[\sum_{i=1}^{m} x_i]}{m \sum_{i=1}^{m} x_i^2 - [\sum_{i=1}^{m} x_i]^2} \qquad (5.62)$$

que sustituidos en la ecuación 5.59 dan la aproximación polinomial de primer grado que **mejor ajusta** la información tabulada. Este polinomio puede usarse a fin de aproximar valores de la función para argumentos no conocidos en la tabla.

Ejemplo 5.15

En la tabla siguiente se presentan los alargamientos de un resorte correspondientes a fuerzas de diferente magnitud que lo deforman

Puntos	1	2	3	4	5
Fuerza (kgf) x	0	2	3	6	7
Longitud del resorte (m) y	0.120	0.153	0.170	0.225	0.260

Determine por mínimos cuadrados el mejor polinomio de primer grado (recta) que represente la función dada.

SOLUCIÓN

Para facilitar los cálculos y evitar errores en los mismos, primero se construye la siguiente tabla

Puntos	Fuerza x_i	Longitud y_i	x_i^2	$x_i y_i$
1	0	0.120	0	0.000
2	2	0.153	4	0.306
3	3	0.170	9	0.510
4	6	0.225	36	1.350
5	7	0.260	49	1.820

$$\sum x_i = 18 \quad \sum y_i = 0.928 \quad \sum x_i^2 = 98 \quad \sum x_i y_i = 3.986$$

Los valores de las sumatorias de la última fila se sustituyen en el sistema de ecuaciones 5.62 y se obtiene

$$a_0 = 0.11564 \text{ y } a_1 = 0.019434, \text{ de donde}$$

$$p(x) = 0.11564 + 0.019434\,x\,.$$

El grado del polinomio no tiene relación con el número de puntos usados y debe seleccionarse de antemano con base en consideraciones teóricas que apoyan el fenómeno estudiado, el diagrama de dispersión (puntos graficados en el plano x-y) o ambos.

El hecho de tener la **mejor recta** que aproxima la información, no significa que la información esté bien aproximada; quizá convenga aproximarla con una parábola o una cúbica.

Para encontrar el polinomio de segundo grado $p_2(x) = a_0 + a_1x + a_2x^2$ que mejor aproxime la tabla, se minimiza

$$\sum_{i=1}^{m} [a_0 + a_1x_i + a_2x_i^2 - f(x_i)]^2 \tag{5.63}$$

donde los parámetros a_0, a_1 y a_2 se obtienen al resolver el sistema de ecuaciones lineales que resulta de derivar parcialmente e igualar a cero la función por minimizar con respecto a cada uno. Dicho sistema queda

$$\begin{aligned} m\,a_0 + a_1 \sum_{i=1}^{m} x_i + a_2 \sum_{i=1}^{m} x_i^2 &= \sum_{i=1}^{m} f(x_i) \\ a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 + a_1 \sum_{i=1}^{m} x_i^3 &= \sum_{i=1}^{m} f(x_i)x_i \\ a_0 \sum_{i=1}^{m} x_i^2 + a_1 \sum_{i=1}^{m} x_i^3 + a_2 \sum_{i=1}^{m} x_i^4 &= \sum_{i=1}^{m} f(x_i)x_i^2, \end{aligned} \tag{5.64}$$

cuya solución puede obtenerse por alguno de los métodos vistos en el capítulo 3.

Ejemplo 5.16

El calor específico Cp (cal/K gmol) del Mn_3O_4 varía con la temperatura de acuerdo con la siguiente tabla

Puntos	1	2	3	4	5	6
T (K)	280	650	1000	1200	1500	1700
Cp (cal/K gmol)	32.7	45.4	52.15	53.7	52.9	50.3

Aproxime esta información con un polinomio por el método de mínimos cuadrados.

SOLUCIÓN

El calor específico aumenta con la temperatura hasta el valor tabulado de 1200 K, para disminuir posteriormente en valores más altos de temperatura. Esto sugiere utilizar un polinomio con curvatura en vez de una recta, por ejemplo uno de segundo grado, que es el más simple.
Para facilitar el cálculo de los coeficientes del sistema de ecuaciones 5.64, se construye la siguiente tabla

Puntos i	T x_i	Cp y_i	x_i^2	x_i^3	x_i^4	$y_i x_i$	$y_i x_i^2$
1	280	32.7	0.78×10^5	0.022×10^9	0.062×10^{11}	9156	2.56×10^6
2	650	45.4	0.42×10^6	0.275×10^9	1.785×10^{11}	29510	19.18×10^6
3	1000	52.15	1.00×10^6	1.000×10^9	1.000×10^{12}	52150	52.15×10^6
4	1200	53.7	1.44×10^6	1.728×10^9	2.074×10^{12}	64440	77.33×10^6
5	1500	52.9	2.25×10^6	3.375×10^9	5.063×10^{12}	79350	119.03×10^6
6	1700	50.3	2.89×10^6	4.900×10^9	8.350×10^{12}	85510	145.37×10^6
ΣTOTALES	6330	287.15	8.08×10^6	11.3×10^9	166.7×10^{11}	320116	415.62×10^6

Los coeficientes se sustituyen en el sistema de ecuaciones 5.64 y se obtiene:

$$6\,a_0 + 6330\,a_1 + 8.08\times10^6\,a^2 = 287.15$$

$$6330\,a_0 + 8.08\times10^6\,a_1 + 11.30\times10^9\,a^2 = 320116$$

$$8.08\times10^6\,a_0 + 11.30\times10^9\,a_1 + 166.70\times10^{11}\,a^2 = 415.62\times10^6$$

cuya solución por el método de eliminación Gaussiana arroja

$$a_0 = 22.4066,\ a_1 = 0.0458,\ a_2 = -1.694\times10^{-5},$$

que forman la aproximación polinomial siguiente

$$Cp\ (T) \approx p_2(T) = 22.4066 + 0.0458\ T - 1.694\times10^{-5}\ T^2.$$

NOTA

Muchas de las calculadoras de mano cuentan con un programa interno para obtener esta aproximación; por otro lado, puede usarse un pizarrón electrónico para los cálculos (sumatorias, solución de ecuaciones, etcétera).

Ejemplo 5.16

Use la aproximación polinomial de segundo grado obtenida en el ejemplo anterior para aproximar el calor específico del Mn_3O_4 a una temperatura de 800 K.

SOLUCIÓN

Con la sustitución de T=800 K en el polinomio de aproximación se tiene

$$\begin{aligned} Cp\ (800) \approx p_2(800) &= 22.4066 + 0.0458(800) - 1.694\times10^{-5}\ (800)^2. \\ &= 48.2 \text{ cal/K gmol.} \end{aligned}$$

En caso de querer aproximar una función dada en forma tabular con un polinomio de grado más alto, n por ejemplo, el procedimiento es el mismo; esto es, minimizar la función

$$\sum_{i=1}^{m} [a_0 + a_1x_i + a_2x_i^2 + \ldots + a_nx_i^n - f(x_i)]^2,$$

lo cual se obtiene derivándola parcialmente con respecto a cada coeficiente a_j, $(0\leq j\leq n)$ e igualando a cero cada una de estas derivadas. Con esto se llega al sistema lineal

$$\begin{aligned}
ma_0 &+ a_1 \Sigma x &+ a_2 \Sigma x^2 &+ \ldots + a_n \Sigma x^n = \Sigma y \\
a_0 \Sigma x &+ a_1 \Sigma x^2 &+ a_2 \Sigma x^3 &+ \ldots + a_n \Sigma x^{n+1} = \Sigma xy \\
a_0 \Sigma x^2 &+ a_1 \Sigma x^3 &+ a_2 \Sigma x^4 &+ \ldots + a_n \Sigma x^{n+2} = \Sigma x^2 y \\
&\cdot & &\cdot \\
&\cdot & &\cdot \\
&\cdot & &\cdot \\
a_0 \Sigma x^n &+ a_1 \Sigma x^{n+1} &+ a_2 \Sigma x^{n+2} &+ \ldots + a_n \Sigma x^{n+n} = \Sigma x^n y
\end{aligned}$$

donde se han omitido, los subíndices i, de x y y, así como los límites de las sumatorias que van de 1 hasta m para simplificar su escritura.

ALGORITMO 5.5 Aproximación con mínimos cuadrados

Para obtener los N+1 coeficientes del polinomio óptimo de grado N que pasa entre M parejas de puntos, proporcionar los

DATOS: El grado del polinomio de aproximación N, el número de parejas de valores (X(I), FX(I), I = 1, 2, ..., M).

RESULTADOS: Los coeficientes A(0), A(1), ..., A(N) del polinomio de aproximación.

PASO 1. Hacer J = 0
PASO 2. Mientras J≤(2*N-1), repetir los pasos 3 a 5.
  PASO 3. Si J ≤ N Hacer SS(J) = 0. De otro modo continuar.
  PASO 4. Hacer S(J) = 0
  PASO 5. Hacer J = J + 1
PASO 6. Hacer I = 1
PASO 7. Mientras I ≤ M, repetir los pasos 8 a 15.
  PASO 8. Hacer XX = 1
  PASO 9. Hacer J = 0
  PASO 10. Mientras J ≤ (2*N-1), repetir los pasos 11 a 14.
    PASO 11. Si J≤N hacer SS(J)=SS(J)+XX*FX(I). De otro modo continuar.
    PASO 12. Hacer XX = XX*X(I)
    PASO 13. Hacer S(J) = S(J) + XX
    PASO 14. Hacer J = J + 1
  PASO 15. Hacer I = I + 1
PASO 16. Hacer $B(0,0)$ = M
PASO 17. Hacer I = 0
PASO 18. Mientras I≤N, repetir los pasos 19 a 24.
  PASO 19. Hacer J = 0
  PASO 20. Mientras J≤N, repetir los pasos 21 y 22.
    PASO 21. Si I ≠ 0 y J ≠ 0. Hacer $B(I,J)$ = S(J-1+I)
    PASO 22. Hacer J = J + 1
  PASO 23. Hacer $B(I,N+1)$ = SS(I)
  PASO 24. Hacer I = I + 1
PASO 25. Resolver el sistema de ecuaciones lineales B **a** = **ss** de orden N+1 con alguno de los algoritmos del capítulo 3.
PASO 26. IMPRIMIR A(0), A(1), ..., A(N) y TERMINAR.

SECCIÓN 5.9 APROXIMACIÓN MULTILINEAL CON MÍNIMOS CUADRADOS

Con frecuencia se tienen funciones de más de una variable; esto es, $f(u,v,z)$. Si se sospecha una funcionalidad lineal en las distintas variables; es decir, si se piensa que la función

$$y = a_0 + a_1 u + a_2 v + a_3 z$$

puede ajustar los datos de la tabla siguiente

Puntos	u	v	z	y
1	u_1	v_1	z_1	$f(u_1, v_1, z_1)$
2	u_2	v_2	z_2	$f(u_2, v_2, z_2)$
3	u_3	v_3	z_3	$f(u_3, v_3, z_3)$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
m	u_m	v_m	z_m	$f(u_m, v_m, z_m)$

se puede aplicar el método de los mínimos cuadrados para determinar los coeficientes a_0, a_1, a_2 y a_3 que mejor aproximen la función de varias variables tabulada. El procedimiento es análogo al descrito anteriormente y consiste en minimizar la función

$$\sum_{i=1}^{m} [(a_0 + a_1 u_i + a_2 v_i + a_3 z_i) - y_i]^2$$

que derivada parcialmente con respecto de cada coeficiente por determinar: a_0, a_1, a_2 y a_3 e igualada a cero cada una, queda

$$\frac{\partial}{\partial a_0} \sum_{i=1}^{m} [(a_0 + a_1 u_i + a_2 v_i + a_3 z_i) - y_i]^2 = 2 \sum_{i=1}^{m} (a_0 + a_1 u_i + a_2 v_i + a_3 z_i - y_i)\, 1 = 0$$

$$\frac{\partial}{\partial a_1} \sum_{i=1}^{m} [(a_0 + a_1 u_i + a_2 v_i + a_3 z_i) - y_i]^2 = 2 \sum_{i=1}^{m} (a_0 + a_1 u_i + a_2 v_i + a_3 z_i - y_i)\, u_i = 0$$

$$\frac{\partial}{\partial a_2}\sum_{i=1}^{m}[(a_0+a_1u_i+a_2v_i+a_3z_i)-y_i]^2=2\sum_{i=1}^{m}(a_0+a_1u_i+a_2v_i+a_3z_i-y_i)\,v_i=0$$

$$\frac{\partial}{\partial a_3}\sum_{i=1}^{m}[(a_0+a_1u_i+a_2v_i+a_3z_i)-y_i]^2=2\sum_{i=1}^{m}(a_0+a_1u_i+a_2v_i+a_3z_i-y_i)\,z_i=0$$

ecuaciones que rearregladas generan el sistema algebraico lineal siguiente

$$\begin{aligned} m\,a_0 + a_1\,\Sigma\,u + a_2\,\Sigma\,v + a_3\,\Sigma\,z &= \Sigma y \\ a_0\,\Sigma\,u + a_1\,\Sigma\,u^2 + a_2\,\Sigma\,uv + a_3\,\Sigma\,uz &= \Sigma uy \\ a_0\,\Sigma\,v + a_1\,\Sigma\,vu + a_2\,\Sigma\,v^2 + a_3\,\Sigma\,vz &= \Sigma vy \\ a_0\,\Sigma\,z + a_1\,\Sigma\,zu + a_2\,\Sigma\,zv + a_3\,\Sigma\,z^2 &= \Sigma zy \end{aligned} \tag{5.65}$$

en las incógnitas a_0, a_1, a_2 y a_3. Para simplificar la escritura se han omitido los índices i, de u, v, y z y los límites de las sumatorias, que van de 1 hasta m.

Ejemplo 5.17

A partir de un estudio experimental acerca de la estabilización de arcilla muy plástica, se observó que el contenido de agua para moldeo con densidad óptima dependía linealmente de los porcentajes de cal y puzolana mezclados con la arcilla. Se tuvieron así los resultados que se dan abajo. Ajuste una ecuación de la forma

$$y = a_0 + a_1u + a_2v$$

a los datos de dicha tabla.

Agua (%) y	Cal (%) u	Puzolana (%) v
27.5	2.0	18.0
28.0	3.5	16.5
28.8	4.5	10.5
29.1	2.5	2.5
30.0	8.5	9.0
31.0	10.5	4.5
32.0	13.5	1.5

SOLUCIÓN

El sistema lineal por resolver es una modificación del sistema de ecuaciones 5.65 para una función y de dos variables u y v

$$n\,a_0 + a_1 \Sigma u + a_2 \Sigma v = \Sigma y$$

$$a_0 \Sigma u + a_1 \Sigma u^2 + a_2 \Sigma uv = \Sigma uy$$

$$a_0 \Sigma v + a_1 \Sigma vu + a_2 \Sigma v^2 = \Sigma vy$$

Con objeto de facilitar el cálculo del sistema anterior se construye la siguiente tabla

i	u_i	v_i	y_i	u_i^2	u_iv_i	v_i^2	u_iv_i	v_iy_i
1	2.0	18.0	27.5	4.00	36.00	324.00	55.00	495.00
2	3.5	16.5	28.0	12.25	57.75	272.25	98.00	462.00
3	4.5	10.5	28.8	20.25	47.25	110.25	129.60	302.40
4	2.5	2.5	29.1	6.25	6.25	6.25	72.75	72.75
5	8.5	9.0	30.0	72.25	76.50	81.00	255.00	270.00
6	10.5	4.5	31.0	110.25	47.25	20.25	325.50	139.50
7	13.5	1.5	32.0	182.25	20.25	2.25	432.00	48.00
Σ Totales:	45.0	62.5	206.4	407.5	291.25	816.25	1367.85	1789.65

Los coeficientes se sustituyen en el sistema de ecuaciones y al aplicar alguno de los métodos del capítulo 3, se obtiene

$$a_0 = 28.69, \qquad a_1 = 0.2569, \qquad a_2 = 0.09607$$

al sutituir estos valores se tiene

$$y = 28.69 + 0.2569\,u + 0.09607\,v$$

Ejercicios

5.1 A continuación se presentan las presiones de vapor del cloruro de magnesio.

Puntos	0	1	2	3	4	5	6	7
P (mmHg)	10	20	40	60	100	200	400	760
T (°C)	930	988	1050	1088	1142	1316	1223	1418

Calcule la presión de vapor correspondiente a T=1000 °C.

SOLUCIÓN

Como la información no está regularmente espaciada en los argumentos (T), pueden usarse diferencias divididas o polinomios de Lagrange para la interpolación.

Con los polinomios de Lagrange de segundo grado se tiene

$$p_2(x) = f(x_0)\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + f(x_1)\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}$$

$$+ f(x_2)\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}$$

Al tomar las presiones como valores de la función $f(x)$, las temperaturas como los argumentos x, seleccionar los puntos (0), (1) y (2) y sustituir los valores se obtiene

$$p_2(1000) = 10\frac{(1000-988)(1000-1050)}{(930-988)(930-1050)} + 20\frac{(1000-930)(1000-1050)}{(988-930)(988-1050)} +$$

$$+ 40\frac{(1000-930)(1000-988)}{(1050-930)(1050-988)} = 23.12\ mm\text{Hg} \approx 23\ mm\text{Hg}$$

5.2 Dada la tabla

Puntos	0	1	2	3	4
x_i	1.00	1.35	1.70	1.90	3.00
$f(x_i)$	0.00000	0.30010	0.53063	0.64185	1.09861

Construya una tabla de diferencias divididas para aproximar $f(x)$ en $x = 1.50$; utilice un polinomio de Newton de segundo grado.

SOLUCIÓN

A continuación se da la tabla de diferencias divididas

Puntos	x_i	$f(x_i)$	Diferencias divididas			
			Primeras	Segundas	Terceras	Cuartas
0	1.00	0.00000				
			0.85743			
1	1.35	0.30010		−0.28396		
			0.65866		0.10832	
2	1.70	0.53063		−0.18647		−0.03049
			0.55610		−0.04735	
3	1.90	0.64185		−0.10835		
			0.41524			
4	3.00	1.09861				

Se pueden seleccionar los puntos (0), (1) y (2) para el polinomio de interpolación o bien (1), (2) y (3). Se escoge el segundo conjunto de puntos, ya que están más cerca de 1.5 que el primero; sin embargo, al querer emplear la fórmula

$$p_2(x) = f[x_0] + (x-x_0)f[x_0\, x_1] + (x-x_0)(x-x_1)f[x_0, x_1, x_2]$$

y la tabla construida se deberá tener cuidado, ya que el valor x_0 de la fórmula en realidad corresponderá a x_1 de la tabla; x_1 de la fórmula, a x_2 de la tabla, etc. Consecuentemente $f[x_0]$, $f[x_0, x_1]$ y $f[x_0, x_1, x_2]$ de la fórmula corresponderán a $f[x_1]$, $f[x_1,x_2]$ y $f[x_1, x_2, x_3]$ de la tabla respectivamente (véase la línea diagonal de la tabla).

Con la sustitución de valores queda

$$p_2(1.5) = 0.30010+(1.5-1.35)(0.65866)+(1.5-1.35)(1.5-1.7)(-0.18647)$$
$$= 0.40449$$

Una solución alterna es construir la tabla de diferencias de modo que quede como punto (0) el más cercano a 1.5 (1.35 en este caso), entre los puntos restantes se elige como punto (1) el más cercano a 1.5 (1.70 en este caso), etc. Abajo se muestra cómo queda esta tabla

Puntos	x_i	$f(x_i)$	Diferencias divididas			
			Primeras	Segundas	Terceras	Cuartas
0	1.35	0.30010				
			0.65866			
1	1.70	0.53063		−0.18647		
			0.55610		0.10846	
2	1.90	0.64185		−0.22443		−0.030567
			0.71320		0.05802	
3	1.00	0.00000		−0.14900		
			0.54930			
4	3.00	1.09861				

con la cual puede usarse el polinomio

$$p_2(x) = f[x_0] + (x-x_0) f[x_0, x_1] + (x-x_0)(x-x_1) f[x_0, x_1, x_2]$$

directamente, ya que ahora los subíndices de los argumentos de la fórmula y de la tabla se corresponden. Sustituyendo valores se tiene

$$\begin{aligned} p_2(1.5) &= 0.30010+(1.5-1.35)(0.65866)+(1.5-1.35)(1.5-1.70)(-0.18647) \\ &= 0.40449 \end{aligned}$$

Obsérvese que el valor interpolado es el mismo que se tuvo anteriormente.

5.3 Las densidades de las soluciones acuosas del ácido sulfúrico varían con la temperatura y la concentración de acuerdo con la tabla

	T (°C)			
C (%)	10	30	60	100
5	1.0344	1.0281	1.0140	0.9888
20	1.1453	1.1335	1.1153	1.0885
40	1.3103	1.2953	1.2732	1.2446
70	1.6923	1.6014	1.5753	1.5417

a) Calcule la densidad a una concentración de 40% y una temperatura de 15 °C.
b) Calcule la densidad a 30 °C y concentración de 50%.
c) Calcule la densidad a 50 °C y 60% de concentración.
d) Calcule la temperatura a la cual una solución al 30% tiene una densidad de 1.215.

SOLUCIÓN

a) La temperatura se toma como el argumento x y las densidades (a 40%) como el valor de la función $f(x)$.
Con una interpolación lineal entre las densidades a 10 °C y 30 °C se tiene:

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

$$d(15) \approx \frac{15-30}{10-30}\ 1.3103 + \frac{15-10}{30-10}\ 1.2953 = 1.3066$$

b) Se toman ahora las concentraciones como argumentos x y las densidades (a 30°C) como los valores funcionales; luego, mediante una interpolación lineal entre las concentraciones a 40% y 70% queda:

$$d(50) = \frac{50-70}{40-70}\ 1.2953 + \frac{50-40}{70-40}\ 1.6014 = 1.3973$$

c) La densidad se aproxima a 50°C, utilizando primero la fila de 40% de concentración y después la fila de 70% de concentración. Con estas densidades obtenidas a 50°C se aproxima la densidad a 60% de concentración.

Primer paso

Aproximación de la densidad a 40% y 50°C.

$$d \approx \frac{50-60}{30-60}\ 1.2953 + \frac{50-30}{60-30}\ 1.2732 = 1.2806$$

Segundo paso

Aproximación de la densidad a 70% y 50°C.

$$d \approx \frac{50-60}{30-60}\ 1.6014 + \frac{50-30}{60-30}\ 1.5753 = 1.5840$$

Tercer paso

Aproximación de la densidad a 60% y 50°C usando los valores obtenidos en los pasos anteriores

$$d \approx \frac{60-70}{40-70}\ 1.2806 + \frac{60-40}{70-40}\ 1.5840 = 1.4829$$

d) En este caso es necesario interpolar los valores de la densidad a 30% de concentración a diferentes temperaturas, para después interpolar la temperatura que corresponda a una densidad de 1.215.

Primer paso

Aproximación de la densidad a 30% y 10°C.

$$d \approx \frac{30 - 20}{40 - 20} 1.1453 + \frac{30 - 40}{20 - 40} 1.3103 = 1.2278$$

Aproximación de la densidad a 30% y 30°C

$$d \approx \frac{30 - 20}{40 - 20} 1.1335 + \frac{30 - 40}{20 - 40} 1.2953 = 1.2144$$

Como la densidad dato (1.215) está entre estos dos valores obtenidos, la temperatura estará también entre 10°C y 30°C; por lo que interpolando linealmente entre estos dos valores de densidad (que ahora es el argumento x) se tiene:

Segundo paso

Aproximación de la temperatura a la que una solución con 30% de concentración tiene una densidad de 1.215

$$T = \frac{1.215 - 1.2144}{1.2278 - 1.2144} 10 + \frac{1.215 - 1.2278}{1.2144 - .2278} 30 \approx 29.1 \text{ °C}$$

5.4 Elabore un programa para leer una tabla de m pares de valores e interpolar o extrapolar, utilizando el polinomio de Newton de grado n en diferencias divididas. Pruebe este programa con los datos del ejercicio 5.1.

SOLUCIÓN

En el disco se encuentra el programa 5.3 que lee *a*) el número de pares de valores (M); *b*) el grado del polinomio interpolante (N); *c*) el argumento que se desea interpolar (XINT), y *d*) los pares de valores (X(1),FX(1),X(2),FX(2),...,X(M),FX(M)). Con esta información primero llama al subprograma TABLA que elabora la tabla de diferencias divididas. Con los valores resultantes y el argumento donde se quiere aproximar el valor de la función y el grado del polinomio interpolante, llama al subprograma INTERPOLA que realiza los cálculos de interpolación.

El resultado es

PARA XINT = 1000.0000 FXINT = 23.1201

5.5 Elabore un programa que lea una tabla de m (seleccionado por el usuario) pares de valores, y que interpole o extrapole con el polinomio de Lagrange de orden $m-1$.

SOLUCIÓN

En el disco se encuentra el programa 5.4 donde se leen M pares de valores X(I) y FX(I) de una tabla y el valor por interpolar XINT.
Resultado:

PARA XINT = 1000.0000 FXINT = 23.1201

5.6 Con el programa 5.4 y la tabla de valores del ejercicio 5.1, calcule la presión de vapor del cloruro de magnesio a las siguientes temperaturas

a) 800 °C (extrapolación) *b*)950 °C (interpolación)

c) 1098 °C (interpolación) *d*)1500 °C (extrapolación)

SOLUCIÓN

PARA XINT = 800.0000 FXINT = 18.1702 con los puntos (0),(1) y (2)

PARA XINT = 950.0000 FXINT = 12.4972 con los puntos (0),(1) y (2)

PARA XINT =1098.0000 FXINT = 65.5236 con los puntos (2),(3) y (4)

PARA XINT =1500.0000 FXINT =1156.1016 con los puntos (5),(6) y (7)

5.7 Con la información del ejercicio 5.2 estime el error cometido R_2 (1.5), aproxime $f(x)$ en x=1.5 con un polinomio de tercer grado y estime el error correspondiente R_3(1.5).

SOLUCIÓN

El valor obtenido con un polinomio de segundo grado (Ejer. 5.2) es

$$p_2(1.5) = 0.40449$$

Al usar la ecuación 5.41 y los valores de la segunda tabla de diferencias divididas (Ejer. 5.2), se tiene

$$R_2(x) \approx (x-x_0)(x-x_1)(x-x_2)\, f[x_0, x_1, x_2, x_3]$$

$$\approx (1.5\text{-}1.35)(1.5\text{-}1.7)(1.5\text{–}1.9)(0.10846) = 0.00130$$

Para aproximar $f(x)$ en $x = 1.5$ con un polinomio de tercer grado se adiciona R_2 (1.5) al valor p_2(1.5), se obtiene

$$p_3(1.5) = 0.40449 + 0.00130 = 0.40579$$

y la estimación del error en esta interpolación es:

$$R_3(x) = (x-x_0)(x-x_1)(x-x_2)(x-x_3)\, f[x_0, x_1, x_2, x_3, x_4]$$

$$= (1.5\text{-}1.35)(1.5\text{-}1.7)(1.5\text{-}1.9)(1.5\text{-}1.0)(\text{-}0.030567)$$

$$= 0.00018$$

Obsérvese que $R_3(1.5)$ es menor que $R_2(1.5)$, por lo que el polinomio de tercer grado da mejor aproximación a esta interpolación que el de segundo grado.

5.8 Para calibrar un medidor de orificio se miden la velocidad v de un fluido y la caída de presión ΔP. Los datos experimentales se dan a continuación y se buscan los mejores parámetros a y b de la ecuación que represente estos datos:

$$v = a\,(\Delta P)^b \tag{1}$$

donde: v = velocidad promedio (pie/s)
ΔP = caída de presión (mm Hg)

i	1	2	3	4	5	6	7	8	9	10	11
v_i	3.83	4.17	4.97	6.06	6.71	7.17	7.51	7.98	8.67	9.39	9.89
ΔP_i	30.0	35.5	50.5	75.0	92.0	105.0	115.0	130.0	153.5	180.0	199.5

SOLUCIÓN

Este problema puede resolverse mediante el método de mínimos cuadrados de la siguiente manera

Se aplican logaritmos a la ecuación 1 y se tiene

$$ln\, v = ln\, a + b\, ln\, (\Delta) \tag{2}$$

al definir $y = ln\, v$; $a_0 = ln\, a$; $a_1 = b$; $x = ln\, (\Delta P)$ y sustituir en la ecuación 2 queda

$$y = a_0 + a_1 x \tag{3}$$

ecuación de una línea recta.

Si se calculan los parámetros a_0 y a_1 de la recta (Ec.3) con el método de mínimos cuadrados, se obtienen (indirectamente) los mejores valores a_0 y b que representan los datos experimentales.

Para calcular a_0 y a_1 se construye la siguiente tabla para que los cálculos sean más eficientes (puede usarse una hoja de cálculo electrónica o un pizarrón electrónico)

Puntos i	v_i	ΔP_i	y_i $\ln v_i$	x_i $\ln \Delta P_i$	x_i^2 $(\ln \Delta P_i)^2$	$y_i x_i$ $\ln v_i \ln \Delta P_i$
1	3.83	30.0	1.34286	3.40120	11.56816	4.56734
2	4.17	35.5	1.42792	3.56953	12.74154	5.09700
3	4.97	50.5	1.60342	3.92197	15.38185	6.28857
4	6.06	75.0	1.80171	4.31749	18.64072	7.77886
5	6.71	92.0	1.90360	4.52179	20.44658	8.60768
6	7.17	105.0	1.96991	4.65396	21.65934	9.16788
7	7.51	115.0	2.01624	4.74493	22.51436	9.56692
8	7.98	130.0	2.07694	4.86753	23.69285	10.10957
9	8.67	153.5	2.15987	5.03370	25.33814	10.87214
10	9.39	180.0	2.23965	5.19296	26.96683	11.63041
11	9.89	199.5	2.29581	5.29581	28.04560	12.13545
Totales			20.83364	49.52087	226.99598	95.82182

Los valores de las sumatorias se sustituyen en el sistema de ecuaciones 5.62 y se tiene

$$a_0 = \frac{\begin{vmatrix} 20.83364 & 49.52087 \\ 95.82182 & 226.99598 \end{vmatrix}}{\begin{vmatrix} 11.0 & 49.52087 \\ 49.52087 & 226.99598 \end{vmatrix}} = -0.35904$$

$$a_1 = \frac{\begin{vmatrix} 11.0 & 20.83364 \\ 49.52087 & 95.82182 \end{vmatrix}}{\begin{vmatrix} 11.0 & 49.52087 \\ 49.52087 & 226.99598 \end{vmatrix}} = 0.50046$$

Ecuación resultante

$$y = -0.35904 + 0.50046\,x$$

De donde:

$$\ln a = -0.35904 \qquad \text{y} \qquad a = 0.69835$$

$$b = 0.50046$$

Con estos valores, la ecuación que representa los datos experimentales queda

$$v = 0.69835\ (\Delta P)^{0.50046}$$

5.9 Al medir la velocidad (con un tubo de Pitot) en una tubería circular de diámetro interior de 20 cm, se encontró la siguiente información

v (cm/s)	600	550	450	312	240
r (cm)	0	3	5	7	8

donde r es la distancia en cm medida a partir del centro del tubo.

a) Obtenga la curva $v = f(r)$ que aproxima estos datos experimentales.
b) Calcule la velocidad en el punto $r = 4$ cm.

SOLUCIÓN

a) Se asume que en la experimentación hay errores, de tal modo que se justifica usar una aproximación por mínimos cuadrados. Por otro lado se sabe que el perfil de velocidades en una tubería generalmente es de tipo parabólico, por lo que se ensayará un polinomio de segundo grado

$$v(r) = a_0 + a_1 r + a_2 r^2$$

Al construir la tabla que proporcione los coeficientes del sistema de ecuaciones 5.64 se tiene

Puntos	v	r	r^2	r^3	r^4	vr	vr^2
1	600	0	0	0	0	0	0
2	550	3	9	27	81	1650	4950
3	450	5	25	125	625	2250	11250
4	312	7	49	343	2401	2184	15288
5	240	8	64	512	4096	1920	15360
Totales	2152	23	147	1007	7203	8004	46848

Estos valores se sustituyen en el sistema de ecuaciones 5.64 particularizado para un polinomio de segundo grado y se tiene

$$\begin{aligned} 5\,a_0 + 23\,a_1 + 147\,a_2 &= 2152 \\ 23\,a_0 + 147\,a_1 + 1007\,a_2 &= 8004 \\ 147\,a_0 + 1007\,a_1 + 7203\,a_2 &= 46848 \end{aligned}$$

Al resolver para los parámetros a_0, a_1 y a_2 y sustituirlos en el polinomio propuesto queda

$$v(r) = 601.714 - 3.667\, r - 5.347\, r^2$$

b) Con la sustitución $r = 4$, se obtiene

$$v(4) = 503.89 \text{ cm/s}$$

Obsérvese que la distribución de velocidades sólo se presenta del centro a la pared del tubo, ya que es simétrica.

5.10 El porcentaje de impurezas que se encuentra, a varias temperaturas y tiempos de esterilización en una reacción asociada con la fabricación de cierta bebida, está representado por los datos siguientes

Tiempo de esterilización (min)	Temperatura °C x_1		
x_2	75	100	125
15	14.05	10.55	7.55
	14.93	9.48	6.59
20	16.56	13.63	9.23
	15.87	11.75	8.78
25	22.41	18.55	15.93
	21.66	17.98	16.44

Estime los coeficientes de regresión lineal en el modelo

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1^2 + a_4 x_2^2 + a_5 x_1 x_2$$

SOLUCIÓN

Si bien el modelo no es lineal, puede transformarse en lineal con los siguientes cambios de variables

$$x_3 = x_1^2,\ x_4 = x_2^2,\ x_5 = x_1 x_2$$

que sustituidos en el modelo propuesto dan

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 + a_5 x_5$$

cuyos parámetros, siguiendo el criterio de los mínimos cuadrados, pueden obtenerse a partir del sistema

$$
\begin{aligned}
n\,a_0 + a_1\Sigma x_1 + a_2\Sigma x_2 + a_3\Sigma x_3 + a_4\Sigma x_4 + a_5\Sigma x_5 &= \Sigma y \\
a_0\Sigma x_1 + a_1\Sigma x_1^2 + a_2\Sigma x_1x_2 + a_3\Sigma x_1x_3 + a_4\Sigma x_1x_4 + a_5\Sigma x_1x_5 &= \Sigma x_1y \\
a_0\Sigma x_2 + a_1\Sigma x_2x_1 + a_2\Sigma x_2^2 + a_3\Sigma x_2x_3 + a_4\Sigma x_2x_4 + a_5\Sigma x_2x_5 &= \Sigma x_2y \\
a_0\Sigma x_3 + a_1\Sigma x_3x_1 + a_2\Sigma x_3x_2 + a_3\Sigma x_3^2 + a_4\Sigma x_3x_4 + a_5\Sigma x_3x_5 &= \Sigma x_3y \\
a_0\Sigma x_4 + a_1\Sigma x_4x_1 + a_2\Sigma x_4x_2 + a_3\Sigma x_4x_3 + a_4\Sigma x_4^2 + a_5\Sigma x_4x_5 &= \Sigma x_4y \\
a_0\Sigma x_5 + a_1\Sigma x_5x_1 + a_2\Sigma x_5x_2 + a_3\Sigma x_5x_3 + a_4\Sigma x_5x_4 + a_5\Sigma x_5^2 &= \Sigma x_5y
\end{aligned}
$$

Ahora, los valores de la tabla que se dan arriba se disponen así

Puntos	1	2	3	4	5	6	7	...
x_1	75	75	75	75	75	75	100	...
x_2	15	15	20	20	25	25	15	...
y	14.05	14.93	16.56	15.87	22.41	21.66	10.55	...
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

Se continúa adicionando las filas necesarias: x_3, x_4, x_5, x_1^2, x_2^2, x_1x_2, ... y sumando los totales de cada una para conseguir los coeficientes y el vector de términos independientes del sistema. Dichos cálculos dan como resultado el siguiente sistema de ecuaciones

$$
\begin{bmatrix}
18 & 1800 & 360 & 187500 & 7500 & 36000 \\
1800 & 187500 & 36000 & 20250000 & 750000 & 3750000 \\
360 & 36000 & 7500 & 3750000 & 162000 & 750000 \\
187500 & 20250000 & 3750000 & 2254687500 & 78125000 & 405000000 \\
7500 & 750000 & 162000 & 78125000 & 3607500 & 16200000 \\
36000 & 3750000 & 750000 & 405000000 & 16200000 & 78125000
\end{bmatrix}
\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix}
=
\begin{bmatrix} 251.94 \\ 24170.0 \\ 5287.9 \\ 2420850.5 \\ 115143.0 \\ 508702.5 \end{bmatrix}
$$

cuya solución por medio de alguno de los métodos del capítulo 3 es

$$
a_0 = 56.4264, \qquad a_1 = -0.362597, \qquad a_2 = -2.74767
$$

$$
a_3 = 0.00081632, \qquad a_4 = 0.0816, \qquad a_5 = 0.00314
$$

se sustituyen en el modelo y resulta

$$y=56.4264-0.362597x_1-2.74767x_2+0.00081632x_1^2+0.0816x_2^2+0.00314x_1x_2$$

Una vez obtenidos los coeficientes, puede estimarse el porcentaje de impurezas correspondiente a un tiempo de esterilización y una temperatura dados; por ejemplo, a un tiempo de 19 min y una temperatura de 80°C se tiene un porcentaje de impurezas de:

$$y = 56.4264 - 0.362597(80) - 2.74767(19) + 0.00081632(80)^2 + 0.0816(19)^2 + 0.00314(80)(19) = 14.67$$

Problemas

5.1 La densidad del carbonato neutro de potasio en solución acuosa varía con la temperatura y la concentración de acuerdo con la tabla siguiente

T (°C) / c (%)	0	40	80	100
4	1.0381	1.0276	1.0063	0.9931
12	1.1160	1.1013	1.0786	1.0663
20	1.1977	1.1801	1.1570	1.1451
28	1.2846	1.2652	1.2418	1.2301

a) Calcule la densidad a 40 °C y 15% de concentración
b) Calcule la densidad a 50 °C y 28% de concentración
c) Calcule la densidad a 90 °C y 25% de concentración
d) Calcule la concentración que tiene una solución de densidad 1.129 a una temperatura de 60°C

Utilice interpolaciones cuadráticas en todos los incisos.

5.2 Los datos de presión-temperatura-volumen para el etano se muestran en la tabla siguiente, donde la temperatura (T) está en °C, la presión (P) en atmósferas y el volumen específico (1/V) en moles/litro.

	P						
T	1	2	4	6	8	9	10
25	20.14	32.84	—	—	—	—	—
75	24.95	43.80	68.89	85.95	104.38	118.32	139.23
150	31.89	59.31	106.06	151.38	207.66	246.57	298.02
200	36.44	69.38	130.18	194.53	276.76	332.56	—
250	40.87	79.16	153.59	237.38	345.38	—	—

Calcule el volumen específico en moles/litro para una presión de 7 atmósferas y una temperatura de 175 °C.

5.3 Dados

Puntos	0	1	2
x	x_0	x_1	x_2
$f(x)$	$f(x_0)$	$f(x_1)$	$f(x_2)$

a) Encuentre los coeficientes a_0, a_1, a_2, del polinomio de segundo grado que pasa por estos tres puntos, por el método de Lagrange.

b) Realice el mismo proceso que en (a) pero ahora empleando el método de aproximación polinomial simple.

c) Demuestre que los polinomios en los incisos (a) y (b) son el mismo, pero escrito en diferente forma.

5.4 Dada una función $y = f(x)$ en forma tabular, a menudo se desea encontrar un valor de x correspondiente a un valor dado de y; este proceso, llamado **interpolación inversa,** se lleva a cabo en la forma ya vista, pero intercambiando los papeles de x y y. Dada la siguiente tabla

Puntos	0	1	2	3	4	5	6
x	0.0	2.5	5.0	7.5	10.0	12.5	15.0
y	10.00	4.97	2.47	1.22	0.61	0.30	0.14

donde y es la amplitud de la oscilación de un péndulo largo, en cm y x es el tiempo medido en min desde que empezó la oscilación.
Encuentre el polinomio de aproximación de Lagrange de segundo grado que pasa por los puntos (1), (2) y (3) y el valor de x correspondiente a $y = 2$ cm.

5.5 Sea $z(x) = \prod_{j=0}^{n} (x - x_j)$. Demuestre que el polinomio 5.22 puede escribirse en la forma

$$p_n(x) = z(x) \sum_{i=0}^{n} \frac{f(x_i)}{(x - x_j) z'(x_i)}$$

5.6 Use las ideas dadas en el problema anterior para demostrar que

$$\sum_{i=0}^{n} L_i(x) = 1 \qquad \text{para toda } x.$$

Sugerencia: Considere que la expresión dada corresponde al polinomio de aproximación por polinomios de Lagrange de $f(x)=1$ (un polinomio de grado cero).

5.7 Demuestre que el polinomio de aproximación de Lagrange de primer grado puede escribirse en notación de determinantes así

$$p_{0,1}(x) = \frac{1}{x_1 - x_0} \begin{vmatrix} p_0(x) & (x_0 - x) \\ p_1(x) & (x_1 - x) \end{vmatrix}$$

donde $p_0(x) = f(x_0)$ y $p_1(x) = f(x_1)$ y los subíndices 0 y 1 de $p(x)$ se refieren a los puntos (0) y (1) por donde pasa el polinomio de aproximación.
Demuestre también que para el caso del polinomio de aproximación de Lagrange de segundo grado que pasa por los puntos (0), (1) y (2)

$$p_{0,1,2}(x) = \frac{1}{x_2 - x_1} \begin{vmatrix} p_{0,1}(x) & (x_1 - x) \\ p_{0,2}(x) & (x_2 - x) \end{vmatrix}$$

5.8 Lo demostrado en el problema anterior es válido, en general, para aproximaciones de tercero, cuarto, ..., n grado. Aitken desarrolló un método para interpolar con este tipo de polinomios y consiste en construir la tabla siguiente

$$\begin{array}{l|lllll|l} x_0 & p_0 & & & & & (x_0-x) \\ x_1 & p_1 & p_{0,1} & & & & (x_1-x) \\ x_2 & p_2 & p_{0,2} & p_{0,1,2} & & & (x_2-x) \\ x_3 & p_3 & p_{0,3} & p_{0,1,3} & p_{0,1,2,3} & & (x_3-x) \\ x_4 & p_4 & p_{0,4} & p_{0,1,4} & p_{0,1,2,4} & p_{0,1,2,3,4} & (x_4-x) \end{array}$$

donde $p_i = f(x_i)$ y x el valor donde se desea interpolar.
Para el cálculo de $p_{0,i}$ se emplea el determinante

$$p_{0,i}(x) = \frac{1}{x_i - x_0} \begin{vmatrix} p_0 & (x_0 - x) \\ p_i & (x_i - x) \end{vmatrix}$$

donde el denominador resulta ser $(x_i-x) - (x_0-x)$.
En cambio para $p_{0,1,i}$ se usa

$$p_{0,1,i}(x) = \frac{1}{x_i - x_1} \begin{vmatrix} p_{0,1} & (x_1 - x) \\ p_{0,i} & (x_i - x) \end{vmatrix}$$

cuyo denominador es $(x_i-x) - (x_1 - x)$.
Se aconseja denotar la abscisa más cercana a x como x_0, la segunda más próxima a x como x_1 y así sucesivamente.
Con ese ordenamiento los valores $p_{0,1}$, $p_{0,1,2}$, $p_{0,1,2,3}$, etc., representan la mejor aproximación al valor buscado $f(x)$ con polinomios de primero, segundo, tercero, ..., n grado.
Con el método descrito, aproxime el valor de la función de Bessel (J_0) dada abajo en $x = 0.8$

Puntos	0	1	2	3
x	0.5	0.7	0.9	1.0
$J_0(x)$	0.9385	0.8812	0.8075	0.7652

5.9 En el método de posición falsa (Capítulo 2) se realiza una interpolación inversa: Dados los puntos $(x_I, f(x_I))$ y $(x_D, f(x_D))$ se encuentra el polinomio $p(x)$ que pasa por esos puntos y luego el valor de x correspondiente a $p(x)=0$. Discuta la interpolación inversa para encontrar raíces de ecuaciones no lineales empleando tres puntos.

5.10 Demuestre que si la función $f(x)$ dada en forma tabular corresponde a un polinomio de grado n, entonces el polinomio de aproximación $p(x)$ de grado mayor o igual a n que pasa por los puntos de la tabla es $f(x)$ misma.
Sugerencia: Con el polinomio $y = 2x + 3$ forme una tabla de valores y tomando dos de esos valores encuentre $p(x)$ y observe que $p(x) = y$; después, tomando 3 valores cualesquiera observe que el $p(x)$ obtenido es nuevamente $y = 2x + 3$. Sólo resta generalizar estos resultados.

5.11 Desarrolle algebraicamente el numerador y el denominador de

$$a_2 = \frac{f[x_2] - f[x_0] - (x_2 - x_0)\dfrac{f[x_1] - f[x_0]}{x_1 - x_0}}{(x_2 - x_0)(x_2 - x_1)}$$

para llegar a

$$a_2 = \frac{\dfrac{f[x_2] - f[x_1]}{x_2 - x_1} - \dfrac{f[x_1] - f[x_0]}{x_1 - x_0}}{x_2 - x_0} = f[x_0, x_1, x_2]$$

5.12 Verifique que para tres puntos distintos cualesquiera de abscisas x_0, x_1 y x_2 se cumple que

$$f[x_0, x_1, x_2] = f[x_2, x_0, x_1] = f[x_1, x_2, x_0]$$

así como con cualquier otra permutación de x_1, x_2, x_0. Esta propiedad de las diferencias de segundo oden es conocida como **simetría respecto a los argumentos** y la cumplen también las diferencias de primer orden (trivial), las de orden 3, etcétera.

5.13 Elabore un subprograma de propósito general para construir la tabla de diferencias divididas de una función tabulada.
Sugerencia: Vea el algoritmo 5.3. Puede usar una hoja de cálculo electrónica.

5.14 Para los valores siguientes

Puntos	0	1	2	3	4	5	6
e	40	60	80	100	120	140	160
p	0.63	1.36	2.18	3.00	3.93	6.22	8.59

donde *e* son los volts y *p* los kilowatts en una curva de pérdida en el núcleo para un motor eléctrico:

a) Elabore una tabla de diferencias divididas.
b) Con el polinomio de Newton en diferencias divididas de segundo grado, aproxime el valor de *p* correspondiente a $e = 90$ volts.

5.15 En la tabla siguiente

i	1	2	3	4
v	120	94	75	62

donde *i* es la corriente y *v* el voltaje consumido por un arco magnético, aproxime el valor de *v* para $i = 3.5$ por un polinomio de Newton en diferencias dividas y compare con el valor dado por la fórmula empírica

$$v = 30.4 + 90.4\, i^{-0.507}$$

5.16 Corrobore que el polinomio de Newton en diferencias divididas puede escribirse en términos de Π, así

$$p_n(x) = p_n(x_0 + sh) = \sum_{k=0}^{n} \Delta^k f(x_0) \prod_{i=0}^{k-1} \frac{s-i}{i+1} \tag{1}$$

Nota: Considere que $\Delta^0 f(x_0) = f(x_0)$. Esta notación es generalmente más útil para programar este algoritmo.

5.17 Con los resultados del problema anterior y con la definición de función binomial siguiente, exprese la ecuación (1) en términos de $\binom{s}{k}$.

$$\binom{s}{k} = \begin{cases} 1 & k = 0 \\ \prod_{i=0}^{k-1} \frac{s-i}{i+1} = \frac{s(s-1)(s-2)\ldots(s-(k-1))}{1(2)(3)\ldots(k)} & k > 0 \end{cases}$$

5.18 Con los siguientes valores

Puntos	0	1	2	3
l/r	140	180	220	240
p/a	12,800	7,500	5,000	3,800

donde *p/a* es la carga en lb/pulg^2 que causa la ruptura de una columna de hierro dulce con extremos redondeados y *l/r* es la razón de la longitud de la columna al mínimo radio de giro de su sección transversal.
Encuentre el polinomio de tercer grado que pasa por estos puntos en sus distintas formas

a) $p_3(x) = a_0 + a_1x + a_2x^2 + a_3x^3$ (aproximación polinomial simple).

b) Forma de Lagrange

c) Aproximación de Newton (en diferencias divididas)

d) Aproximación de Newton en diferencias finitas (hacia delante y hacia atrás)

5.19 En una reacción química, la concentración del producto C_B cambia con el tiempo como se indica en la tabla de abajo. Calcule la concentración C_B cuando $t = 0.82$, usando un polinomio de Newton en diferencias finitas.

C_B	0.00	0.30	0.55	0.80	1.10	1.15
t	0.00	0.10	0.40	0.60	0.80	1.00

5.20 Resuelva el problema 5.13, empleando diferencias finitas; compare los cálculos realizados y los resultados obtenidos en ambos problemas.

5.21 Elabore un diagrama de flujo y codifíquelo para leer n pares de valores x y $f(x)$, calcular e imprimir la tabla de diferencias finitas hacia atrás

Puntos	x_i	$f[x_i]$	$\nabla f[x_i]$	$\nabla^2 f[x_i]$	$\nabla^3 f[x_i]$	...	$\nabla^{n-2} f[x_i]$
0	x_0	$f[x_0]$					
			$\nabla f[x_1]$				
1	x_1	$f[x_1]$		$\nabla^2 f[x_1]$			
			$\nabla f[x_2]$		$\nabla^3 f[x_1]$		
2	x_2	$f[x_2]$		$\nabla^2 f[x_2]$			
			$\nabla f[x_3]$		$\nabla^3 f[x_2]$		
3	x_3	$f[x_3]$		$\nabla^2 f[x_3]$			
.	.	.	.	.	.	...	$\nabla^{n-2} f[x_{n-1}]$
.	.	.	.	.	.		
.	.	.	.	.	$\nabla^3 f[x_{n-1}]$		
				$\nabla^2 f[x_{n-1}]$			
			$\nabla f[x_{n-1}]$				
$n-1$	x_{n-1}	$f[x_{n-1}]$					

$$\nabla f[x] = f[x] - f[x-h]$$

$$\nabla^n f[x] = \nabla^{n-1} f[x] - \nabla^{n-1} f[x-h]$$

5.22 En el caso en que la distancia h entre dos argumentos consecutivos cualesquiera es la misma a lo largo de la tabla, puede usarse la ecuación 5.35 para interpolar en puntos cercanos a x_0 o bien la 5.38 cuando se quiere interpolar en puntos al final de la tabla (véase Sec. 5.5). Si hay que interpolar en puntos centrales de la tabla, resulta conveniente denotar alguno de dichos puntos centrales como x_0, como x_1, x_2, x_3, ... las abscisas mayores que x_0 y como x_{-1}, x_{-2}, x_{-3}, ..., las abscisas menores que x_0. En estas condiciones e introduciendo el operador lineal δ, conocido como **operador en diferencias centrales** y definido sobre $f(x)$ como

$$\delta f(x) = f(x + h/2) - f(x - h/2) \tag{1}$$

y cuya aplicación sucesiva conduce a

$$\delta(\delta f(x)) = \delta^2 f(x) = f(x + h) - 2 f(x) + f(x - h)$$

y en general a

$$\delta^i f(x) = \delta(\delta^{i-1} f(x)) \tag{2}$$

Nótese que $\delta f(x_0)$ no emplea, en general, los valores de la tabla, lo cual constituye una dificultad para su uso. En cambio la segunda diferencia central

$$\delta^2 f(x_k) = f(x_k + h) - 2 f(x_k) + f(x_k - h)$$

incluye sólo valores funcionales tabulados; esto es cierto para todas las diferencias centrales de orden par. A fin de evitar que se requieran valores funcionales no tabulados en la primera diferencia central, puede aplicarse δ a puntos no tabulados, por ejemplo a $f(x_k + h/2)$ con lo cual queda

$$\delta f(x_k + h/2) = f(x_k + h) - f(x_k) = f(x_{k+1}) - f(x_k),$$

donde ya sólo aparecen valores funcionales de la tabla.
En general $\delta^{2i+1} f(x_k + h/2)$ (orden impar) queda en función de ordenadas presentes en la tabla.
Con la notación de diferencias divididas se tiene que

$$\begin{aligned}
\delta f(x_0 + h/2) &= f(x_1) - f(x_0) = h\, f[x_0, x_1] \\
\delta f(x_0 - h/2) &= f(x_0) - f(x_{-1}) = h\, f[x_0, x_{-1}] \\
\delta^2 f(x_1) = \delta f(x_1 + h/2) - \delta f(x_1 - h/2) &= h f[x_1, x_2] - h f[x_0, x_1] \\
&= 2!\, h^2 f[x_0, x_1, x_2]
\end{aligned}$$

y en general

$$\delta^{2i+1} f(x_k + h/2) = h^{2i+1} (2i+1)! f[x_{k-i}, \ldots, x_k, \ldots, x_{k+i}, x_{k+i+1}] \tag{3}$$

$$\delta^{2i+1} f(x_k - h/2) = h^{2i+1}\, (2i+1)! f[x_{k-i-1}, x_{k-i}, \ldots, x_k, \ldots, x_{k+i}] \tag{4}$$

para orden impar y

$$\delta^{2i} f(x_k) = h^{2i}\, (2i)! f[x_{k-i}, \ldots, x_k, \ldots, x_{k+i}] \tag{5}$$

para orden par.
La tabla de diferencias centrales queda entonces

.	.				
.	.				
.	.				
x_{-2}	$f(x_{-2})$				
		$\delta f(x_{-2} + h/2)$			
x_{-1}	$f(x_{-1})$		$\delta^2 f(x_{-1})$		
		$\delta f(x_{-1} + h/2)$		$\delta_3 f(x_{-1} + h/2)$	
x_0	$f(x_0)$		$\delta^2 f(x_0)$		...
		$\delta f(x_0 + h/2)$		$\delta_3 f(x_0 + h/2)$	
x_1	$f(x_1)$		$\delta^2 f(x_1)$		
		$\delta f(x_1 + h/2)$			
x_2	$f(x_2)$				
.	.				
.	.				
.	.				

Note que el argumento permanece constante en cualquier línea horizontal de la tabla. Con esta notación y la aplicación sucesiva de las ecuaciones 3 y 5 con $k = 0$, la 5.29 se transforma en

$$f(x) = f(x_0) + (x-x_0)\frac{\delta f(x_0+h/2)}{1!h} + (x-x_0)(x-x_1)\frac{\delta^2 f(x_0)}{2!h^2} +$$

$$(x-x_0)(x-x_1)(x-x_{-1})\frac{\delta^3 f(x_0+h/2)}{3!h^3} + \ldots \qquad (6)$$

Al emplear el cambio de variable

$$x = x_0 + sh,$$

de donde

$$s = \frac{x - x_0}{h}$$

el polinomio (6) queda

$$p_n(x_0+sh) = f(x_0) + s\,\delta f(x_0+h/2) + \frac{s(s-1)}{2!}\delta^2 f(x_0) +$$

$$\frac{s(s^2-1^2)}{3!}\delta^3 f(x_0+h/2) + \frac{s(s^2-1^2)(s-2)}{4!}\delta^4 f(x_0) +$$

$$\ldots + \frac{s(s^2-1^2)\ldots(s^2-(i-1)^2)(s-i)}{(2i)!}\delta^{2i} f(x_0) \qquad (7)$$

cuando el grado del polinomio es par; si es impar, el último término de la ecuación 7 queda como

$$\ldots + \frac{s(s^2-1^2)\ldots(s^2-i^2)}{(2i+1)!}\delta^{2i+1} f(x_0+h/2)$$

Este polinomio se conoce como la **fórmula hacia delante de Gauss.**

Con la tabla del ejemplo 5.7 construya una tabla de diferencias centrales y mediante la ecuación 7 encuentre por interpolación la presión correspondiente a una temperatura de 76°F.

5.23 Si aproxima la función dada abajo por un polinomio de segundo grado y con éste interpola en $x = 10$, estime el error cometido en esta interpolación.

Puntos	0	1	2	3	4	5	6
x	0	1	6	8	11.5	15	19
$f(x)$	38000	38500	35500	27500	19000	15700	11000

5.24 Demuestre que el término del error para la aproximación polinomial de segundo grado es

$$R_2(x) = (x-x_0)(x-x_1)(x-x_2)\,f[x,x_0,x_1,x_2]$$

5.25 Encuentre una cota inferior y una cota superior del error de interpolación $R_3(x)$ en $x = 6.3$ para la función $f(x) = e^x$ dada en los puntos $x_0=5$, $x_1=6$, $x_2=7$, $x_3=8$ (véase ejemplo 5.11).

5.26 Demuestre que la función dada por $z(x) = |(x-x_0)(x-x_1)|$ con $x_0 \le x \le x_1$ alcanza su valor máximo en $(x_0 + x_1)/2$ y está dado por $(x_1-x_0)^2/4$.

5.27 Con los resultados del problema anterior y la fórmula

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x-x_i),$$

demuestre que el error $R_1(x)$ con $x_0 \le x \le x_1$ correspondiente a una aproximación lineal de $f(x)$ usando como argumento x_0 y x_1 es menor en magnitud (valor absoluto) que $M(x_1-x_0)^2/8$, donde M es el valor máximo de $|f''(x)|$ en $[x_0, x_1]$.

5.28 Los siguientes valores fueron obtenidos de una tabla de distribución binomial

b		*p*				
n	*x*	0.0500	0.1000	0.1500	0.2000	0.2500
3	0	0.8574	0.7290	0.6141	0.5120	0.4219
	1	0.1354	0.2430	0.3251	0.3840	0.4219
	2	0.0071	0.0270	0.0574	0.0960	0.1406
	3	0.0001	0.0010	0.0034	0.0080	0.0156

Al pie de dicha tabla se lee "la interpolación lineal dará valores exactos de *b* a lo más en dos cifras decimales".

Encuentre una aproximación de $f(x; n, p) = f(1;3,0.13)$ exacta en tres cifras decimales. Recuerde que

$$b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

¿Cree usted que si los valores de la tabla son exactos en las cuatro cifras decimales dadas, pueda obtenerse exactitud con cuatro cifras decimales aplicando el método de interpolación?

5.29 En la siguiente tabla, *r* es la resistencia de una bobina en ohms y T la temperatura de la bobina en °C. Por mínimos cuadrados determine el mejor polinomio lineal que represente la función dada.

r	10.421	10.939	11.321	11.794	12.242	12.668
T	10.50	29.49	42.70	60.01	75.51	91.05

5.30 En la tabla

Puntos	0	1	2	3	4	5	6	7	8
v	26.43	22.40	19.08	16.32	14.04	12.12	10.51	9.15	8.00
P	14.70	17.53	20.80	24.54	28.83	33.71	39.25	45.49	52.52

v es el volumen en pie^3 de una lb de vapor y P es la presión en psia. Encuentre los parámetros *a* y *b* de la ecuación

$$P = a\, v^b$$

aplicando el método de mínimos cuadrados.

5.31 Se sabe que el número de pulgadas que una estructura recién construida se hunde en el suelo está dada por

$$y = 3 - 3\,e^{-ax}$$

donde x es el número de meses que lleva construida la estructura. Con los valores

x	2	4	6	12	18	24
y	1.07	1.88	2.26	2.78	2.97	2.99

estime a, usando el criterio de los mínimos cuadrados (véase ejercicio 5.8).

5.32 En el estudio de la constante de velocidad k de una reacción química a diferentes temperaturas, se obtuvieron los datos

T (K)	293	300	320	340	360	380	400
k	8.53×10^{-5}	19.1×10^{-5}	1.56×10^{-3}	0.01	0.0522	0.2284	0.8631

Calcule el factor de frecuencia z y la energía de activación E asumiendo que los datos experimentales siguen la ley de Arrhenius:

$$k = z\,e^{-E/1.98T}$$

5.33 Sieder y Tate* encontraron que una ecuación que relaciona la transferencia de calor de líquidos por dentro de tubos en cambiadores de calor se puede representar con números adimensionales

$$Nu = a\,(\mathrm{Re})^b\,(Pr)^c\,\left(\frac{\mu}{\mu_w}\right)^d$$

Donde *Nu es* el número de Nusselt, *Re* es el número de Reynolds, *Pr* el número de Prandtl y μ y μ_w y las viscosidades del líquido a la temperatura promedio de éste y a la temperatura de la pared del tubo, respectivamente.
Encuentre los valores de *a, b, c* y *d* asumiendo que la tabla siguiente representa datos experimentales para un grupo de hidrocarburos a diferentes condiciones de operación.

Nu	97.45	109.50	129.90	147.76	153.44	168.90	177.65	175.16
Re	10500	12345	15220	18300	21050	25310	28560	31500
Pr	18.2	17.1	16.8	15.3	12.1	10.1	8.7	6.5
μ/μ_w	0.85	0.90	0.96	1.05	1.08	1.15	1.18	1.22

5.34 Elabore un programa de propósito general, para aproximar una función dada en forma tabular por un polinomio de grado n usando el método de mínimos cuadrados.

---

*Sieder y Tate. Ind. and Eng. Chem. 28, 1429 (1936).

5.35 En una reacción gaseosa de expansión a volumen constante, se observa que la presión del reactor (*batch*) aumenta con el tiempo de reacción según se muestra en la tabla de abajo. ¿Qué grado de polinomio (con el criterio de ajuste exacto) aproxima mejor la función P = f(t)?

P (atm)	1.0000	1.0631	1.2097	1.3875	1.7232	2.0000	2.9100
t (min)	0.0	0.1	0.3	0.5	0.8	1.0	1.5

5.36 La aparición de una corriente inducida en un circuito que tiene la constante de tiempo τ está dada por

$$I = 1 - e^{-t/\tau}$$

donde t es el tiempo medido desde el instante en que el interruptor se cierra e I la razón de la corriente en tiempo t al valor total de la corriente dado por la ley de Ohm. Con las mediciones siguientes, estime la constante de tiempo τ de este circuito (consúltese el Ejer. 5.8).

I	0.073	0.220	0.301	0.370	0.418	0.467	0.517
t (seg)	0.1	0.2	0.3	0.4	0.5	0.6	0.7

5.37 Los valores

t	0.0	10.0	27.4	42.1
s	61.5	62.1	66.3	70.3

representan la cantidad s en gr de dicromato de potasio disueltos en 100 partes de agua a la temperatura t indicada en °C. La relación entre estas variables es

$$\log_{10} s = a + b\,t + ct^2$$

Calcule los parámetros a, b y c por el método de mínimos cuadrados.

5.38 Para la tabla de datos que se da abajo, encuentre los parámetros a y b de la ecuación

$$y = a + (0.49-a)\,e^{-b(x-8)}$$

x	10	20	30	40
y	0.48	0.42	0.40	0.39

5.39 Veinte tipos de hojas de acero procesadas en frío tienen diferentes composiciones de cobre y temperaturas de templado. Al medir su dureza resultante se obtuvieron los siguientes valores

Dureza Rockwell 30-T	u Contenido de cobre %	v Temp. de Templado °F
78.9	0.02	1000
65.1	0.02	1100
55.2	0.02	1200
56.4	0.02	1300
80.9	0.10	1000
69.7	0.10	1100
57.4	0.10	1200
55.4	0.10	1300
85.3	0.18	1000
71.8	0.18	1100
60.7	0.18	1200
58.9	0.18	1300

Se sabe que la dureza depende en forma lineal del contenido u de cobre en % y de la temperatura de templado v

$$y = a_0 + a_1 u + a_2 v$$

Determine los parámetros a_0, a_1 y a_2, siguiendo el criterio de los mínimos cuadrados.

CAPÍTULO 6

INTEGRACIÓN Y DIFERENCIACIÓN NUMÉRICA

EN ESTE CAPÍTULO se abordan los temas clásicos de integración y derivación con procesos finitos de aproximación.

INTRODUCCIÓN

Una vez que se ha determinado un polinomio $p_n(x)$* de manera que aproxime satisfactoriamente una función dada $f(x)$ sobre un intervalo de interés, puede esperarse que al diferenciar $p_n(x)$ o integrarla en forma definida, también aproxime satisfactoriamente la derivada o integral definida correspondientes a $f(x)$. Sin embargo, si se observa la figura 6.1 —donde aparece la gráfica de un polinomio $p_n(x)$ que aproxima la curva que representa la función $f(x)$— puede anticiparse que aunque la desviación de $p_n(x)$ y $f(x)$ en el intervalo $[x_0, x_n]$ es pequeña, las pendientes de las curvas que las representan pueden diferir considerablemente; esto es, la diferenciación numérica tiende a ampliar pequeñas discrepancias o errores del polinomio de aproximación.

Por otro lado, en el proceso de integración (véase Fig. 6.2), el valor de

$$\int_{x_0}^{x_n} f(x)\,dx,$$

está dado por el área bajo la curva de $f(x)$, mientras que la aproximación

$$\int_{x_0}^{x_n} p_n(x)\,dx,$$

está dada por el área bajo la curva de $p_n(x)$ y los errores que se cometen en diferentes segmentos del intervalo tienden a cancelarse entre sí o a reducirse. Por esto el error total al integrar $p_n(x)$ entre x_0 y x_n puede ser muy pequeño, aún cuando $p_n(x)$ no sea una buena aproximación de $f(x)$.

En resumen: Si la aproximación polinomial $p_n(x)$ es buena, la integral

$$\int_{x_0}^{x_n} p_n(x)\,dx,$$

*Ya sea por el criterio del ajuste exacto o el de mínimos cuadrados.

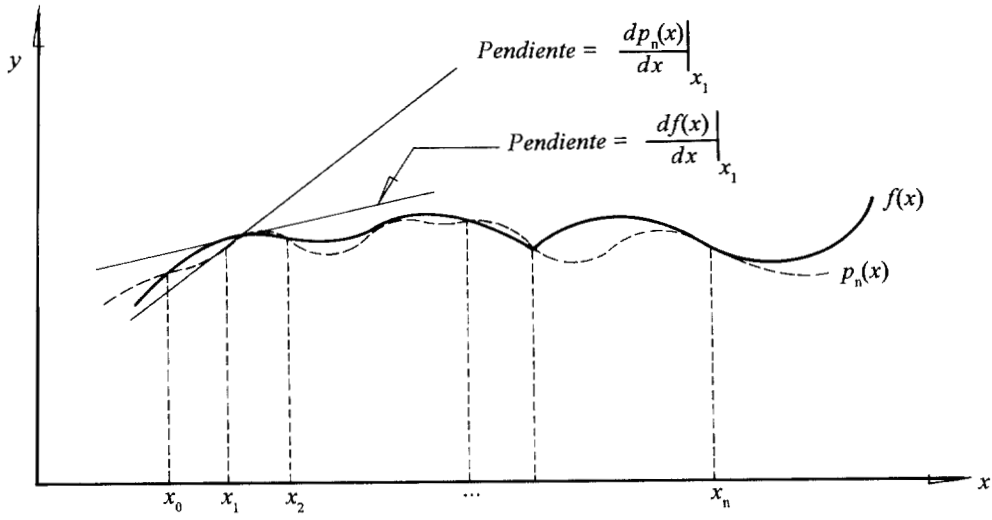


Figura 6.1. Diferenciación del polinomio de aproximación.

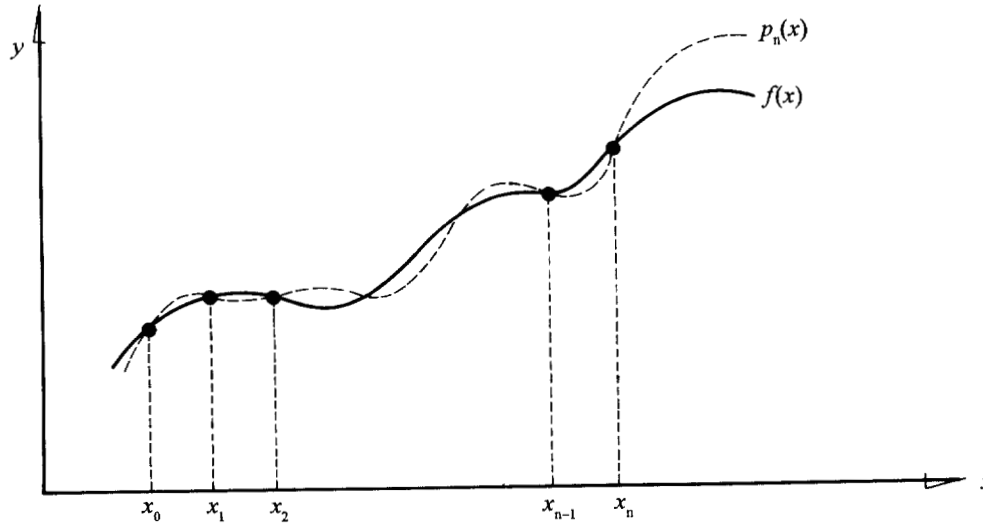


Figura 6.2. Integración del polinomio de interpolación.

puede dar una aproximación excelente de $\int_{x_0}^{x_n} f(x)\,dx$. Por otro lado, $\frac{d}{dx}\,[p_n(x)]$, que da la pendiente de la línea tangente a $p_n(x)$, puede variar en magnitud respecto a $\frac{d}{dx}\,[f(x)]$ significativamente, aunque $p_n(x)$ sea una buena aproximación a $f(x)$. Por tanto, la diferenciación numérica debe tomarse con el cuidado y reservas que lo amerita; particularmente cuando los datos obtenidos experimentalmente puedan tener errores significativos.

Los métodos de integración comúnmente usados pueden clasificarse en dos grupos: los que emplean valores dados de la función $f(x)$ en abscisas equidistantes y que se conocen como **fórmulas de Newton-Cotes**, y aquellos que utilizan valores de $f(x)$ en abscisas desigualmente espaciadas, determinadas por ciertas propiedades de familias de polinomios ortogonales, conocidas como **fórmulas de cuadratura gaussiana.**

6.1 MÉTODOS DE NEWTON-COTES

Para estimar $I = \int_a^b f(x)\,dx$, los métodos de Newton-Cotes funcionan en general en dos pasos

1. Se divide el intervalo $[a,b]$ en n intervalos de igual amplitud cuyos valores extremos son sucesivamente

$$x_i = x_0 + i\left(\frac{b-a}{n}\right),\ i = 0,1,2,\ldots,n \tag{6.1}$$

   Para quedar en la nueva notación $x_0 = a$ y $x_n = b$.

2. Se aproxima $f(x)$ por un polinomio de grado n, $p_n(x)$ y se integra para obtener la aproximación de I.

Es evidente que se obtendrán valores diferentes de I para distintos valores de n, como se muestra a continuación.

Método trapezoidal

En el caso de $n = 1$, el intervalo de integración $[a, b]$ queda tal cual y $x_0 = a$, $x_1 = b$; la aproximación polinomial de $f(x)$ es una línea recta (un polinomio de primer grado $p_1(x)$) y la aproximación a la integral es el área del trapezoide bajo esta línea recta, como se ve en la figura 6.3. Este método de integración se llama **regla trapezoidal.**

Para llevar a cabo la integración $\int_{x_0}^{x_1} p_1(x)\,dx$, es preciso seleccionar una de las formas de representación del polinomio $p_1(x)$, y como $f(x)$ está dada para valores equidistantes de x con distancia h, la elección lógica es una de las fórmulas en diferencias finitas (hacia delante, hacia atrás o centrales)*. Si se eligen las diferencias finitas hacia delante, se tendrá entonces que

*Consúltese el capítulo 5.

$$f(x) \approx p_1(x)$$

donde $p_1(x)$ es, según la ecuación 5.35

$$p_1(x) = p_1(x_0 + \text{sh}) = f(x_0) + \text{s}\ \Delta f(x_0)$$

Se remplaza $p_1(x)$ en la integral y se tiene

$$\int_a^b f(x)\,dx \approx \int_{x_0}^{x_1} [f(x_0) + s\Delta f(x_0)]\,dx \tag{6.2}$$

Para realizar la integración del lado derecho de la ecuación 6.2 es necesario tener a toda la integral en términos de la nueva variable s que, como se sabe, está dada por la expresión

$$x = x_0 + sh,$$

De ésta, la diferencial de x queda en términos de s

$$dx = h\,ds,$$

ya que x_0 y h son constantes.

Para que los límites de integración x_0 y x_1 queden a su vez en términos de s, se sustituyen por x en $x = x_0 + sh$ y se despeja s, lo que da respectivamente

$$x_0 = x_0 + sh \text{ de donde } s = 0$$

$$x_1 = x_0 + sh \text{ de donde } s = 1,$$

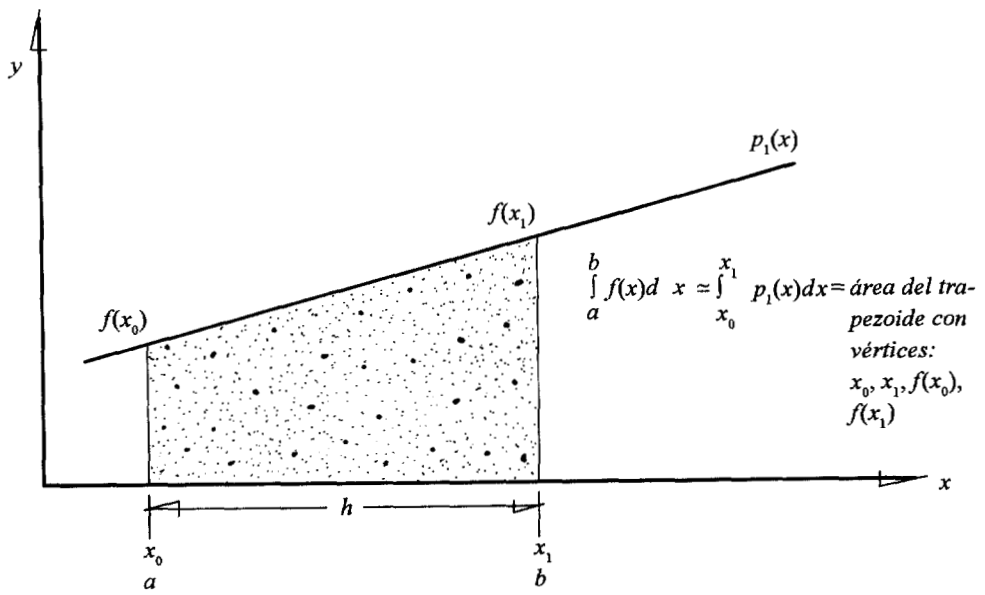


Figura 6.3. Integración numérica por medio de la regla trapezoidal.

y resulta

$$\int_{x_0}^{x_1} [f(x_0) + s\,\Delta f(x_0)]\,dx = \int_0^1 h\,[f(x_0) + s\,\Delta f(x_0)]\,ds$$

Al integrar se tiene

$$h\int_0^1 [f(x_0) + s\,\Delta f(x_0)]\,ds = h\left[sf(x_0) + \frac{s^2}{2}\Delta f(x_0)\right]\Big|_0^1 = h\left[f(x_0) + \frac{\Delta f(x_0)}{2}\right]$$

como $\Delta f(x_0) = f(x_0+h) - f(x_0)$, se llega finalmente a

$$\boxed{\int_a^b f(x)\,dx \approx \frac{h}{2}\,[f(x_0) + f(x_1)]} \qquad (6.3)$$

el algoritmo del método trapezoidal.

Nótese que el lado derecho de la ecuación 6.3 es el área de un trapezoide de altura h y lados paralelos de longitud $f(x_0)$ y $f(x_1)$ (véase Fig. 6.3).

Antes de empezar a resolver ejercicios, es conveniente observar que los métodos vistos y los siguientes sirven también cuando la función $f(x)$ está dada analíticamente y las técnicas estudiadas en el cálculo integral no dan resultado, o bien cuando esta función es imposible de integrar analíticamente. En esos casos, la tabla de puntos se elabora evaluando la función del integrando en abscisas seleccionadas adecuadamente.

Ejemplo 6.1

Uso del algoritmo trapezoidal

a) Aproxime el área A_1 bajo la curva de la función dada por la tabla siguiente, en el intervalo $a = 500$, $b = 1800$.

Puntos	0	1	2	3	4	5
$f(x)$	9.0	13.4	18.7	23.0	25.1	27.2
x	500	900	1400	1800	2000	2200

b) Aproxime $A_2 = \int_0^5 (2 + 3x)\,dx$

c) Aproxime $A_3 = \int_{-2}^4 (1 + 2x + 3x^2)\,dx$

d) Aproxime $A_4 = \int_0^{\pi/2} \operatorname{sen} x\,dx$

SOLUCIÓN

Con la ecuación 6.3 se tiene

$a)\ h = 1800 - 500, \qquad x_0 = 500, x_1 = 1800$

$$A_1 \approx \frac{1300}{2}(9 + 23) = 20800$$

$b)\ h = 5 - 0, x_0 = 0, x_1 = 5$

$$A_2 \approx \frac{5}{2}([2 + 3(0)] + [2 + 3(5)]) = 47.5$$

$c)\ h = 4 - (-2), x_0 = -2, x_1 = 4$

$$A_3 \approx \frac{6}{2}([1 + 2(-2) + 3(-2)^2] + [1 + 2(4) + 3(4)^2]) = 198$$

$d)\ h = \pi/2 - 0, x_0 = 0, x_1 = \pi/2$

$$A_4 \approx \frac{\pi/2}{2}(\text{sen } 0 + \text{sen } \pi/2) = \pi/4.$$

Se deja al lector la comparación y discusión de los resultados obtenidos analíticamente [incisos (*b*), (*c*) y (*d*)].

Método de Simpson

Si $n = 2$; esto es, el intervalo de integración $[a, b]$ se divide en dos subintervalos, se tendrán tres abscisas dadas por la ecuación 6.1 como

$$\begin{aligned} x_0 &= a \\ x_1 &= x_0 + 1\,\frac{(b-a)}{2} = a + \frac{b}{2} - \frac{a}{2} = \frac{1}{2}(b-a), \\ x_2 &= b \end{aligned}$$

Se aproxima $f(x)$ con una parábola [un polinomio de segundo grado $p_2(x)$], y la aproximación a la integral será el área bajo el segmento de parábola comprendida entre $f(x_0)$ y $f(x_2)$ como muestra la figura 6.4. Esto es

$$\int_a^b f(x)\,dx \approx \int_{x_0}^{x_2} p_2(x)\,dx$$

para realizar la integración $\int_{x_0}^{x_2} p_2(x)\,dx$, se usa la fórmula de Newton en diferencias finitas hacia delante para expresar $p_2(x)$ (Ec. 5.35)

$$p_2(x) = p_2(x_0 + sh) = f(x_0) + s\,\Delta f(x_0) + \frac{s(s-1)}{2!}\,\Delta^2 f(x_0)$$

al sustituir $p_2(x)$ y expresar toda la integral en términos de la nueva variable s, queda

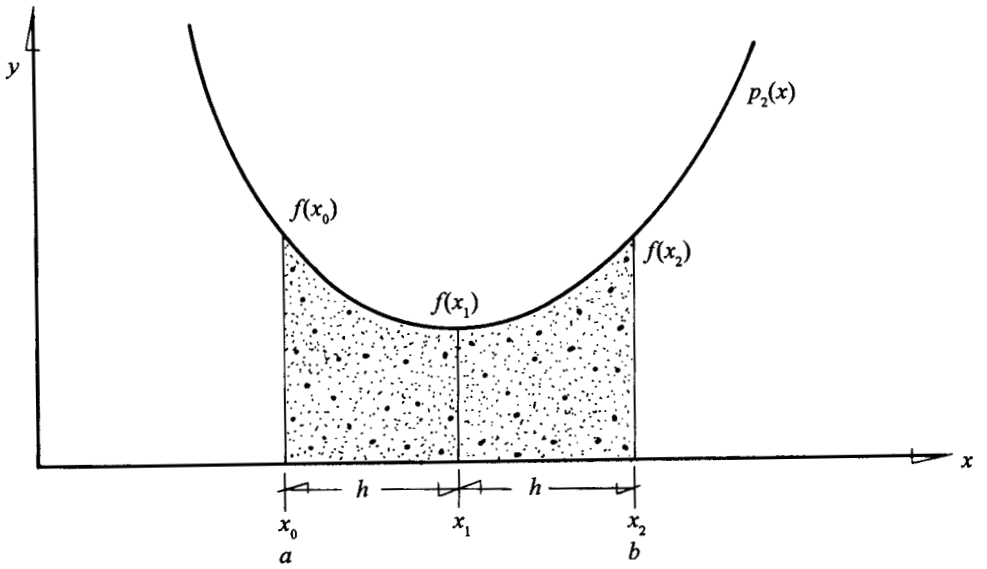


Figura 6.4. Integración numérica mediante la regla de Simpson.

$$\int_a^b f(x)\,dx \approx \int_{x_0}^{x_2} p_2(x)\,dx = h\int_0^2 p_2(x_0 + sh)\,ds$$

$$\int_0^2 p_2(x_0 + sh)\,ds = h\int_0^2 \left[f(x_0) + s\,\Delta f(x_0) + \frac{s(s-1)}{2!}\,\Delta^2 f(x_0)\right] ds$$

$$= h\left[sf(x_0) + \frac{s^2}{2}\,\Delta f(x_0) + \frac{s^3}{3!}\Delta^2 f(x_0) - \frac{s^2}{4}\,\Delta^2 f(x_0)\right]\Big|_0^2$$

$$= h\left[2f(x_0) + 2\Delta f(x_0) + \frac{1}{3}\Delta^2 f(x_0)\right]$$

De la definición de la primera y segunda diferencia hacia adelante se tiene

$$\Delta f(x_0) = f(x_0 + h) - f(x_0) = f(x_1) - f(x_0)$$

y

$$\Delta^2 f(x_0) = f(x_0 + 2h) - 2f(x_0 + h) + f(x_0) = f(x_2) - 2f(x_1) + f(x_0)$$

que sustituidas en la última ecuación dan lugar a

$$\int_a^b f(x)\,d(x) \approx \frac{h}{3}\left[f(x_0) + 4f(x_1) + f(x_2)\right] \qquad (6.4)$$

el algoritmo de Simpson.

Ejemplo 6.2

Con el algoritmo de Simpson aproxime las integrales del ejemplo 6.1.

SOLUCIÓN

Con la ecuación 6.4 se tiene

$$a)\ h = \frac{1800-500}{2} = 650, x_0 = 500, x_1 = x_0 + h = 500 + 650 = 1150, x_2 = 1800$$

$$f(x_0) = 9, f(x_1) = 16.08,\ f(x_2) = 23$$

[$f(x_1)$ se obtiene interpolando con un polinomio de segundo grado en diferencias divididas]

$$A_1 \approx \frac{650}{3}\,[\,9 + 4\,(\,16.08\,) + 23\,] = 20869.33$$

$$b)\ h = \frac{5-0}{2} = 2.5, x_0 = 0, x_1 = 0 + 2.5 = 2.5, x_2 = 5$$

$$A_2 \approx \frac{2.5}{3}\,[\,2 + 3\,(\,0\,) + 4(2 + 3\,(\,2.5\,)\,) + 2 + 3\,(\,5\,)\,] = 47.5$$

$$c)\ h = \frac{4-(-2)}{2} = 3, x_0 = -2, x_1 = -2 + 3 = 1, x_2 = 4$$

$$A_3 = \frac{3}{3}[\,1 + 2\,(-2) + 3\,(-2)^2 + 4\,(\,1 + 2\,(1) + 3\,(1)^2\,) + 1 + 2\,(4) + 3\,(4)^2\,] = 90$$

$$d)\ h = \frac{\frac{\pi}{2} - 0}{2} = \pi/4, x_0 = 0, x_1 = 0 + \pi/4, x_2 = \pi/2$$

$$A_4 \approx \frac{\pi/4}{3}\,(\ \text{sen}\ 0 + 4\ \text{sen}\ \pi/4 + \text{sen}\ \pi/2\ ) = 1.0023$$

Se deja al lector la comparación y discusión de los resultados obtenidos analíticamente [casos de los incisos (*b*), (*c*) y (*d*)] y con los obtenidos en el ejemplo 6.1.

Caso general

A continuación se verá el caso más general, donde el intervalo de integración $[a,b]$ se divide en n subintervalos y da lugar a $n+1$ abscisas equidistantes $x_0, x_1, \ldots, x_n$, con $x_0 = a$ y $x_n = b$ (véase Fig. 6.2). Esta vez el polinomio de interpolación es de n-ésimo grado $p_n(x)$ y se utilizará la representación 5.35 para éste.

La aproximación a la integral $\int_a^b f(x)\,dx$ está dada por

$$\int_a^b f(x)\,dx \approx \int_{x_0}^{x_n} p_n(x)\,dx = h\int_0^n p_n(x_0 + sh)\,ds$$

$$= h\int_0^n \left[f(x_0) + s\,\Delta f(x_0) + \frac{s(s-1)}{2!}\Delta^2 f(x_0) + \frac{s(s-1)(s-2)}{3!}\Delta^3 f(x_0)\right.$$

$$\left. + \ldots + \frac{s(s-1)(s-2)\ldots(s-(n-1))}{n!}\Delta^n f(x_0)\right] ds$$

Con la integración de los cinco primeros términos se tiene

$$h\int_0^n p_n(x_0 + sh)\,ds = h\left[sf(x_0) + \frac{s^2}{2}\Delta f(x_0) + \left(\frac{s^3}{6} - \frac{s^2}{4}\right)\Delta^2 f(x_0)\right.$$

$$+ \left(\frac{s^4}{24} - \frac{s^3}{6} + \frac{s^2}{6}\right)\Delta^3 f(x_0)$$

$$\left. + \left(\frac{s^5}{120} - \frac{s^4}{16} + \frac{11s^3}{72} - \frac{s^2}{8}\right)\Delta^4 f(x_0) + \text{términos faltantes}\right]\Bigg|_0^n$$

Todos los términos son cero en el límite inferior, por lo que

$$\int_a^b f(x)\,dx \approx h\left[nf(x_0) + \frac{n^2}{2}\Delta f(x_0) + \left(\frac{n^3}{6} - \frac{n^2}{4}\right)\Delta^2 f(x_0)\right.$$

$$+ \left(\frac{n^4}{24} - \frac{n^3}{6} + \frac{n^2}{6}\right)\Delta^3 f(x_0) \qquad (6.5)$$

$$\left. + \left(\frac{n^5}{120} - \frac{n^4}{16} + \frac{11n^3}{72} - \frac{n^2}{8}\right)\Delta^4 f(x_0) + \text{términos faltantes}\right]$$

A continuación se dan las fórmulas de Newton-Cotes para integrar cuando $n = 1, 2, 3, 4, 5$ y 6. El lector puede verificarlas sustituyendo el valor seleccionado de n y las diferencias correspondientes en términos de sus valores funcionales en la ecuación 6.5.

$n = 1$

$$\int_{x_0}^{x_1} f(x)\,dx \approx \frac{h}{2}\left[f(x_0) + f(x_1)\right] \qquad \text{trapezoidal}$$

$$
\begin{aligned}
&n = 2 \\
&\int_{x_0}^{x_2} f(x)\,dx \approx \frac{h}{3}\,[f(x_0) + 4f(x_1) + f(x_2)] \qquad \text{Simpson 1/3} \\
&n = 3 \\
&\int_{x_0}^{x_3} f(x)\,dx \approx \frac{3h}{8}\,[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)] \qquad \text{Simpson 3/8} \\
&n = 4 \\
&\int_{x_0}^{x_4} f(x)\,dx \approx \frac{2h}{45}\,[7f(x_0) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 7f(x_4)] \\
&n = 5 \\
&\int_{x_0}^{x_5} f(x)\,dx \approx \frac{5h}{288}\,[19f(x_0) + 75f(x_1) + 50f(x_2) + 50f(x_3) + 75f(x_4) + 19f(x_5)] \\
&n = 6 \\
&\int_{x_0}^{x_6} f(x)\,dx \approx \frac{h}{140}\,[41f(x_0) + 216f(x_1) + 27f(x_2) + 272f(x_3) + 27f(x_4) \\
&\qquad\qquad + 216f(x_5) + 41f(x_6)]
\end{aligned}
\tag{6.6}
$$

Métodos compuestos de integración

Algunas veces el intervalo de integración es tan amplio, que resulta conveniente dividirlo en subintervalos y aproximar cada uno por medio de un polinomio.

Método trapezoidal compuesto

Por ejemplo, en vez de aproximar la integral de $f(x)$ en $[a, b]$ por una recta (véase Fig. 6.5.*a*), conviene dividir $[a,b]$ en n subintervalos y aproximar cada uno por un polinomio de primer grado (véase Fig. 6.5 *b*). Una vez hecho esto, se aplica la fórmula trapezoidal a cada subintervalo y se obtiene el área de cada trapezoide, de tal modo que la suma de todas ellas da la aproximación al área bajo la curva $f(x)$. Esto es

$$
\mathrm{I} = \int_a^b f(x)\,dx \approx \int_{a=x_0}^{x_1} p_1(x)\,dx + \int_{x_1}^{x_2} p_2(x)\,dx + \ldots + \int_{x_{n-1}}^{x_n=b} p_n(x)\,dx
$$

donde $p_i(x)$ es la ecuación de la recta que pasa por los puntos $(x_{i-1}, f(x_{i-1}))$, $(x_i, f(x_i))$. Con la ecuación 6.3 se tiene

$$
\begin{aligned}
\mathrm{I} \approx \frac{x_1 - x_0}{2}\,[f(x_0) + f(x_1)] + \frac{x_2 - x_1}{2}\,[f(x_1) + f(x_2)] + \ldots \\
+ \frac{x_n - x_{n-1}}{2}\,[f(x_{n-1}) + f(x_n)]
\end{aligned}
\tag{6.7}
$$

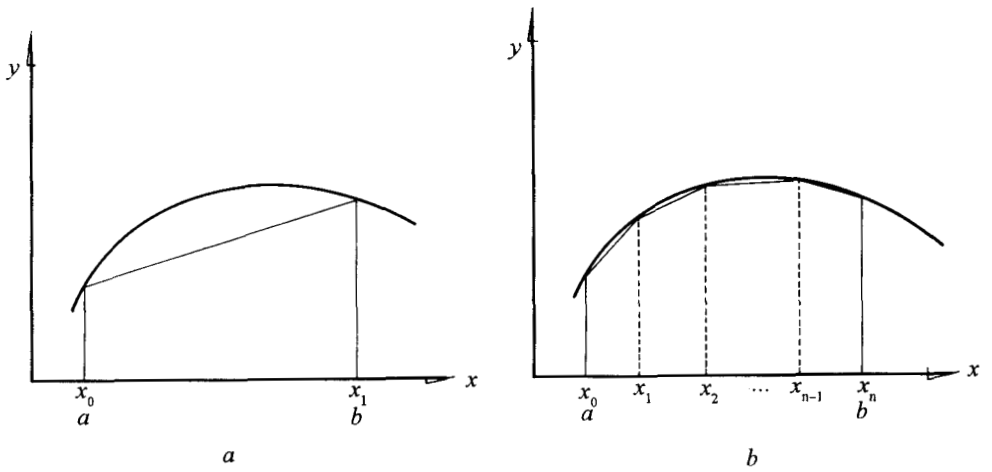


Figura 6.5. Integración por el método trapezoidal compuesto.

Si todos los subintervalos son del mismo tamaño h, esto es, si $x_{i+1} - x_i = h$, para i=0, 1, ..., (n-1), entonces la ecuación 6.7 puede anotarse

$$I \approx \frac{h}{2}[f(x_0) + 2f(x_1) + 2f(x_2) + \ldots + 2f(x_{n-1}) + f(x_n)],$$

que puede escribirse con la notación de sumatoria

$$\boxed{I \approx \frac{h}{2}\left[f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n)\right]} \tag{6.8}$$

Ejemplo 6.3

Mediante el algoritmo trapezoidal compuesto, aproxime el área bajo la curva de la siguiente función dada en forma tabular, entre $x = -1$ y $x = 4$.

Puntos	0	1	2	3	4	5
x	-1	0	1	2	3	4
$f(x)$	8	10	10	20	76	238

SOLUCIÓN

Si se toman todos los puntos de la tabla, se puede aplicar cinco veces el método trapezoidal. Como todos los intervalos son del mismo tamaño ($h=1$), se usa la ecuación 6.8 directamente

$$A \approx \frac{1}{2}[8 + 2(10 + 10 + 20 + 76) + 238] = 239$$

Compárese este resultado con la solución analítica (los datos de la tabla corresponden a la función $f(x) = x^4 - 2x^2 + x + 10$).

ALGORITMO 6.1 Método trapezoidal compuesto

Para aproximar el área bajo la curva de una función analítica $f(x)$ en el intervalo $[a, b]$, proporcionar la función por integrar F(x) y los

DATOS: El número de trapecios N, el límite inferior A y límite superior B.
RESULTADOS: El área aproximada AREA.

```
PASO 1.  Hacer X = A
PASO 2.  Hacer S = 0
PASO 3.  Hacer H = (B - A)/N
PASO 4.  SI N = 1, ir al paso 10. De otro modo continuar.
PASO 5.  Hacer I = 1
PASO 6.  Mientras I ≤ N-1, repetir los pasos 7 a 9.
           PASO 7.  Hacer X = X + H
           PASO 8.  Hacer S =S + F(X)
           PASO 9.  Hacer I = I + 1
PASO 10. Hacer AREA = H/2 * (F(A) + 2*S + F(B))
PASO 11. IMPRIMIR AREA y TERMINAR.
```

Método de Simpson compuesto

Como para cada aplicación de la regla de Simpson se requieren dos subintervalos, a fin de aplicarla n número de veces, deberá dividirse el intervalo $[a, b]$ en un número de subintervalos igual a $2n$ (véase Fig. 6.6).

Cada par de subintervalos sucesivos se aproxima por un polinomio de segundo grado (parábola) y se integra usando la ecuación 6.4, de tal manera que la suma de las áreas bajo cada segmento de parábola sea la aproximación a la integración deseada. Esto es

$$I = \int_a^b f(x)\,dx \approx \int_{a=x_0}^{x_2} p_1(x)\,dx + \int_{x_2}^{x_4} p_2(x)\,dx + \ldots + \int_{x_{n-2}}^{x_n=b} p_n(x)\,dx$$

donde $p_i(x)$, $i = 1,2,\ldots, n$, es el polinomio de segundo grado que pasa por tres puntos consecutivos.

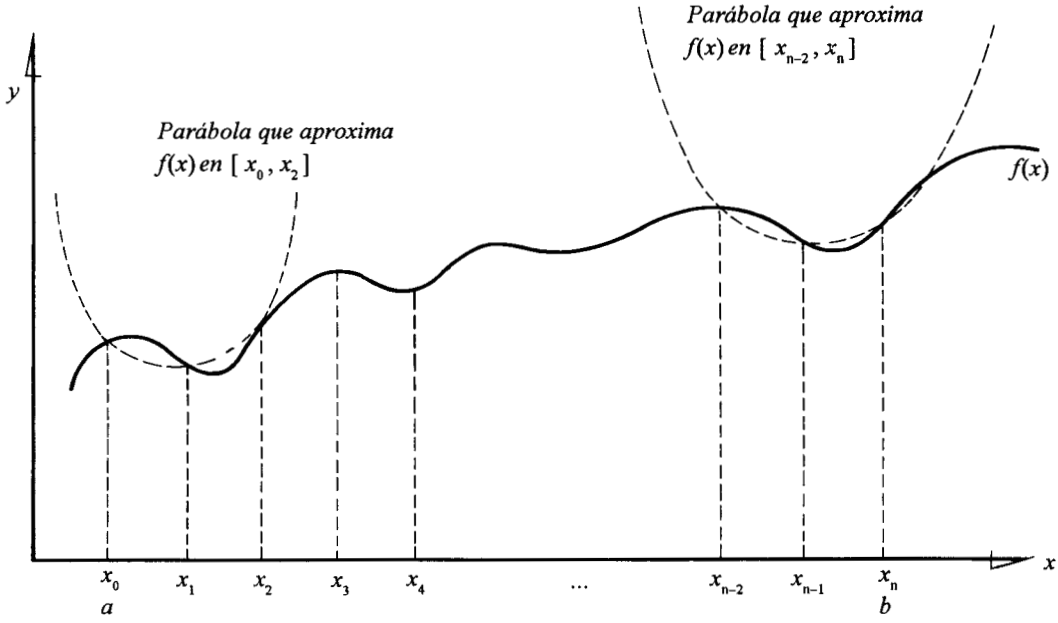


Figura 6.6. Integración por el método de Simpson compuesto.

Al sustituir la ecuación 6.4 en cada uno de los sumandos se tiene

$$I \approx \frac{h_1}{3}[f(x_0) + 4f(x_1) + f(x_2)] + \frac{h_2}{3}[f(x_2) + 4f(x_3) + f(x_4)] + \ldots + \frac{h_n}{3}[f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)] \qquad (6.09)$$

donde

$$h_1 = x_1 - x_0 = x_2 - x_1$$
$$h_2 = x_3 - x_2 = x_4 - x_3$$
$$\vdots$$
$$h_n = x_{n-1} - x_{n-2} = x_n - x_{n-1}$$

Si $h_1 = h_2 = \ldots = h_n$, la ecuación (6.9) queda como sigue

$$I \approx \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)] + \frac{h}{3}[f(x_2) + 4f(x_3) + f(x_4)] + \ldots + \frac{h}{3}[f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)]$$

que usando la notación de sumatoria queda de la siguiente manera

$$I \approx \frac{h}{3}\left[f(x_0) + 4\sum_{\substack{i=1 \\ \Delta i=2}}^{n-1} f(x_i) + 2\sum_{\substack{i=2 \\ \Delta i=2}}^{n-2} f(x_i) + f(x_n)\right] \tag{6.10}$$

donde Δi significa el incremento de i.

Ejemplo 6.4

Mediante el algoritmo de Simpson de integración, aproxime el área bajo la curva del ejemplo 6.3.

SOLUCIÓN

Con los puntos dados de la tabla, se puede aplicar la regla de Simpson en dos ocasiones; por ejemplo, una vez con los puntos (0), (1) y (2) y otra con los puntos (2), (3) y (4). Como la integración debe hacerse de $x = -1$ a $x = 4$, se integra entre los puntos (4) y (5) con el método trapezoidal y la suma será la aproximación buscada:

a) Método de Simpson aplicado dos veces: $h_1 = h_2 = h_3 = h_4 = 1$, entonces puede usarse la ecuación 6.10

$$A_1 \approx \frac{1}{3}[8 + 4(10 + 20) + 2(10) + 76] = 74.666$$

b) Método trapezoidal aplicado a los puntos (4) y (5)

$$A_2 \approx \frac{1}{2}(76 + 238) = 157$$

por lo tanto, la aproximación al área es

$$A \approx 74.666 + 157 = 231.666$$

Compare este resultado con el obtenido en el ejemplo 6.3 y el resultado de la solución analítica (la función tabulada es $f(x) = x^4 - 2x^2 + x + 10$).

Ejemplo 6.5

Encuentre la integral aproximada de la función

$$\frac{1}{\sqrt{2\pi}}\, e^{-x^2/2},$$

que da lugar a la curva normal tipificada, entre los límites –1 y 1.

a) Utilice la regla trapezoidal con varios trapezoides y compare con el resultado (0.682) obtenido de tablas.
b) Use la regla de Simpson varias veces y compare con el resultado (0.682) obtenido de tablas.

SOLUCIÓN

a) Con $n = 1$, $h = \dfrac{1-(-1)}{1} = 2$

$$I \approx \frac{2}{2\sqrt{2\pi}} [f(x_0) + f(x_1)] = \frac{1}{\sqrt{2\pi}} [0.606 + 0.606] = 0.484$$

El error relativo, tomando el valor de tablas como verdadero, es

$$E_{n=1} = \frac{|0.484 - 0.682|}{0.682} = 0.29 \quad \text{o} \quad 29\%$$

Con $n = 2$, $h = \dfrac{1-(-1)}{2} = 1$

$$I \approx \frac{1}{2\sqrt{2\pi}} [f(x_0) + 2f(x_1) + f(x_2)] = \frac{1}{2\sqrt{2\pi}} [0.606 + 2(1) + 0.606] = 0.64$$

$$E_{n=2} = \frac{|0.64 - 0.682|}{0.682} = 0.0587 \quad \text{o} \quad 5.87\%$$

Con $n = 4$, $h = \dfrac{1-(-1)}{4} = 0.5$

$$I \approx \frac{0.5}{2\sqrt{2\pi}} [f(x_0) + 2f(x_1) + 2f(x_2) + 2f(x_3) + f(x_4)]$$

$$= \frac{0.5}{2\sqrt{2\pi}} [0.606 + 2(0.882) + 2(1) + 2(0.882) + 0.606] = 0.672$$

$$E_{n=4} = \frac{|0.672 - 0.682|}{0.682} = 0.0147 \quad \text{o} \quad 1.47\%$$

b) Con $n = 2$, $h = \dfrac{1-(-1)}{2} = 1$

$$I \approx \frac{1}{3\sqrt{2\pi}} [f(x_0) + 4f(x_1) + f(x_2)] = \frac{1}{3\sqrt{2\pi}} [0.606 + 4(1) + 0.606] = 0.693$$

$$E_{n=2} = \frac{|0.693 - 0.682|}{0.682} = 0.0162 \quad \text{o} \quad 1.62\%$$

$$\text{Con } n = 4, h = \frac{1-(-1)}{4} = 0.5$$

$$I \approx \frac{0.5}{3\sqrt{2\pi}} [f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + f(x_4)]$$

$$= \frac{0.5}{3\sqrt{2\pi}} [0.606 + 4(0.882) + 2(1) + 4(0.882) + 0.606] = 0.683$$

$$E_{n=4} = \frac{|0.683 - 0.682|}{0.682} = 0.0015 \quad \text{o} \quad 0.15\%$$

ALGORITMO 6.2 Método de Simpson compuesto

Para aproximar el área bajo la curva de una función analítica $f(x)$ en el intervalo $[a,b]$, proporcionar la función por integrar F(X) y los

DATOS: El número (par) de subintervalos N, el límite inferior A y el límite superior B.
RESULTADOS: El área aproximada ÁREA.

PASO 1. Hacer S1 = 0
PASO 2. Hacer S2 = 0
PASO 3. Hacer X = A
PASO 4. Hacer H = (B–A)/N
PASO 5. Si N = 2, ir al paso 13. De otro modo continuar.
PASO 6. Hacer I = 1
PASO 7. Mientras I ≤ N/2-1, repetir los pasos 8 a 12.
  PASO 8. Hacer X = X + H
  PASO 9. Hacer S1 = S1 + F(X)
  PASO 10. Hacer X = X+H
  PASO 11. Hacer S2 = S2 + F(X)
  PASO 12. Hacer I = I + 1
PASO 13. Hacer X = X + H
PASO 14. Hacer S1 = S1 + F (X)
PASO 15. Hacer ÁREA = H/3 * (F(A) + 4*S1 + 2*S2 + F(B))
PASO 16. IMPRIMIR ÁREA y TERMINAR.

Ejemplo 6.6

Elabore un subprograma para integrar la función del ejemplo 6.5 con el método trapezoidal compuesto, usando sucesivamente 1, 2, 4, 8, 16, 32, 64, ..., 1024 subintervalos y calcule sus correspondientes errores relativos en porciento. Con los resultados obtenidos elabore una gráfica, error porcentual contra número de subintervalos y discútala.

SOLUCIÓN

El programa 6.1 que se encuentra en el disco fue diseñado para usar el subprograma TRAPECIOS en la integración de

$$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

en el intervalo [–1,1], usando N=1,2,4,...,1024 subintervalos sucesivamente.
Los resultados obtenidos para cada valor de N son

APROXIMACION DEL AREA BAJO LA FUNCION

F (X) = 1 / SQR (2*3.14159) * EXP (–X^2/2)

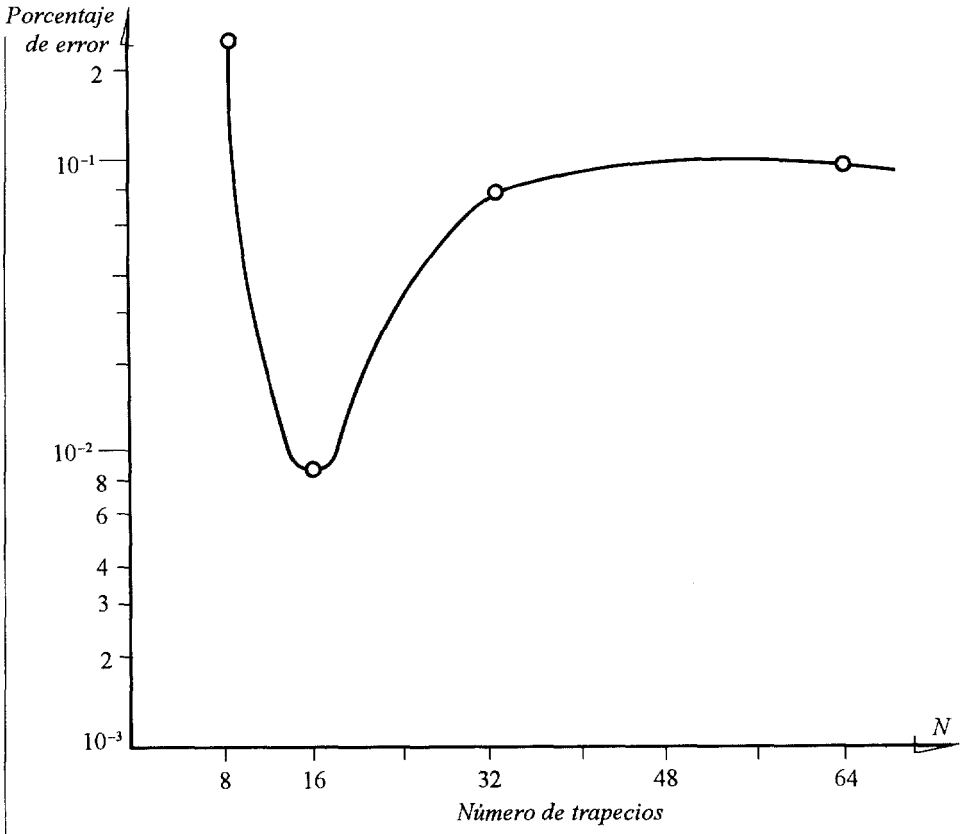
DESDE X = N	–1.00 HASTA X = APROXIMACIÓN AL AREA	1.00 ERROR EN 1%
1	0.484	29.041
2	0.641	6.024
4	0.673	1.390
8	0.680	0.269
16	0.682	0.009
32	0.683	0.078
64	0.683	0.095
128	0.683	0.100
256	0.683	0.101
512	0.683	0.101
1024	0.683	0.101

EL VALOR VERDADERO DE LA INTEGRAL ES: 0.682

El error relativo porcentual se calculó utilizando como valor verdadero el encontrado en tablas, que es 0.682. Se traza la gráfica y se hacen los comentarios correspondientes.

Comentarios

El eje de las ordenadas está en escala logarítmica.
El error obtenido por el programa es básicamente la suma de dos tipos de errores, el de truncamiento (debido a la aproximación de la función en cada subintervalo por una línea recta) y el de redondeo (por el tamaño de la palabra de memoria de la computadora en que se realizan los cálculos).



El error de truncamiento disminuye al aumentar el número de subintervalos y teóricamente tiende a cero cuando N tiende a infinito. Por otro lado, el error de redondeo crece al aumentar el número de subintervalos (debido al aumento del número de cálculos). En la gráfica se ve que el error global disminuye al incrementar el número de subintervalos hasta llegar a un mínimo para N = 16, para después aumentar debido a que el peso del error de redondeo empieza a dominar.

En general, cuando se recurre a una integración numérica, no se tiene el resultado verdadero y resulta conveniente integrar con un número n de subintervalos y luego con el doble. Si los resultados no difieren considerablemente, puede aceptarse como bueno cualquiera de los dos.

Ejemplo 6.7

Elabore un subprograma para integrar una función analítica por el método de Simpson compuesto, usando sucesivamente 2, 4, 8, 16, ..., 2048 subintervalos. Compruébela con la función del ejemplo 6.5.

SOLUCIÓN

Ver programa 6.2 en el disco.

Análisis del error de truncamiento en la aproximación trapezoidal

Considérese el i-ésimo trapezoide de una integración trapezoidal compuesta o sucesiva, con abscisas x_{i-1} y x_i. La distancia entre estas abscisas es $h = (b - a)/n$. Sea además $F(x)$ la primitiva del integrando $f(x)$, es decir $dF(x)/dx = f(x)$. Con esto la integral de la función $f(x)$ en el intervalo $[x_{i-1}, x_i]$ queda dada por

$$I_i = \int_{x_{i-1}}^{x_i} f(x)\,dx = F(x_i) - F(x_{i-1}) \tag{6.11}$$

Por otro lado, la aproximación de I_i usando el método trapezoidal es

$$T_i = \frac{h}{2}\,[f(x_{i-1}) + f(x_i)] \tag{6.12}$$

En ausencia de errores de redondeo puede definirse el error de truncamiento para este trapezoide particular así

$$E_i = T_i - I_i \tag{6.13}$$

Para continuar con este análisis, $f(x)$ se expande en serie de Taylor alrededor de $x = x_i$, de modo de obtener $f(x_{i-1})$

$$f(x_{i-1}) = f(x_i) + (x_{i-1} - x_i)\,f'(x_i) + \frac{(x_{i-1} - x_i)^2}{2!}\,f''(x_i) + \ldots$$

como $h = x_i - x_{i-1}$

$$f(x_{i-1}) = f(x_i) - hf'(x_i) + \frac{h^2}{2!}\,f''(x_i) + \ldots \tag{6.14}$$

se sustituye la ecuación 6.14 en la 6.12

$$T_i = \frac{h}{2}\,[2f(x_i) - h\,f'(x_i) + \frac{h^2}{(2!)}\,f''(x_i) + \ldots]$$

$$T_i = hf(x_i) - \frac{h^2}{2}\,f'(x_i) + \frac{h^3}{2(2!)}\,f''(x_i) + \ldots \tag{6.15}$$

En forma análoga puede obtenerse

$$F(x_{i-1}) = F(x_i) - hF'(x_i) + \frac{h^2}{2!}\,F''(x_i) - \frac{h^3}{3!}\,F'''(x_i) + \ldots$$

cuya sustitución en la ecuación 6.11 produce

$$I_i = hF'(x_i) - \frac{h^2}{2!}\,F''(x_i) + \frac{h^3}{3!}\,F'''(x_i) - \ldots$$

como

$$\begin{aligned} f(x) &= F'(x) \\ f'(x) &= F''(x) \\ f''(x) &= F'''(x) \\ &\vdots \end{aligned}$$

y al sustituir se obtiene:

$$I_i = hf(x_i) - \frac{h^2}{2!} f'(x_i) + \frac{h^3}{3!} f''(x_i) - \ldots \tag{6.16}$$

Al remplazar las ecuaciones 6.15 y 6.16 en la (6.13)

$$E_i = \left[hf(x_i) - \frac{h^2}{2!} f'(x_i) + \frac{h^3}{2(2!)} f''(x_i) + \ldots\right]$$

$$- \left[hf(x_i) - \frac{h^2}{2!} f'(x_i) + \frac{h^3}{3!} f''(x_i) - \ldots\right]$$

$$E_i = \left(\frac{1}{4} - \frac{1}{6}\right) h^3 f''(x_i) + \text{términos en } h^4, h^5, \text{ etcétera.}$$

Considerando que h es pequeña ($h << 1$), los términos en h^4, h^5, etc., pueden despreciarse, de modo que el error de truncamiento del i-ésimo trapezoide queda dado aproximadamente así

$$E_i \approx \frac{h^3}{12} f''(x_i) \tag{6.17}$$

Si además $|f''(x)| \leq M$ para $a \leq x \leq b$, entonces

$$| E_i | \leq \frac{h^3}{12} M,$$

de donde el error de truncamiento usando n trapezoides en la integración de $f(x)$ en $[a,b]$ queda dado por

$$| E_T | \leq \frac{nh^3}{12} M = nh \frac{h^2}{12} M = (b - a) \frac{h^2}{12} M \tag{6.18}$$

Por tanto, el error de truncamiento en el método trapezoidal es proporcional a h^2 lo cual, para fines de análisis, suele expresarse así

$$\int_a^b f(x)\, dx = \frac{h}{2} \left[f(x_0) + 2 \sum_{i=1}^{n-1} f(x_i) + f(x_n)\right] + O(h^2) \tag{6.19}$$

y se dice que es una fórmula que genera aproximaciones del orden $O(h^2)$ (véase Ec. 6.8).

Extrapolación de Richardson. Integración de Romberg

Con el nombre de **extrapolación de Richardson** se conoce un conjunto de técnicas que generan mejores aproximaciones a los resultados buscados o aproximaciones equivalentes a métodos de alto orden, a partir de las aproximaciones obtenidas con algún método de bajo orden y pocos cálculos. Estas técnicas están basadas en el análisis del error de truncamiento, cuya aplicación a la integración numérica se presenta a continuación.

Supóngase que el error de truncamiento de cierto algoritmo de aproximación de

$$I = \int_a^b f(x)\, dx$$

se expresa

$$E = c h^r f^{(r)}\ (\xi)$$

donde c es independiente de h, r es un entero positivo y ξ un punto desconocido de (a,b). Luego de obtener dos aproximaciones de I emplear h_1 y h_2, llamar a dichas aproximaciones I_1 y I_2 respectivamente y despreciar errores de redondeo, se puede escribir

$$I - I_1 = c h_1^r f^{(r)}(\xi_1)$$
$$I - I_2 = c h_2^r f^{(r)}(\xi_2)$$

Estas dos últimas ecuaciones se dividen miembro a miembro y como $f^{(r)}(\xi_1)$ y $f^{(r)}(\xi_2)$ son prácticamente iguales, se tiene

$$\frac{I - I_1}{I - I_2} = \frac{c h_1^r f^{(r)}(\xi_1)}{c h_2^r f^{(r)}(\xi_2)}$$

de donde

$$\boxed{I = \frac{h_1^r I_2 - h_2^r I_1}{h_1^r - h_2^r}} \tag{6.20}$$

Si en particular $h_2 = h_1/2$, la ecuación 6.20 se simplifica a

$$I \approx \frac{2^r I_2 - I_1}{2^r - 1} \tag{6.21}$$

Este proceso, conocido como integración de Romberg, es efectivo cuando $f^{(r)}(x)$ no varía bruscamente en (a,b) y no cambia de signo en dicho intervalo. En estos casos, las ecuaciones 6.20 y 6.21 permiten obtener una mejor aproximación a I a partir de I_1 y I_2 sin repetir el proceso de integración y con cálculos breves.

En el método trapezoidal (véase Ec. 6.18); por ejemplo, $r = 2$ y la ecuación 6.21 toma la forma

$$I = \frac{2^2 I_2 - I_1}{2^2 - 1} = \frac{4 I_2 - I_1}{3}$$

Para sistematizar la integración de Romberg en la aproximación trapezoidal, denótense por $I_k^{(0)}$ las aproximaciones de I obtenidas empleando 2k trapezoides (véase tabla 6.1). Ahora, para obtener mejores aproximaciones de I mediante $I_k^{(0)}$ y $I_{k+1}^{(0)}$, se aplica la extrapolación de Richardson

$$I \approx \frac{2^2 I_{k+1}^{(0)} - I_k^{(0)}}{2^2 - 1}$$

Este resultado se denota como $I_k^{(1)}$ y se genera la cuarta columna de la tabla 6.1. Estos valores sirven para producir una segunda extrapolación y obtener una mejor aproximación de I. Con el empleo de $I_k^{(1)}$ y $I_{k+1}^{(1)}$ se llega a

$$I \approx \frac{2^4 I_{k+1}^{(1)} - I_k^{(1)}}{2^4 - 1},$$

que se denota como $I_k^{(2)}$, con lo que se genera la quinta columna de la tabla 6.1. Este proceso puede continuar en tanto cada iteración responda al algoritmo

$$I_k^{(m)} = \frac{4^m I_{k+1}^{(m-1)} - I_k^{(m-1)}}{4^m - 1}; m = 1, 2, 3, \ldots \tag{6.22}$$

Cuando los valores de $I_k^{(0)} \rightarrow I$ al crecer k, los valores de la diagonal superior de la tabla convergen* a I.

Para entender mejor esto, a continuación se resuelve y analiza un ejemplo.

k	Número de trapezoides 2^k	Aproximación trapezoidal	Primera Extrapolación	Segunda Extrapolación	...
0	1	$I_0^{(0)}$			
1	2	$I_1^{(0)}$	$I_0^{(1)}$		
2	4	$I_2^{(0)}$	$I_1^{(1)}$	$I_0^{(2)}$	
3	8	$I_3^{(0)}$	$I_2^{(1)}$	$I_1^{(2)}$	...
4	16	$I_4^{(0)}$	$I_3^{(1)}$	$I_2^{(2)}$	
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	

Tabla 6.1. Aplicación del método de Romberg.

*Ralston, A. *Introducción al análisis numérico.* Limusa-Willey, S.A. [1970] p. 149-152.

Ejemplo 6.8

Encuentre una aproximación de la integral

$$\int_0^1 \text{sen } \pi x \, dx,$$

empleando 1, 2, 4, 8 y 16 trapezoides.

Con los resultados obtenidos y la ecuación 6.22, obtenga mejores aproximaciones. Compare los valores obtenidos con el valor calculado analíticamente: 0.6366197.

SOLUCIÓN

Con el programa del ejemplo 6.6 se obtienen los valores

k	2^k	$I_k^{(0)}$
0	1	0.0
1	2	0.5
2	4	0.6035534
3	8	0.6284174
4	16	0.6345731

Nótese que $I_k^{(0)}$converge al valor analítico al aumentar k; sin embargo, emplear aún más subintervalos implica aumentar los errores de redondeo y un considerable incremento en el número de cálculos.

En cambio, si se aplica la ecuación 6.22 con $m = 1$, se obtiene sucesivamente

$$I_1^{(1)} = \frac{4^1 (0.5) - 0}{4^1 - 1} = 0.6666667$$

$$I_2^{(1)} = \frac{4^1 (0.6035534) - 0.5}{4^1 - 1} = 0.6380712$$

$$I_3^{(1)} = \frac{4^1 (0.6284174) - 0.6035534}{4^1 - 1} = 0.6367054$$

$$I_4^{(1)} = \frac{4^1 (0.6345731) - 0.6284174}{4^1 - 1} = 0.6366250$$

Obsérvese que con estos breves cálculos se obtienen mejores aproximaciones de la integral.

Al aplicar la ecuación 6.22 con $m = 2$ y los valores de arriba

$$I_1^{(2)} = \frac{4^2 (0.6380712) - 0.6666667}{4^2 - 1} = 0.6361648$$

$$I_2^{(2)} = \frac{4^2 (0.6367054) - 0.6380712}{4^2 - 1} = 0.6366143$$

$$I_3^{(2)} = \frac{4^2 (0.6366250) - 0.63667054}{4^2 - 1} = 0.6366196$$

Al continuar con $m = 3$ y $m = 4$ se obtienen la sexta y séptima columnas de la tabla de resultados

k	2^k	$I_k^{(0)}$	$I_k^{(1)}$	$I_k^{(2)}$	$I_k^{(3)}$	$I_k^{(4)}$
0	1	0.0000000				
1	2	0.5000000	0.6666667			
2	4	0.6035534	0.6380712	0.6361648		
3	8	0.6284174	0.6367054	0.6366143	0.6366214	
4	16	0.6345731	0.6366250	0.6366196	0.6366197	0.6366197

El valor $I_4^{(4)}$ es el valor analítico de la integral.

El método de Romberg puede emplearse sucesivamente hasta que dos elementos consecutivos de una fila $I_k^{(m)}$, $I_k^{(m+1)}$ coincidan hasta cierta cifra decimal; esto es,

$$| I_k^{(m)} - I_k^{(m+1)} | \leq \text{EPS}$$

Además, puede generarse otra columna y ver si

$$| I_{k+1}^{(m+2)} - I_k^{(m+2)} | \leq \text{EPS}$$

con lo que se evita la posibilidad de que dos elementos consecutivos de una fila coincidan entre sí pero no con el valor de la integral que se está aproximando.

Utilice estos criterios para resolver el ejemplo 6.8 con EPS $= 10^{-6}$.

SECCIÓN 6.2 CUADRATURA DE GAUSS

Gauss investigó y encontró que es factible disminuir el error en la integración cambiando la localización de los puntos sobre la curva de integración $f(x)$. El investigador desarrolló su propio método, conocido como **Cuadratura de Gauss**, el cual se describe a continuación.

En la figura 6.7 se tiene la curva de la función $f(x)$ que se desea integrar entre los límites a y b. La parte (a) de la figura muestra cómo se integraría usando un trapezoide: uniendo el punto A de coordenadas $(a, f(a))$ con el punto B $(b, f(b))$ mediante un segmento de recta $p_1(x)$. Esto forma un trapezoide de base $h = (b-a)$, cuya área es

$$T = \frac{h}{2}\,[f(a) + f(b)]\,,$$

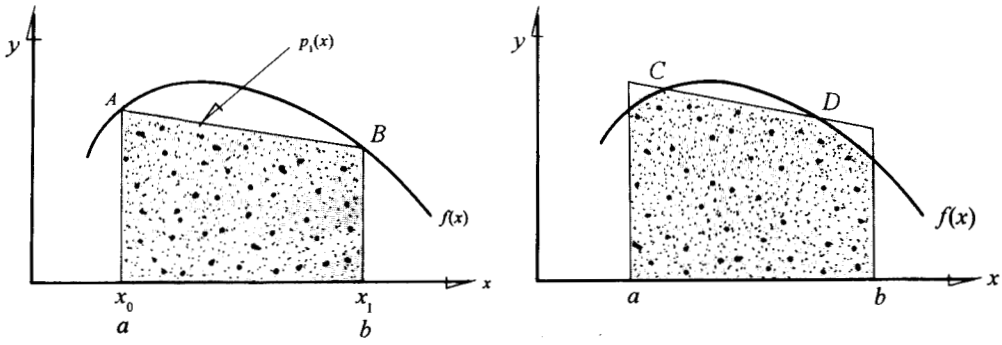
y que podría escribirse como*

$$T = w_1 f(a) + w_2 f(b), \tag{6.23}$$

donde $w_1 = w_2 = \frac{h}{2}$.

El área del trapezoide calculada T, aproxima el área bajo la curva $f(x)$.

Por otro lado, en la aplicación de la cuadratura de Gauss, en lugar de tomar los dos puntos A y B en los extremos del intervalo, se escogen dos puntos interiores C y D (véase la parte b de la Fig. 6.7).



a) Método trapezoidal b) Método de Gauss con dos puntos

Figura 6.7. Desarrollo del método de integración de Gauss usando dos puntos a partir del método trapezoidal.

*De hecho, cualquiera de las fórmulas de integración desarrolladas en las secciones anteriores puede ponerse en la forma

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{n} w_i f(x_i)$$

donde, por ejemplo, la regla de Simpson aplicada una vez tendría $w_1 = w_3 = h/3$ y $w_2 = 4h/3$ (véase Ec. 6.4).

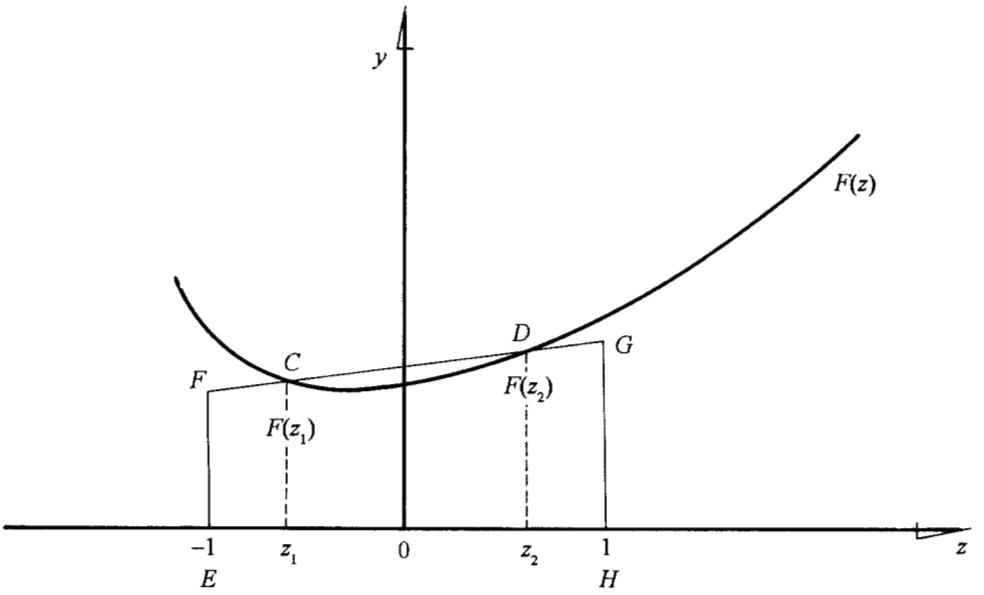


Figura 6.8. Derivación del método de integración de Gauss.

Se traza una línea recta por estos dos puntos, se extiende hasta los extremos del intervalo y se forma el trapezoide sombreado. Parte del trapezoide queda por encima de la curva y parte por abajo. Si se escogen adecuadamente los puntos C y D, cabe igualar las dos zonas de modo que el área del trapezoide sea igual al área bajo la curva; el cálculo del área del trapezoide resultante da la integral *exacta*. El método de Gauss consiste esencialmente en seleccionar los puntos C y D **adecuados**. La técnica se deduce a continuación.

Considérese primero, sin que esto implique perder generalidad, que se desea integrar la función mostrada en la figura 6.8 entre los límites −1 y +1*. Los puntos C y D se escogen sobre la curva y se forma el trapezoide con vértices E, F, G, y H.

Sean las coordenadas del punto C $(z_1,F(z_1))$ y las del punto D $(z_2,F(z_2))$. Motivado por la fórmula trapezoidal (Ec. 6.3), Gauss se propuso desarrollar una fórmula del tipo

$$A = w_1 F(z_1) + w_2 F(z_2) \tag{6.24}$$

ya que esto simplificaría relativamente el cálculo del área. El problema, planteado de esta manera, consiste en encontrar los valores de z_1, z_2, w_1 y w_2. Entonces hay cuatro parámetros, por determinar y, por tanto, cuatro condiciones que se pueden imponer. Éstas se eligen de manera que el método dé resultados exactos cuando la

*Si los límites son distintos, se hace un cambio de variable para pasarlos a −1 y +1.

función por integrar sea alguna de las cuatro siguientes o combinaciones lineales de ellas

$$F(z) = 1$$
$$F(z) = z$$
$$F(z) = z^2$$
$$F(z) = z^3$$

Los valores exactos de integrar estas cuatro funciones entre –1 y +1 son

$$I_1 = \int_{-1}^{1} 1\, dz = z \Big|_{-1}^{1} = 1 - (-1) = 2$$

$$I_2 = \int_{-1}^{1} z\, dz = \frac{z^2}{2} \Big|_{-1}^{1} = \frac{1^2}{2} - \frac{(-1)^2}{2} = 0$$

$$I_3 = \int_{-1}^{1} z^2\, dz = \frac{z^3}{3} \Big|_{-1}^{1} = \frac{1^3}{3} - \frac{(-1)^3}{3} = \frac{2}{3}$$

$$I_4 = \int_{-1}^{1} z^3\, dz = \frac{z^4}{4} \Big|_{-1}^{1} = \frac{1^4}{4} - \frac{(-1)^4}{4} = 0$$

Suponiendo que una ecuación de la forma 6.24 funciona exactamente, se tendría el siguiente sistema de ecuaciones

$$I_1 = w_1(1) + w_2(1) = 2$$

$$I_2 = w_1 z_1 + w_2 z_2 = 0$$

$$I_3 = w_1 z_1^2 + w_2 z_2^2 = 2/3$$

$$I_4 = w_1 z_1^3 + w_2 z_2^3 = 0$$

De la primera ecuación se tiene que $w_1 + w_2 = 2$; nótese también que si

$$w_1 = w_2$$

y

$$z_1 = -z_2,$$

se satisfacen la segunda y la cuarta ecuaciones. Entonces se elige

$$w_1 = w_2 = 1$$

y

$$z_1 = -z_2$$

y al sustituir en la tercera ecuación se obtiene

$$z_1^2 + (-z_1)^2 = 2/3$$

o bien

$$z_1^2 = 1/3$$

de donde

$$z_1 = \pm \frac{1}{\sqrt{3}} = \pm\ 0.57735\ldots$$

y queda entonces

$$z_1 = -\ 0.57735\ldots$$
$$z_2 = 0.57735\ldots$$

con lo que se tiene la fórmula

$$\begin{aligned} \int_{-1}^{1} F(z)\, dz &= w_1 F(z_1) + w_2 F(z_2) \\ &= F(-0.57735\ldots) + F(0.57735\ldots) \end{aligned} \tag{6.26}$$

que, salvo porque se tiene que calcular el valor de la función en un valor irracional de z, es tan simple como la regla trapezoidal; además, trabaja perfectamente para funciones cúbicas, mientras que la regla trapezoidal lo hace sólo para líneas rectas.

En páginas anteriores se comentó que para integrar en un intervalo distinto de $[-1, 1]$, se requiere un cambio de variable a fin de pasar del intervalo de integración general $[a, b]$ a $[-1, 1]$ y así aplicar la ecuación 6.26; por ejemplo, si se desea obtener

$$\int_0^5 e^{-x}\, dx$$

se puede cambiar a $z = \frac{2}{5}x - 1$, de modo que si $x = 0, z = -1$ y si $x = 5, z = 1$. El resto de la integral se pone en términos de la nueva variable z y se encuentra que

$$e^{-x} = e^{-5(z+1)/2}$$

y

$$dx = d\left(\frac{5}{2}(z+1)\right) = \frac{5}{2}\, dz$$

entonces la integral queda

$$\int_0^5 e^{-x}\, dx = \frac{5}{2} \int_{-1}^{1} e^{-5(z+1)/2}\, dz,$$

de modo que las condiciones de aplicación del método de Gauss quedan satisfechas.

Al resolver se tiene

$$\frac{5}{2} \int_{-1}^{1} e^{-5(z+1)/2}\, dz \approx \frac{5}{2} \left[w_1 F(-0.57735\ldots) + w_2 F(0.57735\ldots) \right]$$

$$= \frac{5}{2} \left[(1)\, e^{-5(-0.57735+1)/2} \right.$$

$$\left. + (1)\, e^{-5(0.57735+1)/2} \right] = 0.91752$$

Esto es

$$\int_0^5 e^{-x}\,dx \approx 0.91752$$

El valor exacto de esta integral es 0.99326.

En general, si se desea calcular $\int_a^b f(x)\,dx$ aplicando la ecuación 6.26, se cambia el intervalo de integración con la siguiente fórmula*

$$z = \frac{2x - (a+b)}{b-a} \tag{6.27}$$

ya que si $x = a$, $z = -1$; y si $x = b$, $z = 1$

El integrando $f(x)\,dx$ en términos de la nueva variable queda

$$f(x) = F\left(\frac{b-a}{2}z + \frac{a+b}{2}\right)$$

y

$$dx = d\left(\frac{b-a}{2}z + \frac{a+b}{2}\right) = \frac{b-a}{2}\,dz$$

Por lo que la integral queda finalmente como

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\int_{-1}^{1} F\left(\frac{b-a}{2}z + \frac{a+b}{2}\right)dz$$

$$\approx \frac{b-a}{2}\left[F\left(\frac{b-a}{2}(-0.57735) + \frac{a+b}{2}\right) + F\left(\frac{b-a}{2}(0.57735) + \frac{a+b}{2}\right)\right] \tag{6.28}$$

Una cuestión importante es que el método de Gauss puede extenderse a tres o más puntos; por ejemplo, si se escogen tres puntos *no equidistantes* en el segmento de la curva $F(z)$ comprendida entre (–1 y 1, se podría pasar una parábola por los tres como en la regla de Simpson, excepto en que dichos puntos se escogerían de modo que minimicen o anulen el error. Similarmente es factible elegir cuatro puntos y una curva cúbica, cinco puntos y una curva cuártica, etc. En general, el algoritmo tiene la forma

$$\int_{-1}^{1} F(z)\,dz = A \approx w_1 F(z_1) + w_2 F(z_2) + w_3 F(z_3) + \ldots + w_n F(z_n) \tag{6.29}$$

donde se han calculado los valores de w_i y z_i por usar y la tabla 6.2 da valores hasta para seis puntos.

*Sólo es aplicable cuando los límites de integración a y b son finitos.

Con dos puntos, el método de Gauss está diseñado para obtener exactitud en polinomios cúbicos; con tres, se tendrá exactitud en polinomios de cuarto grado y así sucesivamente.

Los coeficientes y abscisas dadas en la tabla 6.2 sirven para integrar sobre todo el intervalo de interés, o bien puede dividirse el intervalo en varios subintervalos (como en los métodos compuestos de integración) y aplicar el método de Gauss a cada uno de ellos.

Número de puntos	Coeficientes w_i	Abscisas z_i
2	$w_1 = w_2 = 1.0$	$-z_1 = z_2 = 0.5773502692$
3	$w_2 = 0.88888\ldots$	$-z_1 = z_3 = 0.7745966692$
	$w_1 = w_3 = 0.55555\ldots$	$z_2 = 0.0$
4	$w_1 = w_4 = 0.3478548451$	$-z_1 = z_4 = 0.8611363116$
	$w_2 = w_3 = 0.6521451549$	$-z_2 = z_3 = 0.3399810436$
5	$w_1 = w_5 = 0.2369268851$	$-z_1 = z_5 = 0.9061798459$
	$w_2 = w_4 = 0.4786286705$	$-z_2 = z_4 = 0.5384693101$
	$w_3 = 0.56888\ldots$	$z_3 = 0.0$
6	$w_1 = w_6 = 0.1713244924$	$-z_1 = z_6 = 0.9324695142$
	$w_2 = w_5 = 0.3607615730$	$-z_2 = z_5 = 0.6612093865$
	$w_3 = w_4 = 0.4679139346$	$-z_3 = z_4 = 0.2386191861$

TABLA 6.2. Coeficientes y abscisas en el método de la cuadratura de Gauss Legendre.

Ejemplo 6.9

Integre la función $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ en el intervalo (–0.8, 1.5) por cuadratura de Gauss.

SOLUCIÓN

a) Con dos puntos
Cambio de límites de la integral con la ecuación

$$z = \frac{2x - (a+b)}{b-a} = \frac{2x - 0.7}{2.3}$$

Si $x = -0.8, z = -1$; si $x = 1.5, z = 1$

Con el cambio de la función en términos de la nueva variable z queda

$$I = \frac{1}{\sqrt{2\pi}} \int_{0.8}^{1.5} e^{-x^2/2}\, dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} \left[\frac{1.5 - (-0.8)}{2} \right] e^{-\left[\frac{1.5-(-0.8)}{2} z + \frac{-0.8+1.5}{2} \right]^2/2}\, dz$$

$$= \frac{2.3}{2\sqrt{2\pi}} \int_{-1}^{1} e^{-(2.3z+0.7)^2/8}\, dz$$

De la tabla 6.2 $w_1 = w_2 = 1.0$; $-z_1 = z_2 = 0.5773502692$
Al evaluar la función del integrando en z_1, z_2

$$F\,(\,0.5773502692\,) = e^{-[\,2.3\,(0.5773502692\,)\,+\,0.7\,]^2/8} = 0.5980684$$
$$F\,(\,-0.5773502692\,) = e^{-[\,2.3\,(-.5773502692\,)\,+\,0.7\,]^2/8} = 0.9519115$$

Se aplica la ecuación 6.29

$$I = \frac{2.3}{2\sqrt{2\pi}} [\,1\,(\,0.5980684\,) + 1\,(0.9519115\,)\,] = 0.711105$$

b) Con tres puntos

De la tabla 6.2

$$w_1 = w_3 = 0.55555..., w_2 = 0.88888...$$
$$-z_1 = z_3 = 0.7745966692, z_2 = 0.0$$

Al evaluar la función del integrando en z_1, z_2 y z_3 y emplear la ecuación 6.29 se tiene

$$I \approx \frac{2.3}{2\sqrt{2\pi}} [\,0.55555...\,(\,0.4631...\,) + 0.88888...\,(\,0.9405...\,)$$
$$+ 0.55555...\,(\,0.8639...\,)\,] = 0.721825$$

Ejemplo 6.10

Halle $\int_0^{2\pi} \text{sen } x\, dx$, por el método de la cuadratura de Gauss utilizando tres puntos.

SOLUCIÓN

Se cambian variable y límites de integración con la expresión

$$z = \frac{2x-(a+b)}{b-a}$$

como $a = 0, b = 2\pi$, entonces

$$z = \frac{2x - 2\pi}{2\pi} = \frac{x}{\pi} - 1 = \frac{x - \pi}{\pi}$$

se despeja $x : x = \pi z + \pi$ de donde $dx = \pi\, dz$
Se sustituye en la integral

$$\int_o^{2\pi} \text{sen } x\, dx = \int_{-1}^{1} \text{sen } (\pi z + \pi)\, \pi\, dz = \pi \int_{-1}^{1} \text{sen } (\pi z + \pi)\, dz$$

Con el empleo de la ecuación 6.29 con $n = 3$ y los valores de la tabla 6.2 queda

$$\begin{aligned} A \approx \pi \{&0.55555...[\text{sen}(\pi(-0.7745966692) + \pi)\,] \\ &+ 0.88888...[\text{sen}(\pi(0) + \pi)] \\ &+ 0.55555...[\text{sen}(\pi(0.7745966692) + \pi)]\} \end{aligned}$$

Se deja al lector la comparación de este resultado con la solución analítica.

La expresión 6.29 puede ponerse en forma más general y adecuada para programarla, así

$$\int_a^b f(x)\, dx = \frac{b-a}{2} \sum_{i=1}^{n} w_i F \left[\frac{(b-a)z_i + b + a}{2} \right] \tag{6.30}$$

la cual puede deducirse de los ejemplos resueltos (ver problema 6.21).

A continuación se presenta un algoritmo para la cuadratura de Gauss-Legendre.

ALGORITMO 6.3 Cuadratura de Gauss-Legendre

Para aproximar el área bajo la curva de una función analítica $f(x)$ en el intervalo $[a, b]$, proporcionar la función a integrar F(X) y los

DATOS: El número de puntos (2, 3, 4, 5, ó 6) por utilizar: N, el límite inferior A y el límite superior B.
RESULTADOS: El área aproximada ÁREA.

PASO 1. Hacer (NP(I),I=1,2,...,5) = (2,3,4,5,6)
PASO 2. Hacer (IAUX(I),I=1,2,..., 6) = (1,2,4,6,9,12)
PASO 3. Hacer (Z(I),I=1,2,...,11) = (0.577350269,0.0,0.774596669, 0.339981044,0.861136312,0.0,0.538469310, 0.906179846,0.238619186,0.661209387,0.932469514)

PASO 4. Hacer (W(I),I=1,2,...,11) = (1.0,0.888888888,0.555555555, 0.652145155,0.347854845,0.568888888,0.478628671, 0.236926885,0.467913935,0.360761573,0.171324493)
PASO 5. Hacer I = 1
PASO 6. Mientras I ≤ 5, repetir los pasos 7 y 8.
  PASO 7. Si N=NP(I), ir al paso 10. De otro modo continuar.
  PASO 8. Hacer I = I + 1
PASO 9. IMPRIMIR "N NO ES 2,3,4,5, o 6" y TERMINAR.
PASO 10. Hacer S = 0
PASO 11. Hacer J = IAUX(I)
PASO 12. Mientras J ≤ IAUX(I+1) - 1, repetir los pasos 13 a 17.
  PASO 13. Hacer ZAUX = (Z(J) * (B - A) + B + A) /2
  PASO 14. Hacer S = S + F (ZAUX) * W(J)
  PASO 15. Hacer ZAUX = (-Z(J) * (B - A) + B + A) /2
  PASO 16. Hacer S = S + F(ZAUX) * W(J)
  PASO 17. Hacer J = J + 1
PASO 18. Hacer AREA = (B - A) /2 * S
PASO 19. IMPRIMIR AREA y TERMINAR.

Ejemplo 6.11

Elabore un programa que integre funciones analíticas con la cuadratura de Gauss-Legendre usando 2, 3, 4, 5 ó 6 puntos, mismos que usted eligirá. Pruebe el programa con la funcion del ejemplo 6.8.

SOLUCIÓN

La expresión general de Gauss-Legendre para integrar es

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2} \sum_{i=1}^{n} w_i\, F\left[\frac{(b-a)z_i + b + a}{2}\right]$$

donde w_i, z_i, $i = 1,2,...,n$ están dados en la tabla 6.2.

En el disco se encuentra el programa 6.3 solicitado.

SECCIÓN 6.3 INTEGRALES MÚLTIPLES

Cualquiera de las técnicas de integración estudiadas en este capítulo es modificable, de modo que se puede aplicar en la aproximación de integrales dobles o triples. A continuación se ilustra el método de Simpson 1/3 en la solución de integrales dobles

a) $\int_0^{\pi} \int_0^3 y \operatorname{sen} x\, dx\, dy$ y *b)* $\int_1^3 \int_0^4 e^{x+y}\, dx\, dy$

Para la integral del inciso (a), se divide el intervalo $[a,b] = [0,3]$ en $n = 6$ subintervalos iguales, con lo que la amplitud de cada subintervalo es igual a

$$h_1 = \frac{3 - 0}{6} = 0.5$$

y se aplica la regla de Simpson compuesta a la integral interna, manteniendo constante la variable y (nótese que se está integrando en el eje x).

$$\int_0^{\pi} \int_0^3 y \operatorname{sen} x \, dx \, dy \approx \int_0^{\pi} \frac{h_1}{3} [\, y \operatorname{sen} 0 + 4y (\operatorname{sen} 0.5 + \operatorname{sen} 1.5 + \operatorname{sen} 2.5) +$$

$$2y (\operatorname{sen} 1 + \operatorname{sen} 2) + y \operatorname{sen} 3 \,] \, dy$$

$$\approx \int_0^{\pi} 1.9907 \, y \, dy$$

Ahora se integra en el eje y. El intervalo $[c, d] = [0,\pi]$ se divide en $m = 8$ subintervalos, por ejemplo, y queda

$$h_2 = \frac{\pi - 0}{8} = \frac{\pi}{8}$$

y

$$1.9907 \int_0^{\pi} y dy \approx 1.9907 \frac{h_2}{3} [0 + 4 (\frac{\pi}{8} + \frac{3\pi}{8} + \frac{5\pi}{8} + \frac{7\pi}{8}) +$$

$$2 (\frac{2\pi}{8} + \frac{4\pi}{8} + \frac{6\pi}{8}) + \frac{8\pi}{8}]$$

$$\approx 9.82373$$

entonces se tiene

$$\int_0^{\pi} \int_0^3 y \operatorname{sen} x \, dx \, dy \approx 9.82373$$

Obsérvese que se ha efectuado una integración repetida; esto es, se ha integrado siguiendo el proceso

$$\int_0^{\pi} \int_0^3 y \operatorname{sen} x \, dx \, dy \quad \int_0^{\pi} \left(\int_0^3 y \operatorname{sen} x \, dx \right) dy$$

En la primera integración y se mantuvo constante. Es importante recordar que la integración iterada puede llevarse a cabo con respecto a y primero y después respecto a x, pero intercambiando los límites de integración. Esto se indica

$$\int_0^3 \int_0^{\pi} y \operatorname{sen} x \, dy \, dx ,$$

y el resultado es el mismo (véase problema 6.30).

Para la integral (b), el intervalo $[a,b] = [0,4]$ se divide en $n = 4$ subintervalos, de donde

$$h_1 = \frac{4 - 0}{4} = 1$$

$$\int_1^3 \int_0^{-4} e^{x+y}\, dx\, dy = \int_1^3 \frac{1}{3} [e^{0+y} + 4\,(e^{1+y} + e^{3+y}) + 2e^{2+y} + e^{4+y}]\, dy$$

cuya integración por Simpson 1/3 con $m = 6$ en el eje y da

$$h_2 = \frac{3 - 1}{6} = \frac{1}{3}$$

$$\int_1^3 \int_0^4 e^{x+y}\, dx\, dy = \frac{1}{3}\,\frac{1}{3(3)} [e^1 + 4(e^{4/3} + e^2 + e^{6/3}) + 2(e^{5/3} + e^{7/3}) + e^3] +$$

$$\frac{1}{3}\,\frac{1}{3(3)} [e^{0.5+1} + 4\,(e^{0.5+4/3} + e^{0.5+2} + e^{0.5+8/3}) +$$

$$2\,(e^{0.5+5/3} + e^{0.5+7/3}) + e^{0.5+3}] +$$

$$\frac{1}{3}\,\frac{1}{3(3)} [e^{1.5+1} + 4\,(e^{1.5+4/3} + e^{1.5+2} + e^{1.5+8/3}) +$$

$$2\,(e^{1.5+5/3} + e^{1.5+7/3}) + e^{1.5+3}] +$$

$$\frac{1}{3}\left(\frac{1}{3} [e^{1+1} + 4\,(e^{1+4/3} + e^{1+2} + e^{1+8/3}) +\right.$$

$$2\,(e^{1+5/3} + e^{1+7/3}) + e^{1+3}]\Big) +$$

$$\frac{1}{3}\left(\frac{1(0.5)}{3} [e^{2+1} + 4\,(e^{2+4/3} + e^{2+2} + e^{2+8/3}) +\right.$$

$$\left.2\,(e^{2+5/3} + e^{2+7/3}) + e^{2+3}]\right)$$

$$\approx 935.53 \text{ (El resultado analítico es 930.853)}$$

En general, la integración de una función $f(x,y)$ sobre una región R del plano $x - y$ dada así: $\{ (x,y): a \le x \le b, c \le y \le d \}$ por el método de Simpson 1/3 es

$$\int_c^d \int_a^b f(x,y)\, dx\, dy = \int_c^d \left[\int_a^b f(x,y)\, dx \right] dy$$

$$\approx \int_c^d \left(\frac{h_1}{3} \left[f(x_0,y) + 4 \sum_{\substack{i=1 \\ \Delta\, i=2}}^{n-1} f(x_i,y) + 2 \sum_{\substack{i=2 \\ \Delta\, i=2}}^{n-2} f(x_i,y) + f(x_n,y) \right] \right) dy$$

donde $h_1 = \dfrac{b-a}{n}$. Desarrollando se tiene

$$\int_c^d \int_a^b f(x,y)\, dx\, dy \approx \frac{h_1}{3} \int_c^d f(x_0,y)\, dy + \frac{4h_1}{3} \sum_{\substack{i=1 \\ \Delta\, i=2}}^{n-1} \int_c^d f(x_i,y)\, dy$$

$$+ \frac{2h_1}{3} \sum_{\substack{i=2 \\ \Delta\, i=2}}^{n-2} \int_c^d f(x_i,y)\, dy + \frac{h_1}{3} \int_c^d f(x_n,y)\, dy,$$

e integrando nuevamente por Simpson 1/3 con $h_2 = \dfrac{d-c}{m}$

$$\int_c^d \int_a^b f(x,y)\, dx\, dy \approx$$

$$\frac{h_1}{3}\,\frac{h_2}{3} \left[f(x_0,y_0) + 4 \sum_{\substack{j=1 \\ \Delta\, j=2}}^{m-1} f(x_0,y_j) + 2 \sum_{\substack{j=2 \\ \Delta\, j=2}}^{m-2} f(x_0,y_j) + f(x_0,y_m) \right]$$

$$+ \frac{4h_1}{3}\,\frac{h_2}{3} \sum_{\substack{i=1 \\ \Delta\, i=2}}^{n-1} \left[f(x_i,y_0) + 4 \sum_{\substack{j=1 \\ \Delta\, j=2}}^{m-1} f(x_i,y_j) + 2 \sum_{\substack{j=2 \\ \Delta\, j=2}}^{m-2} f(x_i,y_j) + f(x_i,y_m) \right]$$

$$+ \frac{2h_1}{3}\,\frac{h_2}{3} \sum_{\substack{i=2 \\ \Delta\, i=2}}^{n-2} \left[f(x_i,y_0) + 4 \sum_{\substack{j=1 \\ \Delta\, j=2}}^{m-1} f(x_i,y_j) + 2 \sum_{\substack{j=2 \\ \Delta\, j=2}}^{m-2} f(x_i,y_j) + f(x_i,y_m) \right]$$

$$+ \frac{h_1}{3}\,\frac{h_2}{3} \left[f(x_n,y_0) + 4 \sum_{\substack{j=1 \\ \Delta\, j=2}}^{m-1} f(x_n,y_j) + 2 \sum_{\substack{j=2 \\ \Delta\, j=2}}^{m-2} f(x_n,y_j) + f(x_n,y_m) \right]$$

$$\approx \left\{ h_1 \frac{h_2}{9} \left[f(x_0,y_0) + f(x_0,y_m) + f(x_n,y_0) + f(x_n,y_m) + \right.\right.$$

$$+ 4 \sum_{\substack{j=1 \\ \Delta j=2}}^{m-1} \left(f(x_0, y_j) + f(x_n, y_j) \right) + 2 \sum_{\substack{j=2 \\ \Delta j=2}}^{m-2} \left(f(x_0, y_j) + f(x_n, y_j) \right)$$

$$+ 4 \sum_{\substack{i=1 \\ \Delta i=2}}^{n-1} \left(f(x_i, y_0) + 4 \sum_{\substack{j=1 \\ \Delta j=2}}^{m-1} f(x_i, y_j) + 2 \sum_{\substack{j=2 \\ \Delta j=2}}^{m-2} f(x_i, y_j) + f(x_i, y_m) \right)$$

$$+ 2 \sum_{\substack{i=2 \\ \Delta i=2}}^{n-2} \left(f(x_i, y_0) + 4 \sum_{\substack{j=1 \\ \Delta j=2}}^{m-1} f(x_i, y_j) + 2 \sum_{\substack{j=2 \\ \Delta j=2}}^{m-2} f(x_i, y_j) + f(x_i, y_m) \right) \Bigg] \tag{6.31}$$

Igualmente puede emplearse la cuadratura de Gauss para integrales dobles o triples. Asi, en el caso general

$$\int_c^d \int_a^b f(x, y)\, dx\, dy$$

primero se cambian los intervalos de x y y a $[-1,1]$ con las fórmulas

$$t = \frac{2x - (a+b)}{b-a} \quad \text{y} \quad u = \frac{2y - (c+d)}{d-c}$$

y asimismo se cambian dx, dy y $f(x,y)$ a términos de las nuevas variables t y u. Para esto, se despeja x

$$x = \frac{(b-a)t}{2} + \frac{(b+a)}{2} \quad \text{de donde} \quad dx = \frac{(b-a)}{2}\, dt$$

y después y

$$y = \frac{(d-c)}{2} + \frac{(c+d)}{2} \quad \text{de donde} \quad dy = \frac{(d-c)}{2}\, du$$

Se sustituye

$$\int_c^d \int_a^b f(x,y)\, dx dy = \frac{(b-a)(d-c)}{4} \int_{-1}^{1} \int_{-1}^{1} f\left[\frac{(b-a)}{2} t + \frac{(a+b)}{2}, \frac{(d-c)}{2} u + \frac{(c-d)}{2} \right] dt du \tag{6.32}$$

a la cual cabe aplicar la fórmula 6.30. Para ilustrar esto, a continuación se resuelve la integral

$$\int_2^3 \int_0^4 e^{x+y}\, dx\, dy$$

empleando dos puntos.

Primero se sustituye e integra respecto a t

$$\int_1^3 \int_0^4 e^{x+y}\,dx\,dy \approx \frac{(4-0)(3-1)}{4} \int_{-1}^{1} \int_{-1}^{1} e^{(2t+2)+(u+2)}\,dt\,du$$

$$\approx 2 \int_{-1}^{1} \left[e^{2(-0.57735)+4+u} + e^{2(0.57735)+4+u} \right] du$$

Ahora se integra respecto a u

$$\int_1^3 \int_0^4 e^{x+y}\,dx\,dy \approx 2 \left[e^{2.84530-0.57735} + e^{2.84530+0.57735} + e^{5.15470-0.57735} + e^{5.15470+0.57735} \right] = 892.335$$

Para mayor sencillez y facilidad de programación, es conveniente emplear la fórmula

$$\int_c^d \int_a^b f(x,y)\,dx\,dy \approx \frac{(b-a)(d-c)}{4} \sum_{j=1}^{m} \sum_{i=1}^{n} w_j w_i F\left[\frac{b-a}{2} t_i + \frac{b+a}{2}, \frac{d-c}{2} u_j + \frac{c+d}{2}\right] \tag{6.33}$$

donde n y m son los números de puntos por usar en los ejes x y y, respectivamente. Su aplicación a la integral del inciso (a) empleando tres puntos en ambos ejes conduce a

$$\int_1^3 \int_0^4 e^{x+y}\,dx\,dy \approx \frac{(4-0)(3-1)}{4} \sum_{j=1}^{3} \sum_{i=1}^{3} w_j w_i F(2t_i + 2, u_j + 2)$$

donde w_1, w_2, y w_3 y $t_1 = u_1 = z_1, t_2 = u_2 = z_2$ y $t_3 = u_3 = z_3$ están dados en la tabla 6.2.

Al sustituir valores se tiene

$$\int_1^3 \int_0^4 e^{x+y}\,dx\,dy \approx 934.39$$

La solución analítica de esta integral es 930.85

Hasta ahora solo se han visto integraciones dobles sobre regiones R rectangulares. No obstante, también pueden resolverse integrales del tipo

$$\int_c^d \int_{a(y)}^{b(y)} f(x,y)\,dx\,dy$$

o del tipo

$$\int_a^b \int_{c(x)}^{d(x)} f(x,y)\,dy\,dx$$

cuyas regiones R_1 y R_2 quedan dadas como se muestra en la figura 6.9 *a* y *b*.

A continuación se resuelve por el método de Simpson 1/3 la integral

$$\int_0^2 \int_{x^2}^{2x} (x^3 + 4y)\,dy\,dx$$

que representa el área sombreada de la figura 6.10.

El intervalo $[a, b] = [0, 2]$ se divide en, por ejemplo, dos subintervalos y queda $h_1 = (2 - 0)/2 = 1$; el tamaño de paso en el eje y varía con x de acuerdo con la expresión

$$h_2(x) = \frac{d(x) - c(x)}{m}$$

donde m es el número de subintervalos en que se divide el eje y.

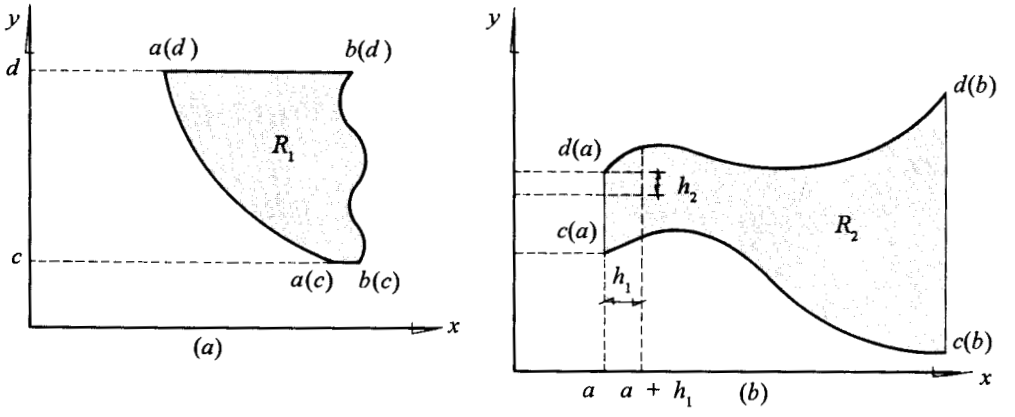


Figura 6.9. Regiones no rectangulares de integración.

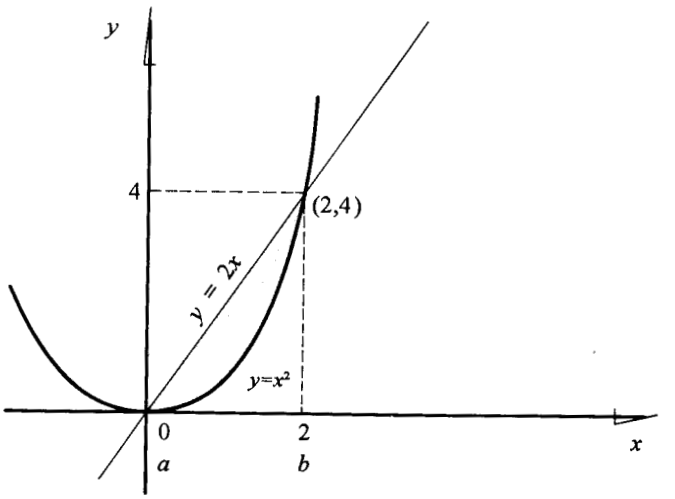


Figura 6.10. Región de integración delimitada por una recta y una parábola.

Si se hace $m = 2$, se tiene

$$\int_0^2 \int_{x^2}^{2x} (x^3 + 4y)\,dy\,dx \approx \int_0^2 \left(\frac{h_2(x)}{3} [x^3 + 4x^2 + 4(x^3 + 4(x^2 + h_2(x))) + x^3 + 4(2x)] \right) dx$$

$$\approx \int_0^2 \frac{h_2(x)}{3} [6x^3 + 20x^2 + 8x + 16h_2(x)]\,dx$$

$$= \frac{h_1}{3} \left(\frac{h_2(0)}{3} [6(0)^3 + 20(0)^2 + 8(0) + 16h_2(0)] \right.$$

$$+ \frac{4h_2(0+1)}{3} [6(1)^3 + 20(1)^2 + 8(1) + 16h_2(1)]$$

$$\left. + \frac{h_2(2)}{3} [6(2)^3 + 20(2)^2 + 8(2) + 16h_2(2)] \right)$$

ya que

$$h_2(0) = \frac{2(0) - 0^2}{2} = 0, \quad h_2(1) = \frac{2(1) - 1^2}{2} = 0.5, \text{ y}$$

$$h_2(2) = \frac{2(2) - 2^2}{2} = 0,$$

$$\int_0^2 \int_{x^2}^{2x} (x^3 + 4y)\,dy\,dx \approx 9.33$$

Si se divide el intervalo $[a, b]$ en cuatro subintervalos y se mantiene $m = 2$. Se tiene $h_1 = (2 - 0)/4 = 0.5$. Entonces, la integración queda como sigue

$$\int_0^2 \int_{x^2}^{2x} (x^3 + 4y)\,dy\,dx \approx \frac{0.5}{3} \left(\frac{4h_2(0.5)}{3} [6(0.5)^2 + 20(0.5)^3 + 8(0.5) + 16h_2(0.5)] \right.$$

$$+ \frac{2h_2(1)}{3} [6(1)^3 + 20(1)^2 + 8(1) + 16h_2(1)]$$

$$\left. + \frac{4h_2(1.5)}{3} [6(1.5)^3 + 20(1.5)^2 + 8(1.5) + 16h_2(1.5)] \right)$$

ya que $h_2(0) = 0, h_2(2) = 0, h_2(0.5) = \dfrac{2(0.5) - 0.5^2}{2} = 0.375$, y

$$h_2(1) = 0.5, h_2(1.5) = \frac{2(1.5) - 1.5^2}{2} = 0.375$$

Al sustituir valores se obtiene

$$\int_0^2 \int_{x^2}^{2x} (x^3 + 4y)\,dy\,dx \approx 10.583$$

El valor analítico es 10.666.

ALGORITMO 6.4 Integración doble por Simpson 1/3

Para aproximar $\int_a^b \int_{c(x)}^{d(x)} f(x,y)\,dy\,dx$, proporcionar las funciones C(X), D(X) y F(X, Y) y los

DATOS: El número N de subintervalos a usar en el eje x, el número M de subintervalos por emplear en el eje y, el límite inferior A y el límite superior B.
RESULTADOS: El área aproximada AREA.

PASO 1. Hacer H1 = (B - A)/N
PASO 2. Hacer S=(F(A,C(A)) + F(A,D(A))) * (D(A) - C(A))/M
+ (F(B,C(B)) + F(B,D(B))) * (D(B) - C(B))/M
PASO 3. Hacer S1 = 0; S2 = 0; Y1 = C(A); Y2 = C(B)
PASO 4. Hacer J = 1
PASO 5. Mientras J ≤ M - 1, repetir los pasos 6 a 9.
PASO 6. Hacer H2A = (D(A) - C(A))/M;
Y1 = Y1 + H2A;
S1 = S1 + H2A*F(A,Y1);
H2B = (D(B)–C(B))/M;
Y2 = Y2 + H2B; S1 = S1 + H2B*F(B,Y2)
PASO 7. SI J = M - 1, ir al paso 9. De otro modo continuar.
PASO 8. Hacer Y1 = Y1 + H2A;
S2 = S2 + H2A*F(A,Y1);
Y2 = Y2 + H2B; S2 = S2 + H2B*F(B,Y2)
PASO 9. Hacer J = J + 2
PASO 10. Hacer S3 = 0; S6 = 0; S7 = 0; X = A
PASO 11. Hacer I = 1
PASO 12. Mientras I ≤ N - 1, repetir los pasos 13 a 16.
PASO 13. Hacer X = X + H1; H2 = (D(X) - C(X))/M;
S3 = S3 + H2*F(X,C(X));
S6 = S6 + H2*F(X,D(X))
PASO 14. SI I = N-1, ir al paso 16. De otro modo continuar.
PASO 15. Hacer X = X + H1; H2 = (D(X)-C(X))/M;
S7 = S7 + 2*(F(X,C(X)) + F(X,D(X)))
PASO 16. Hacer I = I + 2
PASO 17. Hacer S4 = 0; S5 = 0; S8 = 0; S9 = 0; X = A - H1
PASO 18. Hacer I = 1
PASO 19. Mientras I ≤ N - 1, repetir los pasos 20 a 31.
PASO 20. Hacer X = X + 2*H1;
Y1 = C(X); Y2 = C(X+H1)
HA = (D(X)-C(X))/M;
HB = (D(X+H1)–C(X+H1))/M
PASO 21. Hacer J = 1
PASO 22. Mientras J ≤ M - 1, repetir los pasos 23 a 30.

```
                PASO 23.  Hacer Y1 = Y1 + HA;
                          S4 = S4 + HA*F(X,Y1)
                PASO 24.  SI I = N-1, ir al paso 26. De otro
                          modo continuar.
                PASO 25.  Hacer Y2 = Y2+HB;
                          S8 = S8 + HB*F(X+H1,Y2)
                PASO 26.  SI J = M-1, ir al paso 30. De otro
                          modo continuar.
                PASO 27.  Hacer Y1 = Y1 + HA;
                           S5 = S5 + HA*F(X,Y1)
                PASO 28.  SI I = N-1, ir al paso 30. De otro
                          modo continuar.
                PASO 29.  Hacer Y2 = Y2+HB;
                          S9=S9+ HA *F(X+H1,Y2)
                PASO 30.  Hacer J = J + 2
        PASO 31.  Hacer I = I + 2
PASO 32.  Hacer AREA=H1/9 * (S + 4*(S1 + S3 + S6 + S9) +
          2*(S2 + S7) + 16 * S4 + 8*(S5 + S8))
PASO 33.  IMPRIMIR AREA y TERMINAR.
```

SECCIÓN 6.4 DIFERENCIACIÓN NUMÉRICA

En la introducción del capítulo 5 se comentó que cuando se va a practicar una operación en una función tabulada, el camino es aproximar la tabla por alguna función y efectuar la operación en la función aproximante. Así se procedió en la integración numérica y así se procederá en la diferenciación numérica; esto es, se aproximará la función tabulada $f(x)$ y se diferenciará la aproximación $p_n(x)$.

Si la aproximación es polinomial y con el **criterio de ajuste exacto***, la diferenciación numérica consiste simplemente en diferenciar la fórmula del polinomio interpolante que se utilizó. Sea en general

$$f(x) = p_n(x) + R_n(x)$$

y la aproximación de la primera derivada queda entonces

$$\frac{df(x)}{dx} \approx \frac{dp_n(x)}{dx}$$

o en general

$$\frac{d^n f(x)}{dx^n} \approx \frac{d^n p_n(x)}{dx^n} \tag{6.36}$$

*Si la aproximación es por mínimos cuadrados, la diferenciación numérica consistirá en diferenciar el polinomio que mejor ajuste la información tabulada.

Al diferenciar la fórmula fundamental de Newton dada arriba se tiene

$$\frac{d^n f(x)}{dx^n} = \frac{d^n p_n(x)}{dx^n} + \frac{d^n R_n(x)}{dx^n} \tag{6.37}$$

donde $\frac{d^n R_n(x)}{dx^n}$ es el error que se comete al aproximar $\frac{d^n f(x)}{dx^n}$ por $\frac{d^n p_n(x)}{dx^n}$

Si las abcisas dadas $x_0, x_1, ..., x_n$ están espaciadas regularmente por intervalos de longitud h, entonces $p_n(x)$ puede escribirse en términos de diferencias finitas. Al sustituir $f[x_0]$, $f[x_0, x_1]$ etcétera en la ecuación 5.29 en términos de diferencias finitas (véase Ec. 5.35), se obtiene

$$p_n(x) = f[x_0] + (x - x_0)\frac{\Delta f[x_0]}{h} + (x - x_0)(x - x_1)\frac{\Delta^2 f[x_0]}{2!h^2} + \ldots$$

$$+ (x - x_0)(x - x_1)\ldots(x - x_{n-1})\frac{\Delta^n f[x_0]}{n!h^n}$$

y se tendrá

$$\frac{df(x)}{dx} \approx \frac{dp_n(x)}{dx} = \frac{df[x_0]}{dx} + \frac{d\left[(x-x_0)\frac{\Delta f[x_0]}{h}\right]}{dx} + \frac{d\left[(x-x_0)(x-x_1)\frac{\Delta^2 f[x_0]}{2!h^2}\right]}{dx}$$

$$+ \ldots + \frac{d\left[(x-x_0)(x-x_1)\ldots(x-x_{n-1})\frac{\Delta^n f[x_0]}{n!h^n}\right]}{dx} \tag{6.38}$$

Se desarrollan algunos de los primeros términos y se tiene

$$\frac{df(x)}{dx} \approx \frac{dp_n(x)}{dx} = \frac{\Delta f[x_0]}{h} + (2x - x_0 - x_1)\frac{\Delta^2 f[x_0]}{2!h^2}$$

$$+ [3x^2 - 2(x_0 + x_1 + x_2)x + (x_0x_1 + x_0x_2 + x_1x_2)]\frac{\Delta^3 f[x_0]}{3!h^3} \tag{6.39}$$

Selecciónese ahora un valor particular para n; por ejemplo, tómese $n = 1$, es decir que se aproxime la función tabulada $f(x)$ por una línea recta. Entonces

$$p_n(x) = p_1(x) = f[x_0] + (x - x_0)\frac{\Delta f[x_0]}{h},$$

y la primera derivada de $f(x)$ queda aproximada por

$$\frac{df(x)}{dx} \approx \frac{dp_1(x)}{dx} = \frac{\Delta f[x_0]}{h} = f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

$$\boxed{\frac{df(x)}{dx} \approx \frac{f(x_1) - f(x_0)}{h}} \tag{6.40}$$

y, como es de esperarse

$$\frac{d^2 f(x)}{dx^2} \approx \frac{d^2 p_1(x)}{dx^2} = 0$$

y así cualquier otra derivada superior de $f(x)$ quedará aproximada por cero.

Geométricamente esto equivale a tomar como primera derivada la pendiente de la recta que une los dos puntos de la curva $f(x)$ de abscisas x_0 y x_1 (véase Fig. 6.11).

La primera derivada de $f(x)$ en todo el intervalo $[x_0, x_1]$ queda aproximada por el valor constante $(f(x_1) - f(x_0))/h$, el cual es muy diferente del valor verdadero $df(x)/dx$ en general.

Si ahora $n = 2$; es decir, aproximando la función tabulada $f(x)$ por un polinomio de segundo grado, se tiene

$$p_n(x) \approx p_2(x) = f[x_0] + (x - x_0)\frac{\Delta f[x_0]}{h} + (x - x_0)(x - x_1)\frac{\Delta^2 f[x_0]}{2!h^2}$$

y la primera derivada de $f(x)$ queda aproximada por

$$\frac{df(x)}{dx} \approx \frac{dp_2(x)}{dx} = \frac{\Delta f[x_0]}{h} + (2x - x_0 - x_1)\frac{\Delta^2 f[x_0]}{2!h^2}$$

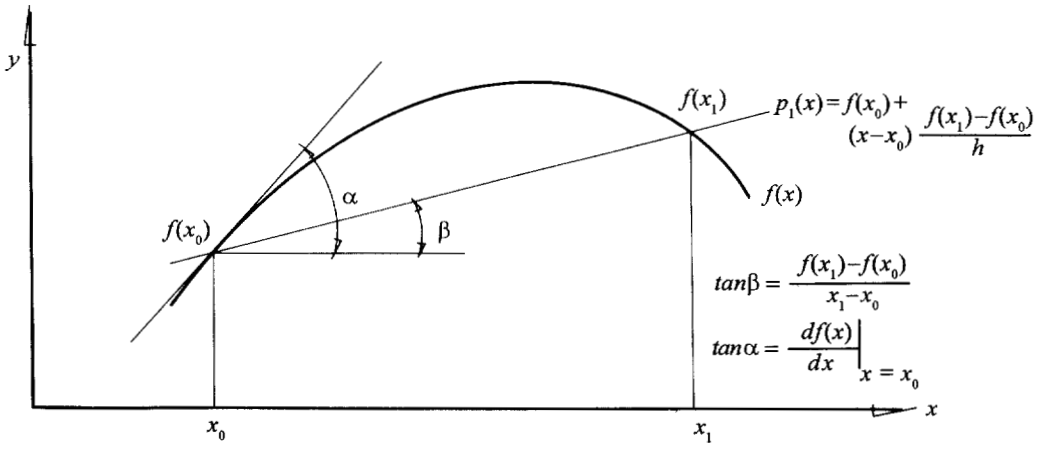


Figura 6.11. Aproximación lineal de la primera derivada.

Se desarrollan las diferencias hacia adelante y se tiene

$$\frac{df(x)}{dx} \approx \left(\frac{2x-x_0-x_1-2h}{2h^2}\right) f(x_0) + \left(\frac{2x_0-4x+2x_1+2h}{2h^2}\right) f(x_1) + \left(\frac{2x-x_0-x_1}{2h^2}\right) f(x_2) \tag{6.41}$$

La segunda derivada puede calcularse derivando una vez más con respecto a x, o sea

$$\frac{d^2f(x)}{dx^2} \approx \frac{d^2p_2(x)}{dx^2} = \frac{\Delta^2 f[x_0]}{h^2} = 2f[x_0, x_1, x_2]$$

$$\frac{d^2f(x)}{dx^2} \approx \frac{1}{h^2} f(x_0) - \frac{2}{h^2} f(x_1) + \frac{1}{h^2} f(x_2) \tag{6.42}$$

De igual modo se obtiene las distintas derivadas para $n > 2$.

Como se dijo al inicio de esta sección, el error cometido al aproximar $\frac{d^n f(x)}{dx^n}$ por $\frac{d^n p_n(x)}{dx^n}$ esta dado por $\frac{d^n R_n(x)}{dx^n}$, donde a su vez $R_n(x)$ está dado por la ecuación 5.41

$$R_n(x) = \left(\prod_{i=0}^{n} (x - x_i)\right) f[x, x_0, x_1, \ldots, x_n]$$

que quedaría más compacta si se denota por $\psi(x)$ a $\prod_{i=0}^{n} (x - x_i)$, es decir

$$R_n(x) = \psi(x) f[x, x_0, x_1, \ldots, x_n] \tag{6.43}$$

En este punto es importante recordar que hay una estrecha **relación entre las diferencias divididas y las derivadas**. En general, esta relación está dada así

$$f[x_0, x_1, \ldots, x_n] = \frac{d^n f(\xi)}{n!\, dx^n} \text{ con } \xi \in (\min x_i, \max x_i)\ 0 \leq i \leq n$$

esto es, ξ es un valor de x desconocido, del cual sólo se sabe que está entre los valores menor y mayor de los argumentos. Se sustituye en la ecuación 6.43

$$R_n(x) = \psi(x) \frac{d^{n+1} f(\xi_1(x))}{(n+1)!\, dx^{n+1}}, \text{ con } \xi_1(x) \in (\min x, x_i, \max x, x_i)\ 0 \leq i \leq n$$

donde se ha escrito ξ_1 como una función de x, ya que su valor depende del argumento x donde se desee evaluar la derivada.

Su primera derivada es

$$\frac{dR_n(x)}{dx} = \psi(x) \frac{d\left(\frac{d^{n+1}f(\xi_1(x))}{(n+1)!\,dx^{n+1}}\right)}{dx} + \frac{d^{n+1}f(\xi_1(x))}{(n+1)!\,dx^{n+1}} \frac{d\psi(x)}{dx} \tag{6.44}$$

Puede encontrarse que*

$$\frac{d\left(\frac{d^{n+1}f(\xi_1(x))}{(n+1)!\,dx^{n+1}}\right)}{dx} = \frac{d^{n+2}f(\xi_2(x))}{(n+2)!\,dx^{n+2}} \tag{6.45}$$

con $\xi_1(x)$, $\xi_2(x)$ $\epsilon(\min x, x_i, \max x, x_i)$ $0 \leq i \leq n$, donde ξ_2 es una función de x distinta de ξ_1.

Por esto, la ecuación 6.44 puede reescribirse como

$$\frac{dR_n(x)}{dx} = \psi(x) \frac{d^{n+2}f(\xi_2(x))}{(n+2)!\,dx^{n+2}} + \frac{d^{n+1}f(\xi_1(x))}{(n+1)!\,dx^{n+1}} \frac{d\psi(x)}{dx} \tag{6.46}$$

con $\xi_1(x)$, $\xi_2(x)$ $\epsilon(\min x, x_i, \max x, x_i)$ $0 \leq i \leq n$.

En particular, para $x = x_i$ [cuando x es una de las abscisas de la tabla de $f(x)$] el **error de truncamiento** dado por la ecuación 6.46 se simplifica, ya que $\psi(xi) = (x_i - x_0)(x_i - x_1) \ldots (x_i - x_i) \ldots (x_i - x_n) = 0$. Entonces

$$\frac{dR_n(x_i)}{dx} = \frac{d^{n+1}f(\xi_1(x_i))}{(n+1)!\,dx^{n+1}} \frac{d\psi(x)}{dx}\Big|_{x_i}$$

$$= \frac{d^{n+1}f(\xi_1(x_i))}{(n+1)!\,dx^{n+1}} \prod_{\substack{j=0 \\ j\neq i}}^{n} (x_i - x_j) \qquad \xi_1(x_1) \,\epsilon\, (\min x_i, \max x_i) \quad 0 \leq i \leq n \tag{6.47}$$

En los ejercicios (al final de este capítulo) se demuestra que

$$\frac{d\psi(x)}{dx}\Big|_{x_i} \prod_{\substack{j=0 \\ j\neq i}}^{n} (x_i - x_j)$$

*Pizer, M.S. Numerical Computing and Mathematical Analysis. S.R.A. (1975), 315-317.

Por ejemplo, la ecuación 6.41 puede escribirse en términos del error como sigue

$$\frac{df(x)}{dx}\Big|_{x_0} = \left(\frac{2x_0 - x_0 - x_1 - 2h}{2h^2}\right) f(x_0) + \left(\frac{2x_0 - 4x_0 + 2x_1 + 2h}{2h^2}\right) f(x_1)$$

$$+ \left(\frac{2x_0 - x_0 - x_1}{2h^2}\right) f(x_2) + (x_0 - x_1)(x_0 - x_2)\frac{d^3 f(x)}{3!\,dx^3}\Big|_{\xi}$$

o

$$\frac{df(x)}{dx}\Big|_{x_0} = \frac{1}{2h}[-3f(x_0) + 4f(x_1) - f(x_2)] + \frac{h^2}{3}\frac{d^3 f(x)}{dx^3}\Big|_{\xi} \quad (6.48)$$

con $\xi \in (\min x_i, \max x_i)$, i=0, 1, 2

y en la misma forma

$$\frac{df(x)}{dx}\Big|_{x_1} = \frac{1}{2h}[f(x_2) - f(x_0)] - \frac{h^2}{6}\frac{d^3 f(x)}{dx^3}\Big|_{\xi} \quad (6.49)$$

con $\xi \in (\min x_i, \max x_i)$, i=0, 1, 2

y

$$\frac{df(x)}{dx}\Big|_{x_2} = \frac{1}{2h}[f(x_0) - 4f(x_1) + 3f(x_2)] + \frac{h^2}{3}\frac{d^3 f(x)}{dx^3}\Big|_{\xi} \quad (6.50)$$

con $\xi \in (\min x_i, \max x_i)$, i=0, 1, 2

Obsérvese que el término del error para la derivada en x_1 es la mitad del término del error para la derivada en x_0 y x_2. Esto es así porque en la primera derivada se utilizan valores de la función a ambos lados de x_1.

En la diferenciación numérica, el error de truncamiento puede ser muy grande. Si por ejemplo $f^{(n+2)}(x)/(n+2)!$ y $f^{(n+1)}(x)/(n+1)!$ son de la misma magnitud, lo cual es común, el primer término de la ecuación 6.46 tiene aproximadamente la misma magnitud que el error de interpolación*; entonces puede decirse que el error de la aproximación de la derivada es, generalmente, mayor que el error de interpolación en

$$\frac{d^{n+1} f(\xi_1(x))}{(n+1)!\,dx^{n+1}} \quad \frac{d\psi(x)}{dx}$$

que es el segundo término de la ecuación 6.46. Además, cuando $x = x_i$, la ecuación 6.47 muestra que el error en la derivada en x_i tiene la misma forma que el error de interpolación (ver nota al pie de esta página), salvo que los polinomios factores son distintos.

Se ha discutido solamente el error de la aproximación de la primera derivada; pero todo lo visto también es aplicable a derivadas de mayor orden.

---

*Recuérdese que el error de interpolación es $\prod_{i=0}^{n}(x - x_i)\frac{f^{(n+1)}(\xi)}{(n+1)!}$ donde ξ depende de x de un modo desconocido.

Ejemplo 6.12

La ecuación de Van der Walls para un gmol de CO_2 es

$$\left(P + \frac{a}{v^2}\right)(v - b) = RT$$

donde

$$a = 3.6 \times 10^{-6} \text{ atm cm}^6 / \text{gmol}^2$$

$$b = 42.8 \text{ cm}^3 / \text{gmol}$$

$$R = 82.1 \text{ atm cm}^3 / (\text{gmol K})$$

Si T = 350 K, se obtiene la siguiente tabla de valores

Puntos	0	1	2	3
P (*atm*)	13.782	12.577	11.565	10.704
v (cm^3)	2000	2200	2400	2600

Calcule $\partial P/\partial v$ cuando $v = 2300 \text{ cm}^3$ y compárelo con el valor de la derivada analítica.

SOLUCIÓN

Al usar la ecuación 6.41 con los puntos (0), (1) y (2) se obtiene

$$\frac{\partial P}{\partial v} = \frac{2v - v_0 - v_1 - 2h}{2h^2} P_0 + \frac{2v_0 - 4v + 2v_1 + 2h}{2h^2} P_1 + \frac{2v - v_0 - v_1}{2h^2} P_2 ; \text{ con } h = 200$$

$$= \frac{2(2300) - 2000 - 2200 - 2(200)}{2(200)^2} 13.782$$

$$+ \frac{2(2000) - 4(2300) + 2(2200) + 2(200)}{2(200)^2} 12.577$$

$$+ \frac{2(2300) - 2000 - 2200}{2(200)^2} 11.565 = -0.00506$$

La derivada analítica es

$$\frac{\partial P}{\partial v} = \frac{-RT}{(v-b)^2} + \frac{2a}{v^3} = \frac{-82.1(350)}{(2300-42.8)^2} + \frac{2(3.6 \times 10^{-6})}{2300^3} = -0.005048$$

Nótese que la aproximación es muy buena (error relativo = –0.24%) a pesar de haber aplicado un polinomio de segundo grado para aproximar la ecuación de Van der Waals que, como se sabe, es un polinomio de tercer grado en v.

Ejemplo 6.13

Obtenga la primera derivada del polinomio general de Lagrange (ecuaciones 5.22 y 5.23).

SOLUCIÓN

De la ecuación $p_n(x) = \sum_{i=0}^{n} f(x_i) \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}$

Se deriva con respecto a x

$$\frac{dp_n(x)}{dx} = \sum_{i=0}^{n} f(x_i) \frac{d}{dx} \left[\prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \right]$$

Se hace

$$y = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}$$

y se toman logaritmos en ambos lados, con lo que se tiene

$$\ln y = \ln \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} = \sum_{\substack{j=0 \\ j \neq i}}^{n} \ln \frac{x - x_j}{x_i - x_j}$$

ya que el logaritmo de un producto es igual a la suma de los logaritmos de los factores.

Ambos miembros se derivan con respecto a x

$$\frac{d}{dx}(\ln y) = \frac{1}{y} \frac{dy}{dx} = \sum_{\substack{j=0 \\ j \neq i}}^{n} \frac{d}{dx} \left(\ln \frac{x - x_j}{x_i - x_j} \right) = \sum_{\substack{j=0 \\ j \neq i}}^{n} \frac{1}{x - x_j}$$

se despeja dy/dx

$$\frac{dy}{dx} = y \sum_{\substack{j=0 \\ j \neq i}}^{n} \frac{1}{x - x_j}$$

se sustituye y en el lado derecho

$$\frac{dy}{dx} = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \sum_{\substack{j=0 \\ j \neq i}}^{n} \frac{1}{x - x_j}$$

y finalmente

$$\frac{dp_n(x)}{dx} = \sum_{i=0}^{n} f(x_i) \left[\prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j} \sum_{\substack{j=0 \\ j\neq i}}^{n} \frac{1}{x - x_j} \right]$$

Obsérvese que esta ecuación no sirve para evaluar la derivada en una de las abscisas de la tabla, ya que significaría dividir entre cero en la sumatoria dentro del paréntisis. Sin embargo, manipulado algebraicamente, el lado derecho puede escribirse en la forma

$$\frac{dp_n(x)}{dx} = \sum_{i=0}^{n} \left[\frac{f(x_i)}{\prod_{\substack{j=0 \\ j\neq i}}^{n} (x_i - x_j)} \sum_{\substack{k=0 \\ k\neq i}}^{n} \prod_{\substack{j=0 \\ j\neq k,i}}^{n} (x - x_j) \right] \quad (6.51)$$

La cual ya no tiene la limitante mencionada.

Ejemplo 6.14

En una reacción química $A + B \xrightarrow{k}$ Productos, la concentración del reactante A es una función de la presión P y la temperatura T. La siguiente tabla presenta la concentración de A en gmol/l como función de estas dos variables

	T (K)			
P (kg/cm^2)	273	300	325	360
1	0.99	0.97	0.96	0.93
2	0.88	0.82	0.79	0.77
8	0.62	0.51	0.48	0.45
15	0.56	0.49	0.46	0.42
20	0.52	0.44	0.41	0.37

Calcule la variación de la concentración de A con la temperatura a P = 8 Kg/cm^2 y T = 300 K, usando un polinomio de segundo grado.

SOLUCIÓN

Lo que se busca es en sí $\frac{\partial C_A}{\partial T}\Big|_{T=300,\,P=8}$ que se puede evaluar con la ecuación 6.51. Al desarrollarla para $n = 2$ se tiene

$$\frac{dp_2(x)}{dx} = \frac{(2x-x_1-x_2)f(x_0)}{(x_0-x_1)(x_0-x_2)} + \frac{(2x-x_0-x_2)f(x_1)}{(x_1-x_0)(x_1-x_2)} + \frac{(2x-x_0-x_1)f(x_2)}{(x_2-x_0)(x_2-x_1)}$$

donde $f(x)$ representa a C_A y x a T; de tal modo que sustituyendo los tres puntos enmarcados de la tabla queda

$$\frac{dp_2(x)}{dx} = \frac{\partial C_A}{\partial T}\bigg|_{\substack{T=300\\P=8}} = \frac{(2(300)-300-325)(0.62)}{(273-300)(273-325)} + \frac{(2(300)-273-325)(0.51)}{(300-273)(300-325)}$$
$$+ \frac{(2(300)-273-300)(0.48)}{(325-273)(325-300)} = -0.0026 \frac{\text{gmol}}{1\,\text{K}}$$

Ejemplo 6.15

Obtenga la primera y segunda derivadas evaluadas en $x = 1$ para la siguiente función tabulada

Puntos	0	1	2	3	4
x	–1	0	2	5	10
$f(x)$	11	3	23	143	583

SOLUCIÓN

Al construir la tabla de diferencias divididas se tiene

			Diferencias divididas	
Puntos	x	$f(x)$	Primeras	Segundas
0	–1	11		
			–8	
1	0	3		6
			10	
2	2	23		6
			40	
3	5	143		6
			88	
4	10	583		

Obsérvese que un polinomio de segundo grado puede representar exactamente la función (ya que la segunda diferencia dividida es constante).
El polinomio de Newton de segundo grado en diferencias divididas es

$$p_2(x) = f[x_0] + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2]$$

que al derivarse da

$$\frac{dp_2(x)}{dx} = f[x_0, x_1] + (2x - x_0 - x_1) f[x_0, x_1, x_2]$$

y al derivarlo nuevamente se obtiene

$$\frac{d^2p_2(x)}{dx^2} = 2f[x_0, x_1, x_2]$$

con la sustitución de valores finalmente resulta

$$\frac{dp_2(1)}{dx} = -8 + (2(1) - (-1) - 0)(6) = 10 \text{ y } \frac{d^2p_2(1)}{dx^2} = 12$$

ALGORITMO 6.5 [illegible] de Lagrange

Para obtener una aproximación a la primera derivada de una función tabular $f(x)$ en un punto x, proporcionar los

DATOS: El grado N del polinomio de Lagrange por usar, las (N+1) parejas de valores (X(I), FX(I), I=0, 1, 2, ..., N) y el punto XD en que se desea la evaluación.

RESULTADOS: Aproximación a la primera derivada en XD:DP.

```
PASO 1.  Hacer DP = 0
PASO 2.  Hacer I = 0
PASO 3.  Mientras I ≤ N, repetir los pasos 4 a 21.
           PASO 4.  Hacer P = 1
           PASO 5.  Hacer J = 0
           PASO 6.  Mientras J ≤ N, repetir los pasos 7 a 8.
                      PASO 7.  SI I < >J
                                 Hacer P = P * (X(I) - X(J))
                      PASO 8.  Hacer J = J + 1
           PASO 9. Hacer S = 0
           PASO 10. Hacer K = 0
           PASO 11. Mientras K ≤ N, repetir los pasos 12 a 19.
                      PASO 12. SI I < > K, realizar los pasos 13 a 18.
                      PASO 13.  Hacer P1 = 1
                      PASO 14.  Hacer J = 0
                      PASO 15.  Mientras J ≤ N, repetir los
                                pasos 16 a 17.
                                PASO 16.  SI J < > I y J < > K
                                          Hacer P1 = P1 *
                                          (XD - X(J))
                                PASO 17.  Hacer J = J + 1
                      PASO 18.  Hacer S = S + P1
```

```
                PASO 19.  Hacer K = K + 1
          PASO 20.  Hacer DP = DP + FX(I)/P * S
          PASO 21.  Hacer I = I + 1
PASO 22.  IMPRIMIR DP y TERMINAR.
```

Ejercicios

6.1 La siguiente tabla representa el gasto instantáneo de petróleo crudo en un oleoducto (en miles de libras por hora). El flujo se mide a intervalos de 12 minutos.

Hora	6:00	6:12	6:24	6:36	6:48	7:00	7:12	7:24
Gasto	6.2	6.0	5.9	5.9	6.2	6.4	6.5	6.8

Hora	7:36	7:48	8:00	8:12
Gasto	6.9	7.1	7.3	6.9

¿Cuál es la cantidad de petróleo bombeado en 2 horas y 12 minutos?

Calcule el gasto promedio en ese periodo.

SOLUCIÓN

El petróleo bombeado se calcula multiplicando el gasto por el tiempo; pero como el gasto es variable, se aplica la integral siguiente

$$W = \int_0^{2.2} G\, dt \text{ lb de petróleo}$$

Integral que se puede aproximar por la regla del trapezoide (véase la Ec. 6.8).

$$I = \frac{h}{2}\left(f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n)\right)$$

en donde

$$h = \frac{2.2}{11} = 0.2$$

$f(x_i)$ = gastos en lb/hr a cada intervalo.

Al sustituir valores queda

$$W = \frac{0.2}{2}\,[6.2+2(6.0+5.9+5.9+6.2+6.4+6.5+6.8+6.9+7.1+7.3) + 6.9]$$

$$= 14.31$$

Este valor se multiplica por 1000, ya que la tabla muestra los valores del gasto en miles de libras por hora.

El gasto promedio se calcula directamente

$$W_{\text{prom}} = \frac{W}{t} = \frac{14310}{2.2} = 6500 \text{ lb/hr}$$

6.2 En el interior de un cilindro de aluminio se tiene una resistencia eléctrica que genera una temperatura $T_1 = 1200$ °F. En la superficie exterior del cilindro circula un fluido que mantiene su temperatura a $T_2 = 300$ °F. Calcule la cantidad de calor transferido al fluido por unidad de tiempo.

Datos adicionales

$R_1 = 2$ pulg, $R_2 = 12$ pulg

$L = 12$ pulg.

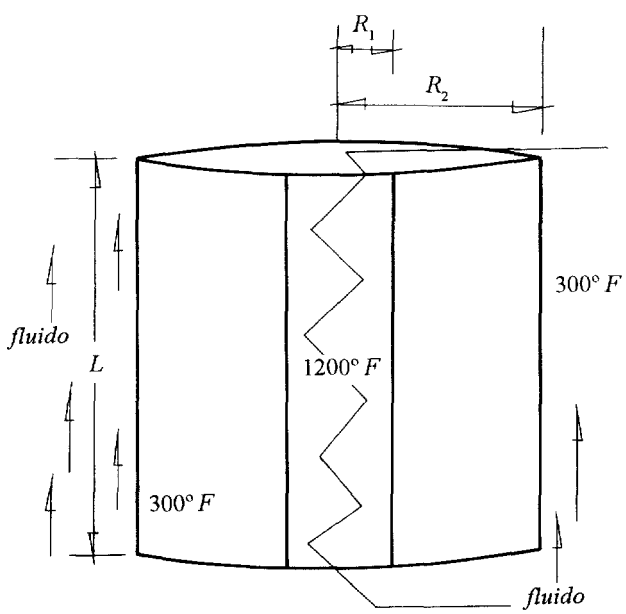
La conductividad térmica del aluminio varía con la temperatura según la tabla

k BTU/(hr pie^2 (°F/pie))	165	150	130	108
T ° F	1200	900	600	300

SOLUCIÓN

Se asume un régimen permanente y se modela el proceso con la ecuación de Fourier

$$q = -k\,A\,\frac{dT}{dr}$$



donde:

q = calor transferido al fluido en BTU/hr.
k = conductividad térmica del aluminio en $\frac{\text{BTU}}{\text{hr pie}^2\,(\,^\circ\text{F/pie}\,)}$
A = área de transferencia de calor en pie^2
T = temperatura en °F.
r = distancia radial a partir del centro del cilindro en pie.

Al separar variables, integrar y aplicar límites, la ecuación de Fourier queda

$$\int_{R_1}^{R_2} \frac{dr}{A} = -\frac{1}{q} \int_{T_1}^{T_2} k \, dT$$

Al sustituir el área de transmisión de calor A en función de la distancia radial r : A = $2\pi rL$ e integrar analíticamente el lado izquierdo se tiene

$$\frac{1}{2\pi L} \ln\left(\frac{R_2}{R_1}\right) = -\frac{1}{q} \int_{T_1}^{T_2} k \, dT$$

Sin embargo, hay que integrar numéricamente el lado derecho, ya que $k = f(T)$ está dada en forma discreta (tabulada). Así que, despejando q, se tiene

$$q = -\frac{\int_{T_1}^{T_2} k \, dT}{\frac{1}{2\pi L} \ln\left(\frac{R_2}{R_1}\right)},$$

y al integrar con la regla trapezoidal el numerador y sustituir valores se obtiene

$$q = -\frac{-124950}{\frac{1}{2\pi\,(12/12)} \ln\left(\frac{12}{2}\right)} = 438163.7 \text{ BTU/hr.}$$

6.3 Evalúe el coeficiente de fugacidad ϕ del butano a 40 atm y 200 °C con la cuadratura de Gauss-Legendre con dos puntos. El coeficiente de fugacidad está dado por la ecuación :

$$\ln \phi = \int_0^P \frac{z-1}{P} dP$$

y la relación de la presión con el factor de compresibilidad z se determinó experimentalmente y se da en la tabla

Puntos	1	2	3	4	5	6	7	8
P (atm)	5	8	15	19	25	30	35	40
z	0.937	0.887	0.832	0.800	0.781	0.754	0.729	0.697

Se sabe también que $\lim_{P \to 0} \frac{z-1}{P} = -0.006 \text{ atm}^{-1}$

SOLUCIÓN

La expresión de Gauss-Legendre para dos puntos queda*

$$\int_a^b f(t)\,dt \approx \frac{b-a}{2}\left[w_1 f\left(\frac{x_1(b-a)+b+a}{2}\right) + w_2 f\left(\frac{x_2(b-a)+b+a}{2}\right)\right]$$

donde $w_1 = w_2 = 1$; $x_1 = 0.5773502692$; $x_2 = -0.5773502692$

Con el cálculo de los argumentos de la función f se tiene

$$\frac{x_1(b-a)+b+a}{2} = \frac{0.5773502692\,(40-0)+40+0}{2} = 31.547$$

$$\frac{x_2(b-a)+b+a}{2} = \frac{-0.5773502692\,(40-0)+40+0}{2} = 8.453$$

El cálculo del factor de compresibilidad z a los valores de P = 31.547 y P = 8.453 se realiza por interpolación.

Con los puntos (6), (7) y (8) de la tabla y alguno de los métodos del capítulo 5 se obtiene $z(31.547) = 0.746$, y con los puntos (1), (2) y (3) se obtiene $z(8.453) = 0.881$.

Con los valores de z y P en los dos puntos, se calcula el valor de la función por integrar

$$\frac{z-1}{\mathrm{P}} = \frac{0.746-1}{31.457} = -0.00805$$

$$\frac{z-1}{\mathrm{P}} = \frac{0.881-1}{8.453} = -0.01408$$

Se sustituyen valores en la ecuación de Gauss-Legendre y se tiene:

$$\ln \phi = \int_0^{40} \frac{z-1}{\mathrm{P}}\,d\mathrm{P} = \frac{40-0}{2}\left[1(-0.00805) + 1(-0.01408)\right] = -0.4426$$

de donde $\phi = 0.6424$

Obsérvese que basta tener el valor experimental de z a las presiones de 8.453 y 31.547, que en este ejemplo se determinaron por interpolación. Es importante señalar que procediendo a la inversa; es decir, calculando los valores de las presiones a las que se requiere el valor de z y después determinando experimentalmente dichos valores, se ahorra un considerable número de experimentos (2 contra 8 en este caso). Esto constituye una de las ventajas más importantes del método de la cuadratura de Gauss-Legendre.

6.4 Encuentre el **centro de masa** de una lámina rectangular de $2\pi \times \pi$, suponiendo que la densidad en un punto P(x, y) de la lámina está dado por $\rho(x,y) = e^{-(x^2+y^2)/2}$.

*Véase el problema 6.21 al final de este capítulo.

SOLUCIÓN

Por definición, los momentos de inercia con respecto al eje x y y, respectivamente, están dados por :

$$M_x = \iint_R x\rho(x,y)\,dx\,dy, \quad M_y = \iint_R y\rho(x,y)\,dx\,dy$$

y el centro de masa de la lámina es el punto $(\bar{x}, \bar{y})$ tal que

$$\bar{x} = \frac{M_x}{M}, \quad \bar{y} = \frac{M_y}{M}$$

donde $M = \iint_R \rho(x,y)\,dx\,dy$

$$M = \iint_R \rho(x,y)\,dx\,dy$$

Para facilitar las integraciones, la lámina se pone como se muestra en la figura 6.12, con lo que

$$R = \{(x,y) : 0 \le x \le 2\pi, 0 \le y \le \pi\}$$

Primero se obtiene M con el método de cuadratura de Gauss empleando tres puntos

$$M = \int_0^{\pi} \int_0^{2\pi} e^{-(x^2+y^2)/2}\,dx\,dy = 1.56814$$

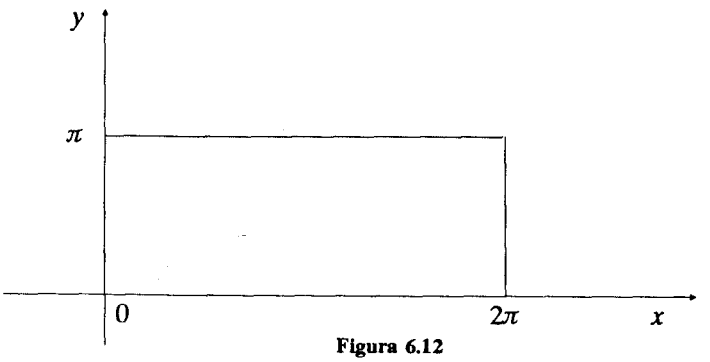


Figura 6.12

Después se calculan M_x y M_y, donde

$$M_x = \int_0^{\pi} \int_0^{2\pi} x\, e^{-(x^2+y^2)/2}\,dx\,dy \approx 1.2556$$

$$M_y = \int_0^{\pi} \int_0^{2\pi} y\, e^{-(x^2+y^2)/2}\,dx\,dy \approx 1.24449$$

Finalmente

$$\bar{y} \approx \frac{1.24449}{1.56814} = 0.7936$$

$$\bar{x} \approx \frac{1.2556}{1.56814} = 0.8007 \ ,$$

el centro de masa es el punto del primer cuadrante (0.8007, 0.7936).

6.5 Las integrales del tipo $\int_0^\infty e^{-x} f(x)\, dx$ se conocen como **integrales impropias** y se pueden aproximar, si su límite existe, por la **cuadratura de Gauss-Laguerre**

$$\int_0^\infty e^{-x} f(x)\, dx \approx \sum_{i=1}^{n} H_i f(x_i) \qquad (1)$$

donde x_i es la i-ésima raíz del polinomio de Laguerre $L_n(x)$ y

$$H_i = -\frac{[(n-1)!]^2}{L_n'(x_i)\, L_{n-1}(x_i)}; \quad i=1,2,\ldots,n$$

Se dan a continuación los primeros polinomios de Laguerre

$$L_0(x) = 1 \ ; \ L_1(x) = 1-x \ ; \ L_2(x) = 2-4x + x^2$$
$$L_3(x) = 6-18x + 9x^2-x^3 \ ; \ L_4(x) = 24-96x + 72x^2-16x^3 + x^4$$
$$L_5(x) = 120 -600x + 600x^2 - 200x^3 + 25x^4 - x^5$$

y la ecuación

$$L_{i+1}(x) = (1 + 2i - x)\, L_i(x) - i^2 L_{i-1}(x)$$

que permite obtener el polinomio de Laguerre de grado $i + 1$ en términos de los polinomios de Laguerre de grado i e $i-1$.

Aproxime $\int_0^\infty e^{-x}$ sen $x\, dx$ con $n=2$

SOLUCIÓN

Como $n = 2 : L_2(x) = 2 - 4x + x^2 \ ; \ L_2'(x) = -4 + 2x$

Las raíces de $L_2(x)$ son: $x_1 = 2 - \sqrt{2}$, $x_2 = 2 + \sqrt{2}$
Con la sustitución en H_i se tiene:

$$H_1 = -\frac{[(2-1)!]^2}{(-4+2(2-\sqrt{2}))\,(1-(2-\sqrt{2}))} = \frac{2+\sqrt{2}}{4}$$

$$H_2 = -\frac{[(2-1)!]^2}{(-4+2(2+\sqrt{2}))\,(1-(2+\sqrt{2}))} = \frac{2-\sqrt{2}}{4}$$

y la integral queda entonces

$$\int_0^\infty e^{-x} \operatorname{sen} x\, dx \approx \frac{1}{4}\left[(2+\sqrt{2}\,) \operatorname{sen} (2-\sqrt{2}\,) + (2-\sqrt{2}\,) \operatorname{sen} (2+\sqrt{2}\,)\right] = 0.43246$$

En general, este proceso de integración puede programarse con una expresión del tipo 6.30

$$\int_0^\infty e^{-ax} f(x)\, dx \approx \frac{1}{a} \sum_{i=1}^{n} w_i f(x_i/a)$$

donde los pesos w_i y las abscisas x_i para $2 \leq n \leq 5$ están dadas en la tabla 6.1

Tabla 6.1. Coeficientes y abscisas para la integración por cuadratura de Gauss-Laguerre.

Número de puntos n	Abscsisas x_i	Coeficientes w_i
2	0.585786	0.853553
	3.414214	0.146447
3	0.415775	0.711093
	2.294280	0.278518
	6.289945	0.0103893
4	0.322548	0.603154
	1.745761	0.357419
	4.536620	0.0388879
	9.395071	0.000539295
5	0.263560	0.521756
	1.413403	0.398667
	3.596426	0.0759424
	7.085810	0.00361176
	12.640801	0.00002337

6.6 De la gráfica de un diagrama de Moliere del amoniaco se obtienen los siguientes datos de temperatura (T) contra presión (P) a entalpía constante (H = 700 BTU/Lb).

Puntos	0	1	2	3	4
T (°F)	175	200	225	250	275
P (*psia*)	100	270	450	640	850

Calcule el coeficiente de Joule-Thompson a una presión de 270 psia

a) Mediante la derivada de la fórmula generalizada del polinomio de Lagrange del ejemplo 6.13 con los primeros cuatro puntos.
b) Mediante la derivada analítica de una curva empírica polinomial de segundo grado calculada con mínimos cuadrados usando todos los puntos.

SOLUCIÓN

El coeficiente de Joule-Thompson está definido como la derivada parcial de la temperatura con respecto a la presión a entalpía constante, o sea

$$\mu = \left(\frac{\partial T}{\partial P}\right)_H$$

a) La fórmula del ejemplo 6.13 se desarrolla para $n = 3$ y se obtiene

$$\frac{dp_3(x)}{dx} = \left[3x^2 - 2(x_1+x_2+x_3)x + (x_1x_2+x_1x_3+x_2x_3)\right] \frac{f(x_0)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)}$$

$$+ \left[3x^2 - 2(x_0+x_2+x_3)x + (x_0x_2+x_0x_3+x_2x_3)\right] \frac{f(x_1)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}$$

$$+ \left[3x^2 - 2(x_0-x_1-x_3)x + (x_0x_1+x_0x_3+x_1x_3)\right] \frac{f(x_2)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)}$$

$$+ \left[3x^2 - 2(x_0-x_1-x_2)x + (x_0x_1+x_0x_3+x_1x_2)\right] \frac{f(x_3)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}$$

donde x representa la presión y $p_3(x)$ la temperatura. Al sustituir x por 270 psia así como los valores de los puntos dados, se obtiene

$$\mu = \left(\frac{\partial T}{\partial P}\right)_H \approx 0.1429 \text{ °F / psia}$$

b) El sistema de ecuaciones 5.64 se resuelve usando los cinco puntos de la tabla a fin de obtener los coeficientes del polinomio de segundo grado que mejor aproxima la función tabulada

$$a_0 = 159.5134, \qquad a_1 = 0.156799, \qquad a_2 = -0.2453 \times 10^{-4}$$

que sustituidos dan

$$T(P) \approx 159.5134 + 0.156799\,P - 0.2453 \times 10^{-4}\,P^2$$

cuya derivada es

$$\left(\frac{\partial T}{\partial P}\right)_H \approx 0.156799 - 2(0.2453 \times 10^{-4})\,P$$

que evaluada en P = 270 resulta

$$\left(\frac{\partial T}{\partial P}\right)_H \approx 0.1436 \text{ °F/psia}$$

6.7 Encuentre la primera derivada numérica de xe^x en el punto $x = 1$, usando un polinomio de aproximación de segundo grado. Estime el error cometido.

SOLUCIÓN

Lo más conveniente es emplear la ecuación 6.49

$$\frac{df(x)}{dx}\Big|_{x_1} = \frac{1}{2h}[f(x_2) - f(x_0)] - \frac{h^2}{6}\frac{d^3f(x)}{dx^3}\Big|_{\xi}; \quad \xi \in (x_0, x_2)$$

También se observa que el término del error en general es proporcional al valor de h, por lo que, si es posible, deberá tomarse muy pequeño. Se tomará en este caso $h = 0.5$. Con esto

$$x_0 = 0.5, \qquad x_1 = 1, \qquad x_2 = 1.5$$
$$f(x_0) = 0.82436, \qquad f(x_1) = 2.71828, \qquad f(x_2) = 6.72253$$

Se sustituye valores

$$\frac{d(xe^x)}{dx}\Big|_{x_1=1} = \frac{1}{2(0.5)}[6.72253 - 0.82436] - \frac{0.5^2}{6}(xe^x + 3e^x)\Big|_{\xi}$$
$$= 5.89817 - 0.04166\,(xe^x + 3e^x)\Big|_{\xi}$$

donde 5.89819 es la aproximación a la primera derivada y el factor $xe^x + 3e^x$ es la derivada de tercer orden de xe^x. Como se desconoce ξ y $xe^x + 3e^x$ es una función creciente, puede calcularse en $x = 1.5$. De esta manera se obtiene el valor máximo posible y, por ende, el valor máximo para el término del error resulta

$$0.04166(1.5\,e^{1.5} + 3\,e^{1.5}) = 0.84032$$

6.8 Demuestre que

$$\frac{d}{dx}\left[\prod_{j=0}^{n}(x - x_j)\right]\Big|_{x=x_i} = \prod_{\substack{j=0 \\ j\neq i}}^{n}(x_i - x_j)$$

SOLUCIÓN

Si $n = 1$ se tiene

$$\frac{d}{dx}[(x-x_0)(x-x_1)]\Big|_{x=x_i} = [(x-x_0)+(x-x_1)]\Big|_{x=x_i}\prod_{\substack{j=0 \\ j\neq i}}^{1}$$

Si $n = 2$ se tiene

$$\frac{d}{dx}[(x-x_0)(x-x_1)(x-x_2)]\Big|_{x=x_i} = \left[(x-x_2)\frac{d}{dx}[(x-x_0)(x-x_1)]\right.$$
$$\left. + (x-x_0)(x-x_1)\right]\Big|_{x=x_i}$$

$$= [(x-x_2)(x-x_0) + (x-x_2)(x-x_1) + (x-x_0)(x-x_1)] \Big|_{x=x_i}$$

$$= \prod_{\substack{j=0 \\ j\neq i}}^{2} (x_i - x_j)$$

y por inducción puede llegarse a

$$\frac{d}{dx}[(x-x_0)(x-x_1)\ldots(x-x_n)]\Big|_{x=x_i} = \Big[(x-x_n)\frac{d}{dx}[(x-x_0)(x-x_1)\ldots(x-x_{n-1})]$$

$$+ (x-x_0)(x-x_1)\ldots(x-x_{n-1})\Big]\Big|_{x=x_i}$$

$$= \prod_{\substack{j=0 \\ j\neq i}}^{2} (x_i - x_j)$$

6.9 En la siguiente tabla

t	0	1	2	3	4	5
T	93.1	85.9	78.8	75.1	69.8	66.7

T representa la temperatura (°C) de una salmuera utilizada como refrigerante y t (min) es el tiempo. Encuentre la velocidad de enfriamiento en los tiempos $t = 2.5$ y $t = 4$ min.

SOLUCIÓN

Al emplear la ecuación 6.41 con $n = 2$ se tiene

$$\frac{dp_2(t)}{dt} = \frac{(2t - t_0 - t_1 - 2h)\,T_0}{2h^2} + \frac{(2t_0 - 4t + 2t_1 + 2h)\,T_1}{2h^2} + \frac{(2t - t_0 - t_1)\,T_2}{2h^2}$$

Se toman $t_0 = 1$, $t_1 = 2$, $t_2 = 3$ y $h=1$ y se tiene, para $t = 2.5$

$$\frac{dp_2(t)}{dt} = \frac{[2(2.5) - 1 - 2 - 2(1)](85.9)}{2(1)^2}$$

$$+ \frac{[2(1) - 4(2.5) + 2(2) + 2(1)](78.8)}{2(1)^2} + \frac{(2(2.5) - 1 - 2)(75.1)}{2(1)^2} = -3.7$$

Al tomar $t_0 = 2, t_1 = 3$ y $t_2 = 4$ se tiene, para $t = 2.5$

$$\frac{dp_2(t)}{dt} = \frac{[2(2.5)-2-3-2(1)](78.8)}{2(1)^2}$$

$$+ \frac{[2(2)-4(2.5)+2(3)+2(1)](75.1)}{2(1)^2} + \frac{(2(2.5)-2-3)(69.8)}{2(1)^2} = -3.7$$

Estos valores confirman que la función tabular se comporta como una parábola (n=2); por lo tanto, el grado seleccionado es adecuado.

Se deja al lector repetir estos cálculos para t=4 min.

6.10 La siguiente tabla muestra las medidas observadas en una curva de imantación del hierro.

β	5	6	7	8	9	10	11	12
μ	1090	1175	1245	1295	1330	1340	1320	1250

En ella β es el número de kilolíneas por cm^2 y μ la permeabilidad. Encuentre la permeabilidad máxima.

SOLUCIÓN

Como la permeabilidad máxima registrada en la tabla es de 1340, correspondiente a β=10, se utilizan los puntos de abscisas $\beta_0 = 9, \beta_1 = 10, \beta_2 = 11$ para obtener un polinomio de segundo grado, por el método de aproximación polinomial simple

$$a_0 + a_1(9) + a_2(9)^2 = 1330$$

$$a_0 + a_1(10) + a_2(10)^2 = 1340$$

$$a_0 + a_1(11) + a_2(11)^2 = 1320$$

Al resolver se tiene $a_0 = -110, a_1 = 295$ y $a_2 = -15$, por lo que el polinomio es

$$\mu = -110 + 295\beta - 15\beta^2$$

Para obtener la permeabilidad máxima, se deriva e iguala con cero este polinomio

$$\frac{d\mu}{d\beta} = 295 - 30\beta = 0$$

Al despejar $\beta = \frac{295}{30} = 9.83333$

de donde

$$\mu_{max} = -110 + 295(9.83333) - 15(9.83333)^2 = 1340.416$$

Problemas

6.1 Emplee la ecuación 6.5 con $n=3$ para obtener la ecuación de Simpson 3/8 (véase Ecs. 6.6).

6.2 Mediante el método de Simpson 3/8 aproxime las integrales del ejemplo 6.1. Compare los resultados con los obtenidos en los ejemplos 6.1. y 6.2.

6.3 Siguiendo las ideas que llevaron a las ecuaciones 6.8 y 6.10, encuentre la ecuación correspondiente a usar la fórmula de Simpson 3/8 sucesivamente.

6.4 Con el algoritmo obtenido en el problema anterior, integre la función

$$\frac{1}{\sqrt{2\pi}} \; e^{-x^2/2}$$

entre los límites $a = -1$ y $b = 1$. Compare el resultado con los valores obtenidos en el ejemplo 6.5.

6.5 En el gasoducto Cactus, Tab. a Reynosa, Tamps. se determina el gasto W (Kg/min) y su contenido de azufre S (en porciento) periódicamente durante el día. Los resultados se presentan en la tabla

t (hr)	0	4	8	12	15	20	22	24
W (kg/min)	20	22	19.5	23	21	20	20.5	20.8
S (%)	0.30	0.45	0.38	0.35	0.30	0.43	0.41	0.40

a) ¿Cuál es el gasto promedio diario?
b) ¿Qué cantidad de gas se bombea en 24 horas?
c) ¿Cuál es el contenido de azufre (%) promedio diario?
d) ¿Qué cantidad de azufre se bombea en 24 horas?

6.6 Integre la función de Bessel de primera especie y orden 1

$$J_1(x) = \sum_{k=0}^{\infty} \frac{(-1)^k (x/2)^{2k+1}}{k!\,(1+k+1)}$$

con el método de Simpson compuesto (aplicado siete veces) en el intervalo [0, 7]; esto es,

$$\int_0^7 J_1(x)\, dx$$

Sugerencia: Consulte las tablas de funciones de Bessel.

6.7 Obtenga

$$\int_1^2 \frac{h^3\, dh}{(1 + h^{1/2})}$$

6.8 Obtenga

$$\int_1^3 x e^{e^x} dx$$

6.9 Elabore un subprograma para integrar una función analítica por el método de Simpson 3/8 compuesto, usando sucesivamente 3, 6, 12, 24, 48, ..., 3072 subintervalos. Compruébelo con la función del ejemplo 6.5.

6.10 De acuerdo con las ideas acerca del análisis del error de truncamiento en la aproximación trapezoidal, analice dicho error en la aproximación de Simpson 1/3.

Sugerencia: La expresión a que debe llegarse es

$$\left| E_T \right| \leq \frac{n}{2} \frac{h^5}{90} \mathrm{M} = nh \frac{h^4}{180} \mathrm{M} = (b-a) \frac{h^4}{180} \mathrm{M},$$

donde $|f^{\mathrm{IV}}(x_i)| \leq \mathrm{M}$ y n es el número de subintervalos en que se divide $[a, b]$; de donde, el error de truncamiento en el método de Simpson 1/3 es proporcional a h^4, lo cual conjuntamente con la ecuación 6.10 se expresa

$$\int_a^b f(x)\,dx = \frac{h}{3}\left[f(x_0) + 4 \sum_{\substack{i=1 \\ \Delta i=2}}^{n-1} f(x_i) + 2 \sum_{\substack{i=2 \\ \Delta i=2}}^{n-2} f(x_i) + f(x_n)\right] + \mathrm{O}(h^4)$$

6.11 Mediante la ecuación 6.18 encuentre una cota para el error de truncamiento al integrar la función $e^{-x^2/2}$ entre los límites [−1, 1], usando 2, 4, 8, 16, ..., 1024 subintervalos.

6.12 Emplee la integración de Romberg a fin de evaluar $\mathrm{I}_2^{(2)}$ para las siguientes integrales definidas

a) $\int_{-1}^1 e^{-x^2/2}\,dx$ b) $\int_0^1 x^3 e^x dx$

c) $\int_2^3 \frac{dx}{x}$ d) $\int_1^3 \ln x\,dx$

6.13 Con el criterio de convergencia siguiente

$$\left| \mathrm{I}_k^{(m)} - \mathrm{I}_{k-1}^{(m+1)} \right| \leq 10^{-3},$$

aproxime las integrales que se dan a continuación

a) $\int_0^1 e^{\cos 2\pi x}\,dx$ b) $\int_1^2 \frac{\cos x}{\sqrt{x}}\,dx$

c) $\int_0^2 e^{x/\pi} \cos kx\,dx$ con $k = 1, 2.$

En el inciso c) compare con la solución analítica.

6.14 Pruebe que en la integración de Romberg con $h_2 = h_1/2$ [Ec. 6.21], $\mathrm{I}_k^{(1)} = \mathrm{S}$, donde S es la aproximación de Simpson compuesta [Ec. 6.10] empleando 2^k subintervalos.

6.15 Estime las siguientes integrales

a) $\int_0^1 \text{sen}\,(\,101\,\pi x\,)\,dx$

b) $\int_0^1 f\,(x)\,dx$ con $f\,(x) = \begin{cases} \dfrac{\text{sen}\,x}{x} & \text{en } x \neq 0 \\ 1 & \text{en } x = 0 \end{cases}$

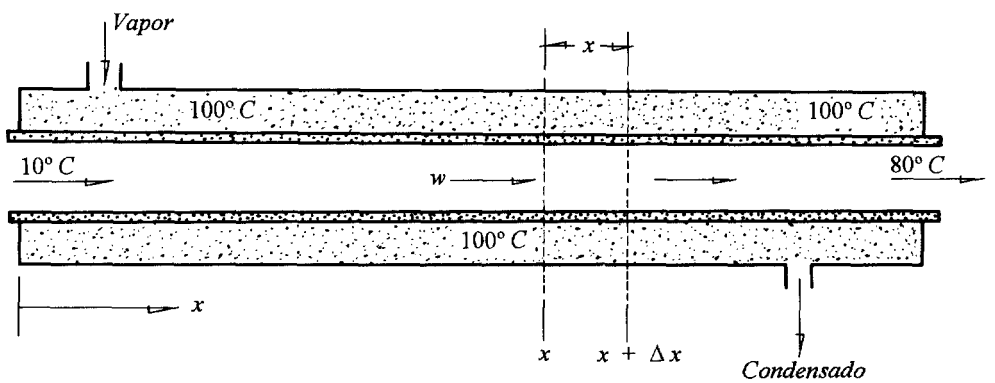
con una aproximación de 10^{-5}
Sugerencia: Emplee como criterio de convergencia

$$\left| I_k^{(m)} - I_k^{(m+1)} \right| \leq 10^{-6}$$

con

$$\left| I_{k+1}^{(m+1)} - I_{k+1}^{(m+2)} \right| \leq 10^{-6}$$

6.16 En un cambiador de calor de tubos concéntricos se calienta nitrobenceno con un gasto w = 3000 lb/hr en el tubo interior con diámetro Di = 3 pulg. Entre los tubos se circula vapor saturado que mantiene la temperatura de los tubos a temperatura constante Ts = 100°C. ¿Cuál será la longitud necesaria de los tubos para calentar el nitrobenceno desde 10 hasta 80°C?



El modelo del proceso se obtiene mediante un balance de calor a regimen permanente en un elemento diferencial de longitud Δx

Acumulación = entrada − salida + transmisión

$$0 = w\,\text{CpT}\,|_x - w\,\text{CpT}\,|_{x+\Delta x} + \text{Di}\,\pi\,\Delta x h\,(\text{Ts} - \text{T})$$

al dividir entre Δx y hacer $\Delta x \to 0$, se obtiene en el límite

$$w\,\text{Cp}\,\frac{dT}{dx} = \text{Di}\,\pi\,h\,(\text{Ts} - \text{T})$$

y al separar variables queda

$$\frac{w}{\mathrm{Di}\,\pi}\int_{10}^{80}\frac{\mathrm{Cp}\,dT}{h(\mathrm{Ts}-\mathrm{T})}=\int_0^{\mathrm{L}}dx=\mathrm{L}.$$

Donde h es el coeficienete de transmisión de calor en BTU/(hr pie °F) y se calcula con la expresión

$$h=0.023\,\mathrm{Re}^{0.8}\,\mathrm{Pr}^{0.33}.$$

Obsérvese que Re, Cp y Pr son funciones de T.

6.17 A principios de siglo, Lord Rayleigh resolvió el problema de la **destilación binaria simple** (una etapa) **por lotes,** con la ecuación que ahora lleva su nombre

$$\int_{L_i}^{L_f}\frac{d\mathrm{L}}{\mathrm{L}}=\int_{x_i}^{xf}\frac{dx}{y-x}$$

donde L son los moles de la mezcla líquida en el hervidor, x las fracciones mol del componente más volátil en la mezcla líquida y y las fracciones mol de su vapor en equilibrio. Los subíndices i y f se refieren al estado inicial y final.
Calcule qué fracción de un lote es necesario destilar en una mezcla binaria para que x cambie de $x_i = 0.7$ a $x_f = 0.4$. La relación de equilibrio está dada por la ecuación

$$y=\frac{\alpha x}{1+(\alpha-1)x}$$

donde α es la **volatilidad relativa** de los componentes y es una función de x según la siguiente tabla (para una mezcla dada)

x	0.70	0.65	0.60	0.55	0.50	0.45	0.40
α	2.20	2.17	2.13	2.09	2.04	1.99	1.94

6.18 La integral $\int_{-\pi}^{\pi}\frac{\operatorname{sen}x}{x}\,dx$ puede presentar serias dificultades.
Estudie cuidadosamente el integrando y aproxime dicha integral empleando alguno de los métodos vistos.

6.19 Ensaye varios métodos de integración numérica para aproximar

$$\int_{-1}^{1}\frac{x^2\,dx}{\sqrt{1-x^2}}$$

6.20 Sea la función $f(x)$ definida en (0, 1) como sigue

$$f(x)=\begin{cases}x & 0\le x\le 0.5\\ 1-x & 0.5\le x\le 1,\end{cases}$$

aproxime numéricamente $\int_0^1 f(x)\,dx$ utilizando

a) El método trapezoidal aplicado una vez en (0, 1)
b) El método trapezoidal aplicado una vez en (0, 0.5) y otra en (0.5, 1)
c) El método de Simpson 1/3 aplicado una vez en (0, 1).

Compare los resultados con el valor analítico y explique las diferencias.

6.21 Demuestre que la expresión general para integrar por Gauss-Legendre puede ponerse en la forma

$$\int_a^b f(t)\,dt \approx \frac{b-a}{2} \sum_{i=1}^{n} w_i\, f\left[\frac{x_i(b-a)+b+a}{2}\right]$$

donde w_i, x_i, $i = 1, 2, ..., n$ dependen de n y están dados en la tabla 6.2

6.22 Use la cuadratura de Gauss con $n = 3$ para aproximar las integrales de los problemas 6.18 y 6.19.

6.23 Dada la función $f(x)$ en forma tabular

x	0	41	56	95	145	180	212	320
$f(x)$	0	1.18	1.65	2.70	3.75	4.10	4.46	5.10

encuentre

$$\int_0^{320} x^2 f(x)\,dx$$

usando la cuadratura de Gauss con varios puntos

6.24 Calcule el cambio de entropía ΔS que sufre un gas ideal a presión constante al cambiar su temperatura de 300 a 380 K. Utilice la cuadratura gaussiana de tres puntos.

$$\int_{T_1}^{T_2} C_p\,\frac{dT}{T}$$

T (K)	280	310	340	370	400
Cp (cal/mol K)	4.87	5.02	5.16	5.25	5.30

6.25 Modifique el programa del ejemplo 6.11 de modo que se puedan integrar funciones dadas en forma discreta o tabular.
Sugerencia: Vea el programa de interpolación de Lagrange en el ejercicio 5.5.

6. 26 Una partícula de masa m se mueve a través de un fluido sujeta a una resistencia R que es función de la velocidad v de m. La relación entre la resistencia R, la velocidad v, y el tiempo t está dada por la ecuación

$$t = \int_{v_0}^{v_f} \frac{m}{R(v)}\,dv$$

Supóngase que $R(v) = -v\sqrt{v}+0.0001$ para un fluido particular. Si $m = 10$ kg y $v_0 = 10$ m/s, aproxime el tiempo requerido para que la partícula reduzca su velocidad a $v_f = 5$ m/s, usando el método de cuadratura de Gauss con dos y tres puntos.

6.27 Aproxime las siguientes integrales usando la cuadratura de Gauss-Laguerre. Consulte el ejercicio 6.5.

a) $\int_0^{\infty} e^{-3x} \ln x\,dx$

b) $\int_0^{\infty} e^{-2x}(\tan x + \operatorname{sen} x)\,dx$

c) $\int_0^{\infty} e^{-x} x^3\,dx$

d) $\int_0^{\infty} e^{-x} 3^x\,dx$

e) $\int_0^{\infty} e^{-3x}\,dx$

6.28 Las integrales del tipo $\int_a^{\infty} f(x)\,dx$, con a >0 se conocen como integrales impropias y se pueden aproximar, si su límite existe, por los métodos de cuadratura haciendo el cambio de variable $t = x^{-1}$. Con este cambio el integrando y los límites pasan a ser

Como $t = x^{-1}, x = t^{-1}$ y $dx = -1/t^2\,dt$.

El integrando queda $f(x)\,dx = (-1/t^2)\,f(t^{-1})\,dt$.

Los límites pasan a

Como $x = a, t = 1/a$

y

$$x = \infty, t = 1/\infty = 0$$

Al sustituir queda

$$\int_a^{\infty} f(x)\,dx = -\int_{1/a}^{0} t^{-2} f(t^{-1})\,dt = \int_0^{1/a} t^{-2} f(t^{-1})\,dt$$

que ya puede aproximarse por los métodos vistos en este capítulo.
Aplique estas ideas para aproximar las siguientes integrales

a) $\int_5^{\infty} x^{-3}\,dx$ *b*) $\int_1^{\infty} x^{-4} \operatorname{sen}(1/x)\,dx$

c) $\int_{10}^{\infty} x^{-2} e^{-3x} \cos(4/x)\,dx$

Utilice el método numérico que considere más conveniente.

6.29 Elabore un algoritmo correspondiente al algoritmo 6.3 usando la cuadratura de Gauss-Laguerre (véase Ejer. 6.5).

6.30 Aproxime las integrales siguientes

a) $\int_0^1 \int_1^2 x\,e^{xy}\,dy\,dx$ b) $\int_1^2 \int_0^1 x\,e^{xy}\,dx\,dy$

c) $\int_1^3 \int_4^8 \operatorname{sen} x \cos y\,dx\,dy$ d) $\int_0^1 \int_0^1 x\,y\,dx\,dy$

empleando el método de Simpson 1/3 dividiendo el intervalo (a, b) del eje x en n (par) subintervalos y el intervalo (c, d) del eje y en m (par) subintervalos. La ecuación por utilizar es

$$\int_c^d \int_a^b f(x,y)\,dx\,dy \approx \frac{h_1 h_2}{9}\Bigg(f(x_0,y_0) + f(x_0,y_m) + f(x_n,y_0) + f(x_n,y_m)$$

$$+ 4 \sum_{\substack{j=1 \\ \Delta j=2}}^{m-1} [f(x_0,y_j) + f(x_n,y_j)] + 2 \sum_{\substack{j=2 \\ \Delta j=2}}^{m-2} [f(x_0,y_j) + f(x_n,y_j)]$$

$$+ 4 \sum_{\substack{i=1 \\ \Delta i=2}}^{n-1} \left[f(x_i, y_0) + 4 \sum_{\substack{j=1 \\ \Delta j=2}}^{m-1} f(x_i, y_j) + 2 \sum_{\substack{j=2 \\ \Delta j=2}}^{m-2} f(x_i, y_j) + f(x_i, y_m) \right]$$

$$+ 2 \sum_{\substack{i=2 \\ \Delta i=2}}^{n-2} \left[f(x_i, y_0) + 4 \sum_{\substack{j=1 \\ \Delta j=2}}^{m-1} f(x_i, y_j) + 2 \sum_{\substack{j=2 \\ \Delta j=2}}^{m-2} f(x_i, y_j) + f(x_i, y_m) \right] \Bigg)$$

6.31 Aproxime las integrales del problema 6.30 empleando la cuadratura de Gauss-Legendre

$$\int_c^d \int_a^b f(x, y)\, dx dy \approx \frac{(b-a)(d-c)}{4} \sum_{j=1}^{m} \sum_{i=1}^{n} w_j\, w_i f\left[\frac{b-a}{2} t_i + \frac{b+a}{2}, \frac{d-c}{2} u_j + \frac{c+d}{2} \right]$$

donde w_i o w_j son los coeficientes w_i dados en la tabla 6.2, t_i o u_j son las abscisas z_i dadas en la tabla 6.2 y n y m son los números de puntos por usar en los ejes x y y, respectivamente.

6.32 Mediante el método de Simpson 1/3 aproxime las integrales

a) $\int_0^{\pi} \int_0^{y} y \text{ sen } x\, dx\, dy$ *b)* $\int_0^1 \int_{\sqrt{x}}^1 dy\, dx$

c) $\int_1^2 \int_x^{x^2} dy\, dx$ d) $\int_0^2 \int_1^{e^y} dx\, dy$

6.33 En el estudio de integrales dobles, un problema típico es demostrar que

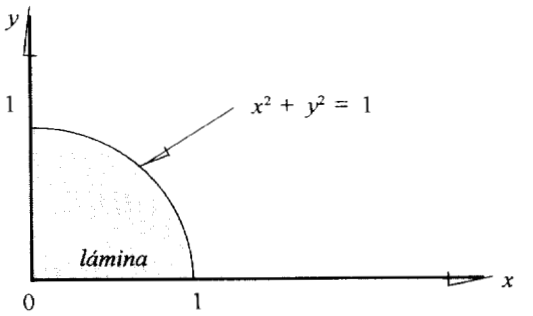
$$\int_0^R e^{-x^2} dx = \int_0^R \int_0^R e^{-x^2-y^2} dy\, dx$$

Demuéstrelo numéricamente con R = 1. Utilice el método de Simpson 1/3.

6.34 Elabore un algoritmo para aproximar integrales dobles empleando el método de cuadratura de Gauss-Legendre.

6.35 Encuentre el centro de masa de una lámina cuya forma se encuentra en la figura adjunta, suponiendo que la densidad en un punto $\rho(x, y)$ de la lámina está dada por

$$\rho(x, y) = x \text{ sen } y$$



6.36 La expresión $\int_c^d \int_a^b dx\, dy$ representa el área del rectángulo cuyos vértices son a, b, c y d (corrobórelo), de modo que $\int_c^d \int_a^b f(x, y)dx\, dy$ representa el volumen del cuerpo cuya base es el rectángulo a, b, c, d y cuya altura para cualquier punto (x, y) dentro de dicho rectángulo es $f(x, y)$. Aproximar el volumen de los siguientes cuerpos

a) $f(x, y) = \operatorname{sen} x + e^{xy}$; $(a, b, c, d) = (0, 4, 1, 3)$

b) $f(x, y) = \operatorname{sen} \pi x \cos \pi y$; $(a, b, c, d) = (0, \pi/4, 0, \pi/4)$

6.37 Emplee las ideas que llevaron a las ecuaciones 6.41 y 6.42 para obtener la aproximación de una función tabulada por un polinomio de tercer grado y su primera y segunda derivadas.

6.38 La ecuación de estado de Redlich-Kwong es

$$[P + \frac{a}{T^{0.5} V(V+b)}](V-b) = RT$$

donde $a = 17.19344$ y $b = 0.02211413$ para el oxígeno molecular.
Si $T = 373.15$ K, se obtiene la siguiente tabla de valores

Puntos	0	1	2	3
P (atm)	30.43853	27.68355	25.38623	23.44122
V (*l/gmol*)	1.0	1.1	1.2	1.3

a) Calcule la dP/dV cuando $V = 1.05$ *l* utilizando las ecuaciones 6.40 y 6.41 y compárelo con el valor de la derivada analítica.
b) Proceda como en el inciso anterior, pero ahora aplique la ecuación 6.51 con $n = 1$ y $n = 2$.

6.39 Calcular $\frac{\partial CA}{\partial P}\Big|_{T=325, P=10}$ utilizando la información del ejemplo 6.14.

6.40 Obtenga la segunda derivada evaluada en $x = 3.7$ para la función que se da enseguida

Puntos	0	1	2	3	4	5
x	1	1.8	3	4.2	5	6.5
$f(x)$	3	4.34536	6.57735	8.88725	10.44721	13.39223

Utilice un polinomio de Newton en diferencias divididas para aproximar $f(x)$.

6.41 Dada la función $f(x) = x e^x + e^{x}$, aproxime $f'(x)$, $f''(x)$ en $x = 0.6$, empleando los valores de $h = 0.4, 0.1, 0.0002$ con $n = 1, 2, 3$ para cada h. Compare los resultados con los valores analíticos.

6.42 Elabore un programa que aproxime la primera derivada de una función dada en forma tabular; usando el algoritmo 6.5.

6.43 Encuentre la primera derivada numérica de $x \ln x$ en el punto $x = 2$, usando un polinomio de aproximación de tercer grado. Estime el error cometido.

6.44 Dada la tabla

x	0.2	0.3	0.4	0.5	0.6
$f(x)$	0.24428	0.40496	0.59673	0.82436	1.09327
$f'(x)$		1.75482	2.08855	2.47308	

calcule $f'(x)$ para $x = 0.3$, 0.4 y 0.5 con $n = 2$ (Ec. 6.41) y compare con los valores analíticos dados en la tabla.

6.45 En la tabla siguiente, x es la distancia en metros que recorre una bala a lo largo de un cañón en t segundos. Encuentre la velocidad de la bala en $x = 3$

x	0	1	2	3	4	5
t	0	0.0359	0.0493	0.0596	0.0700	0.0786

6.46 Dado un circuito con un voltaje E(t) y una inductancia L, la primera ley de Kirchhoff que lo modela es

$$E = L\, di/dt + R\, i$$

donde i es la corriente en amperes y R la resistencia en ohms. La tabla de abajo da los valores experimentales de i correspondientes a varios tiempos t dados en segundos. Si la inductancia L es constante e igual a 0.97 henries y la resistencia es de 0.14 ohms, aproxime el voltaje E en los valores de t dados en la tabla usando la ecuación 6.41.

t	0.95	0.96	0.97	0.98	0.99	1.0
i	0.90	1.92	2.54	2.88	3.04	3.10

6.47 La reacción en fase líquida entre trimetilamina y bromuro de propilo en benceno, se llevó a cabo introduciendo cinco ampolletas con una mezcla de reactantes en un baño a temperatura constante.
Las ampolletas se sacan a varios tiempos, se enfrían para detener la reacción y se analiza su contenido. El análisis se basa en que la sal cuaternaria de amoniaco está ionizada, de aquí que la concentración de los iones bromuro se pueda obtener por titulación. Los resultados obtenidos son

Tiempo (min)	10	35	60	85	110
Conversión (%)	12	28	40	46	52

Calcule la variación de la conversión con respecto al tiempo en los distintos puntos de la tabla.

6.48 La densidad de soluciones de cloruro de calcio ($CaCl_2$) a diferentes temperaturas y concentraciones se presenta en la siguiente tabla

T °C c % peso	−5	0	20	40	80	100
2		1.0171	1.0148	1.0084	0.9881	0.9748
8	1.0708	1.0703	1.0659	1.0586	1.0382	1.0257
16	1.1471	1.1454	1.1386	1.1301	1.1092	1.0973
30		1.2922	1.2816	1.2709	1.2478	1.2359
40			1.3957	1.3826	1.3571	1.3450

Calcule

a) La variación de la densidad con respecto a la temperatura a T = 0°C y c = 40 %

b) La variación de la densidad con respecto a la concentración a T = 0°C y c = 40 %

c) La variación de la densidad con respecto a la temperatura a T = 30°C y c = 10 %

d) La variación de la densidad con respecto a la concentración a T = 30°C y c = 10 %

CAPÍTULO 7

ECUACIONES DIFERENCIALES ORDINARIAS

EN ESTE CAPÍTULO se formula el problema de valor inicial a partir de una situación física sencilla y se estudia gráfica y analíticamente los métodos más utilizados en su solución numérica.

INTRODUCCIÓN

Se llama **ecuación diferencial** aquella ecuación que contiene una variable dependiente y sus derivadas con respecto a una o más variables independientes. Muchas de las leyes generales de la naturaleza se expresan en el lenguaje de las ecuaciones diferenciales; abundan también las aplicaciones en ingeniería, economía, en las mismas matemáticas y en muchos otros campos de la ciencia aplicada.

Esta gran utilidad de las ecuaciones diferenciales es fácil de explicar; recuérdese que si se tiene la función $y = f(x)$, su derivada dy/dx puede interpretarse como la velocidad de cambio de y con respecto a x. En cualquier proceso natural, las variables incluidas y sus velocidadaes de cambio se relacionan entre sí mediante los principios científicos que gobiernan el proceso. El resultado de expresar en símbolos matemáticos estas relaciones, a menudo es una ecuación diferencial.

Se tratará de ilustrar estos comentarios con el siguiente ejemplo.

Supóngase que se quiere conocer cómo varía la altura h del nivel en un tanque cilíndrico de área seccional A cuando se llena con un líquido de densidad ρ a razón de G l/min como se ve en la figura 7.1.

La ecuación diferencial se obtiene mediante un balance de materia (principio universal de continuidad) en el tanque

Acumulación	=	Entrada	-	Salida
(Kg/min)		(Kg/min)		(Kg/min)

donde la acumulación significa la variación de la masa de líquido en el tanque con respecto al tiempo, la cual se expresa matemáticamente como una derivada:

$d(V\rho)/dt$. Lo que entra es $\frac{G}{\rho}$ (Kg/min) y el término de salida es nulo, con lo cual la ecuación de continuidad queda como sigue

$$\frac{d(V\rho)}{dt} = G/\rho$$

Por otro lado, el volumen de líquido V que contiene el tanque a una altura h es* $V = A\,h$. Al sustituir V en la ecuación diferencial de arriba y considerando que la densidad ρ es constante, se llega a

$$A\frac{dh}{dt} = G \tag{7.1}$$

ecuación diferencial cuya solución describe cómo cambia la altura h del líquido dentro del tanque con respecto al tiempo t.

A continuación se enlistan ejemplos de ecuaciones diferenciales.

$$\frac{dy}{dt} = -ky, \tag{7.2}$$

$$m\frac{d^2y}{dt^2} = ky, \tag{7.3}$$

$$\frac{dy}{dx} + 2xy = e^{-x^2} \tag{7.4}$$

$$\frac{d^2y}{dx^2} - 5\frac{dy}{dx} + 6y = 0, \tag{7.5}$$

$$(1-x^2)\frac{d^2y}{dx^2} - 2x\frac{dy}{dx} + p(p+1)y = 0 \tag{7.6}$$

$$x^2\frac{d^2y}{dx^2} + x\frac{dy}{dx} + (x^2 - p^2) = 0 \tag{7.7}$$

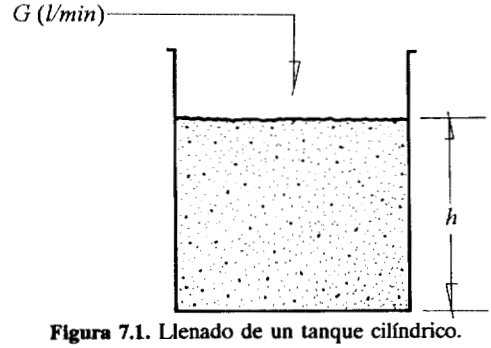


Figura 7.1. Llenado de un tanque cilíndrico.

*El fondo del tanque es plano.

La variable dependiente en cada una de estas ecuaciones es y, y la variable independiente es x o t. Las letras k, m y p representan constantes. Una ecuación diferencial es **ordinaria** si sólo tiene una variable independiente, por lo que todas las derivadas que tiene son ordinarias o totales. Las ecuaciones 7.1 a 7.7 son ordinarias. El **orden** de una ecuación diferencial es el orden de la derivada de más alto orden en ella. Las ecuaciones 7.1, 7.2 y 7.4 son de primer orden, y las demás de segundo.

SECCIÓN 7.1 FORMULACIÓN DEL PROBLEMA DE VALOR INICIAL

La ecuación diferencial ordinaria (EDO) general de primer orden es

$$\frac{dy}{dx} = f(x, y) \tag{7.8}$$

En la teoría de las EDO se establece que su solución general debe contener una constante arbitraria c, de tal modo que la solución general de la ecuación 7.8 es

$$F(x, y, c) = 0 \tag{7.9}$$

La ecuación 7.9 representa una familia de curvas en el plano x-y, obtenida cada una de ellas para un valor particular de c como se muestra en la figura 7.2. Cada una de estas curvas corresponde a una solución particular de la EDO 7.8, y analíticamente dichas constantes se obtienen exigiendo que la solución de esa ecuación pase por algún punto (x_0, y_0); esto es, que

$$y(x_0) = y_0 \tag{7.10}$$

lo cual significa que la variable dependiente y vale y_0 cuando la variable independiente x vale x_0 (véase la curva F_2 de la Fig. 7.2).

En los cursos regulares de cálculo y ecuaciones diferenciales se estudian técnicas **analíticas** para encontrar soluciones del tipo de la ecuación 7.9 a problemas como el de la ecuación 7.8 o mejor aún, a problemas de valor inicial —ecuación 7.8 y condición 7.10, simultáneamente.

En la práctica la gran mayoría de las ecuaciones no pueden resolverse utilizando estas técnicas, y se deberá recurrir a los **métodos numéricos.**

Cuando se usan métodos numéricos no se encuentran soluciones de la forma $F(x, y, c) = 0$, ya que trabajan con números y dan por resultado números. Sin embargo, el propósito usual de encontrar una solución es determinar valores de y (números) correspondientes a valores específicos de x, lo cual es factible con los mencionados métodos numéricos sin tener que encontrar $F(x, y, c) = 0$.

El problema de valor inicial (PVI) por resolver numéricamente queda formulado como sigue

a) Una ecuación diferencial de primer orden (del tipo 7.8)
b) El valor de y en un punto conocido x_0 (condición inicial)
c) El valor x_f donde se quiere conocer el valor de y (y_f)

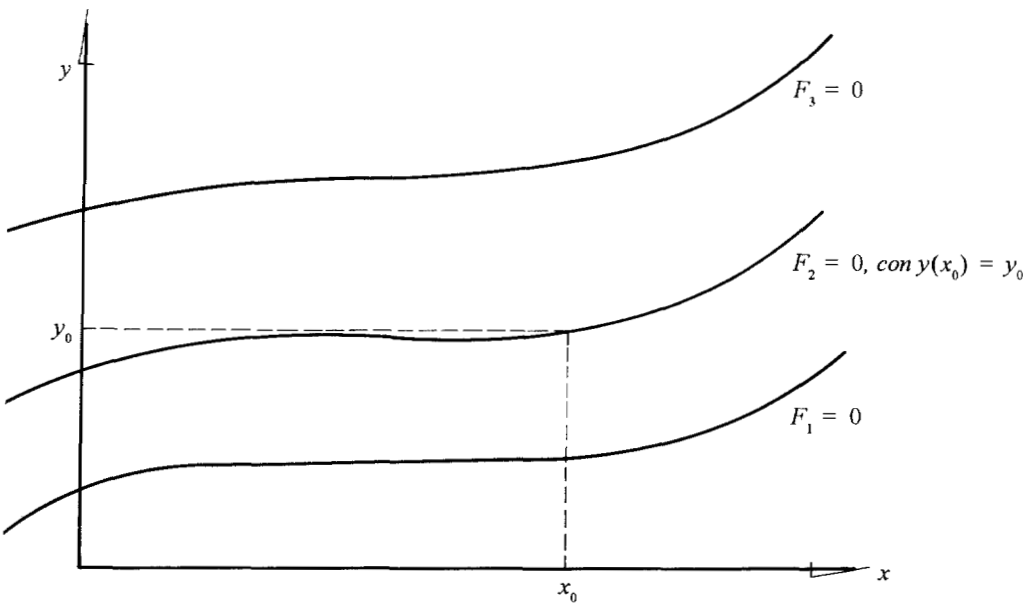


Figura 7.2. Representación gráfica de la solución general, ecuación 7.9.

que en lenguaje matemático quedará así

$$\text{PVI} \quad \begin{cases} \dfrac{dy}{dx} = f(x,y) \\ y(x_0) = y_0 \\ y(x_f) = ? \end{cases} \tag{7.11}$$

Formulado el problema de valor inicial, a continuación se describe una serie de técnicas numéricas para resolverlo.

SECCIÓN 7.2 MÉTODO DE EULER

El método de Euler es el más simple de los métodos numéricos para resolver un problema de valor inicial del tipo 7.11. Consiste en dividir el intervalo que va de x_0 a x_f en n subintervalos de ancho h (véase Fig. 7.3); o sea,

$$h = \frac{x_f - x_0}{n} \tag{7.12}$$

de manera que se obtiene un conjunto discreto de $(n+1)$ puntos*: $x_0, x_1, x_2, \ldots, x_n$ del intervalo de interés $[x_0, x_f]$. Para cualquiera de estos puntos se cumple que

$$x_i = x_0 + i\,h \quad 0 \leq i) \leq n \tag{7.13}$$

Nótese la similitud de este desarrollo con el primer paso de la integración numérica.

La condición inicial $y(x_0) = y_0$ representa el punto $P_0 = (x_0, y_0)$ por donde pasa la curva solución de la ecuación 7.11, la cual por simplicidad se denotará como $F(x) = y$ en lugar de $F(x, y, c_1) = 0$.

Con el punto P_0 se puede evaluar la primera derivada de $F(x)$ en ese punto; a saber

$$F'(x) = \frac{dy}{dx}\Big|_{P_0} = f(x_0, y_0) \tag{7.14}$$

Con esta información se traza una recta, aquella que pasa por P_0 y de pendiente $f(x_0, y_0)$. Esta recta aproxima $F(x)$ en una vecindad de x_0. Tómese la recta como remplazo de $F(x)$ y localícese en ella (la recta) el valor de y correspondiente a x_1. Entonces, de la figura 7.3

$$\frac{y_1 - y_0}{x_1 - x_0} = f(x_0, y_0) \tag{7.15}$$

Se resuelve para y_1

$$y_1 = y_0 + (x_1 - x_0)\, f(x_0, y_0) = y_0 + h\, f(x_0, y_0) \tag{7.16}$$

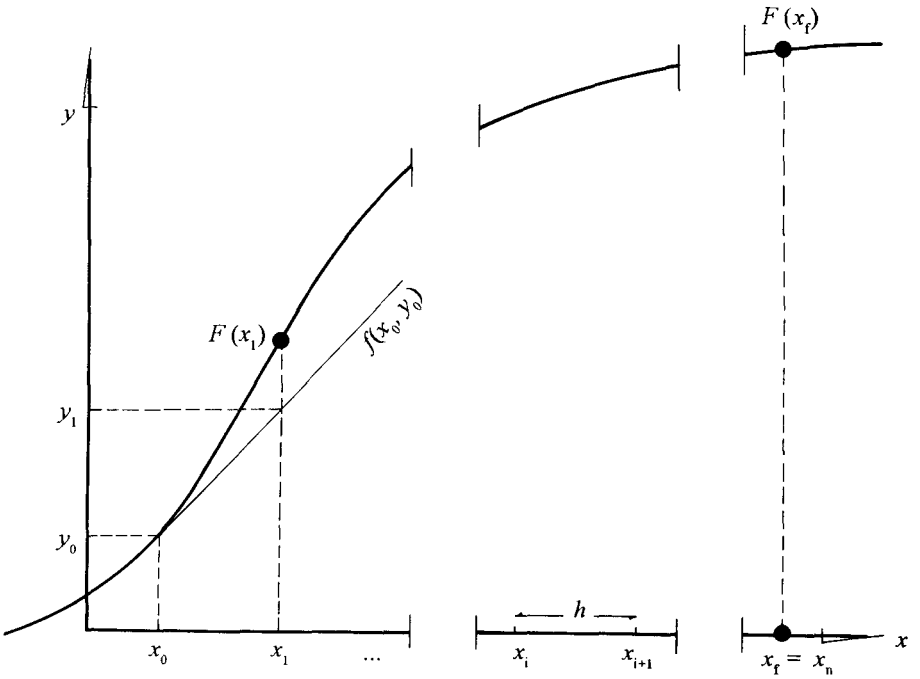


Figura 7.3 Deducción gráfica del método de Euler.

*x_f se convierte en x_n.

Es evidente que la ordenada y_1 calculada de esta manera no es igual a $F(x_1)$, pues existe un pequeño error. No obstante, el valor y_1 sirve para aproximar $F'(x)$ en el punto $P = (x_1, y_1)$ y repetir el procedimiento anterior a fin de generar la sucesión de aproximaciones siguiente

$$\begin{aligned} y_1 &= y_0 + h f(x_0, y_0) \\ y_2 &= y_1 + h f(x_1, y_1) \\ &\vdots \\ y_{i+1} &= y_i + h f(x_i, y_i) \\ &\vdots \\ y_n &= y_{n-1} + h f(x_{n-1}, y_{n-1}) \end{aligned} \tag{7.17}$$

Como se muestra en la figura 7.4, en esencia se trata de aproximar la curva $y = F(x)$ por medio de una serie de segmentos de línea recta.

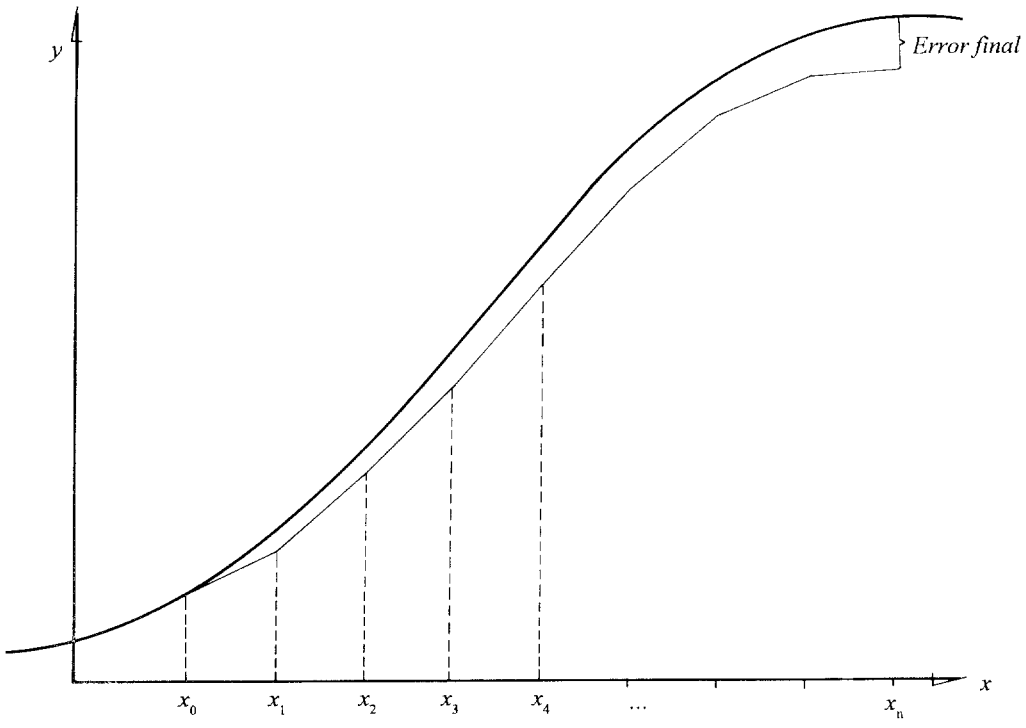


Figura 7.4 Aplicación repetida del método de Euler.

Como la aproximación a una curva mediante una línea recta no es exacta, se comete un error propio del método mismo. De modo similar a otros capítulos, éste se denominará **error de truncamiento.** Dicho error puede disminuirse tanto como se quiera (al menos teóricamente) reduciendo el valor de h, pero a cambio de un mayor número de cálculos y tiempo de máquina y, por consiguiente, de un **error de redondeo** más alto.

Ejemplo 7.1

Resuelva el siguiente

$$\text{PVI} \begin{cases} \dfrac{dy}{dx} = (x - y) \\ y(0) = 2 \\ y(1) = ? \end{cases}$$

mediante el método de Euler.

SOLUCIÓN

Sugerencia: Puede usar un pizarrón electrónico o el GC para seguir los cálculos.

El intervalo de interés para este ejemplo es [0, 1] y al dividirlo en cinco subintervalos se tiene

$$h = \frac{1 - 0}{5} = 0.2$$

con lo cual se generan los argumentos

$$x_0 = 0.0, x_1 = x_0 + h = 0.0 + 0.2 = 0.2$$
$$x_2 = x_1 + h = 0.2 + 0.2 = 0.4$$
$$\vdots$$
$$x_5 = x_4 + h = 0.8 + 0.2 = 1.0$$

Con $x_0 = 0.0$ y $y_0 = 2$ y las ecuaciones 7.17 se obtienen los valores

$$y_1 = y(0.2) = 2 + 0.2[0.0 - 2] = 1.6$$
$$y_2 = y(0.4) = 1.6 + 0.2[0.2 - 1.6] = 1.32$$
$$y_3 = y(0.6) = 1.32 + 0.2[0.4 - 1.32] = 1.136$$
$$y_4 = y(0.8) = 1.136 + 0.2[0.6 - 1.136] = 1.0288$$
$$y_5 = y(1.0) = 1.0288 + 0.2[0.8 - 1.0288] = 0.98304$$

Por otro lado, la solución analítica es 1.10364 (el lector puede verificarla resolviendo analíticamente el PVI); el error cometido es 0.1206 en valor absoluto y 10.92 en porciento.

ALGORITMO 7.1 Método de Euler

Para obtener la aproximación YF a la solución de un problema de valor inicial o PVI (Ec. 7.11), proporcionar la función F(X,Y) y los

DATOS: La condición inicial X0,Y0, el valor XF donde se desea conocer el valor de YF y el número N de subintervalos por emplear.
RESULTADOS: Aproximación a YF : Y0.

PASO 1. Hacer H = (XF - X0)/N
PASO 2. Hacer I = 1
PASO 3. Mientras I ≤ N, repetir los pasos 4 a 6.
  PASO 4. Hacer Y0 = Y0 + H * F(X0,Y0)
  PASO 5. Hacer X0 = X0 + H
  PASO 6. Hacer I = I + 1
PASO 7. IMPRIMIR Y0 y TERMINAR.

SECCIÓN 7.3 METODOS DE TAYLOR

Antes de explicar estos métodos, conviene hacer una acotación al método de Euler.

Puede decirse que el método de Euler utiliza los primeros dos términos de la serie de Taylor para su primera iteración; o sea,

$$F(x_1) \approx y_1 = F(x_0) + F'(x_0)\,(x_1 - x_0) \tag{7.18}$$

donde se señala que y_1 no es igual a $F(x_1)$.

Esto pudo hacer pensar que para encontrar y_2, se expandió de nuevo $F(x)$ en serie de Taylor, como sigue

$$F(x_2) \approx y_2 = F(x_1) + F'(x_1)\,(x_2 - x_1), \tag{7.19}$$

sin embargo, no se dispone de los valores exactos de $F(x_1)$ y $F'(x_1)$ y, rigurosamente hablando, son los que deben usarse en una expansión de Taylor de $F(x)$ —en este caso alrededor de x_1—; por tanto, el lado derecho de la ecuación 7.19 no es evaluable. Por ello, sólo en la primera iteración, para encontrar y_1, se usa realmente una expansión en serie de Taylor de $F(x)$, aceptando desde luego que se tienen valores exactos en la condición inicial $y_0=F(x_0)$. Después de eso, se emplea la ecuación

$$\begin{aligned} y_{i+1} &= y_i + f(x_i, y_i)\,(x_{i+1} - x_i) \\ &= F(x_i) + F'(x_i)\,(x_{i+1} - x_i) \end{aligned} \tag{7.20}$$

que guarda similitud con una expansión en serie de Taylor.

Aclarado este punto, a continuación se aplicará la información acerca de las series de Taylor para mejorar la exactitud del método de Euler y obtener extensiones que constituyen la familia de métodos llamados **algoritmos de Taylor.**

Si se usan tres términos en lugar de dos en la expansión de $F(x_1)$, entonces

$$F(x_1) \approx y_1 = F(x_0) + F'(x_0)(x_1 - x_0) + F''(x_0)\frac{(x_1 - x_0)^2}{2!} \qquad (7.21)$$

Como

$$F''(x) = \frac{dF'(x)}{dx} = \frac{df(x,y)}{dx},$$

y

$$h = x_1 - x_0,$$

la primera iteración (Ec. 7.21) tomaría la forma*

$$y_1 = y_0 + h f(x_0, y_0) + \frac{h^2}{2!}\left.\frac{df(x,y)}{dx}\right|_{x_0, y_0} \qquad (7.22)$$

Ahora cabe pensar que usando una fórmula de iteración basada en la ecuación 7.22 para obtener $y_2, y_3, \ldots, y_n$ mejoraría la exactitud obtenida con la 7.18. Se propone entonces la fórmula

$$\boxed{y_{i+1} = y_i + h f(x_i, y_i) + \frac{h^2}{2!}\left.\frac{df(x,y)}{dx}\right|_{x_i, y_i}} \qquad (7.23)$$

La utilidad de esta ecuación depende de cuán fácil sea la diferenciación de $f(x, y)$. Si $f(x, y)$ es una función sólo de x, la diferenciación con respecto a x es relativamente fácil y la fórmula propuesta es muy práctica.

Si, como es el caso general, $f(x,y)$ es una función de x y y, habrá que usar derivadas totales. La derivada total de $f(x, y)$ con respecto a x está dada por

$$\frac{df(x,y)}{dx} = \frac{\partial f(x,y)}{\partial x} + \frac{\partial f(x,y)}{\partial y}\frac{dy}{dx}$$

Si se aplican las ideas vistas en el método de Euler pero empleando como fórmula la ecuación 7.23, se obtiene el método de Taylor de **segundo orden.** Esto último es indicativo de la derivada de mayor orden que se emplea y de cierta exactitud. Con esta terminología, al método de Euler le correspondería el nombre de **método de Taylor de primer orden.**

---

*La notación $\left.\frac{df(x,y)}{dx}\right|_{x_0, y_0}$ significa la evaluación de la derivada de $f(x, y)$ con respecto a x en el punto (x_0, y_0).

Ejemplo 7.2

Resuelva el PVI del ejemplo 7.1 por el método de Taylor de segundo orden. Puede usar un pizarrón electrónico para seguir los cálculos.

SOLUCIÓN

Al utilizar cinco intervalos de nuevo se tiene

$$h = 0.2, \quad x_0 = 0.0, \quad x_1 = 0.2, \quad x_2 = 0.4,$$
$$x_3 = 0.6, \quad x_4 = 0.8, \quad x_5 = 1.0$$

Se aplica la ecuación 7.23 con $y_0 = 2$ y con

$$\frac{df(x,y)}{dx} = \frac{\partial f(x,y)}{\partial x} + \frac{\partial f(x,y)}{\partial y}(x-y) = 1 - x + y$$

ya que $\frac{dy}{dx} = x - y$

$$y_1 = y(0.2) = y_0 + h(x_0 - y_0) + \frac{h^2}{2!}(1 - x_0 + y_0)$$

$$= 2 + 0.2(0-2) + \frac{0.2^2}{2}(1-0+2) = 1.66$$

$$y_2 = y(0.4) = y_1 + h(x_1 - y_1) + \frac{h^2}{2!}(1 - x_1 + y_1)$$

$$= 1.66 + 0.2(0.2 - 1.66) + \frac{0.2^2}{2}(1 - 0.2 + 1.66) = 1.4172$$

al continuar este procedimiento se llega a

$$y_5 = y(1.0) = 1.11222$$

que da un error absoluto de 0.00858 y un error porcentual de 0.78. Nótese la mayor exactitud y el mayor número de cálculos.

La extensión de esta idea a cuatro, cinco o más términos de la serie de Taylor significaría obtener métodos con mayor exactitud pero menos prácticos, ya que incluirían diferenciaciones complicadas de $f(x, y)$; por ejemplo, si se quisiera usar cuatro términos de la serie, se necesitaría la segunda derivada de $f(x, y)$, la cual está dada por

$$\frac{d^2f(x,y)}{dx^2} = \frac{\partial^2 f(x,y)}{\partial x^2} + 2\frac{dy}{dx}\frac{\partial^2 f(x,y)}{\partial x\,\partial y} + \left(\frac{dy}{dx}\right)^2 \frac{\partial^2 f(x,y)}{\partial y^2}$$

$$+ \frac{\partial f(x,y)}{\partial x}\frac{\partial f(x,y)}{\partial y} + \left(\frac{\partial f(x,y)}{\partial y}\right)^2 \frac{dy}{dx}$$

Las derivadas totales de orden superior al segundo de $f(x, y)$ son aún más largas y complicadas.

Ya que el uso de varios términos de la serie de Taylor presenta serias dificultades, los investigadores han buscado métodos comparables con ellos en exactitud pero más fáciles. De hecho, el patrón para evaluarlos son los métodos derivados de la serie de Taylor; por ejemplo, dado un método, se compara con el derivado de la serie de Taylor que dé la misma exactitud. La derivada de más alto orden en este último proporciona el orden del primero. Un método que diera una exactitud comparable al método de Euler sería de **primer orden**; si diera una exactitud comparativamente igual a usar tres términos de la serie de Taylor, sería de **segundo orden,** y así sucesivamente.

A continuación se estudian métodos de orden dos, tres, etc., en los que no se requieren diferenciaciones de $f(x, y)$.

SECCIÓN 7.4 MÉTODO DE EULER MODIFICADO

En el método de Euler se tomó como válida para todo el intervalo la derivada encontrada en un extremo de éste (véase Fig. 7.5). Para obtener una exactitud razonable se utiliza un intervalo muy pequeño, a cambio de un error de redondeo mayor (ya que se realizarán más cálculos).

El método de Euler modificado trata de evitar este problema utilizando un valor promedio de la derivada tomada en los dos extremos del intervalo, en lugar de la derivada tomada en un solo extremo.

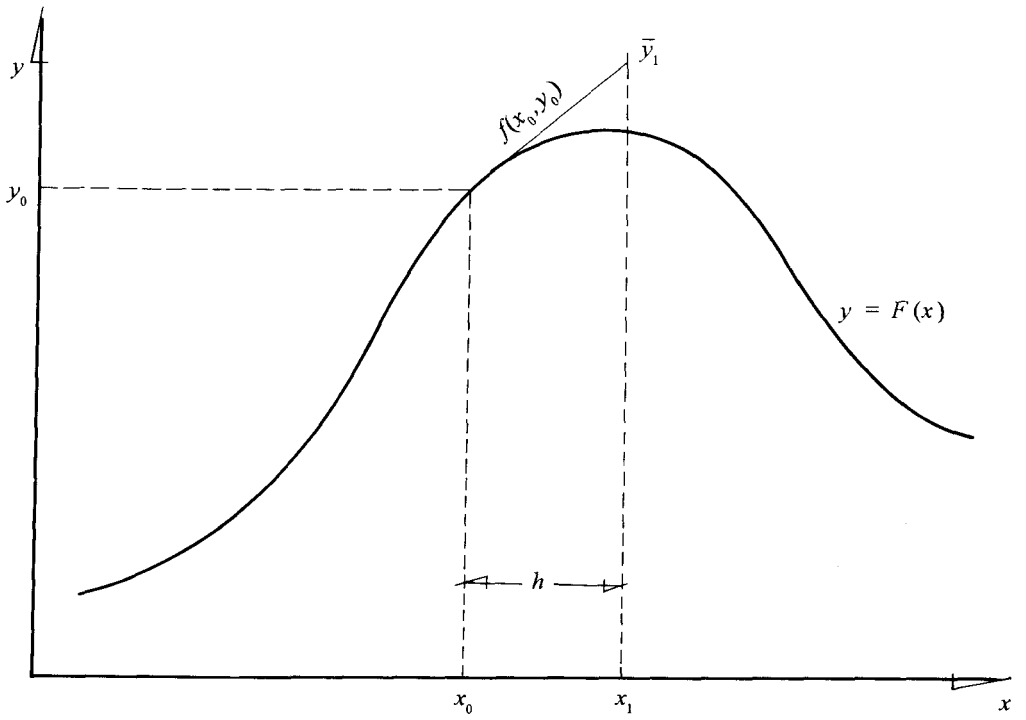


Figura 7.5. Primer paso en el método de Euler modificado.

El método de Euler modificado consta de dos pasos básicos*

1. Se parte de (x_0, y_0) y se utiliza el método de Euler a fin de calcular el valor de y correspondiente a x_1. Este valor de y se denotará aquí como $\bar{y}_1$, ya que sólo es un valor transitorio para y_1. Esta parte del proceso se conoce como paso **predictor.**
2. El segundo paso se llama **corrector**, pues trata de corregir la predicción. En el nuevo punto obtenido $(x_1, \bar{y}_1)$ se evalúa la derivada $f(x_1, \bar{y}_1)$ usando la ecuación diferencial ordinaria del PVI que se esté resolviendo; se obtiene la media aritmética de esta derivada y la derivada en el punto inicial (x_0, y_0)

$$\frac{1}{2}\left[f(x_0, y_0) + f(x_1, \bar{y}_1)\right] = \textit{derivada promedio}$$

Se usa la derivada promedio para calcular un nuevo valor de y_1, con la ecuación 7.17, que deberá ser más exacto que $\bar{y}_1$

$$y_1 = y_0 + \frac{(x_1 - x_0)}{2}\left[f(x_0, y_0) + f(x_1, \bar{y}_1)\right]$$

y que se tomará como valor definitivo de y_1. Este procedimiento se repite hasta llegar a y_n.

El esquema iterativo para este método quedaría en general así

Primero, usando el paso de predicción resulta

$$\bar{y}_{i+1} = y_i + h\, f(x_i, y_i). \tag{7.24a}$$

Una vez obtenida $\bar{y}_{i+1}$ se calcula $f(x_{i+1}, \bar{y}_{i+1})$, la derivada en el punto $(x_{i+1}, \bar{y}_{i+1})$, y se promedia con la derivada previa $f(x_i, y_i)$ para encontrar la derivada promedio

$$\frac{1}{2}\left[f(x_i, y_i) + f(x_{i+1}, \bar{y}_{i+1})\right]$$

Se sustituye $f(x_i, y_i)$ con este valor promedio en la ecuación de iteración de Euler y se obtiene

$$\boxed{y_{i+1} = y_i + \frac{h}{2}\left[f(x_i, y_i) + f(x_{i+1}, \bar{y}_{i+1})\right]} \tag{7.24b}$$

Ejemplo 7.3

Resuelva el PVI del ejemplo 7.1 por el método de Euler modificado.

SOLUCIÓN

Al utilizar nuevamente cinco intervalos para que la comparación de los resultados obtenidos sea consistente con los anteriores, se tiene

*Se omitió la subdivisión de $[x_0, x_f]$ en n subintervalos para dar énfasis a los pasos fundamentales de predicción y corrección.

Primera iteración

Primer paso: $\bar{y}_1 = y_0 + h f(x_0, y_0) = 2 + 0.2(0 - 2) = 1.6$

Segundo paso: $\frac{1}{2}[f(x_0, y_0) + f(x_1, \bar{y}_1)] =$

$$\frac{1}{2}[(0-2) + (0.2-1.6)] = -1.7$$

$$y(0.2) = y_1 = 2 + 0.2(-1.7) = 1.66$$

Segunda iteración

Primer paso: $\bar{y}_2 = y1 + hf(x_1, y_1) = 1.66 + 0.2(0.2\text{-}1.66) = 1.368$

Segundo paso: $\frac{1}{2}[f(x_1, y_1) + f(x_2, \bar{y}_2)]$

$$= \frac{1}{2}[(0.2-1.66) + (0.4-1.368)] = -1.214$$

$$y(0.4) = y_2 = 1.66 + 0.2(-1.214) = 1.4172$$

Al continuar los cálculos se llega a

$$\bar{y}_5 = 1.08509$$

$$y_5 = 1.11222$$

Los resultados obtenidos en este caso son idénticos a los del ejemplo 7.2 en que se utilizó el método de Taylor de segundo orden; por tanto presumiblemente el método de Euler modificado es de segundo orden. Esto se demuestra en la siguiente sección.

ALGORITMO 7.2 Método de Euler modificado

Para obtener la aproximación YF a la solución de un PVI, proporcionar la función F(X,Y) y los

DATOS: La condición inicial X0, Y0, el valor XF donde se desea conocer el valor de YF y el número N de subintervalos por emplear.

RESULTADOS: Aproximación a YF: Y0.

PASO 1. Hacer H = (XF - X0) / N

PASO 2. Hacer I = 1

PASO 3. Mientras I ≤ N, repetir los pasos 4 a 7.

PASO 4. Hacer Y1 = Y0 + H * F(X0,Y0)

PASO 5. Hacer Y0 = Y0 + H/2 * (F(X0,Y0)+F(X0+H,Y1))

PASO 6. Hacer X0 = X0 + H

PASO 7. Hacer I = I + 1

PASO 8. IMPRIMIR Y0 y TERMINAR.

SECCIÓN 7.5 MÉTODOS DE RUNGE-KUTTA

Los métodos asociados con los nombres de Runge (1885), Kutta (1901), Heun (1900) y otros para resolver el PVI (Ec. 7.11) consisten en obtener un resultado que se obtendría al utilizar un número finito de términos de una serie de Taylor de la forma

$$y_{i+1} = y_i + hf(x_i, y_i) + \frac{h^2}{2!} f'(x_i, y_i) + \frac{h^3}{3!} f''(x_i, y_i) + \dots \tag{7.25}$$

con una aproximación en la cual se calcula y_{i+1} de una fórmula del tipo*

$$y_{i+1} = y_i + h\,[\alpha_0 f(x_i, y_i) + \alpha_1 f(x_i + \mu_1 h, y_i + b_1 h) + \alpha_2 f(x_i + \mu_2 h, y_i + b_2 h)$$
$$+ \dots + \alpha_p f(x_i + \mu_p h, y_i + b_p h)\,] \tag{7.26}$$

donde las α, μ, y b se determinan de modo que si se expandiera $f(x_i + \mu_j h, y_i + b_j h)$, con $1 \le j \le p$ en series de Taylor alrededor de (x_i, y_i), se observaría que los coeficientes de h, h^2, h^3, etc. coincidirían con los coeficientes correspondientes de la ecuación 7.25.

A continuación se derivará sólo el caso más simple, cuando $p = 1$, para ilustrar el procedimiento del caso general, ya que los lineamientos son los mismos.

A fin de simplificar y sistematizar la derivación, conviene expresar la ecuación 7.26 con $p = 1$ en la forma

$$y_{i+1} = y_i + h\,[\alpha_0 f(x_i, y_i) + \alpha_1 f(x_i + \mu h, y_i + bh)\,] \tag{7.27}$$

Obsérvese que en esta expresión se evalúa f en (x_i, y_i) y $(x_i + \mu h, y_i + bh)$. El valor $x_i + \mu h$ es tal que $x_i < x_i + \mu h \le x_{i+1}$ para mantener la abcisa del segundo punto dentro del intervalo de interés (véase Fig. 7.6), con lo que $0 < \mu \le 1$.

Por otro lado, b puede manejarse más libremente y expresarse $y_i + bh$, sin pérdida de generalidad, como una ordenada arriba o abajo de la ordenada que da el método de Euler simple

$$y_{i+1} + bh = y_i + \lambda h f(x_i, y_i) = y_i + \lambda k_0 \tag{7.28}$$

con $k_0 = h f(x_i, y_i)$

Queda entonces por determinar α_0, α_1, μ y λ tales que la ecuación 7.27 tenga una expansión en potencias de h cuyos primeros términos, tantos como sea posible, coincidan con los primeros términos de la 7.25.

Para obtener los parámetros desconocidos, se expande primero $f(x_i + \mu h, y_i + \lambda k_0)$ en serie de Taylor (obviamente mediante el desarrollo de Taylor de funciones de dos variables).**

---

*Nótese que en la ecuación 7.26 ya no aparecen diferenciaciones, sólo evaluaciones de $f(x, y)$.

**Spiegel, M.R. *Manual de fórmulas y tablas matemáticas, Schaum.* McGraw Hill. Serie Schaum. (1970), p 113.

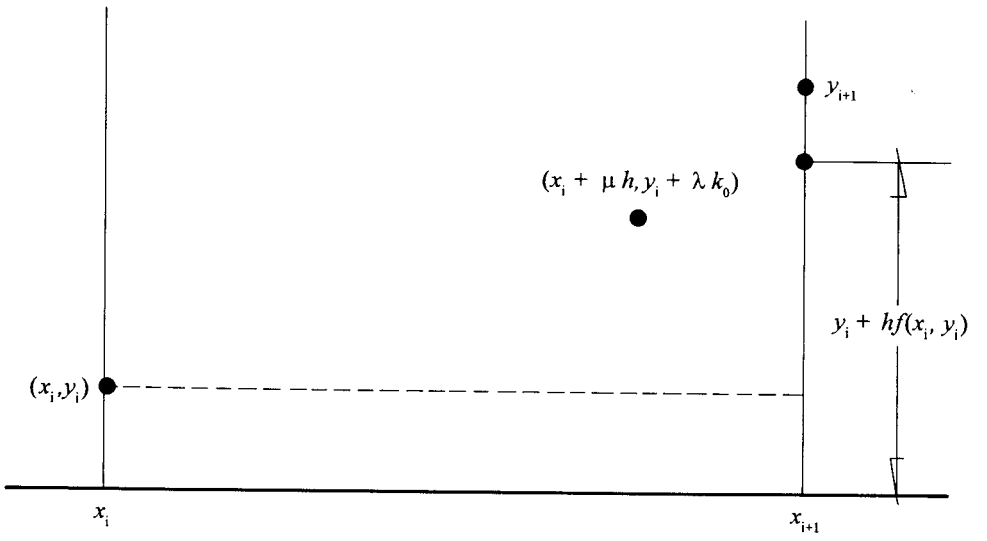


Figura 7.6 Deducción del método de Runge-Kutta.

$$f(x_i + \mu h, y_i + \lambda k_0) = f(x_i, y_i) + \mu h \frac{\partial f}{\partial x} + \lambda k_0 \frac{\partial f}{\partial y} + \frac{\mu^2 h^2}{2!} \frac{\partial^2 f}{\partial x^2} + \\ + \mu h \lambda k_0 \frac{\partial^2 f}{\partial x \partial y} + \frac{\lambda^2 k_0^2}{2!} \frac{\partial^2 f}{\partial y^2} + O(h^3) \tag{7.29}$$

Todas las derivadas parciales son evaluadas en (x_i, y_i).

Se sustituye en la ecuación 7.27

$$y_{i+1} = y_i + \alpha_0 h f(x_i, y_i) + \alpha_1 h \left[f(x_i, y_i) + \mu h \frac{\partial f}{\partial x} + \lambda k_0 \frac{\partial f}{\partial y} + \right. \\ \left. + \frac{\mu^2 h^2}{2!} \frac{\partial^2 f}{\partial x^2} + \mu h \lambda k_0 \frac{\partial^2 f}{\partial x \partial y} + \frac{\lambda^2 k_0^2}{2!} \frac{\partial^2 f}{\partial y^2} + O(h^3) \right]$$

Esta última ecuación se arregla en potencias de h, y queda

$$y_{i+1} = y_i + h(\alpha_0 + \alpha_1) f(x_i, y_i) + h^2 \alpha_1 \left(\mu \frac{\partial f}{\partial x} + \lambda f(x_i, y_i) \frac{\partial f}{\partial y} \right) \\ + \frac{h^3}{2} \alpha_1 \left(\mu^2 \frac{\partial^2 f}{\partial x^2} + 2\mu\lambda f(x_i, y_i) \frac{\partial^2 f}{\partial x \partial y} + \lambda^2 f^2(x_i, y_i) \frac{\partial^2 f}{\partial y^2} \right) + O(h^4) \tag{7.30}$$

Para que los coeficientes correspondientes de h y h^2 coincidan en las ecuaciones 7.25 y 7.30 se requiere

$$\begin{aligned} \alpha_0 + \alpha_1 &= 1 \\ \mu\alpha_1 = \frac{1}{2}, &\qquad \lambda\alpha_1 = \frac{1}{2} \end{aligned} \tag{7.31}$$

Hay cuatro incógnitas para sólo tres ecuaciones y, por tanto, se tiene un grado de libertad en la solución de la ecuación 7.31. Podría pensarse en usar este grado de libertad para hacer coincidir los coeficientes de h^3. Sin embargo, es obvio que esto es imposible para cualquier forma que tenga la función $f(x, y)$. Existe entonces un número infinito de soluciones de la ecuación 7.31, pero quizá la más simple sea

$$\alpha_0 = \alpha_1 = \frac{1}{2}; \qquad \mu = \lambda = 1$$

Esta elección conduce a la fórmula

$$y_{i+1} = y_i + \frac{h}{2}\left[f(x_i, y_i) + f(x_i + h, y_i + hf(x_i, y_i))\right]$$

o bien

$$\boxed{\begin{aligned} &y_{i+1} = y_i + \frac{h}{2}(k_0 + k_1) \\ &\text{con} \\ &k_0 = f(x_i, y_i) \qquad ; \qquad k_1 = f(x_i+h, y_i+hk_0) \end{aligned}} \tag{7.32}$$

conocida como **algoritmo de Runge-Kutta de segundo orden (lo de segundo orden** por coincidir con los primeros tres términos de la serie de Taylor), y que es la fórmula del método de Euler modificado, con dos pasos compactados en uno.

Por ser de orden superior al de Euler, este método proporciona mayor exactitud (véase Ejem. 7.3); por tanto, es posible usar un valor de h no tan pequeño como en el primero. El precio es la evalución de $f(x, y)$ dos veces en cada subintervalo, contra una en el método de Euler.

Las fórmulas de Runge-Kutta de cualquier orden se pueden derivar en la misma forma en que se llega a la ecuación 7.32.

El **método de Runge-Kutta de cuarto orden** (igual que para orden dos, existen muchos métodos de cuarto orden) es una de las fórmulas más usadas de esta familia y está dado como

$$y_{i+1} = y_i + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

donde

$$\begin{aligned} k_1 &= f(x_i, y_i) \\ k_2 &= f(x_i + h/2, y_i + hk_1/2) \\ k_3 &= f(x_i + h/2, y_i + hk_2/2) \\ k_4 &= f(x_i + h, y_i + hk_3) \end{aligned} \tag{7.33}$$

En la ecuación 7.33 hay coincidencia con los primeros cinco términos de la serie de Taylor, lo cual significa gran exactitud sin cálculo de derivadas; pero a cambio, hay que evaluar la función $f(x, y)$ cuatro veces en cada subintervalo.

Ejemplo 7.4

Resuelva el PVI del ejemplo 7.1 por el método de Runge-Kutta de cuarto orden (RK-4). Se recomienda usar un pizarrón electrónico, el GC o el software del libro.

SOLUCIÓN

Al tomar nuevamente cinco subintervalos y emplear la ecuación 7.33 se tiene

Primera iteración

Cálculo de las constantes k_1, k_2, k_3, k_4

$$k_1 = f(x_0, y_0) = (0 - 2) = -2$$

$$k_2 = f(x_0+h/2, y_0+hk_1/2) = [(0+0.2/2) - (2+0.2(-2)/2)] = -1.7$$

$$k_3 = f(x_0+h/2, y_0+hk_2/2) = [(0+0.2/2) - (2+0.2(-1.7)/2)] = -1.73$$

$$k_4 = f(x_0+h, y_o+hk_3) = [(0+0.2) - (2+0.2(-1.73))] = -1.454$$

Cálculo de y_1

$$y(0.2) = y_1 = y_0 + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

$$= 2 + (0.2/6)(-2 + 2(-1.7) + 2(-1.73) - 1.454) = 1.6562$$

Segunda iteración

Cálculo de las constantes k_1, k_2, k_3, k_4

$$k_1 = f(x_1, y_1) = (0.2 - 1.6562) = -1.4562$$

$$k_2 = f(x_1+h/2, y_1+hk_1/2) = [(0.2+0.2/2)-(1.6562+0.2(-1.4562)/2)] = -1.21058$$

$$k_3 = f(x_1+h/2, y+hk_2/2) = [(0.2+0.2/2)-(1.6562+0.2(-1.21058)/2)] = -1.235142$$

$$k_4 = f(x_1+h, y_1+hk_3) = [(0.2+0.2) - (1.6562+0.2(-1.235142))] = -1.0091716$$

Cálculo de y_2

$$y(0.4) = y_2 = y_1 + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

$$= 1.6562 + (0.2/6)(-1.4562+2(-1.21058)+2(-1.235142) -1.0091716) = 1.410972813$$

Con la continuación de este procedimiento se obtiene

$$y(0.6) = y_3 = 1.246450474$$

$$y(0.8) = y_4 = 1.148003885$$

$$y(1.0) = y_5 = 1.103655714$$

que da un error absoluto de 0.00001 y un error porcentual de 0.0009

ALGORITMO 7.3 Método de Runge-Kutta de cuarto orden

Para obtener la aproximación YF a la solución de un PVI, proporcionar la función F(X,Y) y los

DATOS: La condición inicial X0, Y0, el valor XF donde se desea conocer el valor de YF y el número N de subintervalos a emplear.

RESULTADOS: Aproximación a YF : Y0

PASO 1. Hacer H=(XF -X0)/N

PASO 2. Hacer I = 1

PASO 3. Mientras I ≤ N, repetir los pasos 4 a 10.

PASO 4. Hacer K1 = F(X0 , Y0)

PASO 5. Hacer K2 = F(X0+H/2 , Y0 + H * K1/2)

PASO 6. Hacer K3 = F(X0+H/2 , Y0 + H * K2/2)

PASO 7. Hacer K4 = F(X0+H , Y0 + H * K3)

PASO 8. Hacer

Y0 = Y0 + H/6 * (K1 + 2*K2 +2*K3 + K4)

PASO 9. Hacer X0 = X0 + H

PASO 10. Hacer I = I + 1

PASO 11. IMPRIMIR Y0 y TERMINAR.

Los métodos descritos hasta aquí se conocen como **métodos de un solo paso** porque se apoyan y usan el punto (x_i, y_i) para el cálculo de y_{i+1} (por ejemplo los métodos de Taylor). Los métodos de Runge-Kutta además se apoyan en puntos entre x_i y x_{i+1} pero nunca en puntos anteriores a x_i. Sin embargo, si se usa información previa a x_i para el cálculo de y_{i+1}, es posible obtener otras familias de métodos con otras características distintas a las ya vistas. A estos métodos se les llama **métodos de múltiples pasos o métodos de predicción-corrección.**

SECCIÓN 7.6 MÉTODOS DE PREDICCIÓN-CORRECCIÓN

En el esquema iterativo del método de Euler modificado (Sección 7.4) se utiliza la fórmula

$$y_{i+1} = y_i + \frac{h}{2}\left[f(x_i, y_i) + f(x_{i+1}, \bar{y}_{i+1})\right]$$

El segundo término del miembro derecho de esta ecuación recuerda la integración trapezoidal compuesta del capítulo 6.

Para ver mejor esta similitud, recuérdese que la solución analítica de la ecuación diferencial del PVI (Ec. 7.11) es

$$y = F(x)$$

y que

$$F'(x) = f(x, y),$$

e integrando ambos miembros con respecto a x

$$\int F'(x)\, dx = F(x) = \int f(x, y)\, dx$$

A partir de que $F(x)$ es la integral indefinida de $f(x, y)$, se integra $f(x, y)$ entre los límites de x : x_i y x_{i+1} para obtener

$$\begin{aligned} \int_{x_i}^{x_{i+1}} f(x, y)\, dx &= F(x) \Big|_{x_i}^{x_{i+1}} \\ &= F(x_{i+1}) - F(x_i) \approx y_{i+1} - y_i \end{aligned} \tag{7.34}$$

donde y_i y y_{i+1} son aproximaciones a $F(x_i)$ y $F(x_{i+1})$, respectivamente.

Por otro lado, es factible realizar la misma integración pero con una aproximación trapezoidal entre los puntos (x_i, y_i) y $(x_{i+1}, \bar{y}_{i+1})$, donde $\bar{y}_{i+1}$ se obtuvo en el paso de predicción.

$$\int_{x_i}^{x_{i+1}} f(x, y)\, dx \approx \frac{h}{2} [f(x_i, y_i) + f(x_{i+1}, \bar{y}_{i+1})] \tag{7.35}$$

donde h es el ancho del trapezoide

$$h = x_{i+1} - x_i$$

Al igualar las integrales 7.34 y 7.35, se tiene

$$y_{i+1} - y_i = \frac{h}{2} [f(x_i, y_i) + f(x_{i+1}, \bar{y}_{i+1})]$$

o bien

$$y_{i+1} = y_i + \frac{h}{2} [f(x_i, y_i) + f(x_{i+1}, \bar{y}_{i+1})]$$

que da la ecuación de corrección del método de Euler modificado; de esta manera se establece la identificación de este algoritmo y la integración trapezoidal. Esto sugiere a su vez la obtención de esquemas iterativos de solución de PVI por medio de la regla de Simpson u otros métodos de integración numérica que usan un mayor número de puntos.

A continuación se derivará un corrector basado en el método de Simpson 1/3.

La ecuación 7.34 toma ahora la forma

$$\int_{x_{i-1}}^{x_{i+1}} f(x, y)\, dx = F(x_{i+1}) - F(x_{i-1}) \approx y_{i+1} - y_{i-1} \tag{7.36}$$

y la correspondiente a 7.35 queda

$$\int_{x_{i-1}}^{x_{i+1}} f(x,y)\,dx \approx \frac{h}{3}\left[f(x_{i-1},y_{i-1}) + 4f(x_i,y_i) + f(x_{i+1},\bar{y}_{i+1})\right] \tag{7.37}$$

Nótese que se está integrando de x_{i-1} a x_{i+1} ya que se utilizan dos subintervalos para cada integración.

Al igualar 7.36 y 7.37 se llega a la fórmula de corrección

$$\boxed{y_{i+1} = y_{i-1} + \frac{h}{3}\left[f(x_{i-1},y_{i-1}) + 4f(x_i,y_i) + f(x_{i+1},\bar{y}_{i+1})\right]} \tag{7.38}$$

donde nuevamente hay que obtener $\bar{y}_{i+1}$ con un predictor.

Al partir de (x_0, y_0), la ecuación 7.38 tomaría la forma

$$y_2 = y_0 + \frac{h}{3}\left[f(x_0,y_0) + 4f(x_1,y_1) + f(x_2,\bar{y}_2)\right] \tag{7.39}$$

para su primera aplicación. En 7.39 $\bar{y}_2$ es estimada con un predictor, el cual a su vez requiere y_1 y $f(x_1, y_1)$. Así pues, antes de realizar la primera predicción deben evaluarse ciertos valores iniciales [en este caso y_1 y $f(x_1, y_1)$].

En esta evaluación se usa alguno de los métodos ya vistos (los de Runge-Kutta, por ejemplo). Este paso se utiliza sólo una vez en el proceso iterativo y se conoce como **paso de inicialización.**

Es evidente que para la predicción también puede utilizarse un método de los ya estudiados o, como se verá más adelante, puede derivarse un predictor usando las mismas ideas que condujeron a la ecuación 7.39.

Ejemplo 7.5

Resuelva el problema de valor inicial del ejemplo 7.1 utilizando el corrector dado por la ecuación 7.38 y el método Euler modificado como inicializador y como predictor.

SOLUCIÓN

El intervalo se divide otra vez en cinco subintervalos y se tiene

Primera iteración

Inicialización (se toma el valor de y_1 del ejemplo 7.3):

$$y_1 = 1.66$$

Predicción: (se toma el valor de y_2 del ejemplo 7.3) :

$$\bar{y}_2 = 1.4172$$

Corrección: Se utiliza la ecuación 7.39 (puede usar un pizarrón electrónico)

$$y\,(0.4) = y_2 = 2 + \frac{0.2}{3}\,[(0-2) + 4(0.2-1.66) + (0.4-1.4172)]$$

$$= 1.40952$$

Segunda iteración

Predicción

$$\bar{y}_3 = y_2 + \frac{h}{2}\,[f(x_2,y_2) + f(x_2 + h, y_2 + hf(x_2,y_2))]$$

$$= 1.40952 + \frac{0.2}{2}\,[(0.4-1.40952) + [(0.4+0.2) - (1.40952$$

$$+ 0.2(0.4 - 1.40952))]] = 1.2478064$$

Corrección (con la ecuación 7.38)

$$y(0.6) = y_3 = y_1 + \frac{h}{3}\,[f(x_1,y_1) + 4f(x_2,y_2) + f(x_3,\bar{y}_3)]$$

$$= 1.66 + \frac{0.2}{3}\,[(0.2-1.66) + 4(0.4-1.40952)$$

$$+ (0.6 - 1.2478064)] = 1.25027424$$

Tercera iteración.

Predicción

$$\bar{y}_4 = y_3 + \frac{h}{2}\,[f(x_3,y_3) + f(x_3 + h, y_3 + hf(x_3,y_3))]$$

$$= 1.25027424 + \frac{0.2}{2}\,[(0.6-1.25027424) + [(0.6+0.2)$$

$$- (1.25027424 + 0.2(0.6 - 1.25027424))]] = 1.153224877$$

Corrección (con la ecuación 7.38)

$$y(0.8) = y_4 = y_2 + \frac{h}{3}\,[f(x_2,y_2) + 4f(x_3,y_3) + f(x_4,\bar{y}_4)]$$

$$= 1.40952 + \frac{0.2}{3}\,[(0.4-1.40952) + 4(0.6-1.25027424)$$

$$+ (0.8 - 1.153224877)] = 1.145263878$$

Cuarta iteración

Predicción

$$\bar{y}_5 = y_4 + \frac{h}{2}\left[f(x_4, y_4) + f(x_4 + h, y_4 + hf(x_4, y_4))\right]$$

$$= 1.145263878 + \frac{0.2}{2}\left[(0.8 - 1.145263878) + [(0.8 + 0.2)\right.$$

$$\left. - (1.145263878 + 0.2(0.8 - 1.145263878))\right] = 1.10311638$$

Corrección (con la ecuación 7.38)

$$y(1) = y_5 = y_3 + \frac{h}{3}\left[f(x_3, y_3) + 4f(x_4, y_4) + f(x_5, \bar{y}_5)\right]$$

$$= 1.25027424 + \frac{0.2}{3}\left[(0.6 - 1.25027424) + 4(0.8 - 1.145263878)\right.$$

$$\left. + (1 - 1.10311638)\right] = 1.107977831$$

que da un error absoluto de 0.00434 y 0.0393 en porcentaje.

En general, puede obtenerse un corrector de cualquier orden utilizando la fórmula

$$y_{i+1} = y_{i-k} + \int_{x_{i-k}}^{x_{i+1}} f(x, y)\, dx, \qquad k = 0, 1, 2, \ldots \tag{7.40}$$

donde la integración se realiza sustituyendo $f(x, y)$ con un polinomio de grado $k+1$ que pasa por $(x_{i+1}, \bar{y}_{i+1})$, (x_i, y_i), ..., (x_{i-k}, y_{i-k}).

En virtud de que se está utilizando x_{i+1} y las abcisas previas a ésta y a sus espaciamientos regulares, lo más indicado para interpolar $f(x, y)$ es el polinomio de interpolación en su forma de diferencias hacia atrás, dado por la ecuación 5.38 del capítulo 5. La ecuación 7.40 queda entonces:

$$y_{i+1} = y_{i-k} + \int_{x_{i-k}}^{x_{i+1}} p(x_i + sh)\, dx, \tag{7.41}$$

Para la obtención de $p(x_i + sh)$, dada por la ecuación 5.38, se empleó el cambio de variable

$$x = x_i + sh,$$

que permite escribir la ecuación 7.41 en términos de la nueva variable s, ya que

$$\begin{aligned} dx &= h\, ds \\ x_{i+1} &= x_i + sh \qquad \text{de donde } s = 1 \\ x_{i-k} &= x_i + sh \qquad \text{de donde } s = -k \end{aligned} \tag{7.42}$$

Al sustituir se llega a

$$y_{i+1} = y_{i-k} + h \int_{-k}^{1} p(x_i + sh)\, ds$$

o bien

$$y_{i+1} = y_{i-k} + h \int_{-k}^{1} f(x_{i+1}, \bar{y}_{i+1}) + (s-1)\nabla f(x_{i+1}, \bar{y}_{i+1}) +$$

$$+ \frac{(s-1)s}{2!}\nabla^2 f(x_{i+1}, \bar{y}_{i+1}) + \frac{(s-1)s(s+1)}{3!}\nabla^3 f(x_{i+1}, \bar{y}_{i+1})$$

$$+ \ldots + \frac{(s-1)s(s+1)\ldots(s+r-2)}{r!}\nabla^r f(x_{i+1}, \bar{y}_{i+1})\,]\, ds$$

La disimilitud de los coeficientes de las diferencias hacia atrás con los de la ecuación 5.38 se debe a que se está utilizando x_{i+1} como punto base. Si se denota por $f_j = f(x_j, \bar{y}_j)$ para $j = i-k, i-k+1, \ldots, i+1$, la última ecuación queda

$$y_{i+1} = y_{i-k} + h \int_{-k}^{1} [f_{i+1} + (s-1)\nabla f_{i+1} + \frac{(s-1)s}{2!}\nabla^2 f_{i+1} +$$

$$+ \frac{(s-1)s(s+1)}{3!}\nabla^3 f_{i+1} + \ldots + \frac{(s-1)s\ldots(s+r-2)}{r!}\nabla^r f_{i+1}]\, ds \tag{7.43}$$

y al integrar se llega a

$$y_{i+1} = y_{i-k} + h[sf_{i+1} + s(\frac{s}{2} - 1)\nabla f_{i+1} + \frac{s^2(\frac{s}{3} - \frac{1}{2})}{2!}\nabla^2 f_{i+1}$$

$$+ \frac{s^2(\frac{s^2}{4} - \frac{1}{2})}{3!}\nabla^3 f_{i+1} + \frac{(\frac{s^5}{5} + \frac{s^4}{2} - \frac{s^3}{3} - s^2)}{4!}\nabla^4 f_{i+1} + \text{términos restantes}\,]\Big|_{-k}^{1} \tag{7.44}$$

Para $k = 0$, 1, 3 y 5, la ecuación 7.44 da

$k = 0$

$$y_{i+1} = y_i + h[f_{i+1} - \frac{1}{2}\nabla f_{i+1} - \frac{1}{12}\nabla^2 f_{i+1} - \frac{1}{24}\nabla^3 f_{i+1} +$$

$$\text{términos faltantes}\,] \tag{7.44a}$$

$k = 1$

$$y_{i+1} = y_{i-1} + h[2f_{i+1} - 2\nabla f_{i+1} + \frac{1}{3}\nabla^2 f_{i+1} + 0\nabla^3 f_{i+1}$$

$$- \frac{1}{90}\nabla^4 f_{i+1} + \text{términos faltantes}\,] \tag{7.44b}$$

$k = 3$

$$y_{i+1} = y_{i-3} + h\left[4f_{i+1} - 8\nabla f_{i+1} + \frac{20}{3}\nabla^2 f_{i+1} - \frac{8}{3}\nabla^3 f_{i+1} + \frac{14}{45}\nabla^4 f_{i+1} + 0\,\nabla^5 f_{i+1} + \text{términos faltantes}\right] \tag{7.44c}$$

$k = 5$

$$y_{i+1} = y_{i-5} + h\left[6f_{i+1} - 18\nabla f_{i+1} + 27\,\nabla^2 f_{i+1} - 24\,\nabla^3 f_{i+1} + \frac{123}{10}\nabla^4 f_{i+1} - \frac{33}{10}\nabla^5 f_{i+1} + \text{términos faltantes}\right] \tag{7.44d}$$

Independientemente del valor que se elija para k, se debe seleccionar también el orden del corrector, el cual está dado en estas fórmulas por el orden r más uno de la diferencia hacia atrás de más alto orden que se utilice. Por ejemplo, para correctores de cuarto orden cabe emplear, entre otras, las combinaciones

$k = 0, r = 3$

$$y_{i+1} = y_i + h\left[f_{i+1} - \frac{1}{2}\nabla f_{i+1} - \frac{1}{12}\nabla^2 f_{i+1} - \frac{1}{24}\nabla^3 f_{i+1}\right] \tag{7.45a}$$

$k = 1, r = 3$

$$y_{i+1} = y_{i-1} + h\left[2f_{i+1} - 2\nabla f_{i+1} + \frac{1}{3}\nabla^2 f_{i+1} + 0\nabla^3 f_{i+1}\right] \tag{7.45b}$$

Por ejemplo, para el orden sexto se usa

$k = 3, r = 5$

$$y_{i+1} = y_{i-3} + h\left[4f_{i+1} - 8\nabla f_{i+1} + \frac{20}{3}\nabla^2 f_{i+1} - \frac{8}{3}\nabla^3 f_{i+1} + \frac{14}{45}\nabla^4 f_{i+1}\right] \tag{7.45c}$$

Si se desarrollan las diferencias hacia atrás en estas fórmulas, se obtienen versiones de 7.45a, 7.45b y 7.45c más útiles para programar; es decir

$k = 0, r = 3$

$$y_{i+1} = y_i + \frac{h}{24}\left[9f_{i+1} + 19f_i - 5f_{i-1} + f_{i-2}\right] \tag{7.46a}$$

$k = 1, r = 3$

$$y_{i+1} = y_{i-1} + \frac{h}{3}\left[f_{i+1} + 4f_i + f_{i-1}\right] \tag{7.46b}$$

$k = 3, r = 5$

$$y_{i+1} = y_{i-3} + \frac{2h}{45}\left[7f_{i+1} + 32f_i + 12f_{i-1} + 32f_{i-2} + 7f_{i-3}\right] \tag{7.46c}$$

Esta familia de correctores se conoce como **correctores de Adams-Moulton** y uno de los más usados es la ecuación 7.46a, la cual toma la forma

$$y_3 = y_2 + \frac{h}{24}\,[\,9f_3 + 19f_2 - 5f_1 + f_0\,]$$

para su primera aplicación o, regresando a la notación original

$$y_3 = y_2 + \frac{h}{24}\,[\,9f(x_3, \bar{y}_3) + 19f(x_2, y_2) - 5f(x_1, y_1) + f(x_0, y_0)\,] \quad (7.47)$$

donde $y_1, f(x_1, y_1)$; $y_2, f(x_2, y_2)$ deben calcularse previamente por un inicializador y $\bar{y}_3$ por un predictor. No podría emplearse este corrector para calcular, por ejemplo, y_2, ya que tomaría la forma

$$y_2 = y_1 + \frac{h}{24}\,[\,9f(x_2, \bar{y}_2) + 19f(x_1, y_1) - 5f(x_0, y_0) + f(x_{-1}, y_{-1})\,]$$

que requiere información en la abscisa x_{-1} que está fuera del intervalo de interés.

Ejemplo 7.6

Resuelva el PVI del ejemplo 7.1 con el corrector de la ecuación 7.46a.

SOLUCIÓN

El intervalo de interés [0,1] se vuelve a dividir en cinco subintervalos y se usa el método de Runge Kutta de cuarto orden tanto de inicializador como de predictor. Es conveniente utilizar un inicializador y un predictor del mismo orden que el corrector.

Primera iteración

Inicialización con RK-4 (se toman los valores del ejemplo 7.4)

$$y\,(0.2) = 1.656200000 = y_1$$
$$y\,(0.4) = 1.410972813 = y_2$$

Predicción con RK-4 (se toma el valor del ejemplo 7.4)

$$y(0.6) = 1.246450474 = \bar{y}_3$$

Corrección con la ecuación 7.47

$$y_3 = 1.410972813 + \frac{0.2}{24}\,[\,9\,(0.6\text{–}1.246450474\,) +$$

$$19\,(\,0.4 - 1.410972813\,) - 5\,(\,0.2 - 1.6562\,) + (\,0 - 2\,)\,] = 1.246426665$$

Segunda iteración

Predicción con RK-4
Cálculo de las constantes k_1, k_2, k_3 y k_4

$$k_1 = f(x_3, y_3) = (0.6\text{-}1.24642665) = \text{-}0.646426665$$

$$k_2 = f(x_3+h/2, y_3+hk_1/2) = [(0.6+0.2/2)-(1.246426665 + 0.2(-0.646426665)/2)] = -0.481783999$$

$$k_3 = f(x_3+h/2, y_3+hk_2/2) = [(0.6+0.2/2)-(1.246426665 + 0.2(-.481783999)/2)] = -0.498248265$$

$$k_4 = f(x_3+h, y_3+hk_3) = [(0.6+0.2)-(1.246426665 + 0.2(-0.498248265))] = -0.346777012$$

Cálculo de $\bar{y}_4$

$$\bar{y}_4 = y_3 + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

$$= 1.246426665 + \frac{0.2}{6}(-0.646426665 + 2(-0.481783999) + 2(-0.498248265) - 0.346777012) = 1.147984392$$

Corrección con la ecuación 7.46a

$$y_4 = y_3 + \frac{h}{24}[9f(x_4, \bar{y}_4) + 19f(x_3, y_3) - 5f(x_2, y_2) + f(x_1, y_1)]$$

$$= 1.246426665 + \frac{0.2}{24}[9(0.8\text{–}1.147984392) + 19(0.6 - 1.246426665) - 5(0.4 - 1.410972813) + (0.2 - 1.6562)] = 1.147965814$$

Tercera iteración

Predicción con RK-4 $\quad \bar{y}_5 = 1.103624544$

Corrección con (7.46a) $\quad y_5 = 1.103609057$

con un error absoluto de 0.0000292 y un error porcentual de 0.00265.

Métodos de predicción

Anteriormente se habló de una familia de predictores obtenida a partir del mismo principio de integración que se empleó para los métodos de Adams-Moulton. A esta familia, que se deduce a continuación, se le llama **métodos de Adams-Bashforth.**

En general, para obtener un predictor de cualquier orden se utiliza la fórmula 7.40

$$y_{i+1} = y_{i-k} + \int_{x_{i-k}}^{x_{i+1}} f(x,y)\, dx,$$

pero ahora la integración se realiza sustituyendo $f(x, y)$ con un polinomio de grado k que pasa por $(x_i, y_i),\ldots,(x_{i-k}, y_{i-k})$; (véase Fig. 7.7). Obviamente, se utiliza el polinomio de interpolación en su forma de diferencias hacia atrás, pues $x_i, \ldots, x_{i-k}$ están regularmente espaciadas. Entonces, al aplicar la ecuación 5.38 se obtiene

$$y_{i+1} = y_{i-k} + \int_{x_{i-k}}^{x_{i+1}} p(x_i + sh)\, ds,$$

donde los límites de integración y dx en términos de la nueva variable s quedan como en la ecuación 7.42. Por tanto

$$\begin{aligned} y_{i+1} &= y_{i-k} + h \int_{-k}^{1} p(x_i + sh)\, ds, \\ y_{i+1} &= y_{i-k} + h \int_{-k}^{1} \left[f_i + s\nabla f_i + s(s+1)\frac{\nabla^2 f_i}{2!} \right. \\ &\left. + s(s+1)(s+2)\frac{\nabla^3 f_i}{3!} + \ldots + s(s+1)(s+2)\ldots(s+k-1)\frac{\nabla^r f_i}{r!} \right] ds \end{aligned} \tag{7.48}$$

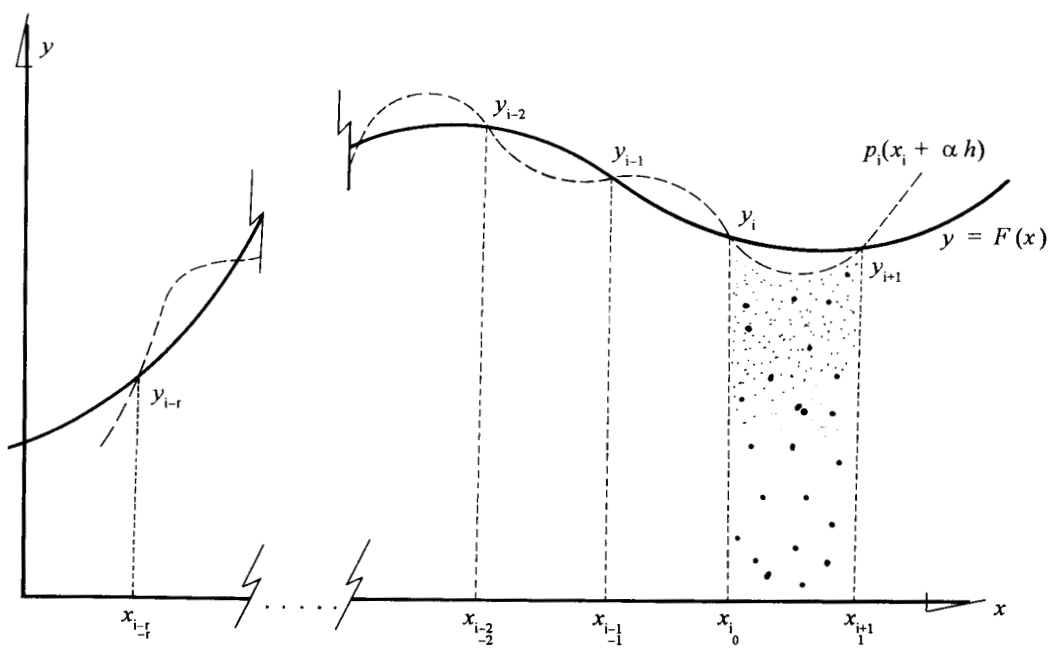


Figura 7.7. Métodos de Adams-Bashforth.

Nótese que ahora el integrando es exactamente la ecuación 5.38, ya que en esta ocasión se está utilizando x_i como punto base. Al integrar la ecuación 7.48 se obtiene

$$y_{i+1} = y_{i-k} + h\left[sf_i + \frac{s^2}{2}\nabla f_i + s^2\left(\frac{s}{3} + \frac{1}{2}\right)\frac{\nabla^2 f_i}{2!} + \right. \tag{7.49}$$

$$\left. s^2\left(\frac{s^2}{4} + s + 1\right)\frac{\nabla^3 f_i}{3!} + s^2\left(\frac{s^3}{5} + \frac{3s^2}{2} + \frac{11s}{3} + 3\right)\frac{\nabla^4 f_i}{4!} + \text{términos faltantes}\right] \Big|_{-k}^{1}$$

La ecuación 7.49 para $k = 0, 1, 2$ y 3 toma las formas

$k = 0$

$$y_{i+1} = y_i + h\left[f_i + \frac{1}{2}\nabla f_i + \frac{5}{12}\nabla^2 f_i + \frac{3}{8}\nabla^3 f_i + \frac{251}{720}\nabla^4 f_i + \text{términos faltantes}\right] \tag{7.50a}$$

$k = 1$

$$y_{i+1} = y_{i-1} + h\left[2f_i + 0\nabla f_i + \frac{1}{3}\nabla^2 f_i + \frac{1}{3}\nabla^3 f_i + \frac{29}{90}\nabla^4 f_i + \text{términos faltantes}\right] \tag{7.50b}$$

$k = 2$

$$y_{i+1} = y_{i-2} + h\left[3f_i - \frac{3}{2}\nabla f_i + \frac{3}{4}\nabla^2 f_i + \frac{3}{8}\nabla^3 f_i + \frac{27}{80}\nabla^4 f_i + \text{términos faltantes}\right] \tag{7.50c}$$

$k = 3$

$$y_{i+1} = y_{i-3} + h\left[4f_i - 4\nabla f_i + \frac{8}{3}\nabla^2 f_i + 0\nabla^3 f_i + \frac{14}{45}\nabla^4 f_i + \text{términos faltantes}\right] \tag{7.50d}$$

La **7.50a** significa la integración aproximada de una función que pasa por los puntos (x_{i-r}, y_{i-r}), (x_{i-r+1}, y_{i-r+1}), ... , (x_i, y_i), donde el subíndice r representa el grado del polinomio que se toma y $r+1$ da el orden del predictor. El intervalo de integración es $[x_i, x_{i+1}]$ (véase Fig. 7.7).

La ecuación 7.50b usa los mismos puntos, pero con intervalo de integración $[x_{i-1}, x_{i+1}]$.

Las fórmulas más usadas de esta familia son

$k = 0, \quad r = 3$

$$y_{i+1} = y_i + h\left[f_i + \frac{1}{2}\nabla f_i + \frac{5}{12}\nabla^2 f_i + \frac{3}{8}\nabla^3 f_i\right] \tag{7.51}$$

$k = 1,\ r = 1$

$$y_{i+1} = y_{i-1} + h\,[2f_i + 0\nabla f_i] \tag{7.52}$$

$k = 3,\ r = 3$

$$y_{i+1} = y_{i-3} + h\left[4f_i - 4\nabla f_i + \frac{8}{3}\nabla^2 f_i + 0\nabla^3 f_i\right] \tag{7.53}$$

$k = 5,\ r = 5$

$$y_{i+1} = y_{i-5} + h\left[6f_i - 12\nabla f_i + 15\nabla^2 f_i - 9\nabla^3 f_i + \frac{33}{10}\nabla^4 f_i + 0\nabla^5 f_i\right] \tag{7.54}$$

cuya apariencia al desarrollar los operadores en diferencias hacia atrás resulta

$k = 0,\ r = 3$

$$y_{i+1} = y_i + \frac{h}{24}[55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}] \tag{7.55}$$

$k = 1, r = 1$

$$y_{i+1} = y_{i-1} + 2hf_i \tag{7.56}$$

$k = 3, r = 3$

$$y_{i+1} = y_{i-3} + \frac{4h}{3}[2f_i - f_{i-1} + 2f_{1-2}] \tag{7.57}$$

$k = 5, r = 5$

$$y_{i+1} = y_{i-5} + \frac{3h}{10}[11f_i - 14f_{i-1} + 26f_{i-2} - 14f_{i-3} + 11f_{i-4}] \tag{7.58}$$

Es importante hacer notar que estas fórmulas son métodos para resolver el PVI (Ec. 7.11).

La ecuación 7.55 toma la forma

$$y_4 = y_3 + \frac{h}{24}[55f(x_3, y_3) - 59f(x_2, y_2) + 37f(x_1, y_1) - 9f(x_0, y_0)] \tag{7.59}$$

para su primera aplicación, y no sería posible determinar con ella un valor de y menor de y_4 (y_3 por ejemplo). Por otro lado, y_1, $f(x_1, y_1)$; y_2, $f(x_2, y_2)$ y y_3, $f(x_3, y_3)$ deberán determinarse con un inicializador.

Con estos métodos y la familia de los Adams-Moulton pueden integrarse esquemas iterativos conocidos como **métodos de predicción-corrección,** que en general funcionan como sigue.

1. Inicialización* (se sugiere uno de la familia de Runge Kutta).
2. Predictor (para corresponder con el inicializador se sugiere usar un predictor del mismo orden).
3. Corrección (se emplea un corrector del mismo orden que el predictor y el inicializador).

Ejemplo 7.7

Resuelva el PVI del ejemplo 7.1 usando como inicializador un RK-4, como predictor la ecuación 7.55 y como corrector la 7.46a.

SOLUCIÓN

El intervalo de interés [0, 1] se divide nuevamente en cinco subintervalos y se tiene

Primera iteración

Inicialización (tómense nuevamente los valores del ejemplo 7.4)

$$y_1 = 1.656200000$$

$$y_2 = 1.410972813$$

$$y_3 = 1.246450474$$

Predicción

$$\bar{y}_4 = 1.246450474 + \frac{0.2}{24}[55(0.6-1.246450474) - 59(0.4 - 1.410972813) + 37(0.2 - 1.6562) - 9(0 - 2)] = 1.148227306$$

Corrección (con la ecuación 7.46a)

$$y_4 = y_3 + \frac{h}{24}[9f(x_4, \bar{y}_4) + 19f(x_3, y_3) - 5f(x_2, y_2) + f(x_1, y_1)]$$

$$= 1.246450474 + \frac{0.2}{24}[9(0.8-1.148227306) + 19(0.6 - 1.246450474) - 5(0.4-1.410972813) + (0.2-1.6562)] = 1.147967635$$

*Recuérdese que este paso sólo se da en la primera iteración.

Segunda iteración

Predicción

$$\bar{y}_5 = y_4 + \frac{h}{24}\,[55f(x_4,y_4) - 59f(x_3,y_3) + 37f(x_2,y_2) - 9f(x_1,y_1)]$$

$$= 1.147967635 + \frac{0.2}{24}\,[55\,(0.8 - 1.147967635) - 59\,(0.6 - 1.246450474)$$

$$+ 37\,(0.4 - 1.410972813) - 9\,(0.2 - 1.6562)] = 1.103819001$$

Corrección (con la ecuación 7.46a)

$$y_5 = y_4 + \frac{h}{24}\,[9f(x_5, \bar{y}_5) + 19f(x_4,y_4) - 5f(x_3,y_3) + f(x_2,y_2)]$$

$$= 1.147967635 + \frac{0.2}{24}\,[9\,(1 - 1.103819001) + 19\,(0.8 - 1.147967635)$$

$$- 5\,(0.6 - 1.246450474) + (0.4 - 1.410972813)] = 1.103596997$$

con un error absoluto de 0.00004 y un error porcentual de 0.0037.

Nótese que aunque el corrector puede emplearse para mejorar y_3 en su primera aplicación (véase Ej. 7.6), el predictor estima a y_4 en su primera aplicación y a partir de ahí se comienza a corregir. Ésta es sólo una de las muchas formas en que se utilizan estos métodos de predicción-corrección.

ALGORITMO 7.4 Método predictor-corrector

(Inicialización con Runge-Kutta de cuarto orden, predicción con la ecuación 7.55 y corrección con la 7.46a).

Para obtener la aproximación YF a la solución de un PVI, proporcionar la función F(X,Y) y los

DATOS: La condición inicial X0, Y0, el valor XF donde se desea conocer el valor de YF y el número N de subintervalos por emplear.

RESULTADOS: Aproximación a YF: Y(4).

PASO 1. Hacer H = (XF - X0)/N
PASO 2. Hacer X(0) = X0
PASO 3. Hacer Y(0) = Y0
PASO 4. Hacer J = 1
PASO 5. Mientras J ≤ 3, repetir los pasos 6 a 9.

```
        PASO 6.  Realizar los pasos 4 a 9 del algoritmo 7.3.
        PASO 7.  Hacer X(J) = X0
        PASO 8.  Hacer Y (J) = Y0
        PASO 9.  Hacer J = J + 1
PASO 10. Hacer I = 4
PASO 11. Mientras I ≤ N, repetir los pasos 12 a 20.
        PASO 12. Hacer Y(4)=Y(3)+H/24*(F(X(3),Y(3))-
                 59*F(X(2),Y(2))+37*F(X(1),Y(1))-9*F(X0), Y(0)))
        PASO 13. Hacer X(4) = X(3) + H
        PASO 14  Hacer Y(4)=Y(3)+H/24*(9*F(X(4), Y(4)) +
                 19*F(X(3),Y(3))-5*F(X(2), Y(2))+F(X(1),Y(1)))
        PASO 15. Hacer J = 0
        PASO 16. Mientras J ≤ 3, repetir los pasos 17 a 19.
                 PASO 17.  Hacer X(J) = X(J+1)
                 PASO 18   Hacer Y (J) = Y (J+1)
                 PASO 19.  Hacer J = J + 1
        PASO 20. Hacer I = I + 1
PASO 21. IMPRIMIR Y(4) y TERMINAR.
```

SECCIÓN 7.7 ECUACIONES DIFERENCIALES ORDINARIAS DE ORDEN SUPERIOR Y SISTEMAS DE ECUACIONES DIFERENCIALES ORDINARIAS

Cuando en el problema de valor inicial aparecen una ecuación diferencial de orden n, n condiciones especificadas en un punto x_0 y un punto x_f donde hay que encontrar el valor de $y(x_f)$, se tiene el problema de valor inicial general (PVIG)

$$\text{PVIG} \quad \begin{cases} \dfrac{d^n y}{dx^n} = f(x, y, y', y'', \ldots, y^{(n-1)}) \\ y(x_0) = y_0, y'(x_0) = y_0', \ldots, y^{(n-1)}(x_0) = y_0^{(n-1)} \\ y(x_f) = ? \end{cases} \tag{7.60}$$

Para resolver la ecuación anterior no se desarrollan nuevos métodos, sino que se emplea una extensión de los estudiados en este capítulo. Para ello necesitaremos primero pasar la ecuación diferencial ordinaria o EDO de la ecuación 7.60 a un **sistema de n ecuaciones diferenciales simultáneas de primer orden** cada una. Esto se logra de la siguiente manera

Sea dada

$$\frac{d^n y}{dx^n} = f(x, y, y', y'', \ldots, y^{(n-1)})$$

Se efectúa el siguiente cambio de variables

$$\begin{aligned} y_1 &= y \\ y_2 &= y' \\ y_3 &= y'' \\ y_4 &= y''' \\ &\vdots \\ y_n &= y^{(n-1)} \end{aligned}$$

Se deriva miembro a miembro la primera y se sustituye en la segunda, con lo que se obtiene

$$y'_1 = y_2$$

Al derivar la segunda y sustituir en la tercera resulta

$$y'_2 = y_3$$

El procedimiento se repite hasta llegar al sistema de n ecuaciones de primer orden siguiente

$$\begin{aligned} y'_1 &= y_2 \\ y'_2 &= y_3 \\ y'_3 &= y_4 \\ &\vdots \\ y'_{n-1} &= y_n \end{aligned}$$

$$y_n' = \frac{d^n y}{dx^n} = f(x, y, y', y'', \ldots, y^{(n-1)}) = f(x, y_1, y_2, y_3, \ldots, y_n)$$

Ejemplo 7.8

Pase la ecuación diferencial ordinaria

$$\frac{d^2y}{dx^2} + \frac{dy}{dx} = x^2 + y^2$$

a un sistema de dos ecuaciones diferenciales ordinarias simultáneas de primer orden.

SOLUCIÓN

Con el despeje de la derivada de segundo orden se tiene

$$\frac{d^2y}{dx^2} = -y' + x^2 + y^2$$

El cambio de variables es

$$y_1 = y \; ; y_2 = y'$$

Al derivar la primera y sustituir en la segunda queda

$$y'_1 = y_2$$

Se deriva la segunda

$$y'_2 = y''$$

y las nuevas variables se sustituyen en la ecuación diferencial, con lo cual resulta

$$y'_1 = y_2$$
$$y'_2 = -y_2 + x^2 + y_1^2$$

Ejemplo 7.9

Una de las ecuaciones diferenciales ordinarias más empleadas en la matemática física es la ecuación de Bessel

$$x^2 y'' + x y' + (x^2 - n^2) y = 0$$

donde n puede tener cualquier valor, pero generalmente toma un valor entero. Escriba esta ecuación como un sistema de ecuaciones diferenciales ordinarias de primer orden.

SOLUCIÓN

La ecuación se pone en la forma normal

$$y'' = -\frac{1}{x} y' + \left(\frac{n^2}{x^2} - 1\right) y$$

Algunas veces es más conveniente para los cálculos computacionales emplear

$$y = y$$
$$y' = z$$

como nuevas variables. Se deriva la segunda y se tiene

$$y'' = z'$$

El sistema queda

$$y' = z$$
$$z' = -\frac{1}{x} z + \left(\frac{n^2}{x^2} - 1\right) y$$

sistema que solo podrá resolverse para valores de x distintos de cero.

En general, una ecuación diferencial ordinaria de n-ésimo orden queda convertida en un sistema de n ecuaciones diferenciales ordinarias simultáneas de la forma general

$$\begin{aligned} y'_1 &= f_1(x, y_1, y_2, \ldots, y_n) \\ y'_2 &= f_2(x, y_1, y_2, \ldots, y_n) \\ &\vdots \\ y'_n &= f_n(x, y_1, y_2, \ldots, y_n) \end{aligned}$$

que puede resolverse aplicando, por ejemplo, alguno de los métodos de Runge-Kutta a cada ecuación, e iterando cada ecuación en turno, tal como en los sistemas de ecuaciones no lineales del capítulo 4, o los métodos de predicción-corrección.

Si se aplica, por ejemplo, el método de Runge-Kutta de cuarto orden a dos ecuaciones simultáneas de la forma

$$\begin{aligned} y' &= f_1(x,y,z) \\ z' &= f_2(x,y,z) \end{aligned}$$

donde sólo se emplea z como nueva variable con el fin de no usar subíndices dobles en las ecuaciones

$$\begin{aligned} y_{i+1} &= y_i + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\ z_{i+1} &= z_i + \frac{h}{6}(c_1 + 2c_2 + 2c_3 + c_4) \end{aligned} \tag{7.61a}$$

las cuales se calculan alternadamente, y las k y c se obtienen de

$$\begin{aligned} k_1 &= f_1(x_i, y_i, z_i) \\ c_1 &= f_2(x_i, y_i, z_i) \\ k_2 &= f_1(x_i + h/2, y_i + hk_1/2, z_i + hc_1/2) \\ c_2 &= f_2(x_i + h/2, y_i + hk_1/2, z_i + hc_1/2) \\ k_3 &= f_1(x_i + h/2, y_i + hk_2/2, z_i + hc_2/2) \\ c_3 &= f_2(x_i + h/2, y_i + hk_2/2, z_i + hc_2/2) \\ k_4 &= f_1(x_i + h, y_i + hk_3, z_i + hc_3) \\ c_4 &= f_2(x_i + h, y_i + hk_3, z_i + hc_3) \end{aligned} \tag{7.61b}$$

calculadas en ese orden.

Ejemplo 7.10

Resuelva el siguiente problema de valor inicial por el método de Runge-Kutta de cuarto orden. (Puede usar el software del libro, el GC o un pizarrón electrónico).

$$\text{PVI} \begin{cases} y'' = -\frac{1}{x}y' + (\frac{1}{x^2} - 1)y \\ y(1) = 1 \\ y'(1) = 2 \\ y(3) = ? \end{cases}$$

Nótese que la EDO es la ecuación de Bessel con $n = 1$ (véase Ejem. 7.9). Al escribir la EDO como un sistema, el PVI queda

$$\text{PVI} \begin{cases} y' = z \\ z' = -\frac{1}{x}z + (\frac{1}{x^2} - 1)y \\ y(1) = 1 \\ z(1) = 2 \\ y(3) = ? \end{cases}$$

SOLUCIÓN

Al dividir el intervalo de interés [1, 3] en ocho subintervalos, el tamaño del paso de integración h es igual a 0.25

Primera iteración (usando la ecuación 7.61a)

Cálculo de las constantes k y c con 7.61b

$$k_1 = f_1(x_0, y_0, z_0) = z_0 = z(1) = 2$$

$$c_1 = f_2(x_0, y_0, z_0) = \frac{-1}{x_0}z_0 + (\frac{1}{x_0^2} - 1)y_0$$

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

$$k_2 = f_1(x_0 + h/2, y_0 + hk_1/2, z_0 + hc_1/2)$$

$$= z_0 + hc_1/2 = 2 + 0.25(-2)/2 = 1.75$$

$$c_2 = f_2(x_0 + h/2, y_0 + hk_1/2, z_0 + hc_1/2)$$

$$= -\frac{1}{x_0 + h/2}(z_0 + hc_1/2) + [\frac{1}{(x_0 + h/2)^2} - 1](y_0 + hk_1/2)$$

$$= -\frac{1}{1 + 0.25/2}(2 + 0.25(-2)/2) + [\frac{1}{(1 + 0.25/2)^2} - 1](1 + 0.25(2)/2)$$

$$= -1.817901235$$

$$k_3 = f_1 (x_0 + h/2, y_0 + hk_2/2, z_0 + hc_2/2) = z_0 + hc_2/2$$

$$= 2 + 0.25(-1.817901235)/2 = 1.772762346$$

$$c_3 = f_2 (x_0 + h/2, y_0 + hk_2/2, z_0 + hc_2/2)$$

$$= -\frac{1}{x_0 + h/2} (z_0 + hc_2/2) + \left[\frac{1}{(x_0 + h/2)^2} - 1\right] (y_0 + hk_2/2)$$

$$= -\frac{1}{1 + 0.25/2} (2 + 0.25(-1.817901235)/2) +$$

$$\left[\frac{1}{(1 + 0.25/2)^2} - 1\right](1 + 0.25(1.75)/2) = -1.831575789$$

$$k_4 = f_1 (x_0 + h, y_0 + hk_3, z_0 + hc_3) = z_0 + hc_3$$

$$= 2 + 0.25 (-1.831575789) = 1.542106053$$

$$c_4 = f_2(x_0 + h, y_0 + hk_3, z_0 + hc_3)$$

$$= -\frac{1}{x_0 + h} (z_0 + hc_3) + \left[\frac{1}{(x_0 + h)^2} - 1\right] (y_0 + hk_3)$$

$$= -\frac{1}{1 + 0.25} (2 + 0.25(-1.831575789))$$

$$+ \left[\frac{1}{(1 + 0.25)^2} - 1\right](1 + 0.25 (1.772762346)) = -1.753233454$$

Cálculo de $y_1 = y(1.25)$ y $z_1 = z(1.25)$ con la ecuación 7.61a

$$y_1 = y_0 + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

$$= 1 + \frac{0.25}{6} [2 + 2(1.75) + 2(1.772762346) + 1.542106053]$$

$$= 1.441151281$$

$$z_1 = z_0 + \frac{h}{6}(c_1 + 2c_2 + 2c_3 + c_4)$$

$$= 2 + \frac{0.25}{6}[-2 + 2(-1.817901235) + 2(-1.831575789)$$

$$-1.753233454] = 1.539492187$$

Segunda iteración

Cálculo de las k y c con la ecuación 7.61b

$$k_1 = f_1 (x_1, y_1, z_1) = z_1 = 1.539492187$$

$$c_1 = f_2(x_1, y_1, z_1) = -\frac{1}{x_1} z_1 + \left(\frac{1}{x_1^2} - 1\right) y_1$$

$$= -\frac{1}{1.25}(1.539492187) + \left(\frac{1}{(1.25)^2} - 1\right)(1.441151281)$$

$$= -1.750408211$$

$$k_2 = f_1(x_1 + h/2, y_1 + hk_1/2, z_1 + hc_1/2) = z_1 + hc_1/2$$

$$= 1.539492187 + 0.25(-1.750408211)/2 = 1.320691161$$

$$c_2 = f_2(x_1 + h/2, y_1 + hk_1/2, z_1 + hc_1/2)$$

$$= -\frac{1}{x_1 + h/2}(z_1 + hc_1/2) + \left[\frac{1}{(x_1 + h/2)^2} - 1\right](y_1 + hk_1/2)$$

$$= -\frac{1}{1.25 + 0.25/2}(1.539492187 + 0.25(-1.750408211)/2)$$

$$+ \left[\frac{1}{(1.25 + 0.25/2)^2} - 1\right](1.441151281 + 0.25(1.539492187)/2)$$

$$= -1.730044025$$

$$k_3 = f_1(x_1 + h/2, y_1 + hk_2/2, z_1 + hc_2/2) = z_1 + hc_2/2$$

$$= 1.539492187 + 0.25(-1.730044025)/2 = 1.323236684$$

$$c_3 = f_2(x_1 + h/2, y_1 + hk_2/2, z_1 + hc_2/2)$$

$$= -\frac{1}{x_1 + h/2}(z_1 + hc_2/2) + \left[\frac{1}{(x_1 + h/2)^2} - 1\right](y_1 + hk_2/2)$$

$$= -\frac{1}{1.25 + 0.25/2}(1.539492187 + 0.25(-1.730044025)/2)$$

$$+ \left[\frac{1}{(1.25 + 0.25/2)^2} - 1\right](1.441151281 + 0.25(1.320691161)/2)$$

$$= -1.71901137$$

$$k_4 = f_1(x_1 + h, y_1 + hk_3, z_1 + hc_3) = z_1 + hc_3$$

$$= 1.539492187 + 0.25(-1.71901137) = 1.109739345$$

$$c_4 = f_2(x_1 + h, y_1 + hk_3, z_1 + hc_3)$$

$$= -\frac{1}{x_1 + h}(z_1 + hc_3) + \left[\frac{1}{(x_1 + h)^2} - 1\right](y_1 + hk_3)$$

$$= -\frac{1}{1.25 + 0.25}\,(1.539492187 + 0.25\,(-1.71901137))$$

$$+ \left[\frac{1}{(1.25 + 0.25)^2} - 1\right](1.441151281 + 0.25\,(1.323236684))$$

$$= -1.724248703$$

Cálculo de $y_2 = y(1.5)$ y $z_2 = z(1.5)$ con la ecuación 7.61a

$$y_2 = y_1 + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

$$= 1.441151281 + \frac{0.25}{6}[1.539492187 + 2(1.320691161) +$$

$$2(1.323236684) + 1.109739345] = 1.771863249$$

$$z_2 = z_1 + \frac{h}{6}(c_1 + 2c_2 + 2c_3 + c_4)$$

$$= 1.539492187 + \frac{0.25}{6}[-1.750408211 + 2(-1.730044025)$$

$$+ 2(-1.71901137) - 1.724248703] = 1.107293533$$

Se continúa calculando en la misma forma y se obtiene

$y(1.75) = 1.994766280$	$z(1.75) = 0.675599895$
$y(2.00) = 2.109754328$	$z(2.00) = 0.245291635$
$y(2.25) = 2.118486566$	$z(2.25) = -0.172076357$
$y(2.50) = 2.026089844$	$z(2.50) = -0.561053191$
$y(2.75) = 1.841680320$	$z(2.75) = -0.905578495$
$y(3.00) = 1.578253875$	$z(3.00) = -1.190934201$

El valor buscado es $y(3) = 1.578253875$

Si en el PVI está dado un sistema de EDO con sus correspondientes condiciones iniciales, el procedimiento es el mismo (consúltense los ejercicios).

A continuación se presenta un algoritmo para el método de Runge-Kutta de cuarto orden con objeto de resolver un sistema de dos ecuaciones diferenciales ordinarias.

ALGORITMO 7.5 Método de Runge-Kutta de cuarto orden para un sistema de dos ecuaciones diferenciales ordinarias

Para aproximar la solución al PVI

$$y' = f_1(x, y, z)$$
$$z' = f_2(x, y, z)$$
$$y(x_0) = y_0, y(x_f) = ?$$
$$z(x_0) = z_0, z(x_f) = ?,$$

proporcionar las funciones F1(X,Y,Z) y F2(X,Y,Z) y los

DATOS: La condición inicial X0, Y0, Z0, el valor XF y el número de N de subintervalos por emplear.
RESULTADOS: La aproximación a los valores Y(XF) y Z(XF)): Y0 y Z0.

PASO 1. Hacer H = (XF - X0)/N
PASO 2. Hacer I = 1
PASO 3. Mientras I ≤ N, repetir los pasos 4 a 15.
PASO 4. Hacer K1=F1 (X0, Y0, Z0)
PASO 5. Hacer C1=F2(X0, Y0, Z0)
PASO 6. Hacer K2=F1(X0+H/2,Y0+H/2*K1,Z0+H/2*C1)
PASO 7. Hacer C2=F2(X0+H/2,Y0+H/2*K1,Z0+H/2*C1)
PASO 8. Hacer K3=F1(X0+H/2,Y0+H/2*K2,Z0+H/2*C2)
PASO 9. Hacer C3=F2(X0+H/2,Y0+H/2*K2,Z0+H/2*C2)
PASO 10. Hacer K4=F1(X0+H,Y0+H*K3,Z0+H*C3)
PASO 11. Hacer C4=F2(X0+H,Y0+H*K3,Z0+H*C3)
PASO 12. Hacer Y0=Y0+H/6*(K1+2*K2+2*K3+K4)
PASO 13. Hacer Z0=Z0+H/6*(C1+2*C2+2*C3+C4)
PASO 14. Hacer X0=X0 + H
PASO 15. Hacer I = I + 1
PASO 16. IMPRIMIR Y0, Z0 y TERMINAR.

Ejercicios

7.1 Un tanque cilíndrico de fondo plano con un diámetro de 1.5 m (Fig. 7.8), contiene un líquido de densidad ρ = 1.5 kg/l a una altura a de 3 m. Se desea saber la altura del líquido dentro del tanque tres minutos después de que se abre completamente la válvula de salida, la cual da un gasto de $0.6A\sqrt{2ga}$ m^3/s, donde A es el área seccional del tubo de salida y es 78.5×10^{-4} m^2 y g=9.81 m/s^2.

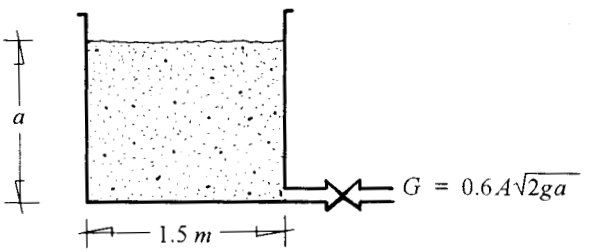


Figura 7.8 Vaciado de un tanque cilíndrico.

SOLUCIÓN

Como se vio en el ejemplo de la sección 7.1, el vaciado o llenado de un tanque cilíndrico se modela haciendo un balance de materia con la siguente expresión

$$\begin{array}{cccccc} \text{Acumulación} & = & \text{Entrada} & - & \text{Salida} \\ \dfrac{dV\rho}{dt} & = & 0 & - & 0.6\text{A}\sqrt{2ga} \end{array}$$

donde

$$\text{V} = \frac{\pi}{4}(1.5)^2 a$$

entonces

$$\frac{\pi}{4}(1.5)^2 \frac{da}{dt} = -0.6\text{A}\sqrt{2g\,a}$$

de donde

$$\frac{da}{dt} = -\frac{2.4\text{A}\sqrt{2g\,a}}{\pi(1.5)^2} = -0.0026653\sqrt{2g\,a}$$

Al considerar como tiempo cero el momento de abrir la válvula y además la altura buscada a un tiempo de 180 s, se llega a

$$\text{PVI}\begin{cases} \dfrac{da}{dt} = -0.0026653\sqrt{2g\,a} \\ a(0) = 3\,m \\ a(180) = ? \end{cases}$$

En virtud de que la exactitud de los resultados que se esperan no es grande, se usa el método de Euler para resolver este PVI.

Los resultados que se obtienen con $h = 30$ s son

tiempo (s)	0	30	60	90	120	150	180
a (m)	3.00	2.39	1.84	1.36	0.95	0.60	0.33

7.2 Calcule el tiempo necesario para que el nivel del líquido dentro del tanque esférico con radio $r = 5$ m mostrado en la figura 7.9 pase de 4 m a 3 m. La velocidad de salida por el orificio del fondo es $v = 4.895\ \sqrt{a}$ m/s, el diámetro de dicho orificio es de 10 cm.

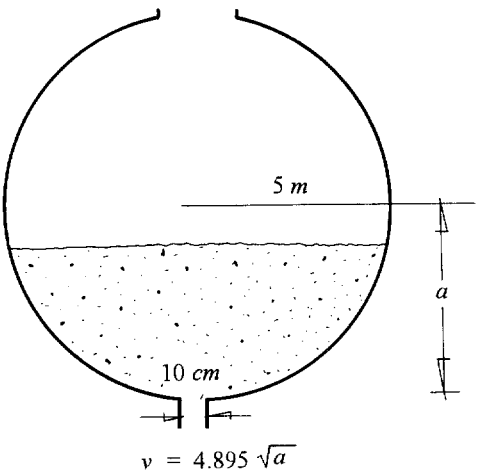


Figura 7.9 Vaciado de un tanque esférico.

SOLUCIÓN

Balance de materia en el tanque

Acumulación	= entrada	−	salida
$\rho \dfrac{dV}{dt}$	= 0	−	$A v \rho$

donde V es el volumen del líquido en el tanque, que en función de la altura está dado por

$$V = \pi \left(5\ a^2 - \frac{a^3}{3}\right)\ \text{m}^3$$

A es el área del orificio de salida

$$A = \frac{\pi}{4}\ (0.1)^2\ \text{m}^2$$

y

$$v = 4.895\ \sqrt{a}\ \text{m/s}$$

Estas cantidades se sustituyen en la primera ecuación y se tiene

$$\pi \frac{d\left(5\ a^2 - \frac{a^3}{3}\right)}{dt} = -\frac{\pi}{4}\ (0.1)^2\ 4.895\ \sqrt{a}$$

se deriva

$$10\ a\ \frac{da}{dt} - \frac{3a^2}{3}\frac{da}{dt} = -\frac{(0.1)^2}{4}\ 4.895\ \sqrt{a}$$

y al despejar se tiene

$$\frac{da}{dt} = \frac{-4.895\ (0.1)^2\ \sqrt{a}}{4\,(10\ a - a^2)}$$

que con la condición inicial y la pregunta forman el siguiente

$$\text{PVI}\ \begin{cases} \dfrac{da}{dt} = -\dfrac{0.122375\ \sqrt{a}}{(10\ a - a^2)} \\ a\,(0) = 4\ \text{m} \\ a\,(?) = 3\ \text{m} \end{cases}$$

Con el método de Euler modificado y un paso de integración h de 10 segundos, se tiene

tiempo (s)	altura a (m)
0	4.0000
10	3.8982
20	3.7968
30	3.6957
40	3.5948
50	3.4941
60	3.3935
70	3.2939
80	3.1924
90	3.0917
100	2.9908

Este último valor de altura puede considerarse como 3 m, por lo que el tiempo necesario para que el nivel del líquido dentro del tanque esférico pase de 4 a 3 m es aproximadamente 100 segundos.

7.3 En un tanque perfectamente agitado se tienen 400 l de una salmuera en la cual están disueltos 25 kg de sal común (NaCl). En cierto momento se hace llegar al tanque un gasto de 80 l/min de una salmuera que contiene 0.5 kg de sal común por litro. Si se tiene un gasto de salida de 80 l/min, determine

a) ¿Qué cantidad de sal hay en el tanque trancurridos 10 minutos?
b) ¿Qué cantidad de sal hay en el tanque trancurrido un tiempo muy grande?

SOLUCIÓN

a) Si se llaman x los kg de sal en el tanque después de t minutos, la acumulación de sal en el tanque está dada por dx/dt y por la expresión

$$\frac{dx}{dt} = \text{masa de sal que entra} - \text{masa de sal que sale}$$

los valores conocidos se sustituyen y se llega a la ecuación

$$\frac{dx}{dt} = 80\ (0.5) - 80\ \left(\frac{x}{400}\right)$$

o

$$\frac{dx}{dt} = 40 - 0.2x$$

que con la condición inicial de que hay 25 kg de sal al tiempo cero, da el siguiente

$$\text{PVI} \begin{cases} \dfrac{dx}{dt} = 40 - 0.2x \\ x(0) = 25 \\ x(10) = ? \end{cases}$$

Como vía de ilustración se utilizará un método de Runge-Kutta de tercer orden cuyo algoritmo está dado por

$$y_{i+1} = y_i + \frac{h}{6}(k_1 + 4k_2 + k_3)$$

con

$$k_1 = f(x_i, y_i)$$

$$k_2 = f(x_i + h/2, y_i + hk_1/2)$$

$$k_3 = f(x_i + h, y_i + 2hk_2 - hk_1)$$

En el disco se encuentra el programa 7.1 para resolver este problema de valor inicial con el algoritmo anotado arriba. El resultado obtenido es

$$x(10) = 176.3 \text{ con un paso de integración } h \text{ de 1 min.}$$

b) La solución se obtiene dejando correr el programa hasta que la cantidad de sal en el tanque no cambie con el tiempo; esto es, hasta que se alcance régimen permanente.

Al dejar correr el programa se obtuvieron los siguientes resultados

CONDICIÓN INICIAL: Y(.00) = 25.0000
PASO DE INTEGRACION H = 1.000
VALOR FINAL DE X = 50.000
SE IMPRIME CADA 2 ITERACIONES

X	Y
2.0000	82.7124
4.0000	121.3920
6.0000	147.3158
8.0000	164.6902
10.0000	176.3348
12.0000	184.1393
14.0000	189.3699
16.0000	192.8755
18.0000	195.2251
20.0000	196.7998
22.0000	197.8552
24.0000	198.5625
26.0000	199.0366
28.0000	199.3543
30.0000	199.5673
32.0000	199.7100
34.0000	199.8056
36.0000	199.8698
38.0000	199.9127
40.0000	199.9415
42.0000	199.9608
44.0000	199.9738
46.0000	199.9824
48.0000	199.9883
50.0000	199.9921

7.4 Se hacen reaccionar isotérmicamente 260 g de acetato de etilo ($CH_3\,COOC_2\,H_5$) con 175 g de hidróxido de sodio (NaOH) en solución acuosa (ajustando el volumen total a 5 litros) para dar acetato de sodio (CH_3COONa) y alcohol etílico (C_2H_5OH), de acuerdo con la siguiente ecuación estequiométrica

$$CH_3COOC_2H_5 + NaOH \xrightarrow{\quad k \quad} CH_3COONa + C_2H_5OH$$

Si la constante de velocidad de reacción k está dada por

$$k = 1.44 \times 10^{-2} \ \frac{1}{mol \ \text{min}}$$

determine la cantidad de acetato de sodio y alcohol etílico presentes 30 minutos después de iniciada la reacción.

SOLUCIÓN

Si x denota el número de moles por litro de acetato de etilo que han reaccionado al tiempo t, entonces la velocidad de reacción dx/dt viene dada por la ley de acción de masas así

$$\frac{dx}{dt} = k\ C_A^1\ \ C_B^1$$

donde C_A y C_B denotan las concentraciones molares de los reactantes acetato de etilo e hidróxido de sodio, respectivamente, al tiempo t y los exponentes son sus coeficientes estequiométricos en la reacción. Entonces

$$C_A = \frac{260\ g}{PM_{CH_3COOC_2H_5}\ 5\,l} - x\ \ \frac{mol}{l}$$

$$C_B = \frac{175\,g}{PM_{NaOH}\ 5l} - x\ \ \frac{mol}{l}$$

Al substituir valores y añadir la condición inicial y la pregunta a la ecuación diferencial resultante, se tiene

$$\text{PVI}\ \begin{cases} \dfrac{dx}{dt} = 1.44 \times 10^{-2}\ (0.59 - x)\,(0.875 - x) \\ x(0) = 0.0 \\ x(30) = ? \end{cases}$$

Al correr el programa 7.2 se obtiene

$$x(30) = 0.169 \text{ con un paso de integración } h \text{ de 1 min.}$$

7.5 Se conecta un inductor (inductancia) de 0.4 henries en serie con una resistencia de 8 ohms, un capacitor de 0.015 farads y un generador de corriente alterna dada por la función 30 sen $5t$ volts para $t \geq 0$ (véase Fig. 7.10).

a) Establezca una ecuación diferencial para la carga instantánea en el capacitor.

b) Encuentre la carga a distintos tiempos.

SOLUCIÓN

a) La caída de voltaje en la resistencia es 8 I, en la inductancia es 0.4 dI/dt y en la capacitancia Q/0.015 = 66.6666 Q

Según las leyes de Kirchhoff

$$8\,I + 0.4\,\frac{dI}{dt} + 66.6666\,Q = 30\ \text{sen}\,5t$$

o

$$0.4\,\frac{d^2Q}{dt^2} + 8\,\frac{dQ}{dt} + 66.6666\ Q = 30\ \text{sen}\ 5t$$

ya que

$$\frac{dQ}{dt} = I$$

y finalmente

$$\frac{d^2Q}{dt^2} + 20\,\frac{dQ}{dt} + 166.6666\,Q = 75 \text{ sen } 5t$$

con las condiciones

$$Q = 0, \quad I = \frac{dQ}{dt} = 0 \text{ a } t = 0$$

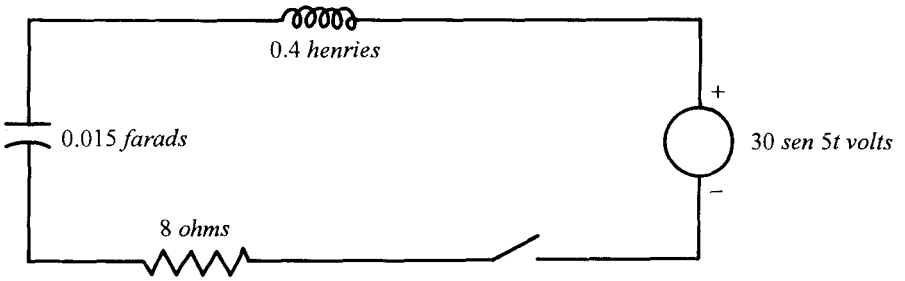


Figura 7.10.

Al pasar a un sistema con el cambio de variable $\frac{dQ}{dt} = z$

$$\text{PVI} \begin{cases} \dfrac{dQ}{dt} = z \\ \dfrac{dz}{dt} = 75 \text{ sen } 5t - 20z - 166.6666\,Q \\ Q(0) = 0 \\ z(0) = 0 \end{cases}$$

b) Al resolver por el método de Runge-Kutta de cuarto orden y usando $h = 0.1$ se tiene

t	Q	$\frac{dQ}{dt}$	t	Q	$\frac{dQ}{dt}$
0.1	0.03093	0.96008	1.1	−0.43060	−0.43375
0.2	0.16949	1.67198	1.2	−0.34174	1.40254
0.3	0.33066	1.33585	1.3	−0.16921	2.02794
0.4	0.42549	0.38455	1.4	−0.04475	2.15684
0.5	0.41473	−0.71114	1.5	0.24775	1.75767
0.6	0.29996	−1.62002	1.6	0.39010	0.92817
0.7	0.11080	−2.11960	1.7	0.43693	−0.12859
0.8	−0.10561	−2.09630	1.8	0.37679	−1.15386
0.9	−0.29609	−1.55950	1.9	0.22440	−1.89662
1.0	−0.41404	−0.64128	2.0	0.01706	−2.17503

7.6 Un proyectil de masa $m = 0.11$ kg se lanza verticalmente hacia arriba con una velocidad inicial $v_0 = 80$ m/s y se va frenando debido a la fuerza de gravedad $F_g = -mg$ y a la resistencia del aire $F_r = -kv^2$, donde $g = 9.8$ m/s y $k = 0.002$ kg/m. La ecuación diferencial para la velocidad v está dada por

$$mv' = -mg - kv^2$$

Encuentre la velocidad del proyectil a diferentes tiempos en su ascenso y el tiempo que tarda en llegar a su altura máxima.

SOLUCIÓN

Al emplear el método de Runge-Kutta de cuarto orden con $h = 0.01$ se tiene

t (s)	v (m/s)
0	80
0.3	53.55
0.6	39.11
0.9	29.76
1.2	23.04
1.5	17.83
1.8	13.55
2.1	9.86
2.4	6.54
2.7	3.46
3.00	0.49
3.01	0.39
3.02	0.30
3.03	0.20
3.04	0.10
3.05	0.002
3.06	-0.10

Dado que al llegar a $t = 3.06$ s, la velocidad es negativa, se toma 3.05 como el lapso que tarda en llegar a su altura máxima.

7.7 Se tienen tres tanques de 1000 litros de capacidad cada uno, perfectamente agitados (véase Fig. 7.11). Los tres recipientes están completamente llenos con una solución cuya concentración es 30 g/l. A partir de cierto momento se alimenta al primer tanque una solución que contiene 50 g/l con un gasto de 300 l/min (hay un arreglo entre los tres recipientes tal que al haber un gasto al primero, la misma cantidad fluye de éste al segundo, del segundo al tercero y de éste afuera del sistema, con lo cual se mantiene constante el volumen en todos ellos).

Calcule la concentración en cada tanque después de 10 minutos de haber empezado a agregar solución al primero.

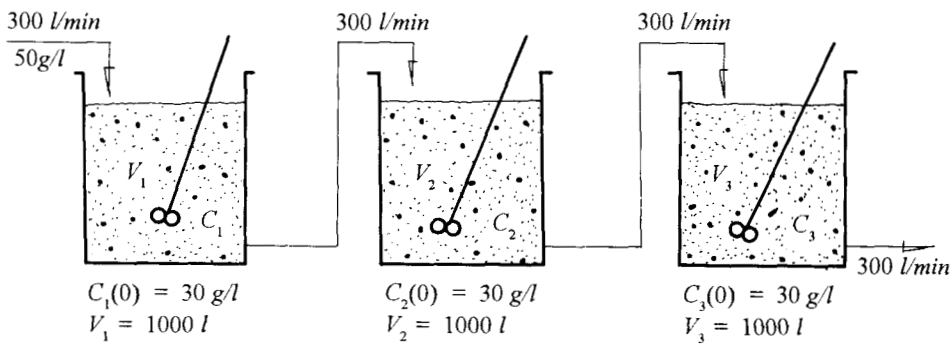


Figura 7.11 Arreglo de tres tanques interconectados.

SOLUCIÓN

Balance de soluto en el primer tanque

$$\text{Acumulación} = \text{entrada} - \text{salida}$$

$$\frac{dC_1 V_1}{dt} = 300(50) - 300\,C_1$$

como $V_1 = 1000$ l y permanece constante

$$\frac{dC_1}{dt} = 15 - 0.3\,C_1 \quad \text{y} \quad C_1(0) = 30 \tag{1}$$

Balance de soluto en el segundo tanque

$$V_2 \frac{dC_2}{dt} = 300\,C_1 - 300\,C_2$$

Como

$$V_2 = 1000\ \text{l}$$

$$\frac{dC_2}{dt} = 0.3(C_1 - C_2) \quad \text{y} \quad C_2(0) = 30 \tag{2}$$

Balance de soluto en el tercer tanque

$$V_3 \frac{dC_3}{dt} = 300\ C_2 - 300\ C_3$$

Como

$$V_3 = 1000\ l$$

$$\frac{dC_3}{dt} = 0.3\,(C_2 - C_3) \qquad \text{y} \qquad C_3(0) = 30 \tag{3}$$

Las ecuaciones 1 a 3, con sus respectivas condiciones iniciales, constituyen un sistema cuya solución representa los valores buscados, esto es,

$$\text{PVI} \begin{cases} \dfrac{dC_1}{dt} = 15 - 0.3C_1 \\ \dfrac{dC_2}{dt} = 0.3\,(C_1 - C_2) \\ \dfrac{dC_3}{dt} = 0.3\,(C_2 - C_3) \\ C_1(0) = 30 \\ C_2(0) = 30 \\ C_3(0) = 30 \\ C_1(10) = ? \\ C_2(10) = ? \\ C_3(10) = ? \end{cases}$$

Se utiliza un paso de integración de 1 min y el algoritmo de RK-4 para sistemas y se tiene

tiempo (min)	C_1	C_2	C_3
0	30.00	30.00	30.00
1	35.18	30.74	30.07
2	39.02	32.44	30.46
3	41.87	34.55	31.25
4	43.98	36.75	32.41
5	45.54	38.85	33.82
6	46.69	40.74	35.89
7	47.55	42.41	37.01
8	48.19	43.83	38.61
9	48.66	45.03	40.13
10	49.00	46.02	41.54

7.8 El mezclado imperfecto en un reactor continuo de tanque agitado se puede modelar como dos o más reactores con recirculación entre ellos, como se muestra en la figura 7.12. En este sistema se lleva a cabo una reacción isotérmica irreversible del tipo A $\xrightarrow{k}$ B de orden 1.8 con respecto al reactante A. Con los datos que se dan abajo, calcule la concentración del reactante A en los reactores (1) y (2) (C_{A1} y C_{A2}, respectivamente) durante el tiempo necesario para alcanzar el régimen permanente. Ensaye varios tamaños de paso de integración y compare los resultados obtenidos en el ejercicio 4.5.

Datos:

$$F = 25 \text{ l/min} \qquad C_{A0} = 1 \text{ mol/l}$$

$$F_R = 100 \text{ l/min} \qquad C_{A1}(0) = 0.0 \text{ mol/l}$$

$$C_{A2}(0) = 0.0 \text{ mol/l} \qquad k = 0.2 \left(\frac{l}{mol}\right)^{0.8} \text{min}^{-1}$$

SOLUCIÓN

Un balance del componente A en cada uno de los reactores da

Acumulación= Entrada − Salida − Reacciona

Reactor 1

$$\frac{dV_1 C_{A1}}{dt} = FC_{A0} + F_R C_{A2} - (F + F_R)C_{A1} - V_1 k C_{A1}^n$$

Reactor 2

$$\frac{dV_2 C_{A2}}{dt} = (F + F_R)C_{A1} - (F + F_R)C_{A2} + V_2 k C_{A2}^n$$

Como V_1 y V_2 son constantes, mediante la sustitución de valores y con las condiciones de operación a tiempo cero, se llega a

$$\text{PVI} \begin{cases} \dfrac{dC_{A_1}}{dt} = 1.25\, C_{A2} + \dfrac{25}{80} - \dfrac{125}{80} C_{A1} - 0.2\, C_{A1}^{1.8} \\ \dfrac{dC_{A_2}}{dt} = \dfrac{125}{20}(C_{A1} - C_{A2}) - 0.2\, C_{A2}^{1.8} \\ C_{A1}(0) = 0.0 \\ C_{A2}(0) = 0.0 \\ C_{A1}(0arp) = ? \qquad \text{rp = régimen permanente} \\ C_{A2}(0arp) = ? \end{cases}$$

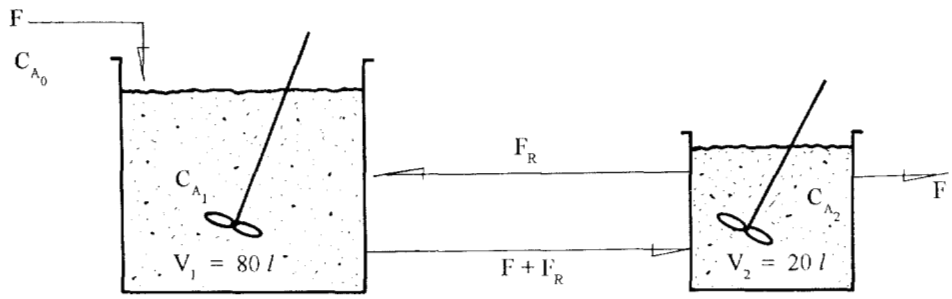


Figura 7.12 Modelación de un reactor con mezclado imperfecto.

Con el programa 7.3 y con un paso de integración de 0.4 minutos, el valor de C_{A2} en la primera iteración resulta negativo (lo cual es imposible) y al efectuar la segunda iteración e intentar calcular el término $C_{A2}^{1.8}$(véase segunda Ec. del PVI) el programa aborta.

Se ensaya ahora un tamaño de paso menor, ya que la constante de velocidad de reacción es alta y es de esperarse que la reacción sea muy rápida y que un paso de 0.4 minutos resulte muy grande. A continuación se dan los resultados para $h = 0.3$ minutos.

```
CONDICIONES INICIALES :
Y1( .00)                =          .000
Y2( .00)                =          .000
PASO DE INTEGRACIÓN H=             .300
VALOR FINAL DE X        =        20.000
SE IMPRIME CADA 5 ITERACIONES

     X                  Y1                 Y2

  1.5000             .3143              .2796
  3.0000             .4839              .4635
  4.5000             .5706              .5528
  6.0000             .6123              .5966
  7.5000             .6321              .6172
  9.0000             .6413              .6268
 10.5000             .6456              .6313
 12.0000             .6476              .6334
 13.5000             .6485              .6343
 15.0000             .6489              .6348
 16.5000             .6491              .6350
 18.0000             .6492              .6351
 19.5000             .6493              .6351
```

Puede observarse que el régimen permanente se alcanza a los 18 minutos. Los valores de las concentraciones a régimen permanente coinciden con los obtenidos en el ejercicio 4.5.

Se probaron además tamaños de paso de 0.25, 0.2 y 0.1 minutos; en cada caso los mismos resultados se obtuvieron que para 0.3 minutos.

7.9 En un reactor de laboratorio continuo tipo tanque perfectamente agitado, se lleva a cabo una reacción química exotérmica cuya temperatura se controla por medio de un líquido que circula por una chaqueta que se mantiene a una temperatura uniforme T_j. Calcule la temperatura T y la concentración C_A de la corriente de salida cuando el reactor trabaja a régimen transitorio y hasta alcanzar el régimen permanente para el caso de una reacción de primer orden. Aplique la siguiente información referida a la figura 7.13

Condiciones iniciales : $C_A(0) = 5$ gmol/l y $T(0) = 300$ K

F = Gasto de alimentación al reactor = 10 ml/s

V = Volumen del reactor = 2000 ml

C_{A0} = Concentración del reactante A en el flujo de alimentación = $5 \frac{\text{gmol}}{\text{l}}$

T_0 = Temperatura del flujo de alimentación = 300 K

ΔH = Calor de reacción = −10000 cal/gmol

U = Coeficiente global de trasmisión de calor = $100 \frac{\text{cal}}{°\text{C s m}^2}$

A = Area de transmisión de calor = 0.02 m^2

k = Constante de velocidad de reacción = $8\times10^{12} \exp(-22500/1.987\ T)\ s^{-1}$

T_j = Temperatura del líquido que circula por la chaqueta = 330 K

Cp = Calor específico de la masa reaccionante = 1 Kcal/Kg°C

ρ = Peso específico de la masa reaccionante = 1 kg/l

SOLUCIÓN

Balance de materia para el reactante A

Acumulación = Entrada − Salida − − Reacciona

$$\frac{dVC_A}{dt} = F\,C_{A0} - F\,C_A - k\,V\,C_A^n$$

Balance de calor

Acumulación = entrada − salida − generado − eliminado

$$\frac{dV\rho CpT}{dt} = F\rho Cp(T_0 - T) - \Delta H k V C_A^n - UA(T - T_j)$$

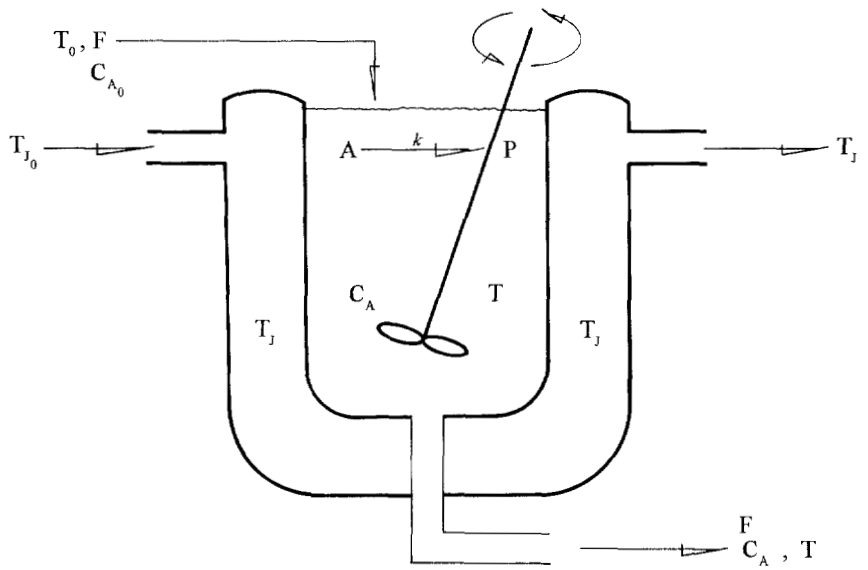


Figura 7.13 Reactor tipo tanque agitado con chaqueta.

Como V, ρ y Cp se consideran constantes, al sustituir valores se tiene

$$\text{PVI}\begin{cases}\dfrac{dC_A}{dt} = 0.005\,(5 - C_A) - 8\times10^{12}\,\exp\,(-22500/1.98T)\,C_A \\ \dfrac{dT}{dt} = 0.005\,(300 - T) + 8\times10^{13}\exp\,(-22500/1.98T)\,C_A - 0.001(T-330) \\ C_A(0) = 5\ \text{gmol/l} \\ T(0) = 300\ \text{K}\end{cases}$$

Al resolver con el programa 7.3, que utiliza el método de Runge-Kutta de tercer orden para un sistema de ecuaciones, se obtienen los resultados

```
SOLUCIÓN DE UN PVI CON UN SISTEMA DE N
ECUACIONES DIFERENCIALES ORDINARIAS DE PRIMER ORDEN
POR EL MÉTODO DE RUNGE-KUTTA DE TERCER ORDEN

CONDICIONES INICIALES:
Y1( .00)                =          5.000
Y2( .00)                =        300.000
PASO DE INTEGRACIÓN H=             20.000
VALOR FINAL DE X        =       3000.000
SE IMPRIME CADA 10 ITERACIONES
```

X	Y1	Y2
.0000	5.0000	300.0000
200.0000	4.6623	306.6382
400.0000	4.3180	310.6624
600.0000	3.9803	313.8187
800.0000	3.6243	316.9112
1000.0000	3.1727	320.8165
1200.0000	2.3743	327.8928
1400.0000	.7730	342.0851
1600.0000	.6438	341.9108
1800.0000	.7104	340.8714
2000.0000	.7314	340.5911
2200.0000	.7359	340.5366
2400.0000	.7366	340.5287
2600.0000	.7367	340.5278
2800.0000	.7367	340.5278
3000.0000	.7367	340.5278

7.10 Encuentre la curva elástica de una viga uniforme con un extremo libre, de longitud L = 5 m y peso constante de w = 300 kg. Determine también la deflexión del extremo libre. Tome EI = 150 000.

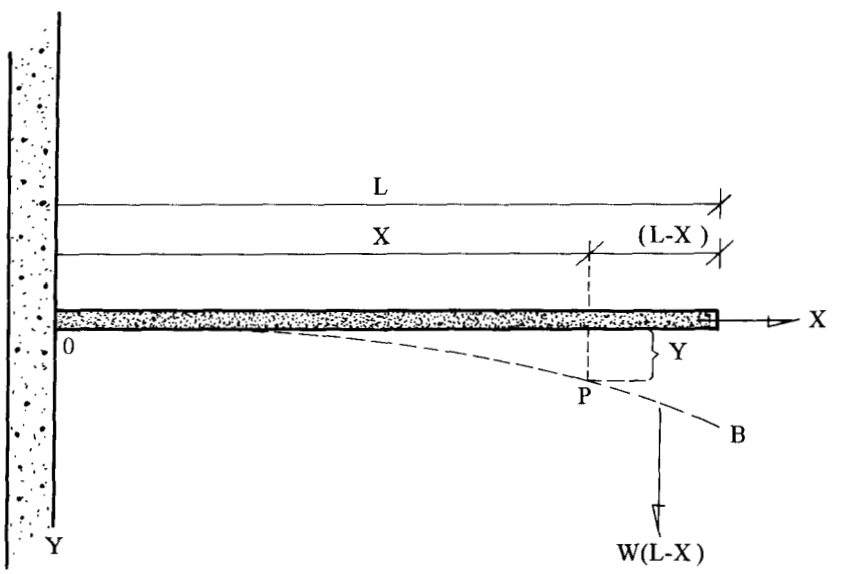


Figura 7.14 Viga empotrada con un extremo libre.

SOLUCIÓN

La figura 7.14 muestra la viga y su curva elástica (línea punteada). Se toma el origen O de un sistema coordenado en el extremo empotrado de la viga y la dirección positiva del eje y hacia abajo.

Sea x un punto cualquiera de la viga. Para calcular el momento de flexión en el punto x, $M(x)$, considere la parte de la viga a la derecha de P y que sólo una fuerza hacia abajo actúa en esa porción, $w(L-x)$, produciendo el momento positivo

$$M(x) = w(L-x)[(L-x)/2] = w(L-x)^2/2$$

En la teoría de vigas, se demuestra que $M(x)$ está relacionado con el radio de curvatura de la curva elástica calculado en x así

$$EI \frac{y''}{[1 + (y')^2]^{3/2}} = M(x) \tag{4}$$

donde E es el módulo de elasticidad de Young y depende del material con que se construyó la viga e I es el momento de inercia de la sección transversal de la viga en x.

Si se asume que la viga se flexiona muy poco, que es el caso general, la pendiente y' de la curva elástica es tan pequeña que

$$1 + (y')^2 \approx 1$$

y la ecuación 4 puede aproximarse por

$$EI\, y'' = M(x) = w(L-x)^2/2$$

Al cambiar de variable en la forma $y' = dy/dx = z$, se obtiene el siguiente

$$\text{PVI} \quad \begin{cases} \dfrac{dy}{dx} = z \\ \dfrac{dz}{dx} = \dfrac{w(L-x)^2}{2EI} \\ y(0) = 0 \\ z(0) = 0 \\ y(5) = ? \end{cases}$$

Con el programa 7.3 y con $h = 0.5$ m se obtiene

x (m)	y(m)
0	0
0.5	0.003
1.0	0.011
1.5	0.023
2.0	0.038
2.5	0.055
3.0	0.074
3.5	0.094
4.0	0.115
4.5	0.135
5.0	0.156

Problemas

7.1 Si al tanque de la figura 7.15, al momento de llegar el nivel de líquido a 0.5 m se hace llegar un gasto de alimentación de 0.04 m^3/s, el nivel del líquido aumentará. Determine el tiempo necesario para que el nivel se recupere nuevamente a 3 m.

7.2 El tiempo que requiere el tanque del ejercicio 7.1 para recuperar su nivel de 0.5 a 3 m con un gasto de alimentación de 0.04 m^3/s es de aproximadamente 432 s. Calcule el gasto de alimentación que se requiere para reducir este tiempo a la mitad.

7.3 Calcule el tiempo necesario para que el nivel del líquido del tanque de la figura 7.15 pase de 6 m a 1 m. El flujo de salida por el orificio del fondo es 3.457 $\sqrt{a}$ l/s.

7.4 Se hace llegar un gasto de alimentación de 7 l/s al tanque de la figura 7.15 cuando la altura del fluido en él es de 5 m. Treinta minutos después, este gasto es interrumpido por falla de la bomba, que se repara y arranca una hora después. Determine el gasto necesario para que el nivel se recupere y se mantenga en 5 m, así como el tiempo necesario para alcanzar ese nivel (régimen permanente). El flujo de salida es 3.457 $\sqrt{a}$ l/s ininterrumpidamente.

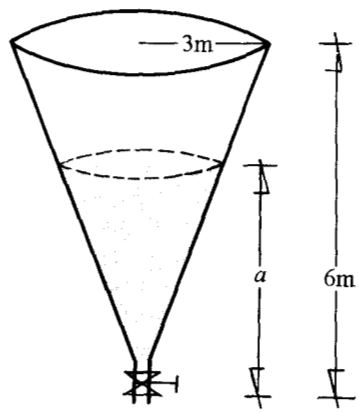


Figura 7.15 Vaciado de un tanque cónico.

7.5 Un tanque perfectamente agitado contiene 400 litros de salmuera en la cual están disueltos 10 kg de sal. Si se hace llegar 1.0 l/min de una salmuera que contiene 2 kg de sal en cada 5 litros y por el fondo se sacan 8 l/min de salmuera, determine la concentración de sal en el tanque a distintos tiempos.

7.6 Se ha encontrado experimentalmente que la constante de velocidad de reacción a volumen constante y a 30°C de la ecuación estequiométrica

$$A + B \xrightarrow{k} S$$

es 0.4967 $(mol/l)^{-1}$ min^{-1}. Determine el tiempo necesario para alcanzar un 90% de conversión del reactivo limitante en cada uno de los casos que se dan abajo, si se mantiene todo el tiempo la mezcla reaccionante a 30°C.

Concentraciones (mol / l)	
C_{A0}	C_{B0}
0.5	1.0
1.0	1.5
1.5	2.0
1.0	1.0
2.0	0.5

7.7 La aplicación de las leyes de Kirchhoff en un circuito cerrado da lugar a sistemas de ecuaciones diferenciales del tipo

$$\frac{dI_1}{dt} = -4\,I_1 + 3I_2 + 6$$

$$\frac{dI_2}{dt} = -2.4\,I_1 + 1.6\,I_2 + 3.6$$

Si se tienen las condiciones iniciales

$$I_1(0) = 0 \qquad I_2(0) = 0$$

Calcule I_1 (3) e I_2(3) con pasos de tiempo 0.05, 0.1, 0.5 y 1.0

7.8 Un capacitor de 0.001 farads está conectado en serie con una fem de 20 volts y una inductancia de 0.4 henries. Si $t = 0$, $Q = 0$, e $I = 0$, encuentre una ecuación para modelar este circuito y use el método de Runge-Kutta de tercer orden para hallar el valor de Q a distintos tiempos (véase Ejer. 7.5).

7.9 Repita el ejercicio 7.8 para $k = 0.0002$. ¿Qué sucede si $k \to 0$?

7.10 Un objeto que pesa 500 kg se coloca en la superficie de un tanque lleno de agua y se suelta ($v_0 = 0$). Las fuerzas que actúan sobre el objeto son la de empuje hacia arriba de 100 kg y la resistencia del agua que es de 30 v, donde v está en m/s. ¿Qué distancia recorre el cuerpo en 5 segundos?

7.11 Las ecuaciones

$$\frac{d^2x}{dt^2} = 0$$

y

$$\frac{d^2y}{dt^2} = -g$$

con $x = 0, y = 0, \frac{dx}{dt} = v_0 \cos\theta_0, \frac{dy}{dt} = v_0 \operatorname{sen}\theta_0$ a $t = 0$

describen la trayectoria de un proyectil disparado con una velocidad inicial v_0 y un ángulo de inclinación θ_0. Aquí x y y son las distancias horizontal y vertical que recorre el proyectil.

Si $\theta_0 = 60$ y $v_0 = 50$ m/s, calcule

a) El tiempo de vuelo del proyectil
b) La distancia que recorre
c) La altura máxima que alcanza

7.12 Si en el ejercicio 7.7 se cambian las condiciones de concentración en los tanques a $C_2(0) = C_1(0) = C_3(0) = 0$, ¿Cuál es el tiempo necesario para alcanzar el régimen permanente?

7.13 Si en el diagrama de la figura 7.16 se toma una corriente de recirculación de 150 l/min a la salida del tanque 3 y se lleva al tanque 2, en tanto el volumen se conserva constante en cada tanque e igual a 1000 litros, determine la concentración en cada tanque 10 minutos después de iniciado el proceso.

7.14 Si en el diagrama del problema anterior se adicionan las corrientes de recirculación mostradas en la figura 7.17, pero conservando la característica de que el volumen en los tres tanques permanece constante, las concentraciones C_1, C_2 y C_3 variarán de manera distinta a como se vio en el ejercicio 7.7.

Con los datos de la figura 7.17 determine las concentraciones en cada uno de los tanques a los diez minutos de haber empezado a agregar solución al primero.

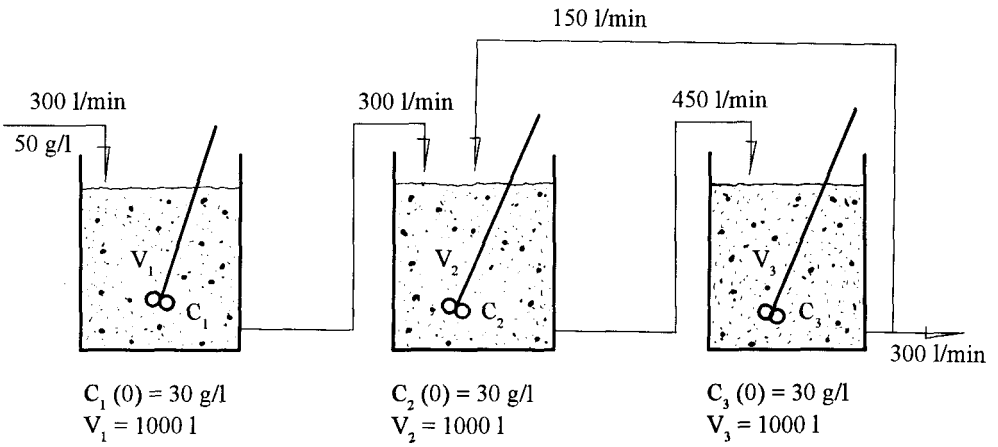


Figura 7.16 Arreglo de tres tanques interconectados.

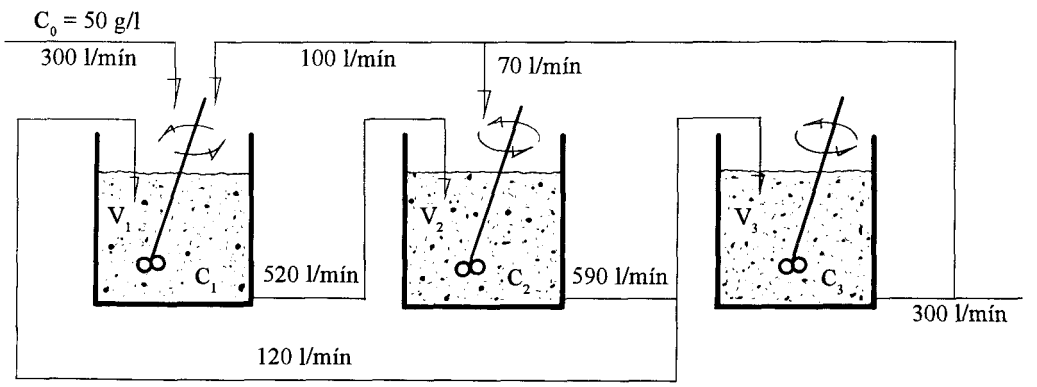


Figura 7.17 Arreglo de tres tanques interconectados con recirculación.

7.15 Repita el ejercicio 7.8 con los siguientes cambios

a) $V_1 = 80$, $V_2 = 20$, $F_R = 10$

b) $V_1 = 80$, $V_2 = 20$, $F_R = 0.1$

c) $V_1 = 50$, $V_2 = 50$, $F_R = 10$

d) $V_1 = 20$, $V_2 = 80$, $F_R = 10$

e) $V_1 = 50$, $V_2 = 50$, $F_R = 200$

7.16 Si en el ejercicio 7.9 la reacción es de segundo orden, calcule la temperatura T y la concentración C_A de la corriente de salida cuando el reactor trabaja a régimen transitorio y hasta alcanzar el régimen permanente.
Utilice

$$k = 1 \times {}^{13}\exp\left(\frac{-23200}{1.987\,T}\right) \quad \frac{l}{\text{gmol}\,s}$$

y la información presentada en el ejercicio 7.9.

7.17 Repita el ejercicio 7.9 utilizando la misma información, con los siguientes cambios

$$T_j = 310,\ 320,\ 340 \text{ y } 350$$

Analice los resultados.

7.18 El término EI del ejercicio 7.10 depende del material de que está construida la viga. Repita el ejercicio para otros materiales, en los que

a) EI = 117187

b) EI = 100000

las demás condiciones se conservan.

7.19 Si en la viga del ejercicio en estudio se aplica además una carga concentrada de 500 kg en el extremo libre, determine el perfil de flexión a lo largo de la viga.

7.20 Se tiene un intercambiador de calor de tubos concéntricos en contracorriente y sin cambio de fase (véase Fig. 7.18). Las ecuaciones que describen el intercambiador de calor en ciertas condiciones de operación son

$$\frac{dT_B}{dx} = 0.03\ (T_s - T_B)$$

$$\frac{dT_s}{dx} = 0.04\ (T_s - T_B)$$

Elabore un programa para calcular T_{B1} y T_{S0} si el intercambiador de calor tiene una longitud de 3 m; use el método de Runge-Kutta de cuarto orden.

7.21 Un tanque cilíndrico de 5 m de diámetro y 11 m de largo aislado con asbesto se carga con un líquido que está a 220°F y el cual se deja reposar durante cinco días. A partir de los datos de diseño del tanque, las propiedades térmicas y físicas del líquido y el valor de la temperatura ambiente, se encuentra la ecuación

$$\frac{dT}{dt} = 0.615 + 0.175 \cos\left(\frac{\pi t}{12}\right) - 0.0114\ T$$

que relaciona la temperatura T del líquido (en°C) con el tiempo t en horas. ¿Cuál es la temperatura final del líquido?

7.22 El radio se desintegra en razón proporcional a la cantidad presente en cada instante. La constante de proporcionalidad es $k = 10^{-2}$ día^{-1}. Si se tienen inicialmente 60 g de radio, calcule la cantidad que hay presente transcurridos cinco días mediante el siguiente esquema de predicción-corrección

$$\bar{y}_{i+1} = y_i + \frac{h}{24}\left(55f_i - 59f_{i-1} - 37f_{i-2} - 9f_{i-3}\right)$$

$$y_{i+1} = y_i + \frac{h}{24}\left(9f_{i+1} + 19f_i - 5f_{i-1} + f_{i-2}\right)$$

7.23 Considere un sistema ecológico simple compuesto solamente de coyotes (y) y correcaminos (x), donde los primeros se alimentan de los segundos (cuando los alcanzan). Los tamaños de las poblaciones cambian de acuerdo con las ecuaciones

$$\frac{dx}{dt} = k_1 x - k_2 xy$$

$$\frac{dy}{dt} = k_3 xy - k_4 y$$

que se pueden entender como sigue

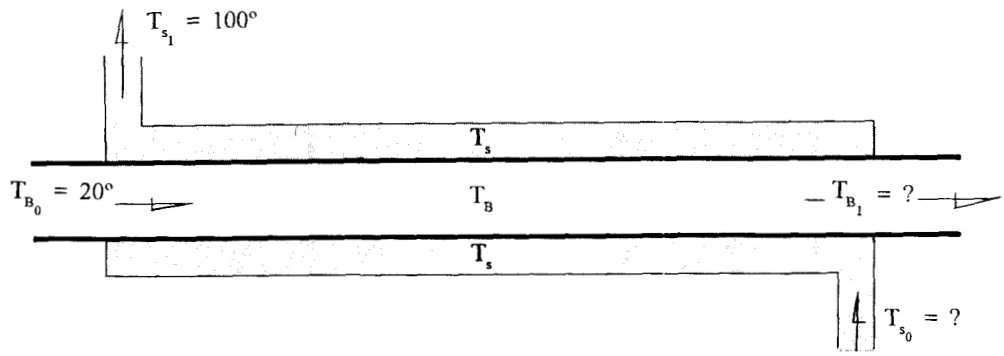


Figura 7.18 Intercambiador de calor de tubos concéntricos en contracorriente.

Si no hay coyotes (y) los correcaminos se reproducen con una velocidad de crecimiento $k_1 x$; si no hay correcaminos, la especie de coyotes desaparece con velocidad $k_4 y$. El término xy representa la interacción de las dos especies y las constantes k_2 y k_3 dependen de la habilidad de los depredadores para atrapar a los correcaminos y de la habilidad de éstos para huir.

Las poblaciones de los coyotes cambian cíclicamente. Calcule el ciclo y su periodo al resolver el modelo con $k_1 = 0.4, k_2 = 0.02, k_3 = 0.001$ y $k_4 = 0.3$. Use $x(0) = 30$ y $y(0) = 3$ como condiciones iniciales.

7.24 Se utilizan dos tanques en serie y provistos de serpentín de enfriamiento por el cual circula agua en contracorriente para enfriar 10000 lb/hr de ácido sulfúrico. Las condiciones de operación se muestran en la figura 7.19. Si en un momento dado fallara el suministro de agua de enfriamiento, ¿cuál será la temperatura del ácido sulfúrico T_2 a la salida del segundo tanque después de una hora? Las ecuaciones que describen el proceso son

$$3600\ T_0 - 3600\ T_1 = 2850\ \frac{dT_1}{dt}$$

$$3600\ T_1 - 3600\ T_2 = 2850\ \frac{dT_2}{dt}$$

donde T_0, T_1 y T_2 están en°C y t en horas.

7.25 Utilice el método de Taylor (elija el orden) para resolver los siguientes problemas de valor inicial (PVI) y compare con las soluciones analíticas

a) $dy/dx = 3x^2$, $y(0) = 0, y(1)=?$ con $h=0.1$

b) $dy/dx = \ln x$, $y(1) = 3, y(2)=?$ con $h=0.2$

c) $dy/dx = 2xy$, $y(1) = 0.5, y(2)=?$ con $h=0.25$

d) $dy/dx = y^2$, $y(0) = 1, y(0.5)=?$ con $h=0.1$

7.26 Resuelva los PVI del problema anterior por el método de Runge-Kutta de segundo orden.

7.27 Resuelva los PVI del problema 7.25 por el método de Runge-Kutta de cuarto orden.

7.28 Resuelva los siguientes PVI con la fórmula 7.55

$$y_{i+1} = y_i + \frac{h}{24}[55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}]$$

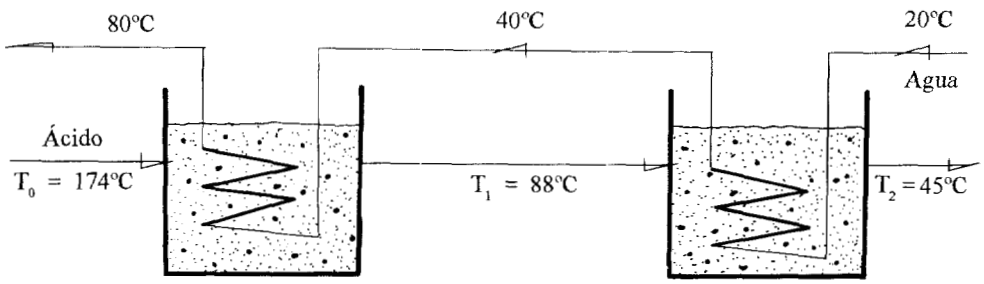


Figura 7.19 Dos tanques interconectados y con serpentín de enfriamiento.

y el método de Runge-Kutta de tercer orden como inicializador

a) $dy/dx + y = 0$, $y(0) = 1$, $y(2)=?$ con $h = 0.25$

b) $dy/dx + 2xy = 2x^3$, $y(0) = 0$, $y(2.5)=?$ con $h = 0.5$

c) $dy/dx + xy = g(x)$, $y(0) = 1$, $y(1)=?$ con $h = 0.25$

donde

x	0.0	0.2	0.4	0.6	0.8	1.0
$g(x)$	0.0	0.19471	0.35868	0.46602	0.49979	0.45465

d) $dy/dx = -y-xy^2$, $y(0) = 1$, $y(-1)=?$ con $h = -0.25$

e) $xdy/dx = 1-y+x^2y^2$, $y(0) = 1$, $y(1.5)=?$ con $h = 0.25$

7.29 Resuelva los PVI del problema anterior con la fórmula

$$y_{i+1} = y_{i-5} + \frac{3h}{10}\,[\,11f_i - 14f_{i-1} + 26f_{i-2} - 14f_{i-3} + 11f_{i-4}\,]$$

con el método de Runge-Kutta de cuarto orden como inicializador.

7.30 Resuelva los PVI del problema 7.28 con los siguientes esquemas de solución

a) Inicialización con Runge-Kutta de tercer orden.
Predicción con la fórmula dada en el ejercicio 7.28.
Corrección con

$$y_{i+1} = y_i + \frac{h}{24}\,[\,9f_{i+1} - 19f_i + 5f_{i-1} - f_{i-2}\,]$$

b) Inicialización con Runge-Kutta de cuarto orden.
Predicción con la fórmula del problema 7.29
Corrección con

$$y_{i+1} = y_{i-3} + \frac{2h}{45}\,[\,7f_{i+1} - 32f_i + 12f_{i-1} + 32f_{i-2} + 7f_{i-3}\,]$$

c) Inicialización con Runge-Kutta de cuarto orden.
Predicción con Adams-Bashford de cuarto orden

$$y_{i+1} = y_i + \frac{h}{720}\,[\,1901f_i - 2774f_{i-1} + 2616f_{i-2} - 1274f_{i-3} + 251f_{i-4}\,]$$

Correción con Adams-Moulton de cuarto orden

$$y_{i+1} = y_i + \frac{h}{720}\,[\,251f_{i+1} - 646f_i - 264f_{i-1} + 106f_{i-2} - 19f_{i-3}\,]$$

7.31 Resuelva el siguiente PVI con el método de Runge-Kutta de segundo orden

$dy/dx = z, \qquad y(0) = 1, \qquad y(1) = ?$

$dz/dx = y, \qquad z(0) = -1, \qquad z(1) = ? \qquad$ con $h = 0.1$

7.32 Resuelva el siguiente PVI con el método de Runge-Kutta de cuarto orden con $h = 0.1$

$$\text{PVI} \quad \begin{cases} \dfrac{d^2y}{dx^2} + 2\dfrac{dy}{dx} + 2y = 0 \\ y(0) = 1 \\ \dfrac{dy}{dx}\Big|_{x=0} = -1 \\ y(1) = ? \end{cases}$$

7.33 Resuelva el PVI del problema 7.31 con el esquema de solución (a) del 7.30.

7.34 Resuelva el PVI del problema 7.32 con el esquema de solución (c) del 7.30.

7.35 Resuelva el siguiente PVI

$$\begin{cases} \dfrac{dy}{dx} = z \\ \dfrac{dz}{dx} = -125y - 20z \\ y(0) = 0 \quad y(1) = ? \\ z(0) = 1 \quad z(1) = ? \end{cases}$$

con el método de Runge-Kutta de cuarto orden usando

$$a)\ h = 0.5$$

$$b)\ h = 0.1$$

Compare los resultado con la solución analítica

$$y = \frac{1}{5} e^{-10x} \operatorname{sen} 5x$$

$$z = e^{-10x} (\cos 5x - 2 \operatorname{sen} 5x)$$

7.36 Escriba las siguientes ecuaciones diferenciales como un sistema de ecuaciones diferenciales ordinarias de primer orden. Pase las condiciones iniciales a términos de las nuevas variables para constituir un PVIG y resuélvalo con los métodos vistos usando los tamaños de paso sugeridos

a) $y'' - 2y' + 2y = e^{2t} \operatorname{sen} t$ $\qquad 0 \le t \le 1$

$y(0) = -0.4; \quad y'(0) = -0.6$ $\qquad h = 0.01$

b) $y'' + 2ty' = e^{t}$ $\qquad 0 \le t \le 2$

$y(0) = 1; \quad y'(0) = -1$ $\qquad h = 0.1$

c) $y''' - 2y'' - y' - 2y = e^{t}$ $\qquad 0 \le t \le 3$

$y(0) = 1; \quad y'(0) = 2; \quad y''(0) = 0$ $\qquad h = 0.2$

7.37 Considere el conjunto de reacciones reversibles

$$A \underset{k_2}{\overset{k_1}{\rightleftharpoons}} B \qquad (1)$$

$$B \underset{k_4}{\overset{k_3}{\rightleftharpoons}} C \qquad (2)$$

Asuma que hay una mol de A solamente al inicio, y tome N_A, N_B y N_C como las moles de A, B, y C presentes respectivamente.

Como la reacción se verifica a volumen constante, N_A, N_B y N_C son proporcionales a las concentraciones. Sean k_1 y k_2 las constantes de velocidad de reacción a derecha e izquierda, respectivamente, de la ecuación 1; igualmente sean k_3 y k_4 aplicables a la 2. La velocidad de desaparición neta de A está dada por

$$\frac{dN_A}{dt} = -k_1 N_A + k_2 N_B$$

y para B

$$\frac{dN_B}{dt} = -(k_2 + k_3) N_B + k_1 N_A + k_4 N_C$$

Determine N_A, N_B y N_C transcurridos 50 minutos del inicio de las reacciones mediante

$k_1 = 0.1 \text{ min}^{-1}$

$k_2 = 0.01 \text{ min}^{-1}$

$k_3 = 0.09 \text{ min}^{-1}$

$k_4 = 0.009 \text{ min}^{-1}$

CAPÍTULO 8

ECUACIONES DIFERENCIALES PARCIALES

Sección 8.1 Obtención de ecuaciones diferenciales parciales a partir de la modelación de fenómenos físicos
Sección 8.2 Aproximación de las ecuaciones diferenciales parciales con ecuaciones de diferencias
Sección 8.3 Solución de problemas de valores en la frontera
Sección 8.4 Convergencia, estabilidad y consistencia
Sección 8.5 Método de Crank-Nicholson
Sección 8.6 Otros métodos para resolver el problema de conducción de calor en una dimensión
Sección 8.7 Tipos de condiciones frontera en procesos físicos y tratamiento de condiciones frontera irregulares

EN ESTE CAPÍTULO se presentará una breve introducción a algunas de las técnicas para aproximar la solución de EDP lineales de segundo orden y con dos variables independientes. Dichas técnicas se estudiarán resolviendo algunos problemas físicos muy conocidos como la ecuación de difusión y la ecuación de onda.

INTRODUCIÓN

Las ecuaciones diferenciales parciales (EDP) involucran una función de más de una variable independiente y sus derivadas parciales. La importancia de este tema radica en que prácticamente en todos los fenómenos que se estudian en ingeniería y otras ciencias, aparecen más de dos variables,* y su modelación matemática conduce frecuentemente a EDP.

Primero se clasificarán las ecuaciones diferenciales parciales lineales atendiendo al siguiente modelo general

$$A(x,y)\frac{\partial^2 U}{\partial x^2} + B(x,y)\frac{\partial^2 U}{\partial x\,\partial y} + C(x,y)\frac{\partial^2 U}{\partial y^2} = F\left(x,y,U,\frac{\partial U}{\partial x},\frac{\partial U}{\partial y}\right) \tag{8.1}$$

en el cual se asume que $A(x, y)$, $B(x, y)$ y $C(x, y)$ son funciones continuas de x y y. Dependiendo de los valores de $A(x, y)$, $B(x, y)$ y $C(x, y)$ en algún punto particular $(x, y) = (a, b)$, la ecuación (8.1) puede ser **elíptica, parabólica** o **hiperbólica**, de acuerdo con las condiciones

$$\begin{aligned} B^2(a,b) - 4\,A(a,b)\,C(a,b) &< 0 \quad \text{Elíptica en } (a,b) \\ B^2(a,b) - 4\,A(a,b)\,C(a,b) &= 0 \quad \text{Parabólica en } (a,b) \\ B^2(a,b) - 4\,A(a,b)\,C(a,b) &> 0 \quad \text{Hiperbólica en } (a,b) \end{aligned} \tag{8.2}$$

*En el análisis del comportamiento de los gases, por ejemplo, se tiene temperatura, presión y volumen; en la transmisión de calor intervienen temperatura, tiempo y direcciones del espacio: x, y, z; etcétera.

Una misma EDP puede ser parabólica en un punto, e hiperbólica en otro, etc. Si en cambio $A(x,y)$, $B(x,y)$ y $C(x,y)$ son constantes, entonces es elíptica, parabólica o hiperbólica completamente (ver Ejer. 8.1).

Algunos ejemplos de estas ecuaciones son

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} = 0$$

$$\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = f(x,y)$$

$$\frac{\partial^2 y}{\partial x^2} = \alpha \frac{\partial^2 y}{\partial t^2}$$

SECCIÓN 8.1 OBTENCIÓN DE ECUACIONES DIFERENCIALES PARCIALES A PARTIR DE LA MODELACIÓN DE FENÓMENOS FÍSICOS

A continuación se presenta la derivación de dos de las ecuaciones en estudio más comunes.

a) Ecuación general de la conducción de calor

Supóngase un cuerpo sólido de conductividad térmica k, peso específico ρ y calor específico C_p independientes de la temperatura T, en el cual fluye calor en las tres dimensiones del espacio y puede generar o absorber calor debido a algún fenómeno, por ejemplo de reacción química.

Al efectuar un balance de calor en un elemento diferencial como el de la figura 8.1 de dimensiones Δx y Δy y Δz se tiene, de acuerdo con la ley de la continuidad

Acumulación de calor	=	calor que entra al elemento diferencial en cada una de las tres dimensiones	−	calor que sale del elemento diferencial en cada una de las tres dimensiones	+	$Q\Delta x\Delta y\Delta z$	(8.3)
(cal/s)		(cal/s)		(cal/s)		(cal/s)	

donde Q puede ser positivo o negativo dependiendo de si el calor es generado o absorbido por unidad de volumen y por unidad de tiempo en el elemento diferencial.

También en la figura 8.1 se esquematiza la entrada q_i y la salida $q_i+\Delta_i$ de los flujos de calor (en cal/s) representados por la ley de Fourier, o sea que son proporcionales a la conductividad térmica k, el área de transmisión y el gradiente de temperatura en dirección de la transmisión. El signo negativo es para obtener flujos de calor positivos (por convención) ya que los gradientes dT/dx, dT/dy y dT/dz son negativos.

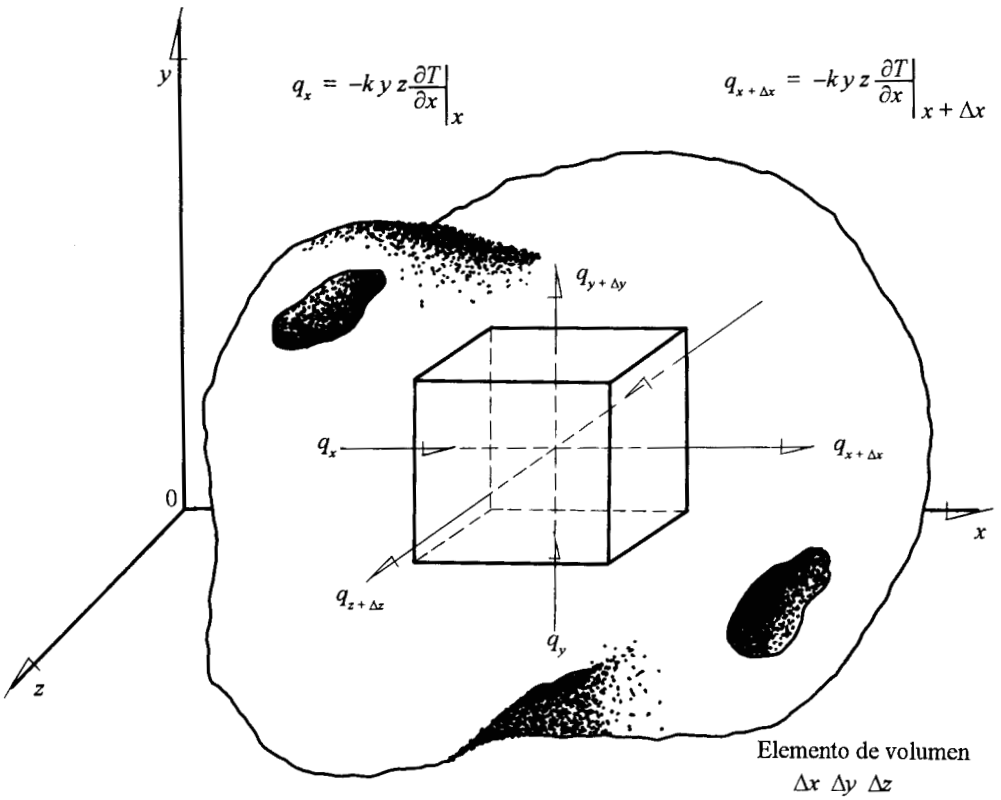


Figura 8.1. Balance de calor en un elemento diferencial de dimensiones Δx, Δy y Δz.

Los flujos de calor se sustituyen en la ecuación (8.3) y se tiene

$$\begin{aligned}\frac{d}{dt}(\Delta x\Delta y\Delta z\rho \mathrm{CpT}) = {} & -k\Delta y\Delta z\left.\frac{d\mathrm{T}}{dx}\right|_x - (-k\Delta y\Delta z\left.\frac{d\mathrm{T}}{dx}\right|_{x+\Delta x}) + \\ & -k\Delta x\Delta z\left.\frac{d\mathrm{T}}{dy}\right|_y - (-k\Delta x\Delta z\left.\frac{d\mathrm{T}}{dy}\right|_{y+\Delta y}) + \\ & -k\Delta x\Delta y\left.\frac{d\mathrm{T}}{dz}\right|_z - (-k\Delta x\Delta y\left.\frac{d\mathrm{T}}{dz}\right|_{z+\Delta z}) + \mathrm{Q}\Delta x\Delta y\Delta z\end{aligned} \tag{8.4}$$

al dividir miembro a miembro entre Δx, Δy, Δz y hacerlos muy pequeños, o sea Δx, Δy, $\Delta z \to 0$, queda

$$\rho\ \mathrm{Cp}\frac{\partial \mathrm{T}}{\partial t} = \lim_{\Delta x \to 0}\left[\frac{\left.\frac{d\mathrm{T}}{dx}\right|_{x+\Delta x} - \left.\frac{d\mathrm{T}}{dx}\right|_x}{\Delta x}\right] + \lim_{\Delta y \to 0}\left[\frac{\left.\frac{d\mathrm{T}}{dy}\right|_{y+\Delta y} - \left.\frac{d\mathrm{T}}{dy}\right|_y}{\Delta y}\right]$$

$$+ \lim_{\Delta z \to 0} k \left[\frac{\left. \frac{dT}{dz} \right|_{z+\Delta z} - \left. \frac{dT}{dz} \right|_{z}}{\Delta z} \right] + Q \tag{8.5}$$

y al aplicar la definición de derivada se obtiene

$$\boxed{\frac{\partial T}{\partial t} = \alpha \left[\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right] + \frac{Q}{\rho \, Cp}} \tag{8.6}$$

donde se ha sustituido a $\frac{k}{\rho Cp}$ con α, la cual se llama **coeficiente de difusividad térmica**, y sus unidades son, por ejemplo,

m^2 / s ya que k [=] cal / (s m °C), Cp [=] cal/(g °C) y ρ [=] g / cm^3.

La ecuación 8.6 se conoce como **ecuación de conducción de calor en régimen transitorio** en tres dimensiones (cartesianas), y es muy empleada en el campo de la ingeniería. También se conoce como **ecuación de difusión**, ya que representa la difusión molecular de masa entre fluidos, cuando la variable dependiente es la concentración C y el coeficiente α representa la **difusividad** $\mathcal{D}$. Así

$$\boxed{\frac{\partial C}{\partial t} = \mathcal{D} \left[\frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2} + \frac{\partial^2 C}{\partial z^2} \right]}$$

donde las unidades pueden ser

$$C \text{ [=] moles/cm}^3, \quad \mathcal{D} \text{ [=] cm}^2\text{/s}, \quad t \text{ [=] s}$$

b) **Ecuación de onda en una dimensión**

Considérese una cuerda (como la de una guitarra) elástica y flexible, la cual se estira y se sujeta en dos puntos fijos en $x = 0$ y $x = L$ sobre el eje de las x (Fig. 8.2*a*). A un tiempo $t = 0$, la cuerda se toma del centro y se eleva verticalmente a una altura $y = h$ (veáse Fig. 8.2*b*). Después se suelta. La descripción del movimiento producido constituye el problema por resolver.

Para simplificarlo, se considera que h es pequeño en comparación con L ($h << L$).

Modelo

Si en un instante dado se tomara una fotografía de la cuerda vibrando, se tendría ésta como en la figura 8.3*a*. El desplazamiento de un punto x de la cuerda en el

tiempo t queda indicado por $y(x, t)$, de igual forma para un punto vecino $x+\Delta x$ y en el mismo tiempo t, su desplazamiento queda indicado por $y(x+\Delta x, t)$.

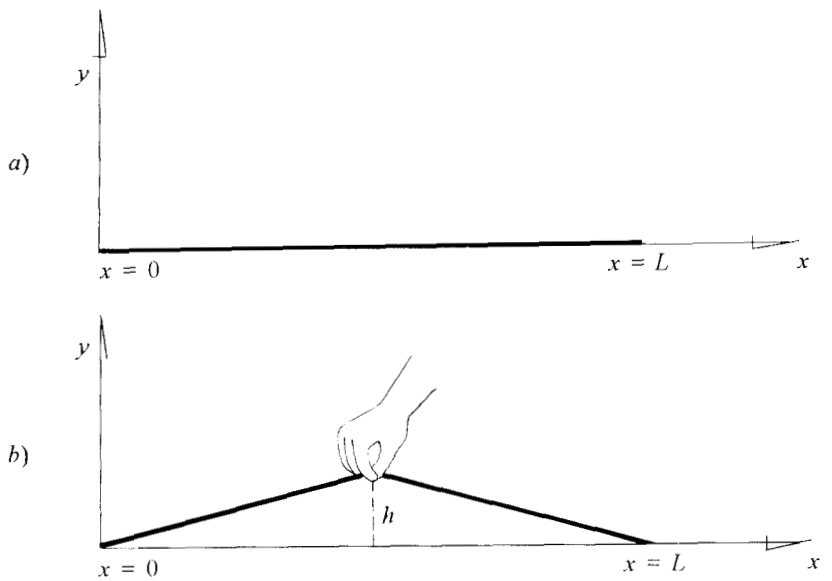


Figura 8.2.

En la figura 8.3*b* se muestra una ampliación del segmento de cuerda ΔS, el cual está sometido a dos tensiones que siempre actúan en la dirección de la tangente a ΔS a izquierda y derecha de la cuerda, o sea $T(x, t)$ y $T(x+\Delta x, t)$ respectivamente. Nótese que la tensión es función de la posición x sobre la cuerda y del tiempo t.

Al hacer una composición de fuerzas sobre el elemento de cuerda ΔS en las direcciones vertical y horizontal se tiene

$$\text{Fuerza vertical neta (hacia arriba)} = T(x+\Delta x, t) \operatorname{sen} \theta_2 - T(x, t) \operatorname{sen} \theta_1$$

$$\text{Fuerza vertical neta (a la derecha)} = T(x+\Delta x, t) \cos \theta_2 - T(x, t) \cos \theta_1$$

La fuerza horizontal neta es cero si se considera que el desplazamiento del punto x de su posición de equilibrio a la posición $y(x, t)$ es vertical.

Por otro lado, la fuerza neta vertical sobre ΔS produce una aceleración definida por la segunda ley de Newton; o sea,

$$\text{Fuerza vertical neta} = T(x+\Delta x, t) \operatorname{sen} \theta_2 - T(x, t) \operatorname{sen} \theta_1$$

$$= \rho \, \Delta S \, \frac{\partial^2 y}{\partial t^2} \tag{8.7}$$

donde ρ es la densidad de la cuerda en unidades de masa/longitud y $\partial^2 y/\partial t^2$ la aceleración de ΔS.

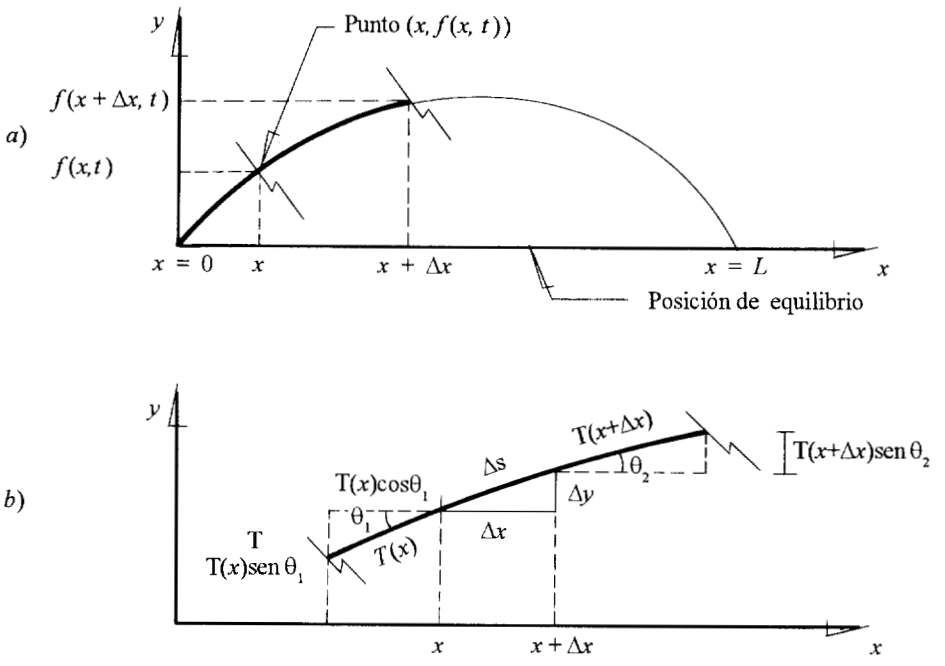


Figura 8.3

Como θ es función de la posición x y el tiempo t, $\theta_1 = \theta(x, t)$ y $\theta_2 = \theta(x+\Delta x, t)$. Estas expresiones se sustituyen en la ecuación 8.7 y al dividir entre Δx queda

$$\frac{T(x+\Delta x,t)\ \operatorname{sen}\theta(x+\Delta x,t) - T(x,t)\ \operatorname{sen}\theta(x,t)}{\Delta x} = \rho\ \frac{\Delta S}{\Delta x}\ \frac{\partial^2 y}{\partial t^2}$$

Para vibraciones cortas θ es pequeño, por lo que $\Delta S/\Delta x \approx 1$ y sen $\theta \approx$ tg θ; por lo que la última ecuación se puede escribir

$$\frac{T(x+\Delta x,t)\ \operatorname{tg}\theta(x+\Delta x,t) - T(x,t)\ \operatorname{tg}\theta(x,t)}{\Delta x} \approx \rho\ \frac{\partial^2 y}{\partial t^2}$$

y haciendo $\Delta x \to 0$

$$\frac{\partial}{\partial x}\ [T(x,t)\ \operatorname{tg}\theta(x,t)] = \rho\ \frac{\partial^2 y}{\partial t^2}$$

Como $tg\ \theta(x,t) = \partial y/\partial x$ y si la tensión T es constante se obtiene

$$\frac{\partial^2 y}{\partial x^2} = \alpha\ \frac{\partial^2 y}{\partial t^2} \tag{8.8}$$

donde $\alpha = \rho/T$.

Puesto que la cuerda permanece sujeta en sus extremos $x = 0$ y $x = L$, el desplazamiento $y(x, t)$ satisface las condiciones siguientes en todo el proceso

$$\begin{aligned} y(0, t) &= 0 \\ & \qquad \text{para } t \geq 0 \\ y(L, t) &= 0 \end{aligned} \tag{8.9}$$

conocidas como **condiciones extremas o frontera** (CF).

Por otro lado, la posición de la cuerda en el momento de soltarse (figura 8.2*b*) puede darse matemáticamente así

$$y(x, 0) = f(x) \tag{810}$$

ecuación que se conoce como **condición inicial** (CI) por describir la condición que se tiene al inicio del proceso.

En resumen, la ecuación 8.8 y las condiciones inicial y de frontera (Ecs. 8.10 y 8.9, respectivamente) constituyen un modelo matemático denotado como **problema de valores en la frontera** (PVF)

$$\text{PVF} \quad \begin{cases} \dfrac{\partial^2 y}{\partial x^2} = \alpha \dfrac{\partial^2 y}{\partial t^2} & \text{(ecuación diferencial parcial)} \\ y(x,0) = f(x) \quad 0 < x < L & \text{(condición inicial)} \\ y(0,t) = 0 & \text{(condición frontera 1)} \\ y(L,t) = 0 \quad t > 0 & \text{(condición frontera 2)} \end{cases}$$

y cuya solución $y(x, t)$ describe la posición de cualquier punto de la cuerda en un tiempo t.

SECCIÓN 8.2 APROXIMACIÓN DE LAS ECUACIONES DIFERENCIALES PARCIALES CON ECUACIONES DE DIFERENCIAS

Generalidades

La expansión de una función $f(x)$ diferenciable en una serie de Taylor alrededor de un punto x_i se estudió en los capítulos 5 y 7 y está definida por

$$f(x_i + a) = f(x_i) + af'(x_i) + \frac{a^2}{2!} f''(x_i) + \frac{a^3}{3!} f'''(x_i) + \ldots \tag{8.11}$$

Esta vez, la utilidad de la serie de Taylor no será estimar el valor de la función $f(x)$ en el punto $x_i + a$, sino aproximar la derivada de la función en x_i a partir de los valores de la función en x_i y en $x_i + a$. Para ello, considérese que $a > 0$ (con esto

la ecuación 8.11 sólo es válida delante del punto x_i) y que a es tan pequeña ($a << 1$) como para despreciar los términos tercero, cuarto, etc., del lado derecho de la expansión, con lo que la derivada $f'(x_i)$ puede aproximarse así

$$\left.\frac{df}{dx}\right|_{x_i} = f'(x_i) \approx \frac{f(x_i + a) - f(x_i)}{a} \tag{8.12}$$

Esta ecuación quedó definida en el capítulo 5 como la aproximación de la primera derivada de $f(x)$ en x_i con **diferencias hacia delante.**

Un resultado similar, válido a la izquierda de x_i se obtendrá restando a de x_i en la ecuación 8.11; esto es,

$$f(x_i - a) = f(x_i) - af'(x_i) + \frac{a^2}{2!}f''(x_i) - \frac{a^3}{3!}f'''(x_i) + \ldots \tag{8.13}$$

y como $a << 1$, puede llegarse a

$$\left.\frac{df}{dx}\right|_{x_i} = f'(x_i) \approx \frac{f(x_i) - f(x_i - a)}{a} \tag{8.14}$$

la aproximación a la primera derivada de $f(x)$ en x_i con **diferencias hacia atrás.**

Si en cambio se resta miembro a miembro la ecuación 8.13 de la 8.11 y se aplican los razonamientos anteriores, se llega a la expresión

$$\left.\frac{df}{dx}\right|_{x_i} = f'(x_i) \approx \frac{f(x_i + a) - f(x_i - a)}{2a} \tag{8.15}$$

conocida como la aproximación a la primera derivada de $f(x)$ en x_i con **diferencias centrales** (nótese que se puede obtener la expresión 8.15 sumando miembro a miembro las ecuaciones 8.12 y 8.14 y luego despejando $f'(x_i)$).

Si en las expansiones 8.11 y 8.13 se desprecian los términos quinto, sexto, etc., del lado derecho y se suman miembro a miembro los términos que quedan, se obtiene

$$f''(x_i) = \left.\frac{d^2f}{dx^2}\right|_{x_i} \approx \frac{f(x_i + a) - 2f(x_i) + f(x_i - a)}{a^2} \tag{8.16}$$

que es la aproximación de la segunda derivada de $f(x)$ en x_i por diferencias centrales.

Las aproximaciones de derivadas no están restringidas a funciones de una sola variable; cuando se tiene una función de dos variables, por ejemplo $T(x, t)$, sus derivadas parciales por definición son como sigue

$$\begin{aligned} \frac{\partial T}{\partial x} &= \lim_{\Delta x \to 0} \frac{T(x + \Delta x, t) - T(x, t)}{\Delta x} \\ \frac{\partial T}{\partial t} &= \lim_{\Delta t \to 0} \frac{T(x, t + \Delta t) - T(x, t)}{\Delta t} \end{aligned} \tag{8.17}$$

Por esto, sus aproximaciones con diferencias hacia delante en el punto (x_i, t_j) quedan

$$\left.\frac{\partial T}{\partial x}\right|_{(x_i,t_j)} \approx \frac{T(x_i + a, t_j) - T(x_i, t_j)}{a} \quad \text{con} \quad a > 0$$
$$\left.\frac{\partial T}{\partial t}\right|_{(x_i,t_j)} \approx \frac{T(x_i, t_j + b) - T(x_i, t_j)}{b} \quad \text{con} \quad b > 0 \tag{8.18}$$

La aproximación con diferencias hacia atrás queda

$$\left.\frac{\partial T}{\partial x}\right|_{(x_i,t_j)} \approx \frac{T(x_i, t_j) - T(x_i - a, t_j)}{a}$$
$$\left.\frac{\partial T}{\partial t}\right|_{(x_i,t_j)} \approx \frac{T(x_i, t_j) - T(x_i, t_j - b)}{b} \tag{8.19}$$

y sumando las correspondientes de 8.18 y 8.19 se obtienen

$$\left.\frac{\partial T}{\partial x}\right|_{(x_i,t_j)} \approx \frac{T(x_i + a, t_j) - T(x_i - a, t_j)}{2a}$$
$$\left.\frac{\partial T}{\partial t}\right|_{(x_i,t_j)} \approx \frac{T(x_i, t_j + b) - T(x_i, t_j - b)}{2b} \tag{8.20}$$

que son las aproximaciones con diferencias centrales a las primeras derivadas parciales de $T(x, t)$.

Las segundas derivadas parciales de $T(x, t)$ quedan aproximadas con diferencias centrales así

$$\left.\frac{\partial^2 T}{\partial x^2}\right|_{(x_i,t_j)} \approx \frac{T(x_i + a, t_j) - 2T(x_i, t_j) + T(x_i - a, t_j)}{a^2}$$
$$\left.\frac{\partial^2 T}{\partial t^2}\right|_{(x_i,t_j)} \approx \frac{T(x_i, t_j + b) - 2T(x_i, t_j) + T(x_i, t_j - b)}{b^2} \tag{8.21}$$

Finalmente, se da la aproximación de la segunda derivada parcial combinada; esto es,

$$\frac{\partial^2 T}{\partial x \partial t} \approx \frac{T(x_i + a, t_j + b) - T(x_i - a, t_j + b) - T(x_i + a, t_j - b) + T(x_i - a, t_j - b)}{4ab} \tag{8.22}$$

cuya deducción se deja al lector como ejercicio.

Es importante observar que las ecuaciones 8.18 a 8.22 se pueden obtener a partir de la expansión de $T(x, t)$ en serie de Taylor, alrededor de (x_i, t_j); esto es, de

$$T(x_i + a, t_j + b) = T(x_i, t_j) + a \left.\frac{\partial T}{\partial x}\right|_{(x_i,t_j)} + b \left.\frac{\partial T}{\partial t}\right|_{(x_i,t_j)}$$

$$+ a^2 \frac{\partial^2 T}{\partial x^2}\Big|_{(x_i, t_j)} + 2ab \frac{\partial^2 T}{\partial x \partial t}\Big|_{(x_i, t_j)} + b^2 \frac{\partial^2 T}{\partial t^2}\Big|_{(x_i, t_j)} + \ldots \tag{8.23}$$

aplicando los mismos razonamientos que condujeron a 8.12 y de 8.14, a 8.16. (véase el problema 8.3 al final del capítulo).

Ecuación de calor en diferencias finitas

Una de las ecuaciones diferenciales parciales más estudiadas es

$$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2} \tag{8.24}$$

que describe la conducción de calor en régimen transitorio en una dimensión, la difusión unidireccional de masa en régimen transitorio, etcétera.

Por ejemplo, puede describir la conducción de calor en una barra aislada longitudinalmente durante cierto periodo, tomado a partir de $t = 0$. La barra se considera suficientemente delgada y de longitud L muy grande en comparación con su grosor. Sean los extremos de la barra tomados como $x = 0$ y $x = L$ (véase Fig. 8.4).

Sean además las condiciones siguientes

$$a)\ T(x,0) = f(x) \qquad 0 < x < L$$

Esta expresión, conocida como **condición inicial**, da el valor de la temperatura T en cualquier punto de la barra al tiempo de inicio $t = 0$.

$$b)\ T(0, t) = g_1(t)$$

$$T(L, t) = g_2(t) \qquad \text{con } t > 0$$

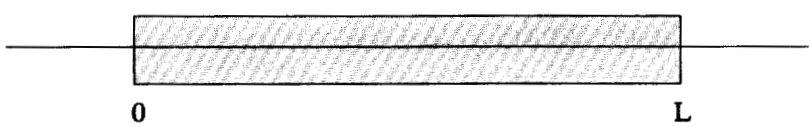


Figura 8.4 Barra aislada longitudinalmente.

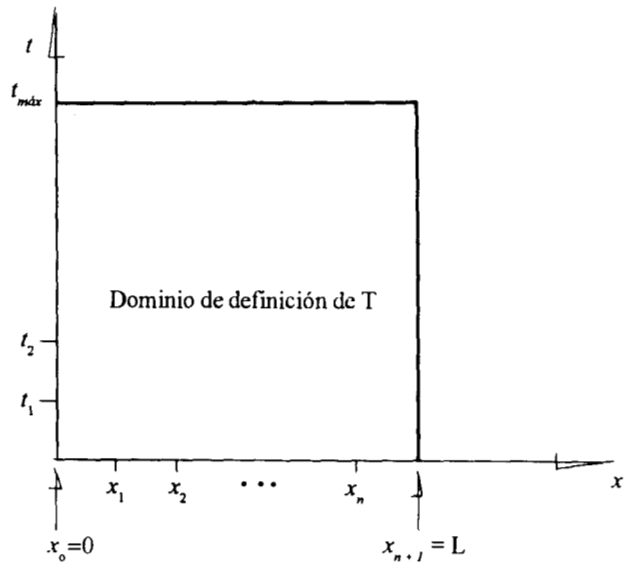


Figura 8.5 Dominio de definición de la función solución del problema de valor en la frontera (PVF) planteado.

Estas expresiones, conocidas como **condiciones frontera**, dan los valores de la temperatura T de la barra en sus extremos a cualquier tiempo t.

La ecuación 8.24 y las condiciones (a) y (b) constituyen un problema de valores en la frontera (PVF). Resolver este problema numéricamente, significa encontrar los valores de T en puntos seleccionados en la barra: $x_1, x_2, \ldots, x_n$ a ciertos tiempos escogidos: $t_1 < t_2 \ldots < t_{máx}$; esto es, calcular

$$\begin{array}{ccc} T(x_1, t_1), & T(x_2, t_1), \ldots, & T(x_n, t_1) \\ T(x_1, t_2), & T(x_2, t_2), \ldots, & T(x_n, t_2) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ T(x_1, t_{máx}), & T(x_2, t_{máx}), \ldots, & T(x_n, t_{máx}) \end{array} \tag{8.25}$$

Para ver esto geométricamente, primero se representa el dominio de definición de T como el rectángulo que se ilustra en el sistema coordenado x-t de la figura 8.5, y los puntos del dominio de definición donde se aproximarán los valores de T son los puntos de cruce de las horizontales $t = t_1, \ldots, t = t_{máx}$ y las verticales $x = x_1, \ldots, x = x_n$, que en adelante se llamarán **nodos** (véase Fig. 8.6).

La ecuación 8.24 es válida en todo el dominio de definición, por lo que evidentemente será válida en cualquier nodo, por ejemplo (x_i, t_j); esto es,

$$\left.\frac{\partial T}{\partial t}\right|_{(x_i, t_j)} = \alpha \left.\frac{\partial^2 T}{\partial x^2}\right|_{(x_i, t_j)}$$

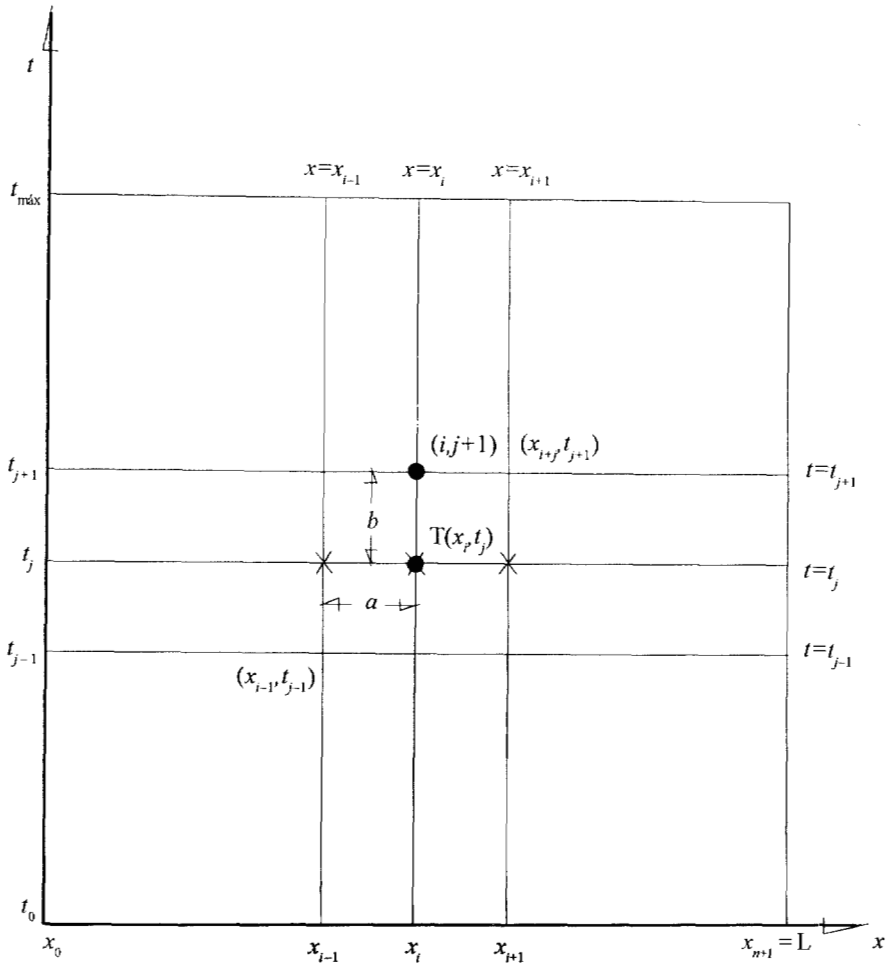


Figura 8.6. Nodos de una red construida en el dominio de definición.

Si se sustituyen ahora las derivadas parciales evaluadas en (x_i, t_j) con sus aproximaciones con diferencias finitas en esta ecuación; por ejemplo, con diferencias finitas hacia delante a $\partial T/\partial t$ y diferencias centrales a $\partial^2 T/\partial x^2$, se obtiene

$$\frac{T_{i,j+1} - T_{i,j}}{b} = \alpha \frac{T_{i-1,j} - 2T_{i,j} + T_{i+1,j}}{a^2} \tag{8.26}$$

Se ha remplazado $T(x_i, t_j)$ con $T_{i,j}$ para simplificar la notación.

Obsérvese además que los nodos marcados con punto negro (•) en la Figura 8.6 son los nodos usados para aproximar $\partial T/\partial t$ y los marcados con una cruz (×) se emplean a fin de aproximar a $\partial^2 T/\partial x^2$.

De manera similar se obtienen las ecuaciones para los demás nodos de la red (o malla), lo que conduce a un conjunto de ecuaciones algebraicas que pueden ser simultáneas o no y que involucran los valores de $T_{i,j}$ que se buscan. Su solución es la misma del problema de valor en la frontera (PVF).

Es oportuno hacer notar que la derivada parcial en el tiempo $\partial T/\partial t$ se pudo aproximar con diferencias hacia atrás o con diferencias divididas centrales.

SECCIÓN 8.3 SOLUCIÓN DE PROBLEMAS DE VALORES EN LA FRONTERA (ecuación de calor unidimensional)

Método explícito

Para ilustrar este método se resuelve el PVF planteado en la sección anterior con los datos siguientes

$$\text{PVF} \begin{cases} \dfrac{\partial T}{\partial t} = \alpha \dfrac{\partial^2 T}{\partial x^2} \\ T(x,0) = 20\ ^\circ\text{F} \qquad 0 < x < L \\ T(0,t) = 100\ ^\circ\text{F} \\ T(1,t) = 100\ ^\circ\text{F} \qquad t > 0 \end{cases}$$

y

$$\alpha = 1 \text{ pie}^2/\text{h}$$
$$L = 1 \text{ pie}$$
$$t_{máx} = 1 \text{ h}$$

SOLUCIÓN

Primero se construye la malla en el dominio de definición dividiendo la longitud de la barra (1 pie) en cuatro subintervalos y el intervalo de tiempo (1 h) en 100 subintervalos.

Las condiciones frontera proporcionan la temperatura en cualquier punto del eje t y de la vertical $x = 1$ a cualquier tiempo, mientras que la condición inicial proporciona la temperatura en cualquier punto del eje horizontal x al tiempo cero.

Cada nodo de la malla queda definido por dos coordenadas (i, j); por ejemplo, el nodo de coordenadas (3,4) representa la temperatura en el punto $x = 0.75$ pies de la barra al tiempo $t = 0.04$ horas, y el nodo (4,1) la temperatura de la barra en $x = 1$ pies (su frontera) y a $t = 0.01$ horas (veáse Fig. 8.7).

Nótese que en el nodo de coordenadas (0,0) (esquina izquierda inferior de la malla), la temperatura debería ser 20°F atendiendo la condición inicial, mientras que la condición frontera T(0, t) establece que debería ser de 100°F.

Los puntos que representan estas características se llaman **puntos singulares**; se acostumbra tomar en ellos un valor de temperatura igual a la media aritmética de las temperaturas sugeridas por la condición inicial y la condición frontera correspondientes. La temperatura tomada para el nodo (0,0) es 60°F. De igual manera se trata el punto (4,0), cuya temperatura también es 60°F.

Hechas estas consideraciones, el segundo paso consiste en aproximar la ecuación diferencial parcial del problema de valor en la frontera en el nodo (1,0) por la ecuación 8.26; entonces queda

$$\frac{T_{1,1} - T_{1,0}}{b} = \alpha \frac{T_{0,0} - 2T_{1,0} + T_{2,0}}{a^2}$$

Los nodos involucrados en esta ecuación están marcados por círculos y cruces en la figura 8.7. De éstos, solamente en el nodo (1,1) la temperatura es desconocida, por lo que puede despejarse; entonces resulta

$$T_{1,1} = \alpha \frac{b}{a^2}(T_{0,0} - 2T_{1,0} + T_{2,0}) + T_{1,0}$$

al sustituir valores queda

$$T_{1,1} = 1 \frac{0.01}{(0.25)^2}(60 - 2(20) + 20) + 20 = 26.4$$

Si ahora se aproxima la ecuación 8.24 en el nodo $(i, j) = (2,0)$, mediante la 8.26, se obtiene

$$\frac{T_{2,1} - T_{2,0}}{b} = \alpha \frac{T_{1,0} - 2T_{2,0} + T_{3,0}}{a^2}$$

Ahora sólo se desconoce la temperatura del punto (2,1), ya que todos los demás están dados por la condición inicial; despejando se tiene

$$T_{2,1} = \alpha \frac{b}{a^2}(T_{1,0} - 2T_{2,0} + T_{3,0}) + T_{2,0}$$

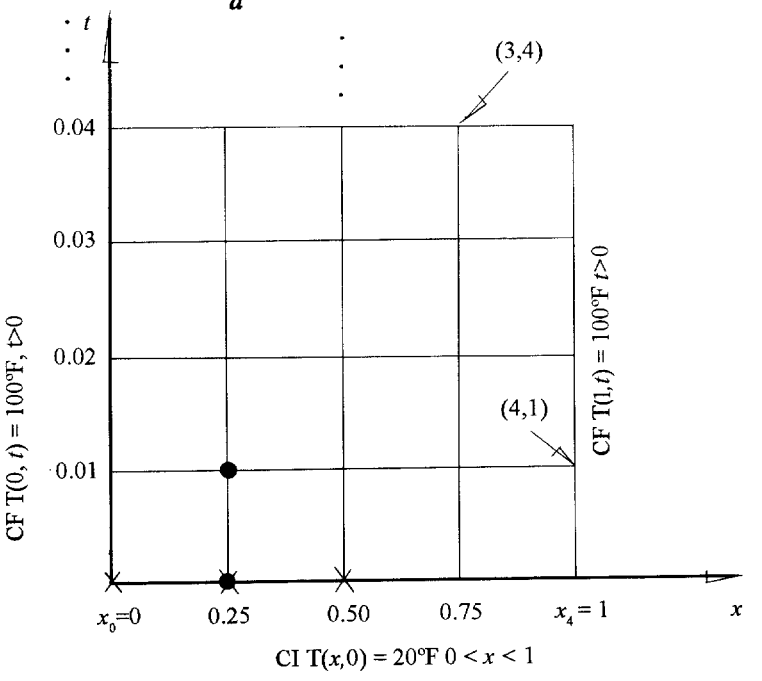


Figura 8.7. Representación de la malla en el dominio de definición.

se sustituyen valores

$$T_{2,1} = 1 \frac{0.01}{(0.25)^2} (20 - 2(20) + 20) + 20 = 20$$

Se repiten las mismas consideraciones y cálculos para el punto (3,0) y se obtiene

$$T_{3,1} = 1 \frac{0.01}{(0.25)^2} (T_{2,0} - 2T_{3,0} + T_{4,0}) + T_{3,0}$$

$$T_{3,1} = 26.4$$

De esta manera se han obtenido aproximaciones a la temperatura en los tres puntos seleccionados de la barra a un tiempo de 0.01 horas. Al momento se tiene la temperatura de todos los nodos de las dos primeras líneas horizontales (filas) de la malla y se procederá, siguiendo el razonamiento anterior, a calcular la temperatura en todos los nodos intermedios de la tercera fila (1,2), (2,2) y (3,2).

Se empieza con el punto $(i, j) = (1,1)$ y se aplica la ecuación 8.26, con lo que se obtiene

$$\frac{T_{1,2} - T_{1,1}}{b} = \alpha \frac{T_{0,1} - 2T_{1,1} + T_{2,1}}{a^2}$$

de la que

$$T_{1,2} = \alpha \frac{b}{a^2} (T_{0,1} - 2T_{1,1} + T_{2,1}) + T_{1,1}$$

con la sustitución de valores queda

$$T_{1,2} = 1 \frac{0.01}{(0.25)^2} (100 - 2(26.4) + 20) + 26.4 = 37.152$$

Al proceder análogamente para los otros puntos se llega a

$$T_{2,2} = 22.048$$

$$T_{3,2} = 37.152$$

Con esto se tiene la temperatura en los tres puntos seleccionados de la barra cuando hayan transcurrido 0.02 horas.

Este procedimiento se repite para la cuarta, quinta, etc. filas, con lo cual se obtienen las temperaturas en los puntos seleccionados de la barra a tiempo $t = 0.03$, $t = 0.04$, etc., hasta llegar al tiempo fijado como $t_{máx} = 1$ hora.

De los cien conjuntos de temperaturas obtenidas, en la tabla 8.1 se muestran sólo algunos para facilitar su presentación y análisis.

Este metódo también se conoce como **método de diferencias hacia delante.**

Discusión de resultados

- Hay simetría en la distribución de temperaturas en la barra debido a que: *a*) la temperatura inicial es constante; *b*) la temperatura es constante e igual en las fronteras, y *c*) las propiedades físicas de la barra son independientes de x y t.

tiempo (hrs)	x (pies)				
	0.00	0.25	0.5	0.75	1.0
0.00	60	20.000	20.000	20.000	60
0.01	100	26.400	20.000	26.400	100
0.02	100	37.152	22.048	37.152	100
0.03	100	44.791	26.881	44.791	100
0.04	100	50.759	32.612	50.759	100
0.05	100	55.734	38.419	55.734	100
0.06	100	60.046	43.960	60.046	100
0.07	100	63.865	49.108	63.865	100
0.08	100	67.285	53.830	67.285	100
0.09	100	70.367	58.136	70.367	100
0.10	100	73.151	62.050	73.151	100
0.20	100	89.968	85.812	89.968	100
0.40	100	98.599	98.018	98.599	100
0.60	100	99.804	99.723	99.804	100
0.80	100	99.973	99.961	99.973	100
1.00	100	99.996	99.995	99.996	100

Tabla 8.1 Resultados de la solución del PVF de conducción de calor en una barra metálica.

- La temperatura en el centro de la barra es un mínimo, de manera que se satisface

$$\left.\frac{dT}{dx}\right|_{x=\frac{1}{2}} = 0$$

(veáse Fig. 8.8), ya que es el punto más alejado de los extremos, los cuales tienen las temperaturas que impulsan el flujo de calor hacia el centro de la barra.

- Nótese que cuando $t = 0.01$ la temperatura en el punto central es igual a la inicial, o sea $T(0.5,0.01) = 20°F$. Esta situación no es congruente con el fenómeno que ocurre, ya que es de esperar que la temperatura cambie después del instante cero. El resultado se debe a que la estimación de la temperatura en un nodo depende de las temperaturas de los nodos en un tiempo previo.
- La temperatura en la barra tiende al régimen permanente a medida que transcurre el tiempo, es decir $T \to 100°F$ cuando $t \to \infty$.

- Solo se encontró la temperatura en tres puntos interiores de la barra; si se desea información de mayor número de puntos interiores, debe construirse una malla más cerrada; es decir, subdividir la longitud L en más subintervalos.

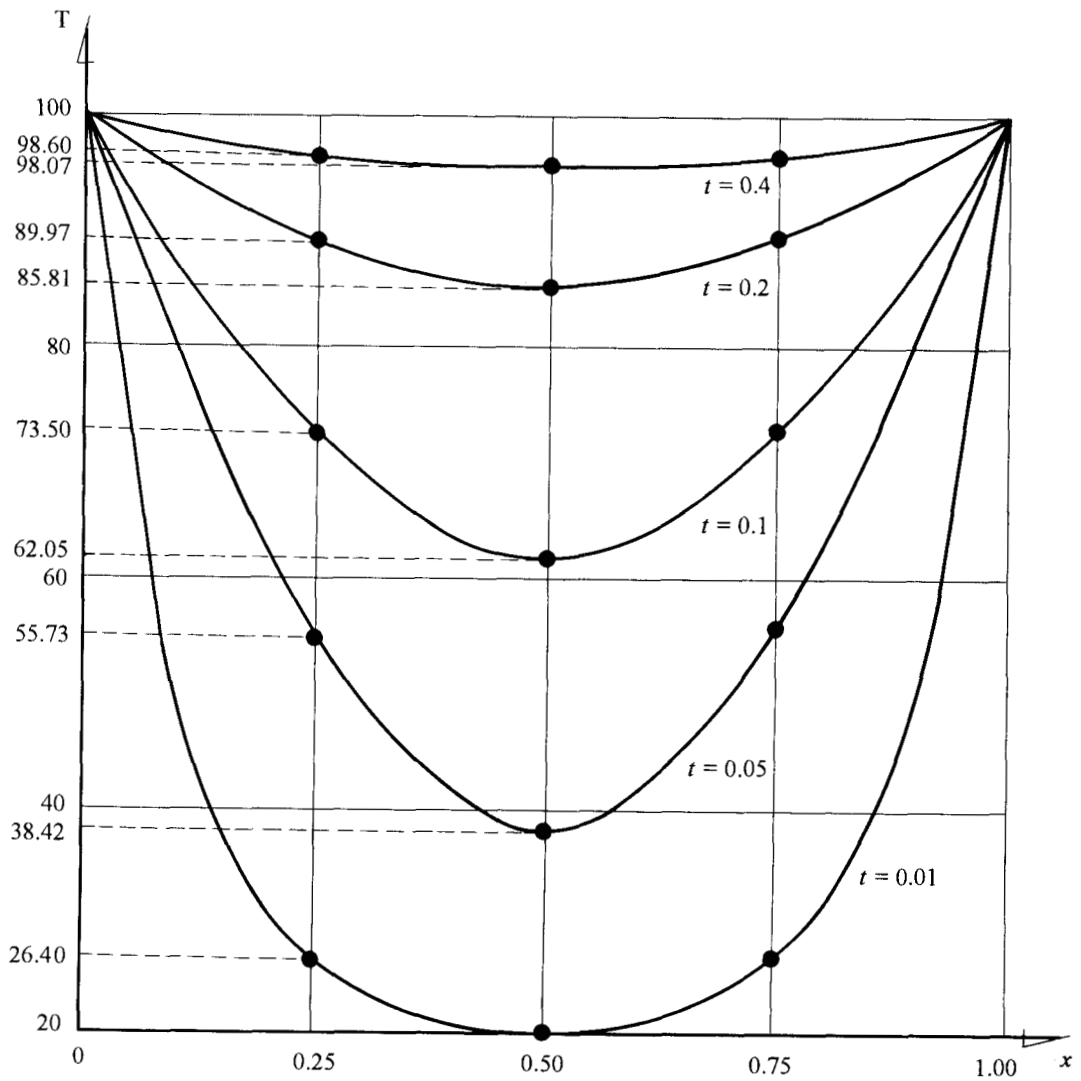


Figura 8.8. Representación gráfica de algunas filas de la tabla 8.1.

ALGORITMO 8.1 Método explícito

Para aproximar la solución al problema de valor en la frontera

$$\begin{cases} \alpha \dfrac{\partial^2 T}{\partial x^2} = \dfrac{\partial T}{\partial t} \\ T(x,0) = f(x) \quad 0 < x < x_F \\ T(0,t) = g_1(t) \\ \qquad\qquad\qquad\qquad t > 0 \\ T(x_F,t) = g_2(t) \end{cases}$$

Proporcionar las funciones CI(X), CF1(T) y CF2(T) y los

DATOS: El número NX de puntos de la malla en el eje x, el número NT de puntos de la malla en el eje t, la longitud total XF del eje x, el tiempo máximo TF por considerar y el coeficiente ALFA de la derivada de segundo orden.

RESULTADOS: Los valores de la variable dependiente T a lo largo del eje x a distintos tiempos t: T.

```
PASO  1  Hacer DX = XF/(NX-1)
PASO  2. Hacer DT = TF/(NT-1)
PASO  3. Hacer LAMBDA = ALFA*DT/DX**2
PASO  4. Hacer I=2
PASO  5. Mientras I ≤ NX-1, repetir los paso 6 y 7.
           PASO 6.  Hacer T(I) = CI(DX*(I-1))
           PASO 7.  Hacer I= I+1
PASO  8. Hacer T(1) = (CI(DX) + CF1(DT))/2
PASO  9. Hacer T(NX) = (CI(XF-DX) + CF2(DT))/2
PASO 10. IMPRIMIR T
PASO 11. Hacer J = 1
PASO 12. Mientras J ≤ NT repetir los paso 13 a 24.
           PASO 13. Hacer I=2
           PASO 14. Mientras I ≤ NX-1, repetir los pasos 15 y 16.
                      PASO 15  Hacer T1(I)= LAMBDA*T(I-1)
                               +(1-2*LAMBDA)*T(I) +
                               LAMBDA*T(I+1)
                      PASO 16. Hacer I = I+1
           PASO 17. Hacer I=2
           PASO 18. Mientras I ≤ NX-1, repetir los pasos 19 y 20.
                      PASO 19. Hacer T(I) = T1(I)
                      PASO 20. Hacer I = I+1
           PASO 21. Hacer T(1) = CF1(DT*J)
           PASO 22. Hacer T(NX) = CF2(DT*J)
           PASO 23. IMPRIMIR T.
           PASO 24. Hacer J = J+1
PASO 25. TERMINAR.
```

Ejemplo 8.1

Calcule la temperatura como una función de x y t en una barra aislada de longitud unitaria (en pies), sujeta a las siguientes condiciones inicial y de frontera

CI	$T(x, 0) = 50 \text{ sen } \pi x$	$0 < x < 1$
CF1	$T(0, t) = 100\ ^\circ F$	$t > 0$
CF2	$T(1, t) = 50\ ^\circ F$	

y con $\alpha = 1$ pie^2 / h.

SOLUCIÓN

El problema de condiciones en la frontera queda establecido como sigue

$$\text{PVF} \begin{cases} \dfrac{\partial T}{\partial t} = \alpha \dfrac{\partial^2 T}{\partial x^2} \text{ (ecuación diferencial parcial, EDP)} \\ T(x, 0) = 50 \text{ sen } \pi x \text{ (condición inicial, CI)} \\ T(0, t) = 100\ ^\circ F \text{ (condición frontera 1, CF1)} \\ T(1, t) = 50\ ^\circ F \text{ (condición frontera 2, CF2)} \end{cases}$$

Ahora se divide la barra en $n = 8$ segmentos o subintervalos, de tal manera que se tiene un total de nueve nodos en cada fila, de los cuales siete son interiores; con esto, $a = 0.125$ pies. El fenómeno se estudiará durante media hora y se dividirá este tiempo de interés en $m = 100$, que da b = 0.005 horas. La malla queda como se ve en la figura 8.9.

Para mayor facilidad del uso de la ecuación 8.26 se despeja el término $T_{i,j+1}$, ya que representa la temperatura desconocida, y se denominará λ el término $\alpha b/a^2$; después de algunas manipulaciones algebraicas, dicha ecuación queda

$$\boxed{T_{i,j+1} = \lambda T_{i-1,j} + (1 - 2\lambda) T_{i,j} + \lambda T_{i+1,j}} \qquad (8.27)$$

cuya aplicación en $(i, j) = (1,0)$ produce

$$T_{1,1} = \lambda T_{0,0} + (1 - 2\lambda) T_{1,0} + \lambda T_{2,0}$$

se calcula el valor de λ

$$\lambda = 1(0.005)/(0.125)^2 = 0.32$$

y se substituyen valores

$$T_{1,1} = 0.32\ (50) + (1-2(0.32))\ (50) \text{ sen } (0.125\pi) + 0.32\ (50) \text{ sen } (0.25\pi)$$
$$= 34.2$$

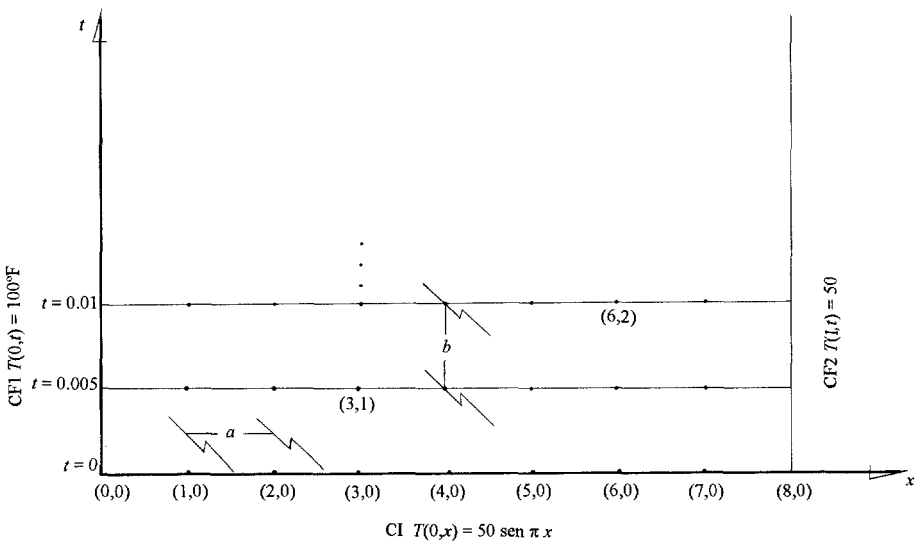


Figura 8.9 Malla del ejemplo 8.1.

Para $(i, j) = (2,0)$ se tiene

$$T_{2,1} = \lambda T_{1,0} + (1 - 2\lambda) T_{2,0} + \lambda T_{3,0}$$

$$T_{2,1} = 0.32\,(50)\,\text{sen}\,(0.125\,\pi) + (1 - 2\,(0.32))\,(50)\,\text{sen}\,(0.250\,\pi)$$

$$+ 0.32\,(50)\,\text{sen}\,(0.375\,\pi) = 33.63$$

Al continuar

$$T_{3,1} = \lambda T_{2,0} + (1 - 2\lambda)\, T_{3,0} + \lambda T_{4,0}$$

$$T_{3,1} = 0.32\,(50)\,\text{sen}\,(0.25\,\pi) + (1 - 2\,(0.32))\,(50)\,\text{sen}\,(0.375\,\pi)$$

$$+ 0.32\,(50)\,\text{sen}\,(0.5\,\pi) = 43.94$$

$$T_{4,1} = \lambda T_{3,0} + (1 - 2\lambda)\, T_{4,0} + \lambda\, T_{5,0}$$

$$T_{4,1} = 0.32\,(50)\,\text{sen}\,(0.375\,\pi) + (1 - 2\,(0.32))\,(50)\,\text{sen}\,(0.5\pi)$$

$$+ 0.32\,(50)\,\text{sen}\,(0.625\,\pi) = 47.56$$

$$T_{5,1} = \lambda\, T_{4,0} + (1-2\lambda)\, T_{5,0} + \lambda\, T_{6,0}$$

$$T_{5,1} = 0.32\,(50)\,\text{sen}\,(0.5\,\pi) + (1-2\,(0.32))\,(50)\,\text{sen}\,(0.625\,\pi) + 0.32\,(50)\,\text{sen}\,(0.75\,\pi) = 43.94$$

$$T_{6,1} = \lambda\, T_{5,0} + (1-2\lambda)\, T_{6,0} + \lambda\, T_{7,0}$$

$$T_{6,1} = 0.32\,(50)\,\text{sen}\,(0.625\,\pi) + (1-2\,(0.32))\,(50)\,\text{sen}\,(0.75\,\pi) + 0.32\,(50)\,\text{sen}\,(0.875\,\pi) = 33.63$$

$$T_{7,1} = \lambda\, T_{6,0} + (1-2\lambda)\, T_{7,0} + \lambda\, T_{8,0}$$

$$T_{7,1} = 0.32\,(50)\,\text{sen}\,(0.75\,\pi) + (1-2\,(0.32))\,(50)\,\text{sen}\,(0.875\,\pi) + 0.32\,(25) = 26.2$$

Estas temperaturas corresponden a puntos discretos sobre la barra a un tiempo igual a 0.005 horas.

Para obtener la temperatura en los mismos puntos de la barra dados arriba, pero ahora a un tiempo de 0.01 h (tercera fila de malla de la fig. 8.9), se aplica nuevamente la ecuación 8.27. De la misma manera se obtienen los valores de temperatura para los tiempos de 0.015, 0.02 h, etc.; o sea, la temperatura en los nodos interiores de las filas 4, 5, etc. Los resultados obtenidos con el programa 8.1 son

t (horas)	x (pies)								
	0.0	0.125	0.25	0.375	0.5	0.625	0.75	0.875	1.0
0.000	50	19.13	35.36	46.19	50.00	46.19	35.36	19.13	25
0.005	100	34.20	33.63	43.94	47.56	43.94	33.63	26.20	50
0.010	100	55.08	37.11	41.80	45.25	41.80	34.55	36.20	50
0.015	100	63.70	44.36	41.40	43.04	40.59	37.40	40.09	50
0.020	100	69.13	49.61	42.88	41.73	40.35	39.28	42.40	50
0.025	100	72.76	53.70	44.66	41.66	40.45	40.62	43.83	50
0.030	100	75.38	56.91	46.59	42.23	40.89	41.59	44.78	50
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
0.050	100	81.19	65.19	53.68	46.92	44.17	44.41	46.68	50
0.100	100	86.98	75.08	65.18	57.81	53.06	50.61	49.86	50
0.150	100	89.73	80.09	71.59	64.57	59.14	55.16	52.28	50
0.200	100	91.32	83.02	75.40	68.67	62.90	58.03	53.83	50
0.300	100	92.86	85.85	79.10	72.67	66.60	60.85	55.36	50
0.400	100	93.42	86.89	80.46	74.14	67.96	61.89	55.92	50
0.500	100	93.63	87.28	80.96	74.68	68.46	62.28	56.13	50

Tabla 8.2 Temperaturas (°F) del ejemplo 8.1.

Discusión de los resultados

- En este caso la distribución de temperaturas con respecto a x no es simétrica, a pesar de que la distribución inicial si lo es, 50 sen (πx). Esto se debe a que la temperatura en los extremos es diferente, lo cual genera un flujo de calor más intenso del extremo izquierdo hacia el centro de la barra, pues el extremo izquierdo tiene la temperatura más elevada.
- La temperatura en el centro de la barra y sus cercanías disminuye en el intervalo $0 < t < 0.05$. Esto se debe a que cuando $t = 0$, la temperatura en el centro de la barra es mayor que en sus vecindades, de tal modo que en los primeros instantes hay flujo de calor del centro de la barra hacia los extremos [más pronunciado hacia el extremo derecho, ya que $T(1, t) < T(0, t)$], razón por la cual la temperatura en la zona central disminuye cierto tiempo.
- Cuando el lapso es amplio ($t > 0.5$), la distribución de la temperatura es casi lineal a lo largo de la barra; es de esperarse que sea lineal cuando $t \to \infty$, ya que la ecuación diferencial parcial se transforma en $d^2T/dx^2 = 0$, pues $dT/dt = 0$; o sea que se alcanza el régimen permanente.

Método implícito

Para ilustrar este método se resolverá nuevamente el ejemplo

$$\text{PVF} \quad \begin{cases} \text{EDP} \quad \dfrac{\partial T}{\partial t} = \alpha \dfrac{\partial^2 T}{\partial x^2} \text{ y } \alpha = 1 \text{ pie}^2/\text{h} \\ \text{C.I. } T(x,0) = 20°\text{F} \quad L = 1 \text{ pie} \\ \text{CF 1 } T(0,t) = 100°\text{F} \quad t_{máx} = 1 \text{ h} \\ \text{CF 2 } T(1,t) = 100 \text{ °F} \end{cases}$$

ya utilizado para mostrar el método explícito.

Primero se obtendrá la ecuación básica del algoritmo.

Se toma el nodo (i, j) de la malla construida sobre el dominio de definición $0 = t_0 < t < t_{máx} = 1$, $0 < x < L = 1$ (Fig. 8.10) y se evalúa la EDP, entonces

$$\left.\frac{\partial T}{\partial t}\right|_{(x_i, t_j)} = \alpha \left.\frac{\partial^2 T}{\partial x^2}\right|_{(x_i, t_j)}$$

Ahora se sustituye $\partial T/\partial t$ en (x_i, t_j) por diferencias hacia atrás, y $\partial^2 T/\partial x^2$ en (x_i, t_j) por diferencias centrales, lo que da

$$\boxed{\frac{T_{i,j} - T_{i,j-1}}{b} = \alpha \frac{T_{i-1,j} - 2T_{i,j} + T_{i+1,j}}{a^2}} \tag{8.28}$$

De acuerdo con la notación de punto negro (•) para los nodos empleados a fin de aproximar a $\partial T/\partial t$ y cruz (×) para aquellos que se usan en la aproximación de $\partial^2 T/\partial x^2$, se tiene el esquema de la figura 8.11.

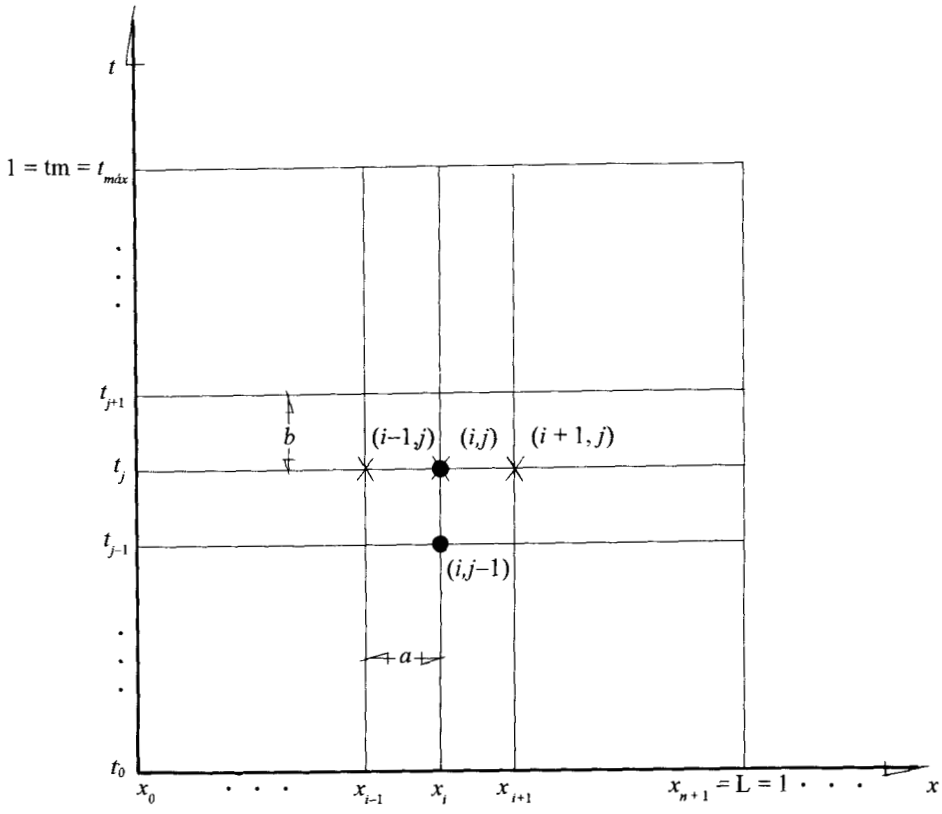


Figura 8.10.

SOLUCIÓN

Se construye la malla con $n = 4$ y $m = 100$, con lo que $a = 0.25$ y $b = 0.01$. Si $(i, j) = (1,1)$, la ecuación 8.28 se aproxima así

$$\frac{T_{1,1,} - T_{1,0}}{b} = \alpha \frac{T_{0,1} - 2T_{1,1} + T_{2,1}}{a^2}$$

La temperatura en los nodos (0,1) y (1,0) está dada por las condiciones frontera e inicial respectivamente; pero se desconoce la temperatura en los nodos (1,1) y (2,1). Entonces se tiene una ecuación con dos incógnitas que rearreglada queda

$$(1 + 2\lambda)\, T_{1,1} - \lambda\, T_{2,1} = T_{1,0} + \lambda\, T_{0,1} \qquad (8.29)$$

donde, como se sabe, $\lambda = \alpha b/a^2$ (que es un parámetro adimensional).

El procedimiento se repite en el nodo (2,1) y la ecuación diferencial parcial queda aproximada por

$$\frac{T_{2,1} - T_{2,0}}{b} = \alpha \frac{T_{1,1} - 2T_{2,1} + T_{3,1}}{a^2}$$

En esta ecuación hay tres incógnitas, $T_{1,1}$, $T_{2,1}$ y $T_{3,1}$; así pues, al rearreglarla queda

$$-\lambda T_{1,1} + (1 + 2\lambda) T_{2,1} - \lambda T_{3,1} = T_{2,0} \tag{8.30}$$

Análogamente para el nodo (3,1), la ecuación diferencial parcial (EDP) queda aproximada por

$$\frac{T_{3,1} - T_{3,0}}{b} = \alpha \frac{T_{2,1} - 2T_{3,1} + T_{4,1}}{a^2}$$

En esta ecuación sólo hay dos incógnitas, que son $T_{2,1}$ y $T_{3,1}$; así pues, al rearreglarla resulta

$$-\lambda T_{2,1} + (1 + 2\lambda) T_{3,1} = T_{3,0} + \lambda T_{4,1} \tag{8.31}$$

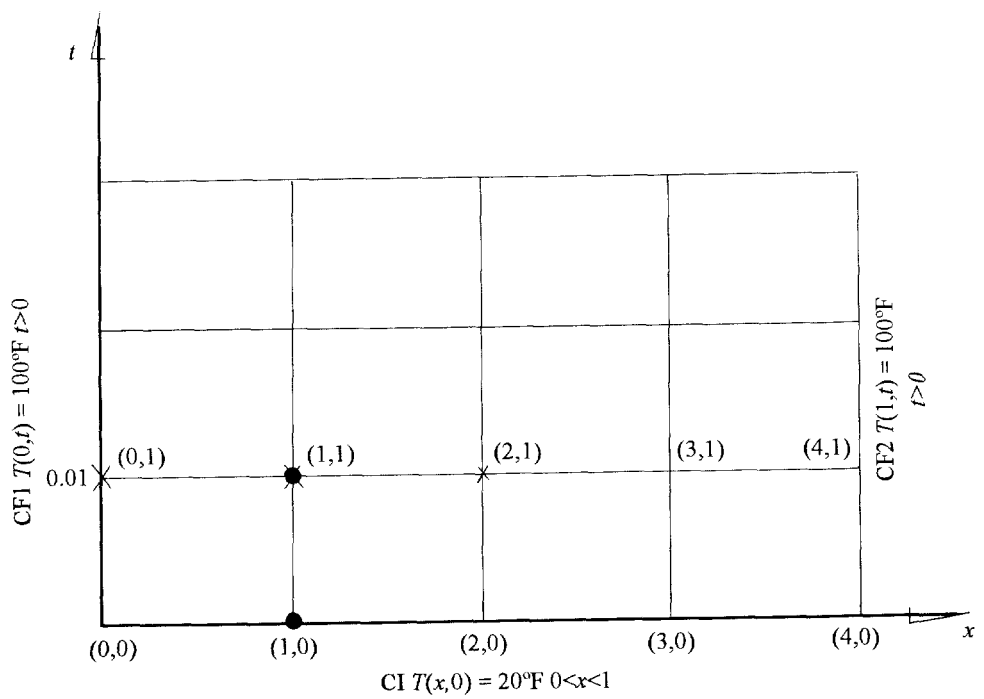


Figura 8.11.

Las ecuaciones 8.29 a 8.31 constituyen un sistema de ecuaciones algebraicas lineales en las incógnitas $T_{1,1}$, $T_{2,1}$ y $T_{3,1}$, que son precisamente las temperaturas que se desea conocer. Esto es

$$\begin{array}{llll} (1+2\lambda)T_{1,1} & -\lambda T_{2,1} & & = \lambda T_{0,1} + T_{1,0} \\ -\lambda T_{1,1} & +(1+2\lambda)T_{2,1} & -\lambda T_{3,1} & = T_{2,0} \\ & -\lambda T_{2,1} & +(1+2\lambda)T_{3,1} & = T_{3,0} + \lambda T_{4,1} \end{array}$$

Con la sustitución de valores

$$\lambda = 0.16, \quad T_{1,0} = T_{2,0} = T_{3,0} = 20\ ^\circ\text{F}, \quad T_{0,1} = T_{4,1} = 100\ ^\circ\text{F}$$

y resolviendo por alguno de los métodos del capítulo 2, se obtiene

$$T_{1,1}, = 29.99, \quad T_{2,1} = 22.42, \quad T_{3,1} = 29.99$$

Obsérvese que estas temperaturas obtenidas para t=0.01 h son diferentes a las obtenidas con el método explícito; además, la temperatura del punto central es distinta de la condición inicial. Esta situación es más congruente con la realidad del fenómeno que ocurre (recuérdese que con el método explícito la temperatura es 20°F). Lo anterior se explica porque para el cálculo se han tomado en cuenta todos los nodos de la primera y segunda filas, excepto los de las esquinas T(0,0) y T(4,0).

Mediante la ecuación 8.28 y los mismos razonamientos para la segunda y tercera filas se llega a

$$\begin{array}{llll} (1+2\lambda)T_{1,2} & -\lambda T_{2,2} & & = \lambda T_{0,2} + T_{1,1} \\ -\lambda T_{1,2} & +(1+2\lambda)T_{2,2} & -\lambda T_{3,2} & = T_{2,1} \\ & -\lambda T_{2,2} & +(1+2\lambda)T_{3,2} & = \lambda T_{4,2} + T_{3,1} \end{array}$$

Al sustituir valores conocidos

$$\lambda = 0.16, \quad T_{0,2} = T_{4,2} = 100, \quad T_{1,1}, = T_{3,1} = 29.99, \quad T_{2,1} = 22.42,$$

y resolver se obtiene:

$$T_{1,2} = 38.02,\ T_{2,2} = 26.2,\ T_{3,2} = 38.02,$$

que son las temperaturas correspondiente a $t = 0.02$ h y a $x = 0.25$, $x = 0.5$, y $x = 0.75$ pies, respectivamente.

Al aproximar la EDP por diferencias divididas en la fila $j+1$ (véase Fig. 8.11) se obtiene el siguiente sistema

$$\begin{array}{llll} (1+2\lambda)T_{1,j+1} & -\lambda T_{2,j+1} & & = \lambda T_{0,j+1} + T_{1,j} \\ -\lambda T_{1,j+1} & +(1+2\lambda)T_{2,j+1} & -\lambda T_{3,j+1} & = T_{2,j} \\ & -\lambda T_{2,j+1} & +(1+2\lambda)T_{3,j+1} & = \lambda T_{4,j+1} + T_{3,j} \end{array}$$

Obsérvese que en todos los casos el sistema por resolver tiene la misma matriz coeficiente, que es tridiagonal y simétrica.

Todo el sistema se soluciona estableciendo y resolviendo secuencialmente los sistemas de tres ecuaciones simultáneas para cada fila a partir de la segunda. Los resultados obtenidos con el programa 8.2 se presentan en la tabla 8.3.

tiempo (hrs)	x (pies)				
	0.00	0.25	0.50	0.75	1.00
0.00	60	20.00	20.00	20.00	60
0.01	100	29.99	22.43	29.99	100
0.02	100	38.02	26.20	38.02	100
0.03	100	44.64	30.67	44.64	100
0.04	100	50.23	35.41	50.23	100
0.05	100	55.04	40.17	55.04	100
0.06	100	59.25	44.80	59.25	100
0.07	100	62.97	49.20	62.97	100
0.08	100	66.29	53.35	66.29	100
0.09	100	69.28	57.21	69.28	100
0.10	100	71.97	60.79	71.97	100
. . .	. . .	. . .	. . .	. . .	. . .
0.20	100	88.62	83.91	88.62	100
0.40	100	98.10	97.32	98.10	100
0.60	100	99.68	99.55	99.68	100
0.80	100	99.95	99.93	99.95	100
1.00	100	99.99	99.99	99.99	100

Tabla 8.3. Resultados de la solución del PVF de conducción de calor en una barra metálica.

En general, si se divide la longitud de la barra en n subintervalos, o sea con n-1 nodos interiores (véase Fig. 8.12), el sistema de n-1 ecuaciones simultáneas con n-1 incógnitas para la fila j+1 queda

$$
\begin{array}{llll}
(1+2\lambda)\,T_{1,j+1} & -\lambda T_{2,j+1} & & = \lambda T_{0,j+1} + T_{1,j} \\
-\lambda T_{1,j+1} & +(1+2\lambda)\,T_{2,j+1} & -\lambda T_{3,j+1} & = T_{2,j} \\
 & -\lambda T_{2,j+1} & +(1+2\lambda)T_{3,j+1} \quad -\lambda T_{4,j+1} & = \lambda T_{3,j} \\
 & & \vdots & \\
 & -\lambda T_{n-3,j+1} & +(1+2\lambda)T_{n-2,j+1} \quad -\lambda T_{n-1,j+1} & = T_{n-2,j} \\
 & & -\lambda T_{n-2,j+1} \quad +(1+2\lambda)T_{n-1,j+1} & = T_{n-1,j} + \lambda T_{n,j+1}
\end{array}
$$

La solución de este sistema corresponde a las temperaturas en los puntos seleccionados de la barra a un tiempo $(j+1)b$.

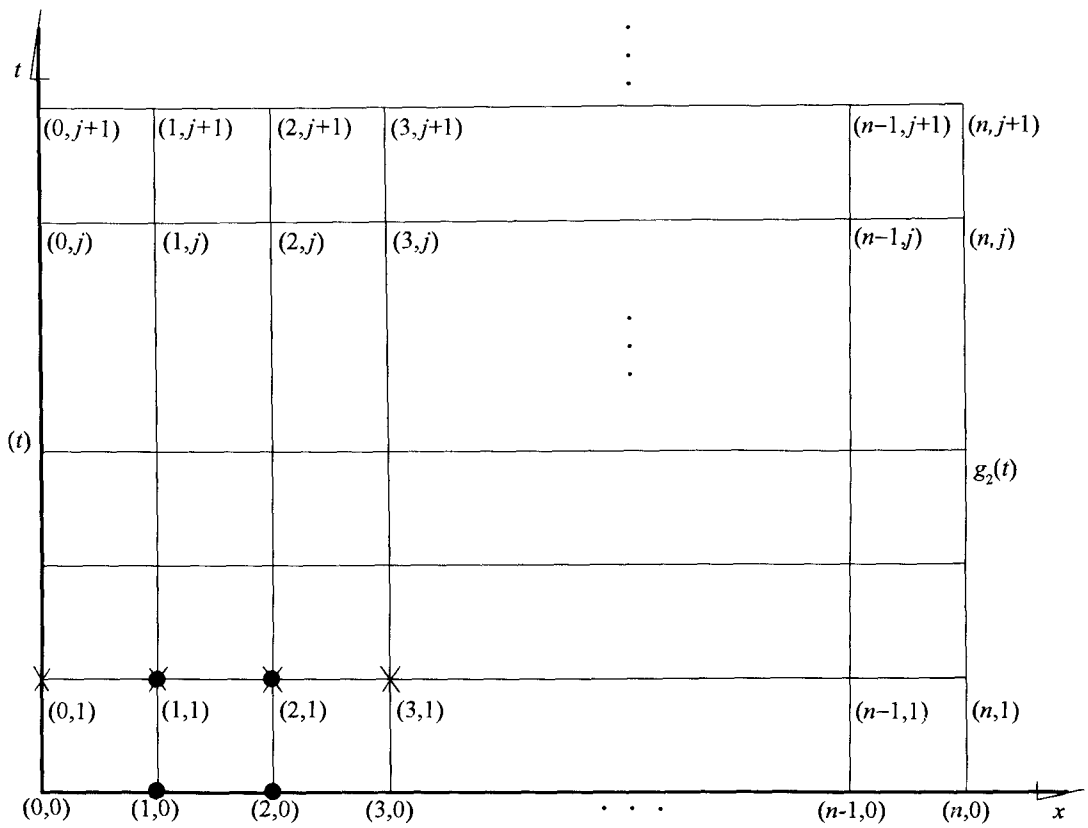


Figura 8.12.

Nótese la simetría de la matriz coeficiente y su característica tridiagonal. Además, los elementos de esta matriz son constantes para cualquier fila (o tiempo) y son

$$\begin{bmatrix} (1+2\lambda) & -\lambda & 0 & & & & & 0 \\ -\lambda & (1+2\lambda) & -\lambda & & & & \cdots & \cdot \\ 0 & -\lambda & (1+2\lambda) & -\lambda & & & & \cdot \\ \cdot & & & & & & & \cdot \\ \cdot & & & & & & & 0 \\ \cdot & & & & -\lambda & (1+2\lambda) & & -\lambda \\ 0 & \cdot & \cdots & & 0 & -\lambda & & (1+2\lambda) \end{bmatrix}$$

ALGORITMO 8.2 Método implícito

Para aproximar la solución al Problema de valor en la frontera

$$\text{PVF} \begin{cases} \alpha \dfrac{\partial^2 T}{\partial x^2} = \dfrac{\partial T}{\partial t} \\ T(x,0) = (x) \quad 0 < x < x_F \\ T(0,t) = g_1(t) \\ \qquad\qquad\qquad\qquad t > 0 \\ T(x_F,t) = g_2(t) \end{cases}$$

Proporcionar las funciones CI(X), CF1(T) y CF2(T) y los

DATOS: El número NX de puntos de la malla en el eje x, el número NT de puntos de la malla en el eje t, la longitud total XF del eje x, el tiempo máximo TF por considerar y el coeficiente ALFA de la derivada de segundo orden.

RESULTADOS: Los valores de la variable dependiente T a lo largo del eje x a distintos tiempos t: T.

PASO 1. Realizar los pasos 1 a 10 del algoritmo 8.1.
PASO 2. Hacer I=1
PASO 3. Mientras I ≤ NX-2, repetir los pasos 4 a 7.
  PASO 4. Hacer A(I) = -LAMBDA
  PASO 5. Hacer B(I) = 1+2*LAMBDA
  PASO 6. Hacer C(I) = -LAMBDA
  PASO 7. Hacer I= I+1
PASO 8. Hacer J = 1
PASO 9. Mientras J ≤ NT, repetir los paso 10 a 24.
  PASO 10. Hacer T(1) = CF1(DT*J)
  PASO 11. Hacer T(NX) = CF2(DT*J)
  PASO 12. Hacer I = 1
  PASO 13. Mientras I ≤ NX-2, repetir los paso 14 a 15.

```
                PASO 14.  Hacer D(I) = T(I+1)
                PASO 15.  Hacer I = I+1
          PASO 16.  Hacer D(1) = D(1)+LAMBDA*T(1)
          PASO 17.  Hacer D(NX-2) = D(NX-2) + LAMBDA*T(NX)
          PASO 18.  Realizar los pasos 1 a 12 del algoritmo
                    3.5 con N = NX-2
          PASO 19.  Hacer I=1
          PASO 20.  Mientras I ≤ NX-2, repetir los pasos 21 y 22.
                    PASO 21.  Hacer T(I+1) = X(I)
                    PASO 22.  Hacer I=I+1
          PASO 23.  IMPRIMIR T
          PASO 24.  Hacer J=J+1
PASO 25.  TERMINAR.
```

SECCIÓN 8.4 CONVERGENCIA, ESTABILIDAD Y CONSISTENCIA

Convergencia

Hasta ahora no se ha analizado la importante pregunta de si los valores obtenidos aproximan "convenientemente" la solución del PVF en los nodos de la malla. En esta sección se contesta parcialmente ese punto.

El **error de discretización** se define en cada nodo como

$$e = T - U,$$

donde U es la solución verdadera del PVF y T la aproximación obtenida con el esquema explícito.

Se dice que un esquema de diferencias es convergente si al hacer $a=\Delta x \to 0$, $b=\Delta t \to 0$ en la malla, el error de discretización e también tiende a cero. Con estas definiciones presentes se demuestra a continuación que una condición suficiente para convergencia del método explícito en la solución de

$$\frac{\partial^2 T}{\partial x^2} = \frac{\partial T}{\partial t} \text{ (adimensionalizando las variables } \lambda = 1)$$

es que $0<(\Delta t/\Delta x^2)<0.5$. Se aceptará que no se cometen errores de redondeo, lo cual es prácticamente imposible, pero aun así, este criterio de convergencia es de uso práctico.

Al expander en serie de Taylor alrededor del nodo (i, j) la solución verdadera U (en la variable t solamente), se obtiene*

$$U_{i,j+1} = U_{i,j} + \Delta t U_t + \frac{(\Delta t^2)}{2!} U_{tt} + O[(\Delta t)^3] \qquad (8.32)$$

*U_t representa la primera derivada de U con respecto a t, U_{tt} la segunda derivada de U con respecto a t, etcétera.

Al expander U en la variable x adelante y atrás del nodo (i, j) se obtienen, respectivamente,

$$U_{i+1,j} = U_{i,j} + \Delta x U_x + \frac{\Delta x^2}{2!} U_{xx} + \frac{\Delta x^3}{3!} U_{xxx} + \frac{\Delta x^4}{4!} U_{xxxx} + \frac{\Delta x^5}{5!} U_{xxxxx} + O[(\Delta x)^6] \tag{8.33}$$

$$U_{i-1,j} = U_{i,j} - \Delta x U_x + \frac{\Delta x^2}{2!} U_{xx} - \frac{\Delta x^3}{3!} U_{xxx} + \frac{\Delta x^4}{4!} U_{xxxx} - \frac{\Delta x^5}{5!} U_{xxxxx} + O[(\Delta x)^6] \tag{8.34}$$

Nótese que en las ecuaciones 8.32 a 8.34 las derivadas se evalúan en el nodo (i, j), cuyas coordenadas son $x = i\Delta x$ y $t = j\Delta t$.

Con la suma de las ecuacioes 8.33 y 8.34 se obtiene

$$U_{i+1,j} + U_{i-1,j} = 2U_{i,j} + \Delta x^2 U_{xx} + \frac{\Delta x^4}{12} U_{xxxx} + O[(\Delta x)^6] \tag{8.35}$$

Al multiplicar por λ

$$\lambda [U_{i+1,j} + U_{i-1,j}] = 2\lambda U_{i,j} + \Delta x^2 \lambda U_{xx} + \frac{\Delta x^4}{12} \lambda U_{xxxx} + \lambda O[(\Delta x)^6] \tag{8.36}$$

Se despeja U_{xx} y se sustituye λ con $\Delta t/\Delta x^2$ en algunos términos y resulta

$$U_{xx} = \frac{\lambda}{\Delta t} [U_{i+1,j} + U_{i-1,j}] - \frac{2\lambda}{\Delta t} U_{i,j} - \frac{\Delta x^2}{12} U_{xxxx} + \frac{\lambda}{\Delta t} O[(\Delta x)^6] \tag{8.37}$$

U_t se despeja de la ecuación 8.32

$$U_t = \frac{1}{\Delta t} [U_{i,j+1} - U_{i,j} - \frac{\Delta t^2}{2!} U_{tt}] - \frac{O[(\Delta t)^3]}{\Delta t} \tag{8.38}$$

Al sustituir las ecuaciones 8.37 y 8.38 en la ecuación diferencial parcial, $U_t = U_{xx}$

$$U_{i,j+1} - U_{i,j} - \frac{\Delta t^2}{2!} U_{tt} - O[(\Delta t)^3] = \lambda U_{i+1,j} + \lambda U_{i-1,j} - 2\lambda U_{i,j} - \frac{\Delta x^2 \Delta t}{12} U_{xxxx} + \lambda O[(\Delta x)^6] \tag{8.39}$$

Se despeja $U_{i,j+1}$

$$U_{i,j+1} = \lambda U_{i-1,j} + (1 - 2\lambda) U_{i,j} + \lambda U_{i+1,j} - \frac{\Delta x^2 \Delta t}{12} U_{xxxx} + \frac{\Delta t^2}{2!} U_{tt} + O[(\Delta t)^3] + \lambda O[(\Delta x)^6] \tag{8.40}$$

Si se hace

$$\frac{Z_{i,j}}{\Delta t} = \frac{\Delta t}{2} U_{tt} - \frac{\Delta x^2}{12} U_{xxxx} + O[(\Delta t)^2] + O[(\Delta x)^4] \qquad (8.41)$$

y se sustituye la ecuación 8.41 en la 8.40

$$U_{i,j+1} = \lambda U_{i-1,j} + (1-2\lambda) U_{i,j} + \lambda U_{i+1,j} + Z_{i,j} \qquad (8.42)$$

Se resta del esquema explícito $T_{i,j+1} = \lambda T_{i-1,j} + (1-2\lambda) T_{i,j} + \lambda T_{i+1,j}$ miembro a miembro la ecuación 8.42

$$\begin{aligned} T_{i,j+1} - U_{i,j+1} = {} & \lambda (T_{i-1,j} - U_{i-1,j}) + (1-2\lambda) (T_{i,j} - U_{i,j}) \\ & + \lambda (T_{i+1,j} - U_{i+1,j}) - Z_{i,j} \end{aligned} \qquad (8.43)$$

Este desarrollo algebraico expresa el error de discretización $e_{i,j+1} = (T_{i,j+1} - U_{i,j+1})$ en función de los errores en los nodos vecinos $e_{i-1,j}$, $e_{i,j}$ y $e_{i+1,j}$ que se usan en el esquema explícito, o sea,

$$e_{i,j+1} = \lambda e_{i-1,j} + (1-2\lambda) e_{i,j} + \lambda e_{i+1,j} - Z_{i,j} \qquad (8.44)$$

Supóngase ahora que $0 < \lambda \leq 0.5$, con lo que los coeficientes λ y $(1-2\lambda)$ son no negativos. Si, por otro lado, se saca el valor absoluto en ambos miembros de la ecuación 8.44 y se aplica la desigualdad del triángulo, se obtiene

$$| e_{i,j+1} | \leq \lambda | e_{i-1,j} | + (1-2\lambda) | e_{i,j} | + \lambda | e_{i+1,j} | + | -Z_{i,j} | \qquad (8.45)$$

Si se llama $e_{máx}(k)$ con $0 \leq k \leq m$ la cota superior de $| e_{i,k} |$ con $1 \leq i \leq n-1$, se denota por $Z_{máx}(k)$ con k —igual que antes— la cota superior de $|-Z_{i,k}|$ con $1 \leq i \leq n-1$ y se substituye $e_{i-1,j}$, $e_{i,j}$, $e_{i+1,j}$, $e_{i,j+1}$ y $Z_{i,j}$ con sus respectivas cotas superiores $e_{máx}(j)$, $e_{máx}(j+1)$ y $Z_{máx}(j)$, simplificando se llega a

$$e_{máx}(j+1) \leq e_{máx}(j) + Z_{máx}(j) \qquad (8.46)$$

Si se analiza esta desigualdad en un periodo $0 \leq t \leq t_{máx} = t_m$ se tiene

$$\begin{aligned} e_{máx}(1) &\leq e_{máx}(0) + Z_{máx}(0) \\ e_{máx}(2) &\leq e_{máx}(1) + Z_{máx}(1) \\ e_{máx}(3) &\leq e_{máx}(2) + Z_{máx}(2) \\ &\vdots \\ e_{máx}(m) &\leq e_{máx}(m-1) + Z_{máx}(m-1) \end{aligned}$$

Al sustituir el término $e_{máx}(1)$ de la segunda desigualdad con el lado derecho de la primera, aquélla permanece e incluso se refuerza, con lo cual queda

$$e_{máx}(2) \leq Z_{máx}(0) + Z_{máx}(1) + e_{máx}(0)$$

Válgase la consideración de que los valores iniciales son exactos, $e_{máx}(0) = 0$.

Este resultado se sustituye por el término $e_{máx}(2)$ de la tercera desigualdad, con lo que

$$e_{máx}\ (3) \leq Z_{máx}\ (0) + Z_{máx}\ (1) + Z_{máx}\ (2)$$

Este procedimiento se repite hasta $e_{máx}(m)$; por tanto

$$e_{máx}\ (m) \leq m\ Z_{máx}\ (m-1)$$

Se tiene

$$e_{máx}\ (m) \leq t_{máx}\ [\ \frac{\Delta t}{2}\ U_{tt} - \frac{\Delta x^2}{12}\ U_{xxxx} + O(\Delta t)^2 + O(\Delta x)^4\ ]$$

recordando que $t_{máx} = m\ \Delta t$ y la ecuación 8.41. De esta desigualdad se deduce que $e_{máx}\ (m)$ tiende a cero si Δx y Δt tienden a cero; y ya que esta deducción se desarrolló para $O < \lambda \leq 0.5$, la conclusión sólo será válida para estos valores de λ; por ello, se constituye como una condición suficiente para convergencia —pero no necesaria— ya que ésta puede ocurrir por otras razones.

Estabilidad

El concepto de estabilidad se refiere a la propiedad de una ecuación de diferencias particular (base de un algoritmo), y significa que cuando $\Delta t \to 0$, el error introducido por cualquier motivo (condiciones iniciales, frontera, redondeo, etc.) está acotado. Lo anterior no significa que la desviación entre la solución verdadera de cierta ecuación diferencial parcial y su aproximación con una ecuación de diferencias sea pequeña, ya que esto está determinado por el concepto de **consistencia.**

Consistencia

Se dice que una ecuación de diferencias tiene consistencia cuando *solamente aproxima* la ecuación diferencial parcial que representa. Aunque esta propiedad parece cumplirse en todos los casos, no es así para algunos esquemas iterativos; por ejemplo, el algoritmo explícito de Dufort-Frankel no es consistente en ciertas circunstancias*.

SECCIÓN 8.5 MÉTODO DE CRANK-NICHOLSON

Además de los métodos vistos para resolver los PVF de las secciones 8.2 y 8.3, existen otros métodos de solución con diferencias. Entre éstos uno de los más importantes por su estabilidad incondicional y alto orden de convergencia** es el algoritmo de Crank-Nicholson.

---

*Dufort, E.C. y Frankel, S.P. "Stability Conditions in the Numerical Treatment of Parabolic Differential Equations", *Math Tables Aids Comput.*, 7, (1953) p 135-152.

**Isaacson, E. y Keller, H.B. *Analysis of Numerical Methods.* John Wiley and Sons, New York, 1966 p. 390, 392, 1082, 1088.

Este método consiste en combinar las aproximaciones de $\partial T/\partial t$ con diferencias hacia delante apoyándose en la fila j y la aproximación con diferencias hacia atrás apoyándose en la fila $j+1$, con lo que se obtiene un algoritmo implícito. Por ejemplo, al aproximar $\partial T/\partial t$ en el nodo (i, j) con diferencias hacia delante y de $\partial^2 T/\partial x^2$ con diferencias centrales (véase Fig. 8.13) se obtiene

$$\frac{T_{i,j+1} - T_{i,j}}{\Delta t} = \alpha \frac{T_{i-1,j} - 2T_{i,j} + T_{i+1,j}}{\Delta x^2} \tag{8.47}$$

Al aproximar $\partial T/\partial t$ en el nodo $(i, j+1)$ con diferencias hacia atrás y a $\partial^2 T/\partial x^2$ con diferencias centrales (véase Fig. 8.13) se llega a

$$\frac{T_{i,j+1} - T_{i,j}}{\Delta t} = \alpha \frac{T_{i-1,j+1} - 2T_{i,j+1} + T_{i+1,j+1}}{\Delta x^2} \tag{8.48}$$

Luego de sumar las ecuaciones 8.47 y 8.48 y rearreglar resulta

$$\boxed{T_{i,j+1} - T_{i,j} = \frac{\lambda}{2} [T_{i-1,j} - 2T_{i,j} + T_{i+1,j} + T_{i-1,j+1} - 2T_{i,j+1} + T_{i+1,j+1}]} \tag{8.49}$$

que es el algoritmo de Crank-Nicholson.

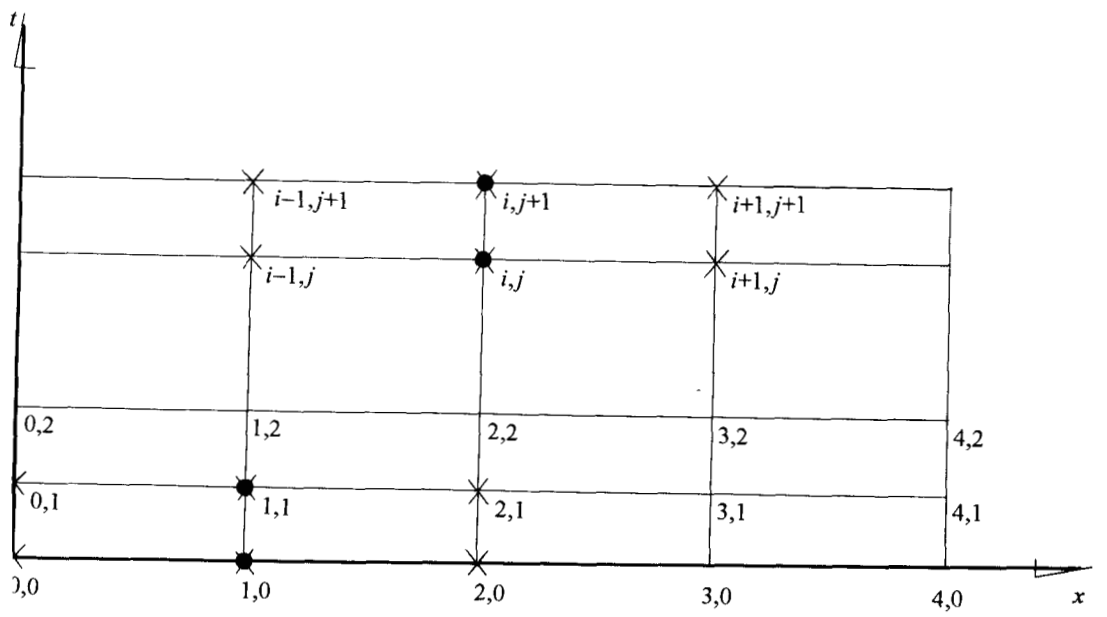


Figura 8.13 Nodos usados en el método de Crank-Nicholson.

Si este algoritmo (Ec. 8.49) se aplica a los nodos (1,0) y (1,1), o sea $i=1$, $j=0$ (véase Fig. 8.13), se tiene

$$T_{1,1} - T_{1,0} = \frac{\lambda}{2}\left[T_{0,0} - 2T_{1,0} + T_{2,0} + T_{0,1} - 2T_{1,1} + T_{2,1}\right] \qquad (8.50)$$

donde los nodos (0,0), (1,0), (2,0) y (0,1) son conocidos a partir de las condiciones inicial y de frontera; en cambio, los nodos (1,1) y (2,1) son incógnitas. Al rearreglar la ecuación 8.50 queda

$$(1+\lambda)T_{1,1} - \frac{\lambda}{2}T_{2,1} = (1-\lambda)T_{1,0} + \frac{\lambda}{2}\left[T_{0,0} + T_{0,1} + T_{2,0}\right] \qquad (8.51)$$

Al aplicar el mismo algoritmo (Ec. 8.49) a los nodos (2,0) y (2,1); es decir, $i=2$, $j=0$, se tiene

$$T_{2,1} - T_{2,0} = \frac{\lambda}{2}\left[T_{1,0} - 2T_{2,0} + T_{3,0} + T_{1,1} - 2T_{2,1} + T_{3,1}\right] \qquad (8.52)$$

donde las incógnitas son $T_{1,1}$, $T_{2,1}$ y $T_{3,1}$, ya que los demás nodos están dados por la condición inicial. Al rearreglar resulta

$$-\frac{\lambda}{2}T_{1,1} + (1+\lambda)T_{2,1} - \frac{\lambda}{2}T_{3,1} = \frac{\lambda}{2}T_{1,0} + (1-\lambda)T_{2,0} + \frac{\lambda}{2}T_{3,0} \qquad (8.53)$$

Análogamente, al aplicar la ecuación 8.49 a los nodos (3,0) y (3,1) es decir $i=3$, $j=0$, queda

$$T_{3,1} - T_{3,0} = \frac{\lambda}{2}\left[T_{2,0} - 2T_{3,0} + T_{4,0} + T_{2,1} - 2T_{3,1} + T_{4,1}\right] \qquad (8.54)$$

donde los nodos desconocidos son solamente (2,1) y (3,1), ya que los otros son conocidos por la condición inicial.

La ecuación 8.54 se rearregla y queda

$$-\frac{\lambda}{2}T_{2,1} + (1+\lambda)T_{3,1} = \frac{\lambda}{2}\left[T_{2,0} + T_{4,0} + T_{4,1}\right] + (1-\lambda)T_{3,0} \qquad (8.55)$$

Las ecuaciones 8.51, 8.53 y 8.55 forman un sistema cuya solución es la temperatura T en los nodos (1,1), (2,1) y (3,1); o sea,

$$\begin{aligned}
(1+\lambda)T_{1,1} - \frac{\lambda}{2}T_{2,1} &= (1-\lambda)T_{1,0} + \frac{\lambda}{2}\left[T_{0,0} + T_{0,1} + T_{2,0}\right] \\
-\frac{\lambda}{2}T_{1,1} + (1+\lambda)T_{2,1} - \frac{\lambda}{2}T_{3,1} &= \frac{\lambda}{2}\left[T_{1,0} + T_{3,0}\right] + (1-\lambda)T_{2,0} \\
-\frac{\lambda}{2}T_{2,1} + (1+\lambda)T_{3,1} &= \frac{\lambda}{2}\left[T_{2,0} + T_{4,0} + T_{4,1}\right] + (1-\lambda)T_{3,0}
\end{aligned} \qquad (8.56)$$

Una vez resuelto el sistema de ecuaciones 8.56 se puede seguir el mismo procedimiento, pero aplicado ahora en los nodos (1,1) (1,2); (2,1) (2,2) y (3,1), (3,2), con lo cual resulta

$$
\begin{aligned}
(1+\lambda)\,T_{1,2} - \frac{\lambda}{2}\,T_{2,2} &= (1-\lambda)\,T_{1,1} + \frac{\lambda}{2}\,[\,T_{0,1} + T_{0,2} + T_{2,1}\,] \\
-\frac{\lambda}{2}\,T_{1,2} + (1+\lambda)T_{2,2} - \frac{\lambda}{2}T_{2,2} - \frac{\lambda}{2}T_{3,2} &= \frac{\lambda}{2}\,[\,T_{1,1} + T_{3,1}\,] + (1-\lambda)\;T_{2,1} \\
-\frac{\lambda}{2}\,T_{2,2} + (1+\lambda)\,T_{3,2} &= \frac{\lambda}{2}\,[T_{2,1} + T_{4,1} + T_{4,2}\,] + (1-\lambda)\,T_{3,1}
\end{aligned}
$$

cuya solución proporciona las temperaturas de los nodos interiores de la segunda fila; o sea, $t=2\Delta t$. Este procedimiento se repite un número m de veces, hasta obtener las temperaturas en ciertos puntos de la barra a lo largo del tiempo, hasta un $t_{máx} = m\ \Delta t$.

Si en lugar de dividir la barra en cuatro subintervalos se dividiera en n subintervalos, se tendrían n-1 nodos interiores, a los que al aplicarse la ecuación 8.49 como en el caso anterior (cuatro subintervalos) se generaría un sistema de n-1 ecuaciones con n-1 incógnitas: $T_{1,1}$, $T_{2,1}$, $T_{3,1}$, ...,$T_{n-1,1}$ (para la primera fila); o sea

$$
\begin{aligned}
(1+\lambda)\,T_{1,1} - \frac{\lambda}{2}\,T_{2,1} &= (1-\lambda)\,T_{1,0} + \frac{\lambda}{2}\,[\,T_{0,0} + T_{0,1} + T_{2,0}\,] \\
-\frac{\lambda}{2}T_{1,1} + (1+\lambda)\,T_{2,1} - \frac{\lambda}{2}T_{3,1} &= \frac{\lambda}{2}[\,T_{1,0} + T_{3,0}\,] + (1-\lambda)\,T_{2,0} \\
&\;\;\vdots \\
-\frac{\lambda}{2}T_{n-3,1} + (1+\lambda)\,T_{n-2,1} - \frac{\lambda}{2}T_{n-1,1} &= \frac{\lambda}{2}[\,T_{n-3,0} + T_{n-1,0}\,] + (1-\lambda)\,T_{n-2,0} \\
-\frac{\lambda}{2}T_{n-2,1} + (1+\lambda)\,T_{n-1,1} &= \frac{\lambda}{2}[\,T_{n-2,0} + T_{n,0} + T_{n,1}\,] + (1-\lambda)\,T_{n-1,0}
\end{aligned}
\qquad (8.57)
$$

Este procedimiento se aplica en las filas j y $j+1$ para tener

$$
\begin{aligned}
(1+\lambda)\,T_{1,j+1} - \frac{\lambda}{2}\,T_{2,j+1} &= (1-\lambda)\,T_{1,j} + \frac{\lambda}{2}[\,T_{0,j} + T_{0,j+1} + T_{2,j}\,] \\
-\frac{\lambda}{2}T_{1,j+1} + (1+\lambda)\,T_{2,j+1} - \frac{\lambda}{2}T_{3,j+1} &= \frac{\lambda}{2}[\,T_{1,j} + T_{3,j}\,] + (1-\lambda)\,T_{2,j} \\
&\;\;\vdots \\
-\frac{\lambda}{2}T_{n-3,j+1} + (1+\lambda)\,T_{n-2,j+1} - \frac{\lambda}{2}T_{n-1,j+1} &= \frac{\lambda}{2}[\,T_{n-3,j} + T_{n-1,j}\,] + (1-\lambda)\,T_{n-2,j} \\
-\frac{\lambda}{2}T_{n-2,j+1} + (1+\lambda)\,T_{n-1,j+1} &= \frac{\lambda}{2}[T_{n-2,j} + T_{n,j} + T_{n,j+1}] + (1-\lambda)\,T_{n-1,j}
\end{aligned}
$$

que en notación matricial queda

$$A\ \mathbf{t}^{(j+1)} = B\,\mathbf{t}^{(j)} + \mathbf{c}$$

donde

$$A = \begin{bmatrix} (1+\lambda) & -\frac{\lambda}{2} & 0 & \cdots & & 0 \\ -\frac{\lambda}{2} & (1+\lambda) & -\frac{\lambda}{2} & & & \cdot \\ 0 & & & & & \cdot \\ & & & & & \cdot \\ \cdot & & & -\frac{\lambda}{2} & (1+\lambda) & -\frac{\lambda}{2} \\ \cdot & & & & & \\ \cdot & & & & & \\ 0 & \cdots & & 0 & -\frac{\lambda}{2} & (1+\lambda) \end{bmatrix}$$

$$\mathbf{t}^{(j+1)} = (T_{1,j+1} \quad T_{2,j+1} \quad T_{3,j+1} \quad \cdots \quad T_{n-1,j+1})^T$$

$$B = \begin{bmatrix} (1-\lambda) & -\frac{\lambda}{2} & 0 & \cdots & & 0 \\ -\frac{\lambda}{2} & (1-\lambda) & -\frac{\lambda}{2} & & & \cdot \\ 0 & & & & & \cdot \\ & & & & & \cdot \\ \cdot & & & -\frac{\lambda}{2} & (1-\lambda) & -\frac{\lambda}{2} \\ \cdot & & & & & \\ \cdot & & & & & \\ 0 & \cdots & & 0 & -\frac{\lambda}{2} & (1-\lambda) \end{bmatrix}$$

$$\mathbf{t}^{(j)} = [T_{1,j} \quad T_{2,j} \quad T_{3,j} \quad \cdots \quad T_{n-1,j}]^T$$

y

$$\mathbf{c} = [\tfrac{\lambda}{2}(T_{0,j} + T_{0,j+1}) \ 0 \ldots 0 \ \tfrac{\lambda}{2}(T_{n,j} + T_{n,j+1})]^T$$

Ejemplo 8.2

Resuelva el siguiente problema por el método de Crank-Nicholson

$$\text{PVF} \quad \begin{cases} \dfrac{\partial T}{\partial t} = \alpha \dfrac{\partial^2 T}{\partial x^2} \\ T(x,0) = 20°F \\ T(0,t) = 100°F \\ T(L,t) = 100°F \end{cases}$$

$$\alpha = 1 \text{ pie}^2 / \text{hr}$$
$$L = 1 \text{ pie}$$
$$t_{máx} = 1 \text{ hr}$$

SOLUCIÓN

Al dividir la longitud de la barra en cuatro subintervalos (véase Fig. 8.13), el primer sistema de ecuaciones por resolver, ya sustituidos los datos, es

$$\begin{bmatrix} 1.16 & -0.08 & 0 \\ -0.08 & 1.16 & -0.08 \\ 0 & -0.08 & 1.16 \end{bmatrix} \begin{bmatrix} T_{1,1} \\ T_{2,1} \\ T_{3,1} \end{bmatrix} = \begin{bmatrix} 31.2 \\ 20 \\ 31.2 \end{bmatrix}$$

Este sistema se resuelve por alguno de los métodos del capítulo 3 y se obtiene

$$T_{1,1} = 28.36 \qquad T_{2,1} = 21.15 \qquad T_{3,1} = 28.36$$

temperaturas que corresponden a un tiempo $t = 0.01$ horas.

Para calcular las temperaturas de la segunda línea se conserva la matriz coeficiente y sólo se varía el vector de términos independientes; o sea

$$\begin{bmatrix} 1.16 & -0.08 & 0 \\ -0.08 & 1.16 & -0.08 \\ 0 & -0.08 & 1.16 \end{bmatrix} \begin{bmatrix} T_{1,2} \\ T_{2,2} \\ T_{3,2} \end{bmatrix} = \begin{bmatrix} 41.5144 \\ 22.3036 \\ 41.5144 \end{bmatrix}$$

El sistema se resuelve para obtener

$$T_{1,2} = 37.47 \qquad T_{2,2} = 24.40 \qquad T_{3,2} = 37.47$$

temperaturas que corresponden a un tiempo $t = 0.02$ horas. Al continuar este procedimiento se obtienen los resultados de la tabla 8.4.*

t (h)	x pies				
	0.0	0.25	0.5	0.75	1.0
0.00	60	20.00	20.00	20.00	60
0.01	100	28.36	21.15	28.36	100
0.02	100	37.47	24.40	37.47	100
0.03	100	44.61	28.99	44.61	100
0.04	100	50.45	34.10	50.45	100
0.05	100	55.38	39.29	55.38	100
0.06	100	59.67	44.32	59.67	100
0.07	100	63.44	49.07	63.44	100
0.08	100	66.81	53.50	66.81	100
0.09	100	69.83	57.59	69.83	100
0.10	100	72.56	61.44	72.56	100
0.20	100	89.28	84.84	89.28	100
0.40	100	98.36	97.68	98.36	100
0.60	100	99.75	99.64	99.75	100
0.80	100	99.96	99.95	99.96	100
1.00	100	99.99	99.99	99.99	100

Tabla 8.4. Resultados de la solución del ejemplo 8.2. Se usó t=0.01 constante y sólo se muestran algunos de los resultados.

Los resultados obtenidos con el método de Crank- Nicholson son —en general— un promedio de los resultados de los métodos explícito e implícito; esto puede explicarse con base en que el método de Crank-Nicholson combina ambos.

*En el ejercicio 8.2 se presenta un programa que aplica el método de Crank-Nicholson.

Enseguida se presenta un algoritmo para este método.

ALGORITMO 8.3 Método de Crank-Nicholson

Para aproximar la solución al

$$\text{PVF} \begin{cases} \text{EDP} \quad \alpha \dfrac{\partial^2 T}{\partial x^2} = \dfrac{\partial T}{\partial t} \\ \text{CI} \quad T(x,0) = f(x) \quad 0 \le x \le x_F \\ \text{CF1} \quad T(0,t) = g_1(t) \\ \text{CF2} \quad T(x_F,t) = g_2(t) \end{cases} \quad t > 0$$

proporcionar las funciones CI(X), CF1(T) y CF2(T) y los

DATOS: El número NX de puntos de la malla en el eje x, el número NT de puntos de la malla en el eje t, la longitud total XF del eje x, el tiempo máximo TF por considerar y el coeficiente ALFA de la derivada de segundo orden.

RESULTADOS: Los valores de la variable dependiente T a lo largo del eje x a distintos tiempos t: T.

PASO 1. Realizar los pasos 1 a 10 del algoritmo 8.1.
PASO 2. Hacer I = 1
PASO 3. Mientras I ≤ NX-2, repetir los pasos 4 a 7.
  PASO 4. Hacer A(I) = LAMBDA
  PASO 5. Hacer B(I) = -2-2*LAMBDA
  PASO 6. Hacer C(I) = LAMBDA
  PASO 7. Hacer I=I+1
PASO 8. Realizar los pasos 8 a 24 del algoritmo 8.2 con los siguientes cambios: En el paso 15 hacerD(I)=
–LAMBDA*T(I)-(2-2*LAMBDA)*T(I+1)-LAMBDA*T(I+2)
En el paso 17 hacer D(1) = D(1) - LAMBDA*T(1)
En el paso 18 hacer D(NX-2) = D(NX-2)-LAMBDA*T(NX)
PASO 9. TERMINAR.

SECCIÓN 8.6 OTROS MÉTODOS PARA RESOLVER EL PROBLEMA DE CONDUCCIÓN DE CALOR EN UNA DIMENSIÓN

Método de Richardson

Este método usa diferencias divididas centrales para aproximar $\partial T/\partial t$ en la ecuación de conducción. De acuerdo con la malla de la figura 8.14 se tiene

$$\frac{T_{i,j+1} - T_{i,j-1}}{2\Delta t} = \alpha \frac{T_{i-1,j} - 2T_{i,j} + T_{i+1,j}}{\Delta x^2} \tag{8.58}$$

Obsérvese que si se conocen las dos primeras filas (la primera podría ser la condición inicial y la segunda se calcula por alguno de los métodos de las secciones anteriores), el método resulta explícito en el nodo $(i, j+1)$; o sea,

$$\boxed{T_{i,j+1} = 2\lambda [T_{i-1,j} - 2T_{i,j} + T_{i+1,j}] + T_{i,j-1}} \tag{8.59}$$

con lo que pueden calcularse la tercera, cuarta, etc., filas.

Método de Dufort-Frankel

Young y Gregory* demuestran que el método de Richardson es poco satisfactorio, ya que presenta considerables problemas de estabilidad; sin embargo, sustituyendo $T_{i,j}$ con $(T_{i,j+1} + T_{i,j-1})/2$ en la ecuación 8.58 se obtiene el método de Dufort-Frankel con mejores propiedades de estabilidad

$$\frac{T_{i,j+1} - T_{i,j-1}}{2\Delta t} = \alpha \frac{T_{i-1,j} - T_{i,j-1} - T_{i,j+1} + T_{i+1,j}}{\Delta x^2}$$

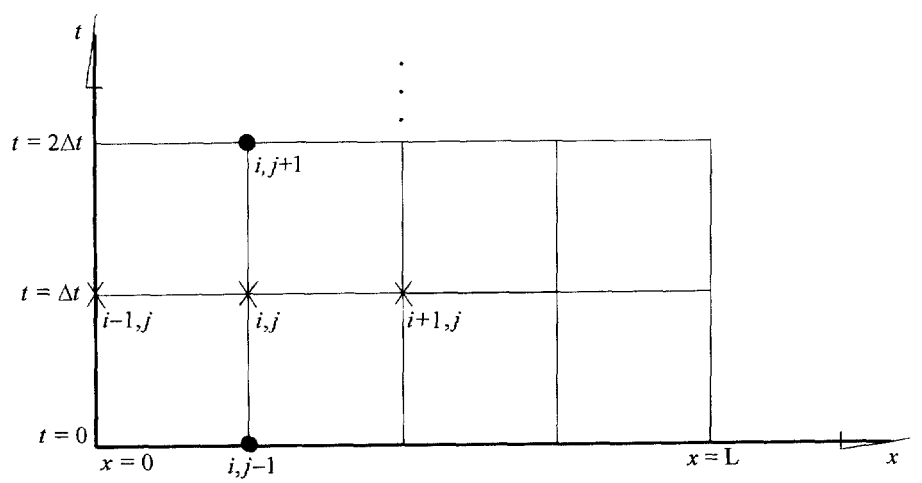


Figura 8.14 Nodos usados en el método de Richardson.

*Young, D.M. y Gregory, R.T. *A Survey of Numerical Mathematics*. Vol. II Addison Wesley (1973); p 1084-1086.

Como en el método de Richardson, si se conocen dos filas el algoritmo resulta explícito para el cálculo de las temperaturas de la siguiente fila; es decir,

$$T_{i,j+1} = \frac{2\lambda}{1+2\lambda}\left[T_{i-1,j} + T_{i+1,j}\right] + \left(\frac{1-2\lambda}{1+2\lambda}\right) T_{i,j-1} \qquad (8.60)$$

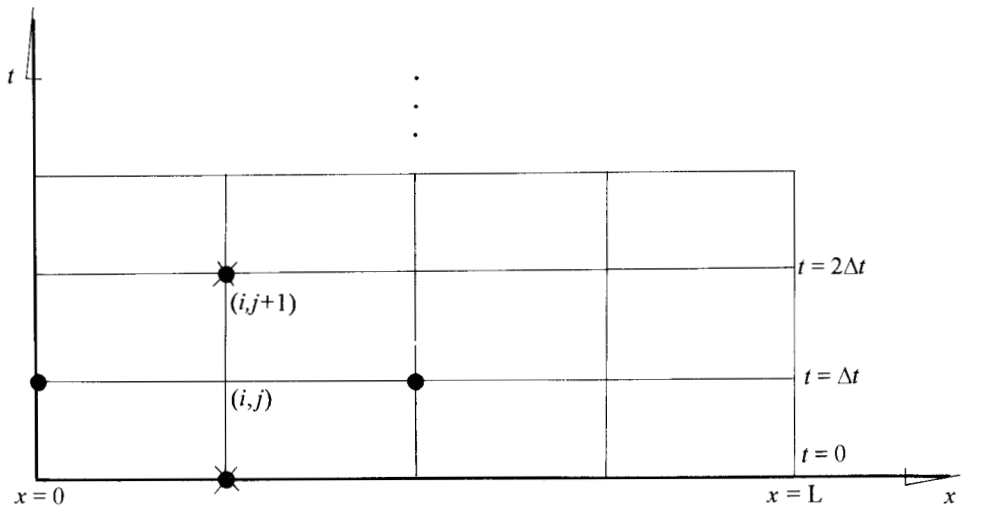


Figura 8.15 Nodos usados en el método de Dufort-Frankel.

Ejemplo 8.3

Mediante el método de Dufort-Frankel, resuelva el ejemplo 8.2 con los mismos valores para Δx, Δt y α.

SOLUCIÓN

La primera fila está dada por las condiciones iniciales y para la segunda fila (t=0.01) se tomarán los resultados obtenidos con el método implícito (véase la tabla 8.3).

Se aplica la ecuación 8.60 para conocer $T_{i,j+1} = T_{1,2}$ y se obtiene

$$T_{1,2} = \frac{2\lambda}{1+2\lambda}\left[T_{0,1} + T_{2,1}\right] + \left(\frac{1-2\lambda}{1+2\lambda}\right) T_{1,0}$$

Al sustituir los valores λ=0.16, $T_{0,1}$ =100, $T_{2,1}$=22.43 y $T_{1,0}$=20 se tiene

$$T_{1,2} = \frac{2(0.16)}{1+2(0.16)}\left[100 + 22.43\right] + \frac{1-2(0.16)}{1+2(0.16)}\left[20\right] = 39.98$$

Con el cálculo del siguiente punto $T_{i,j+1} = T_{2,2}$ queda

$$T_{2,2} = \frac{2\lambda}{1+2\lambda}\left[T_{1,1} + T_{3,1}\right] + \left(\frac{1-2\lambda}{1+2\lambda}\right) T_{2,0}$$

y al sustituir valores se obtiene $T_{2,2} = 24.84$

El algoritmo se aplica en la misma forma para las filas siguientes. Los resultados se presentan en la tabla 8.5.

t (hrs)	x (pies)				
	0.00	0.25	0.50	0.75	1.00
0.00	60	20.00	20.00	20.00	60
0.01	100	29.99	22.43	29.99	100
0.02	100	39.98	24.84	39.98	100
0.04	100	52.34	34.96	52.34	100
0.06	100	61.22	45.29	61.22	100
0.08	100	68.14	54.46	68.14	100
0.10	100	73.72	62.25	73.72	100
0.20	100	89.85	85.38	89.85	100
0.40	100	98.48	97.81	98.48	100
0.60	100	99.77	99.67	99.77	100
0.80	100	99.97	99.95	99.97	100
1.00	100	99.99	99.99	99.99	100

Tabla 8.5 Resultados del ejemplo 8.3.

SECCIÓN 8.7 TIPOS DE CONDICIONES FRONTERA EN PROCESOS FÍSICOS Y TRATAMIENTO DE CONDICIONES FRONTERA IRREGULARES

Dependiendo de las características del proceso físico modelado y de las circunstancias que rodean al proceso de estudio, se tendrán en general tres tipos de condiciones frontera en un PVF.

1. Condiciones de Dirichlet.

Se dan estas condiciones cuando la variable dependiente es conocida en todos los puntos frontera. Los ejemplos de las secciones anteriores tienen este tipo de condiciones frontera.

2. Condiciones de Neumann.

Cuando se conoce la derivada de la variable dependiente en los puntos frontera, se dice que se tienen las condicioes de Neumann. Por ejemplo, el problema de conducción de calor de la barra con condiciones de este tipo, quedaría formulado así

$$\text{PVF} \quad \begin{cases} \text{EDP} \quad \dfrac{\partial T}{\partial t} = \alpha \dfrac{\partial^2 T}{\partial x^2} \\ \text{CI} \quad T(x,0) = f(x) \quad 0 \le x \le < L \\ \text{CF1} \quad \dfrac{dT}{dx}\Big|_{x=0} = g_1(t) \\ \qquad\qquad\qquad\qquad t > 0 \\ \text{CF2} \quad \dfrac{dT}{dx}\Big|_{x=L} = g_2(t) \end{cases}$$

Estas condiciones pueden obtenerse físicamente, por ejemplo aislando térmicamente una frontera, ya que en este caso

$$\frac{dT}{dx}\Big|_{x=L} = 0,$$

es decir, no habría cambio de temperatura en la frontera. O bien si se tiene una frontera en contacto con un fluido (que puede ser aire), la ley de enfriamiento de Newton proporcionaría esta condición

$$\frac{dT}{dx}\Big|_{x=L} = h(T - T_0),$$

donde h es el coeficiente de transmisión de calor y T_0 la temperatura del fluido.

3. Condiciones combinadas

Esta condición aparece cuando se tiene una combinación de las dos anteriores. Nuevamente, el problema de conducción de calor en la barra quedaría formulado con este tipo de condiciones así

$$\text{PVF} \quad \begin{cases} \text{EDP} \quad \dfrac{\partial T}{\partial t} = \alpha \dfrac{\partial^2 T}{\partial x^2} \\ \text{CI} \quad T(x,0) = f(x) \quad 0 \le x \le < L \\ \text{CF1} \quad \dfrac{dT}{dx}\Big|_{x=0} = g_1(t) \\ \qquad\qquad\qquad\qquad t > 0 \\ \text{CF2} \quad T(L,t) = g_2(t) \end{cases}$$

En los ejercicios al final de capítulo, se resuelven problemas con condiciones de Neumann y combinadas.

Fronteras irregulares

Según la geometría del sistema, se pueden tener fronteras irregulares; esto es, casos como el de la figura 8.16.

Si se tienen, por ejemplo, las condicioes frontera de Dirichlet, los valores de la variable dependiente en C y D son conocidos; por tanto, la aproximación de la variable dependiente en el punto P puede hacerse con una interpolación. El caso más simple es una interpolación lineal entre los puntos A y C o entre B y D.

Para la interpolación entre los puntos A y C sería

$$\frac{T_C - T_A}{\Delta x + c\,\Delta x} = \frac{T_P - T_A}{\Delta x}, \text{ de donde } T_P = \frac{T_C - T_A}{1 + c} + T_A \qquad (8.61)$$

donde $0 < c < 1$.

Para la interpolación entre los puntos B y D

$$\frac{T_D - T_B}{\Delta y + d\,\Delta y} = \frac{T_P - T_B}{\Delta y}, \text{ de donde } T_P = \frac{T_D - T_B}{1 + d} + T_B \qquad (8.62)$$

Si se quisiera una aproximación mayor de T_P, cabe promediar los valores obtenidos por medio de las ecuaciones 8.61 y 8.62, o se cierra la malla (con lo que se aumentan los cálculos) y se usa alguna de las ecuaciones 8.61 u 8.62 o bien se toman T_C o T_D como aproximación de T_P, según la que esté más cerca.

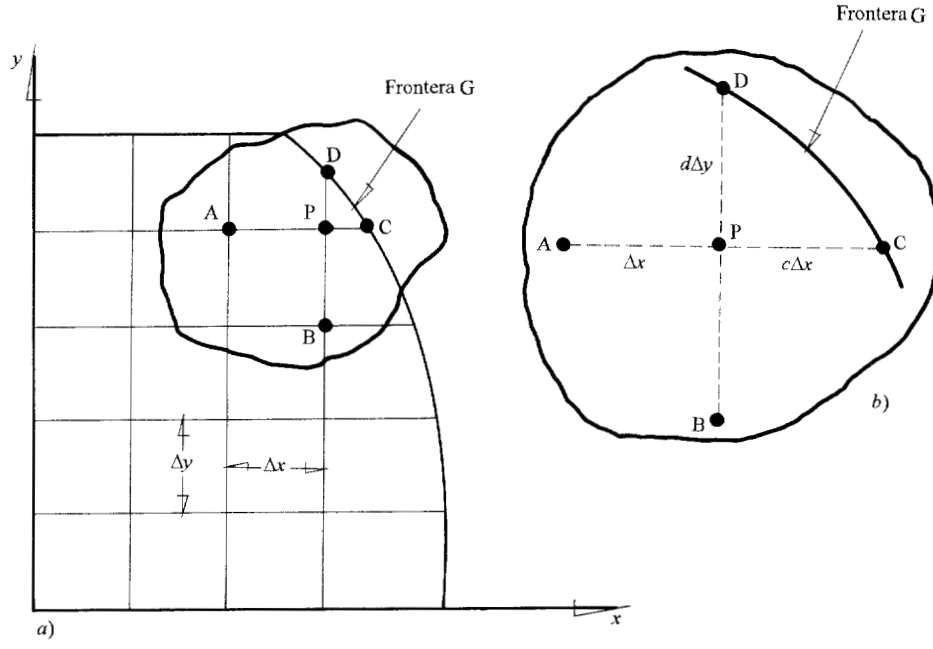


Figura 8.16 a) Malla sobre un dominio con frontera irregular. b) Ampliación de la región con puntos frontera D y C.

Cuando en los puntos de la frontera irregular G (veáse Fig. 8.17) se conoce

$$\frac{\partial T}{\partial N}\Big|_G$$

en vez de T, donde N es el vector normal a la frontera (condiciones de Neumann), el problema de estimar el valor de los puntos cercanos a la frontera se torna un poco más difícil. Supóngase que se tiene una rejilla como en la figura 8.17. Ya que se conoce

$$\frac{\partial T}{\partial N}\Big|_G ,$$

este valor se puede igualar según

$$\frac{\partial T}{\partial N}\Big|_G = \frac{T_P - T_F}{\overline{FP}} \text{ , de donde } T_P = \frac{\partial T}{\partial N}\Big|_G \overline{FP} + T_F \qquad (8.63)$$

Por construcción de la malla

$$\overline{FP} = \Delta x \,/\, \cos\theta \qquad (8.64)$$

y también

$$\frac{T_E - T_A}{\Delta y} = \frac{T_F - T_A}{\Delta y\, tg\theta} \text{ , de donde } T_F = (T_E - T_A)\, tg\theta + T_A \qquad (8.65)$$

Se sustituyen las ecuaciones 8.64 y 8.65 en la 8.63

$$T_P = \frac{\Delta x}{\cos\theta} \frac{\partial T}{\partial N}\Big|_G + (T_E - T_A)\, tg\theta + T_A$$

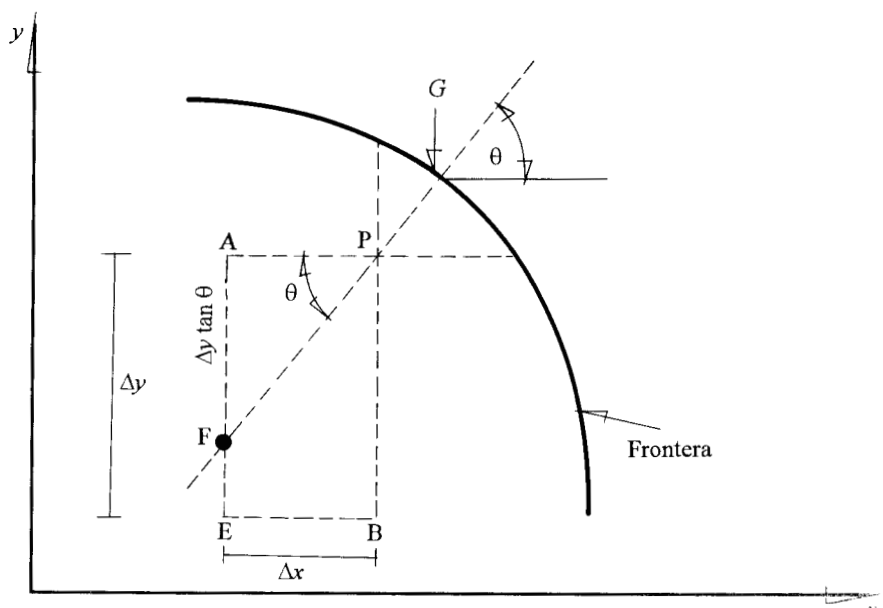


Figura 8.17.

En los problemas por resolver (al final del capítulo) se pide determinar T_P cuando el punto F cae entre los puntos E y B (véase Fig. 8.18).

Por último, si se tienen condiciones frontera combinadas, se aplica alguno de los tratamientos anteriores a cada punto frontera, según corresponda.

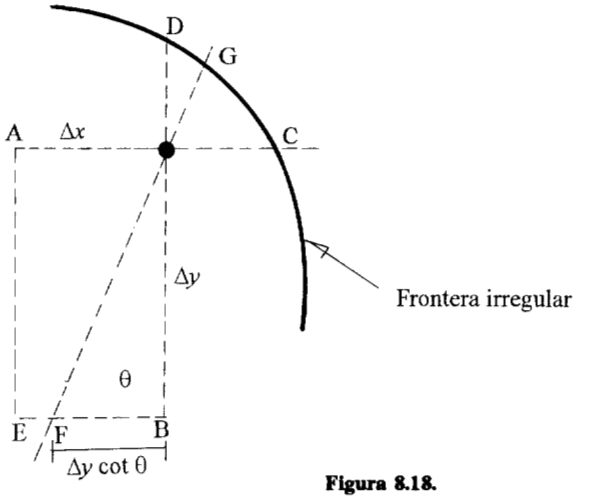


Figura 8.18.

Ejercicios

8.1 Confirme que las siguientes ecuaciones diferenciales parciales

$$a)\ \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0 \qquad \text{ecuación de Laplace}$$

$$b)\ \frac{\partial^2 U}{\partial x^2} = \frac{\partial^2 U}{\partial t^2} \qquad \text{ecuación de onda}$$

$$c)\ \frac{\partial^2 U}{\partial x^2} = \frac{\partial U}{\partial t} \qquad \text{ecuación de difusión}$$

son elíptica, hiperbólica y parabólica, respectivamente, en cualquier punto donde T y U estén definidas.

SOLUCIÓN

a) La función solución T es —en este caso— función de x y y solamente; esto es,

$$T = T(x, y)$$

Al identificar los coeficientes A, B y C de la ecuación de Laplace con los del modelo general (Ec. 8.1), se tiene

$$A = 1, \qquad B = 0, \qquad C = 1$$

Nótese que los tres coeficientes son constantes y, por tanto, independientes del punto (x, y) donde se desee establecer su clasificación. Así pues, en un punto cualquiera (x, y) donde T esté definida, se ve que

$$0^2 - 4\,(1)\,(1) < 0$$

por lo que la ecuación es **elíptica** en todo el dominio de definición de T.

b) De la misma manera que en (a), aunque ahora con U como función de x y t, se tiene

$$A = 1, \qquad B = 0, \qquad C = -1$$

y para un punto cualquiera (x, t) donde U esté definida

$$0^2 - 4(1)(-1) = 4 > 0$$

por lo que la ecuación de onda es **hiperbólica** en el punto (x, t) dado.

c) En este caso para un punto (x, t) donde U esté definida

$$A = 1, \qquad B = 0, \qquad C = 0$$

y

$$0^2 - 4(1)(0) = 0$$

por lo que la ecuación de difusión es **parabólica** en dicho punto (x, t).

8.2 Una loza de gel de agar contiene una concentración uniforme de urea de 2×10^{-4} gmol/cm^3; la loza tiene 3 cm de espesor (véase Fig. 8.19). Determine la concentración de urea en la parte central de la loza después de 2, 4, 6 y 8 horas de inmersión en agua (la urea es soluble en agua).

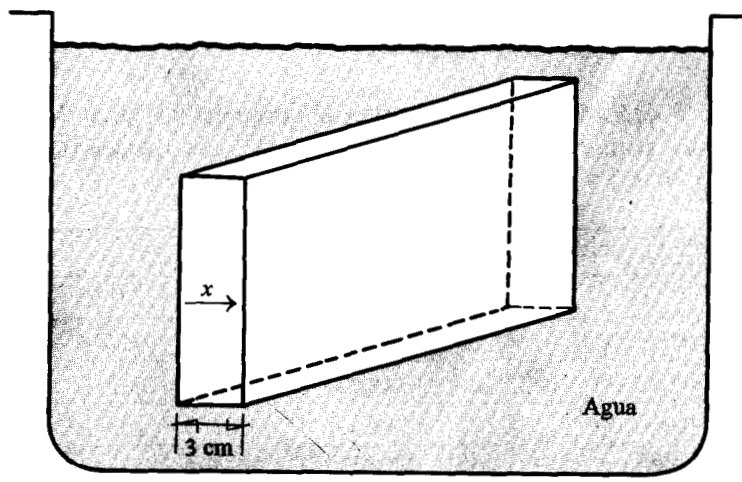


Figura 8.19. Loza de agar sumergida en agua.

SOLUCIÓN

El modelo matemático que permite establecer la concentración está dado por

$$\frac{\partial C}{\partial t} = \mathcal{B} \frac{\partial^2 C}{\partial x^2}$$

donde

C es la concentración de urea en la loza
t el tiempo
x la distancia
$\mathcal{B}$ la constante de difusividad (equivalente a α en el fenómeno de conducción de calor).

Por el problema se sabe que $C = 2\times10^{-4}$ gmol/cm^3, que es la condición inicial (concentración inicial de la urea en la loza).

Por otro lado se puede establecer que

$$\begin{aligned} C(0,t) &= 0 \\ & \qquad t > 0 \\ C(1,t) &= 0 \end{aligned}$$

lo cual físicamente significa que al sumergirse la loza en el agua, la urea de la superficie se disuelve de inmediato y la concentración de las caras (fronteras de la loza) es cero cualquier tiempo después.

El problema de valores en la frontera queda formulado

$$\text{PVF} \quad \begin{cases} \text{EDP} & \dfrac{\partial C}{\partial t} = \mathcal{B} \dfrac{\partial^2 C}{\partial x^2} \\ \text{CI} & C(x,0) = 2\times10^{-4} \quad 0 \le x < L \\ \text{CF1} & C(0,t) = 0 \\ & \qquad t > 0 \\ \text{CF2} & C(1,t) = 0 \end{cases}$$

Si se toma $\mathcal{B} = 1.7\times10^{-2}$ cm^2/h y se aplica el programa 8.3, se obtienen los resultados siguientes para $x = 1.5$ cm (el centro de la loza) transcurridas 2, 4, 6 y 8 horas, con $\Delta x = 0.3$ y $\Delta t = 0.01$.

t (h)	x (cm)					
	0.00	0.30	0.60	0.90	1.20	1.50
0.0	0.000100	0.000200	0.000200	0.000200	0.000200	0.000200
2.0	0.000000	0.000146	0.000191	0.000199	0.000200	0.000200
4.0	0.000000	0.000117	0.000176	0.000195	0.000199	0.000200
6.0	0.000000	0.000099	0.000162	0.000188	0.000197	0.000199
8.0	0.000000	0.000088	0.000149	0.000181	0.000194	0.000197

8.3 Calcule la distribución de temperatura $T(x, t)$ en una barra cilíndrica de vidrio y aislada térmicamente excepto en el plano A (véase Fig. 8.20). Inicialmente la barra está a 20°C y en el instante cero se ajusta con el plano A una placa cuya temperatura es de 100°C y permanece constante durante el tiempo de estudio (tres horas). La barra es lo suficientemente delgada como para despreciar la distribución de temperatura radial y se sabe que para el material vidrio $\alpha = 1.23 \times 10^{-3}$ m^2/h.

SOLUCIÓN

Este problema es semejante al del ejemplo resuelto al inicio de la sección 8.3 con la diferencia de que un extremo está aislado, lo que modifica la condición frontera correspondiente. Un aislamiento térmico "ideal" significa que no hay flujo de calor en dirección alguna y matemáticamente se expresa

$$\frac{\partial T}{\partial x}\Big|_{x=1} = T_x = 0$$

Por lo anterior, el problema de valor en la frontera con condiciones frontera combinadas queda formulado por

$$\text{PVF} \quad \begin{cases} \dfrac{\partial T}{\partial t} = \alpha \dfrac{\partial^2 T}{\partial x^2} \\ T(x,0) = 20\,°C \qquad 0 < x < 1 \\ T(0,t) = 100°C \\ \qquad\qquad\qquad t > 0 \\ T_x(1,t) = 0 \end{cases}$$

Con el empleo del método explícito y la selección de $\Delta x = 0.25$ y $\Delta t = 0.1$,

$$\lambda = \frac{\alpha \Delta t}{\Delta x^2} = 1.968 \times 10^{-3}$$

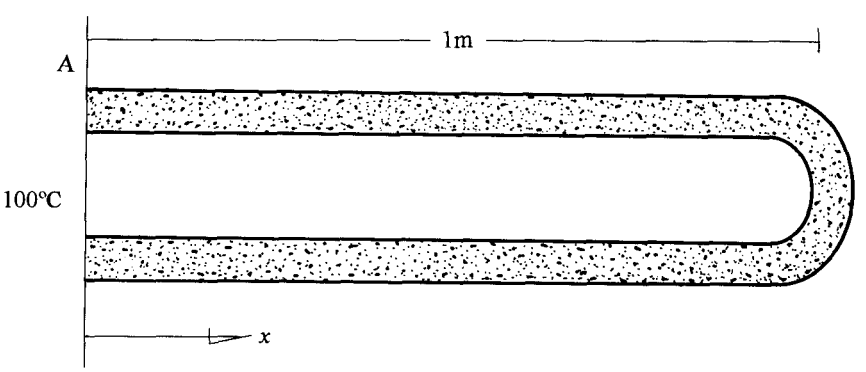


Figura 8.20. Barra cilíndrica de vidrio aislada.

la malla queda como se ilustra en la figura 8.21, y se tiene

Para $i = 1, j = 0$ en la ecuación 8.27

$$T_{1,1} = 0.001968(60) + (1\text{-}2(0.001968))20 + 0.001968(20) = 20.08$$

donde $T_{0,0}$ se aproxima con la media aritmética de los valores límites de $T(0, t)$ cuando $t \to 0$ y $T(x, 0)$ cuando $x \to 0$, que en este caso es la media de 100 y 20°C.

Al aplicar el mismo algoritmo al nodo (2,1) se tiene

$$T_{2,1} = 0.001968(20) + (1\text{-}2(0.001968))20 + 0.001968(20) = 20$$

de igual manera para el nodo (3,1) resulta

$$T_{3,1} = 0.001968(20) + (1\text{-}2(0.001968))20 + 0.001968(20) = 20$$

Nótese que la temperatura del nodo (4,0) es 20°C, ya que la condición inicial lo establece y esa frontera está aislada.

Para el cálculo del nodo (4,1) se usa la condición frontera $T_x = 0$ y su aproximación con diferencias hacia atrás como sigue

$$T_x(1, t) = 0 \approx \frac{T_{4,1} - T_{3,1}}{\Delta x}$$

por lo que $T_{4,1} \approx T_{3,1} \approx 20$ °C

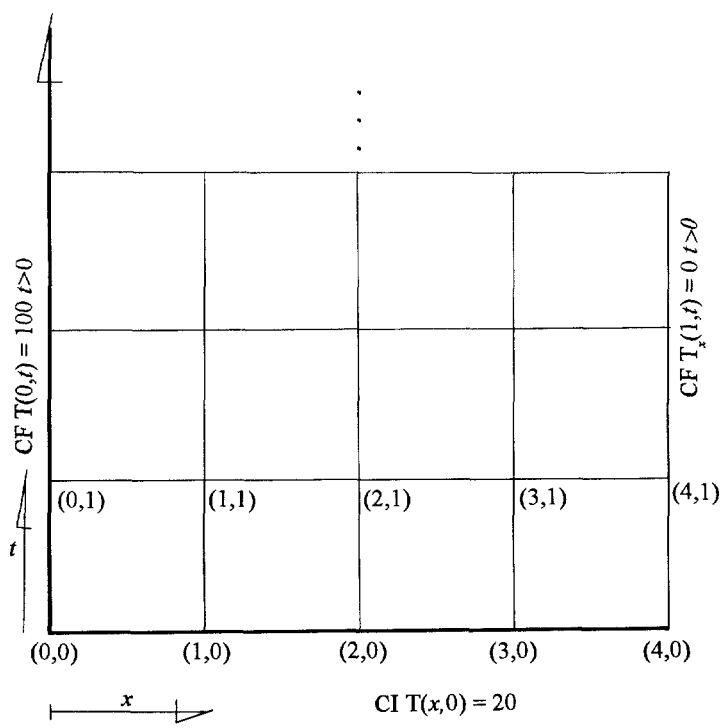


Figura 8.21.

Con este procedimiento se calculan las temperaturas de los nodos de las filas superiores; aquí debe notarse que por la condición frontera $T_x = 0$, la temperatura en el extremo aislado de la barra será aproximadamente igual a la temperatura de la barra en un nodo anterior ($x = 0.75$).

Los resultados de la tabla 8.5, obtenidos con el programa 8.1 muestran lo anterior.

t (en horas)	x (m)				
	0.00	0.25	0.50	0.75	1.00
0.0	60.000	20.000	20.000	20.000	20.000
0.1	100.000	20.079	20.000	20.000	20.000
0.2	100.000	20.236	20.000	20.000	20.000
0.4	100.000	20.548	20.001	20.000	20.000
0.6	100.000	20.858	20.004	20.000	20.000
0.8	100.000	21.166	20.007	20.000	20.000
1.0	100.000	21.471	20.012	20.000	20.000
1.5	100.000	22.223	20.029	20.000	20.000
2.0	100.000	22.961	20.053	20.001	20.001
2.5	100.000	23.685	20.084	20.001	20.001
3.0	100.000	24.395	20.121	20.002	20.002

Tabla 8.5

8.4 Encuentre la distribución de temperatura $T(x,t)$ en una aleta delgada de cobre (véase Fig. 8.22), unida por la cara sombreada a un radiador cuya temperatura constante es 200°F. La función de la aleta es disipar calor por convección a la atmósfera, cuya temperatura es de 68°F. Considere que la aleta está inicialmente a 68°F y que el coeficiente de transmisión de calor h es 30 BTU/(h pie^2 °F).

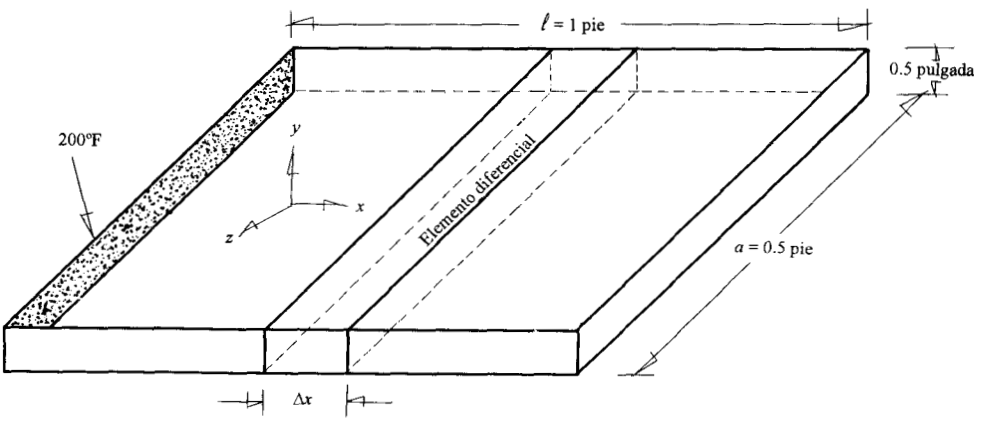


Figura 8.22.

SOLUCIÓN

Al efectuar un balance de calor en un elemento diferencial de la aleta, de dimensioes Δx, $l = 1$ pie y $a = 0.5$ pie se tiene, de acuerdo con la ley de continuidad (Ec. 8.3)

$$(A\,\Delta x\,\rho\,C_p)\,\frac{\partial T}{\partial t} = -k\,A\,\frac{\partial T}{\partial x}\Big|_x - (-k\,A\,\frac{\partial T}{\partial x}\Big|_{x+\Delta x}) - 2\,\Delta x\,(ah)\,(T-68)$$

donde el primer y segundo términos del lado derecho se refieren al calor que entra y que sale, respectivamente, del elemento diferencial por las caras perpendiculares al eje x y de área $A = 0.5(0.5)/12 = 0.020833$ pies^2. En cambio, el tercer término se refiere al calor que sale del elemento diferencial hacia la atmósfera; con el factor 2 de éste se incluyen las dos caras perpendiculares al eje y. Nótese que se ha despreciado el calor que sale por las caras perpendiculares al eje z, ya que la placa es muy delgada y $Q = 0$.

El lado izquierdo de la ecuación representa la acumulación de calor en el elemento diferencial considerado.

Toda la ecuación se divide entre $A\,\Delta x\,\rho\,C_P$ y después se hace que $\Delta x \to 0$, con lo cual

$$\frac{\partial T}{\partial t} = \frac{k}{\rho\,C_P}\,\frac{\partial^2 T}{\partial x^2} - \frac{2\,(ah)}{C_P\,A\rho}\,(T-68)$$

de tal manera que se obtiene el modelo matemático que rige el fenómeno descrito.

Si a este modelo se unen las condiciones

$$T(x,0) = 68°F,$$

que describen la temperatura en las fronteras de la aleta, se tiene un problema de valores en la frontera.

Las propiedades físicas del cobre requeridas para resolver la ecuación se enlistan enseguida.

$$k = 223 \text{ BTU/(h ft}^2\text{ °F/ft)}$$

$$C_P = 0.09 \text{ BTU/lb°F}$$

$$\rho = 560 \text{ lb/ft}^3$$

Para resolver este PVF se ha utilizado el método de Crank-Nicholson, para lo cual se ha modificado el programa 8.3 a fin de incluir el término

$$\frac{2\,ah}{A\rho\,C_P}\,(T-68)$$

El programa resultante (programa 8.4) utiliza $\Delta t = 0.001$ h y la longitud de la aleta (1 pie) se dividió en intervalos de 0.05 cada uno.

En la tabla siguiente se presentan algunos de los resultados obtenidos

t (en hora)	(pies)					
	0.00	0.20	0.40	0.60	0.80	1.00
0.000	134	68.00	68.00	68.00	68.00	68.00
0.001	200	71.28	68.05	68.00	68.00	68.00
0.002	200	81.30	68.42	68.01	68.00	68.00
0.004	200	102.00	71.56	68.20	68.01	68.00
0.006	200	113.86	77.05	68.99	68.07	68.00
0.008	200	121.43	82.53	70.52	68.28	68.00
0.010	200	126.66	87.31	72.48	68.71	68.00
0.015	200	134.51	96.16	77.62	70.51	68.00
0.020	200	138.76	101.85	81.95	72.57	68.00
0.040	200	144.85	111.07	90.42	77.43	68.00
0.060	200	146.16	113.17	92.50	78.71	68.00
0.080	200	146.46	113.66	93.00	79.02	68.00
0.100	200	146.53	113.78	93.11	79.09	68.00

Problemas

8.1 Clasifique las siguientes ecuaciones diferenciales parciales (consúltese el ejercicio 8.1).

a) $\operatorname{sen} x \dfrac{\partial^2 u}{\partial x^2} + y^2 \dfrac{\partial^2 u}{\partial y^2} = 0$

a1) en $0 < x < \pi$, $\quad -\infty < y < \infty$

a2) en $x = 0$, $\quad -\infty < y < \infty$

a3) en $\pi < x < 2\pi$; $\quad -\infty < y < \infty$

b) $y \dfrac{\partial^2 u}{\partial x^2} - x \dfrac{\partial^2 u}{\partial y^2} + \dfrac{\partial u}{\partial x} + y \dfrac{\partial u}{\partial y} = 0$

c) $y \dfrac{\partial^2 u}{\partial x^2} + 2e^{x+y} \dfrac{\partial^2 u}{\partial x \partial y} + e^{2y} \dfrac{\partial^2 u}{\partial y^2} = 0$

d) $\dfrac{\partial^2 u}{\partial x^2} + (1 + y^2) \dfrac{\partial^2 u}{\partial y^2} = 0$

e) $\operatorname{sen}^2 y \dfrac{\partial^2 u}{\partial x^2} - e^{2x} \dfrac{\partial^2 u}{\partial y^2} + 3 \dfrac{\partial u}{\partial x} - 5u = 0$

8.2 ¿En qué regiones la ecuación

$$\frac{\partial^2 u}{\partial x^2} + y \frac{\partial^2 u}{\partial y^2} = 0$$

es hiperbólica, elíptica y parabólica? (consúltese ejercicio 8.1).

8.3 Obtenga las ecuaciones (8.18) a (8.22) a partir de la expansión en serie de Taylor de $T(x, t)$, alrededor del punto (x_i, t_j), aplicando los mismos razonamientos que condujeron a las ecuaciones (8.12) y (8.14) a (8.16).

8.4 Exprese las siguientes ecuaciones diferenciales en términos de diferencias finitas

a) $\dfrac{d^2 y}{dx^2} - y \dfrac{dy}{dx} + 2y = 0$ con diferencias centrales

b) $\operatorname{sen} x \dfrac{\partial^2 u}{\partial x^2} + y^2 \dfrac{\partial^2 u}{\partial y^2} = 0$ con diferencias hacia adelante

c) $y \dfrac{\partial^2 u}{\partial x^2} - x \dfrac{\partial^2 u}{\partial y^2} + \dfrac{\partial u}{\partial x} + y \dfrac{\partial u}{\partial y} = 0$ con diferencias centrales

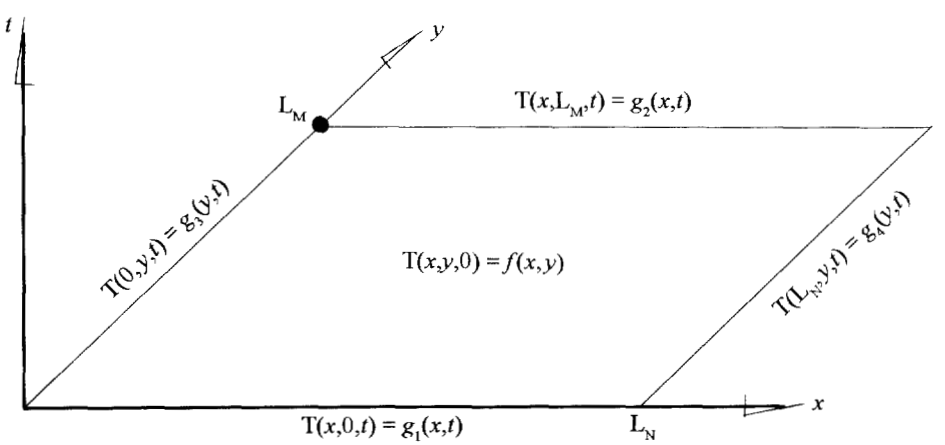
d) $\dfrac{\partial^2 u}{\partial x^2} + (1 + y^2) \dfrac{\partial^2 u}{\partial y^2} = 0$ con diferencias hacia atrás

8.5 La ecuación

$$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2} + \alpha \frac{\partial^2 T}{\partial y^2}$$

describe la conducción de calor en régimen transitorio en dos dimensiones. Exprésela en términos de diferencias finitas. El término de la izquierda en diferencias hacia adelante y los términos de la derecha en diferencias centrales.

8.6 La ecuación del problema 8.5 describe la conducción de calor en una lámina delgada (espesor despreciable) y permite calcular la temperatura en cualquier punto de la lámina a cualquier tiempo en régimen transitorio. Si las condiciones inicial y frontera son



establezca el problema de valor en la frontera, encuentre el algoritmo correspondiente al método explícito y resuelva con $\alpha = 0.01$ y las siguientes condiciones inicial y de frontera

$$\text{CI: } T(x,y,0) = 20\ ^\circ\text{C};\ \leq x \leq 0.1\ \text{m};\ 0 \leq y \leq 0.2\ \text{m}$$

$$\text{CF1: } T(x,0,t) = 100\ ^\circ\text{C};\ 0 \leq x \leq 0.1\ \text{m};\ 0 \leq t \leq 1\ \text{hora}$$

$$\text{CF2: } T(x,0.2,t) = 50\ ^\circ\text{C};\ 0 \leq x \leq 0.1\ \text{m};\ 0 \leq t \leq 1\ \text{hora}$$

$$\text{CF3: } T(0,y,t) = 100\ ^\circ\text{C};\ 0 \leq y \leq 0.2\ \text{m};\ 0 \leq t \leq 1\ \text{hora}$$

$$\text{CF4: } T(0.1,y,t) = 50\ ^\circ\text{C};\ 0 \leq y 0.2\ \text{m};\ 0 \leq t \leq 1\ \text{hora}$$

Nota: Elabore una malla tal que $0 < \lambda \leq 0.5$.

8.7 Resuelva el problema de valor en la frontera del problema 8.6 con el método implícito correspondiente.

8.8 Resuelva el PVF del problema 8.6 con el método de Crank-Nicholson correspondiente.

8.9 Se tiene una solución de urea contenida en un tubo de 1 cm de diámetro interior (véase Fig. 8.24), con un concentración inicial de 0.02 g/litro. Una membrana semipermeable conecta el tubo con un frasco que contiene una solución de urea con 2 g/litro. Otra membrana lo conecta con un reactivo con el cual la urea reacciona para desaparecer instantáneamente.

Si se considera que la difusión de la urea ocurre únicamente en el eje x, calcule la concentración de ésta a lo largo del tubo en los primeros 10 minutos. La difusividad de la urea es $\beta = 0.017\ \text{cm}^2/\text{h}$ (véase Ejer. 8.2).

8.10 Resuelva el problema 8.9 considerando que en el extremo derecho del tubo se tiene un frasco que contiene una solución con 1 g/l de urea en lugar del reactivo. Todas las demás condiciones permanecen.

8.11 Resuelva el siguiente PVF por los métodos explícito e implícito

$$\text{EDP: } \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2}$$

$$\text{CI: } T(x,0) = 20 \operatorname{sen} x \qquad \alpha = 1\ \text{pie}^2/\text{h}$$

$$\text{CF1: } T(0, t) = 100^\circ\text{C} \qquad L = 1\ \text{pie}$$

$$\text{CF2: } T(L, t) = 50^\circ\text{C} \qquad t_{\text{máx}} = 1\ \text{hora}$$

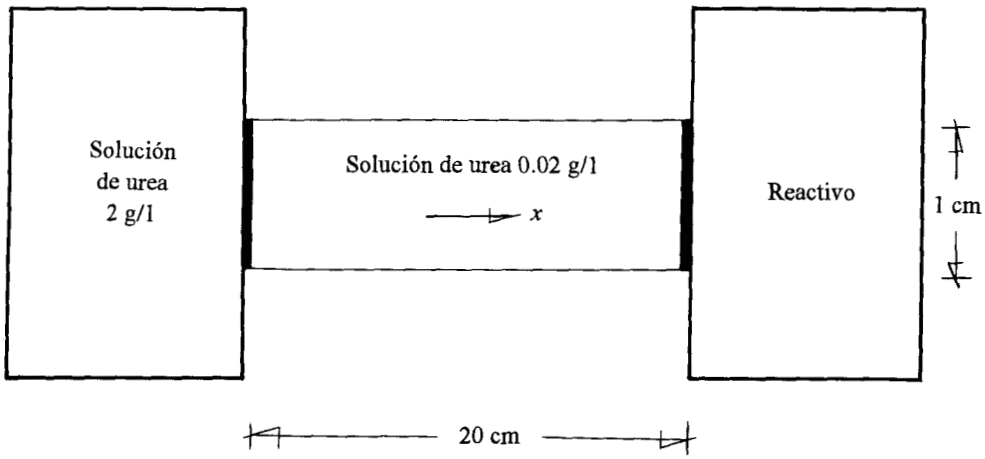


Figura 8.24. Difusión de urea en una solución.

8.12 Resuelva el siguiente PVF por el método de Crank-Nicholson

$$\text{EDP: } \frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2} \qquad \alpha = 1 \text{ pie/h}$$

$$\text{CI: } T(x,0) = 20°C \qquad L = 1 \text{ pie}$$

$$\text{CF1: } T(0, t) = 100°C \qquad 0 < t \leq 12 \text{ minutos}$$

$$T(0, t) = 20°C \qquad 12 < t \leq 60 \text{ minutos}$$

$$\text{CF2: } T(L, t) = 100°C \qquad 0 < t \leq 12 \text{ minutos}$$

$$T(L, t) = 20°C \qquad 12 < t \leq 60 \text{ minutos}$$

8.13 Resuelva el ejercicio 8.3 por el método de Crank Nicholson. Compare resultados.

8.14 Resuelva la EDP del ejercicio 8.4 con las siguientes condiciones

$$\text{CI: } T(x, 0) = (80 - 10x)°F$$

$$\text{CF1: } T(0, t) = 200°F$$

$$\text{CF2: } T(1, t) = 68°F$$

8.15 Si en el ejercicio 8.4 se modifica la geometría de la aleta para tenerla como se muestra en la figura 8.25, plantee y resuelva el PVF resultante.

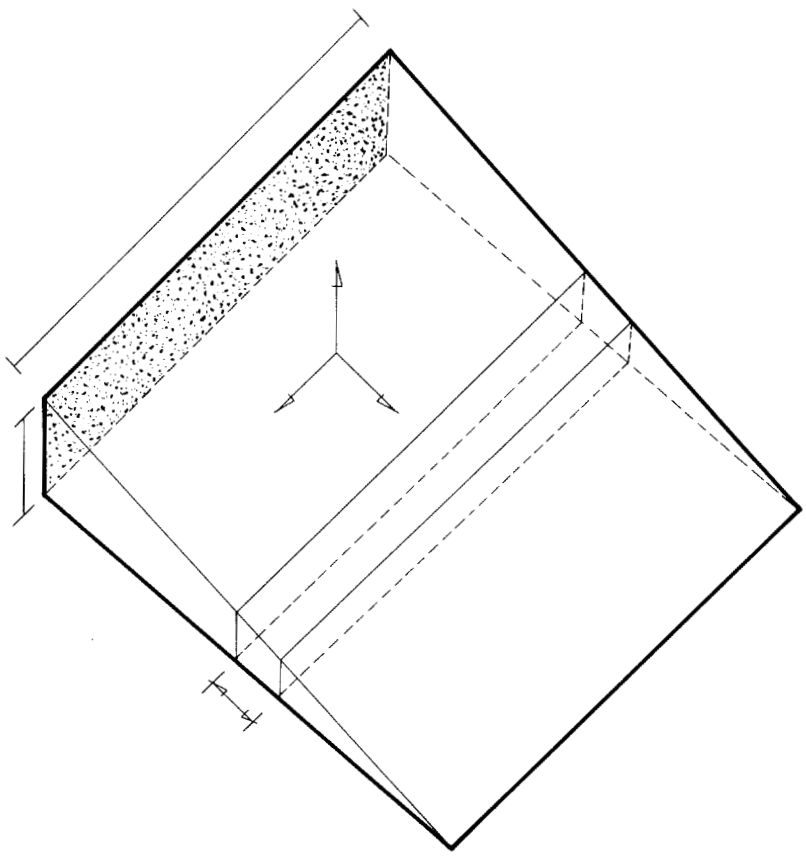


Figura 8.25. Conducción de calor en una aleta triangular.

RESPUESTAS A PROBLEMAS SELECCIONADOS

CAPÍTULO 1

1.2 *a)* $536_{10} = 1030_8 = 1000011000_2$

b) $923_{10} = 1633_8 = 1110011011_2$

c) $1536_8 = 3000_8 = 11000000000_2$

d) $8_{10} = 10_8 = 1000_2$

e) $2_{10} = 2_8 = 10_2$

f) $10_{10} = 12_8 = 1010_2$

g) $0_{10} = 0_8 = 0_2$

1.3 *a)* $777_8 = 111111111_2$

b) $573_8 = 101111011_2$

c) $7_8 = 111_2$

d) $2_8 = 10_2$

e) $10_8 = 1000_2$

f) $0_8 = 0_2$

1.5 *a)* $1000_2 = 8_{10}$

b) $10101_2 = 21_{10}$

c) $111111_2 = 63_{10}$

1.9 *a)* $985.34_{10} \approx 1731.256_8 \approx 1111011001.010101110_2$

b) $10.1_{10} \approx 12.063_8 \approx 1010.000110011_2$

c) $888.222_{10} \approx 1570.1615_8 \approx 1101111000.001100011010_2$

d) $3.57_{10} \approx 3.4436_8 \approx 11.100100011110_2$

e) $977.93_{10} \approx 1721.7341_8 \approx 1111010001.111011100001_2$

$f)$ $0.357_{10} \approx 0.2666_8 \approx 0.010110110110_2$

$g)$ $0.9389_{10} \approx 0.740557_8 \approx 0.1111000001011011111_2$

$h)$ $-0.9389_{10} \approx -0.740557_8 \approx -0.1111000001011011111_2$

1.13 $a)$ 0.19921875

$b)$ −160

$c)$ 9306112

1.14 $a)$ $0.10110100111000111011001 01 \times 2^{1010}$

$b)$ $-0.1111010100101111 \times 2^{100}$

$c)$ $0.11100100011001001 \times 2^{-100}$

$d)$ $0.1111101 \times 2^{1101}$

1.23 La mantisa normalizada más pequeña en binario es 0.10000000 (=1/2 en decimal), no 0.00000001 (2^{-8}) y la mayor es 0.11111111 (≈ 1).
Por esto, los números de máquina positivos deben quedar en el intervalo cerrado [s, L], donde

s = número de máquina positivo más pequeño = $(0.10000000) \times 2^{-64}$

$= 2^{-65} \approx 0.2710505 \times 10^{-19}$

y

L = número de máquina positivo mayor = $+(0.11111111) \times 2^{63}$

$\approx 0.91873437 \times 10^{17}$

El intervalo [s, L] puede dividirse en 128 subintervalos

$[s, 2s), \quad [2s, 2^2s), \quad [2^2s, 2^3s), \quad \ldots, \quad [2^{126}s, 2^{127}s), \quad [2^{127}s, L]$

$E = -64 \quad E = -63 \quad E = -62 \quad \ldots \quad E = +62 \quad E = +63$

donde E es la característica.

Nótese que cada subintervalo es dos veces mas grande que su predecesor. Para cada E hay 2^8 posibles mantisas normalizadas. Por tanto, una computadora con una palabra de 16 bits puede almacenar un total de

$128 \times 2^8 = 32768$ números positivos de máquina en el intervalo [s, L]

1.26 x = -278; y = 248.67

CAPÍTULO 2

2..1 a) $g'(x) = -\dfrac{2}{(x+1)^3}$; $\begin{matrix} x > 0.26 \\ x < -2.26 \end{matrix}$

b) $g'(x) = \dfrac{2(x-1)^{-2/3}}{3(x+1)^{4/3}}$; $x > 1.125$

c) $g'(x) = \cos x$; $x = n\pi$; $n = 0,1,2,\ldots$

d) $g(x) = x + \dfrac{\ln x}{2\tan x} - \dfrac{1}{2}$

$$g'(x) = 1 + \frac{(1/x)\tan x - \ln x \sec^2 x}{2\tan^2 x} \; ; \; x = 3.8 \; ; \; x = 4.2$$

e) $g'(x) = \dfrac{1 + 3x^2}{4\sqrt{6 - x - x^3}}$; $x = 0.8$; $x = 1.2$

f) $g(x) = \dfrac{\sec x}{2}$; $g'(x) = \dfrac{\sec x \tan x}{2}$; $x = 0$; $x = 0.5$

2.3 (del problema 2.1)

a) $\bar{x} \approx 0.46557$ *b*) $\bar{x} \approx 4.87035$ *c*) $\bar{x} = 0.$

d) $\bar{x} \approx 4.09546$ *e*) $\bar{x} \approx 1.$ *f*) $\bar{x} \approx 0.61003.$

2.6 *a*) n multiplicaciones y n sumas.

b) $2n$ multiplicaciones y n sumas.

2.8 *a*) $\bar{x} \approx 3.14619$ *b*) $\bar{x} \approx 0.85261$

c) $\bar{x} \approx 1.02987$ *d*) $\bar{x} \approx 0.20164$

2.9 *a*) $\bar{x} \approx 0.82626$ $\bar{y} \approx 1.12817$

b) $\bar{x} \approx 0.74798$ $\bar{y} \approx 1.11894$

c) $\bar{x} \approx 1.31555$ $\bar{y} \approx 0.32104$ $\bar{z} \approx 1.1362$

d) $\bar{x} \approx 3.82878$ $\bar{y} \approx 0.86662$

2.17 *a*) $\bar{x} \approx 10$ *b*) $\bar{x} \approx 0.66624$

c) $\bar{x} \approx 1.82938$ *d*) $\bar{x} \approx 1.2032$

2.20 Si $X_I = 2$ y $X_D = 4$, $n \approx 11$

2.26 *a*) $\bar{x} \approx 0.25753$ *b*) $\bar{x} \approx 3.83910$ *c*) $\bar{x} \approx -0.56574$

2.28 $\bar{x}_{1,2} = \pm 2\,i$

2.30 $\bar{x}_{1,2} \approx -1.6844 \pm 3.43133\,i$

2.33 *a*) $\bar{x}_1 \approx 1.1$; $\bar{x}_2 \approx 1{,}1$; $\bar{x}_{3,4} \approx 3 \pm 4\,i$

b) $\bar{x}_1 \approx 1.24144$; $\bar{x}_2 \approx 10.01798$

$\bar{x}_3 \approx 2.96396$; $x_4 \approx 0.97661$

c) $\bar{x}_1 = 1$; $\bar{x}_2 = 2$; $\bar{x}_3 = 3$; $\bar{x}_{4,5} = 2 \pm i$

d) $\bar{x}_1 \approx 1.7$; $\bar{x}_{2,3} = 1 \pm i$; $\bar{x}_{4,5} = \pm \sqrt{2}\,i$

2.36 $\bar{V}_{He} \approx 0.62542$ l; $\bar{V}_{H2} \approx 0.62785$ l; $\bar{V}_{02} \approx 0.6106$ l

2.41 $T \approx 105.33$ °C

2.43 $T \approx 102.3$ °C

2.47 $t \approx 3.041$ hrs.

2.48 $f \approx 0.04878$

2.50 $\lambda \approx 0.101$

CAPÍTULO 3

3.15 *a*) $\mathbf{e}_1 = [1, -2, 5, 7, 8, 0.3]^T$

$\mathbf{e}_2 = [-3.0343, 3.0686, 1.8284, -4.2402, 3.7255, -0.3103]^T$

$\mathbf{e}_3 = [-1.0029, 5.8915, 0.4998, 3.4940, -1.8232, 1.3820]^T$

$\mathbf{e}_4 = [5.7399, -3.2717, 0.1600, -0.7681, 2.8280, 38.8973]^T$

$\mathbf{e}_5 = [4.8912, 2.1153, -0.6869, -1.0594, 1.3045, -0.8202]^T$

b) $\mathbf{e}_1 = [4, 2, 1]^T$

$\mathbf{e}_2 = [-0.42857, 0.28571, 1.14286]^T$

$\mathbf{e}_3 = [-1.21212, 3.0303, -1.21212]^T$

c) $\mathbf{e}_1 = [10, \quad -20, \quad 5]^T$

$\mathbf{e}_2 = [1.66667, 1.66667, 3.33333]^T$

$\mathbf{e}_3 = [-1.07143, -0.35714, 0.71429]^T$

d) $\mathbf{e}_1 = [-1, \quad 1, \quad 0, \quad 2]^T$

$\mathbf{e}_2 = [4.33333, \quad 7.66667, \quad 1., \quad -1.66667]^T$

$\mathbf{e}_3 = [1.5, \quad -0.5, \quad -1, \quad 1]^T$

$\mathbf{e}_4 = [0.27322, -0.20036, 0.74681, 0.23679]^T$

3.16 3, 3, 3 y 4, respectivamente.

3.17 Número de reacciones independientes = 8

3.20 $w = 2$ y $w = 3$ para solución única.

$w = 1$ para número infinito de soluciones.

3.26 *a*) $\mathbf{x} = [-0.14114, \quad 1.56229, \quad -1.09371, \quad 0.30210]^T$

b) $\mathbf{x} = [4, \quad 3, \quad 1]^T$

c) $\mathbf{x} = [-8, \quad 31, \quad -20, \quad -34, \quad -11.9]^T$

3.33 $C_{A1} = 0.4507$; $C_{A2} = 0.33803$; $C_{A3} = 0.25352$

3.35 $\mathbf{x} = [0.04, 9.6\text{ E-}4, 2.304\text{ E-}5, 5.5264\text{ E-}7, 1.29525\text{ E-}8]^T$

3.36 $4\ Fe(CrO_2)_2 + 7\ O_2 + 8\ Na_2CO_3 \rightarrow 2\ Fe_2O_3 + 8\ Na_2CrO_4 + 8\ CO_2$

3.49 $\mathbf{p} = [0.69118, -9.9278, 6.41471, -5.32941, 5., -1.35379]^T$

3.50 $\mathbf{x} = [0.89052, \quad 0.99421, \quad 1.07371]^T$

3.54 *a*) $\mathbf{x} = [2, \quad 5.33333, \quad 1.66666]^T$ *b*) $\mathbf{x} = [1, 1, 1]^T$

d) $\mathbf{x} = [2.2872, 6.0449, -0.2532, 4.5294\text{ E-}3, -1.4221\text{ E-}5]^T$

e) $x_1 = -0.75404$; $x_2 = 1.05687$

$x_3 = 1.19697$; $x_4 = -0.14944$

$x_5 = 0.84542$; $x_6 = 0.17786$

$x_7 = -0.80314$; $x_8 = 0.20438$

$x_9 = -0.24447$; $x_{10} = -7.21371$ E-3

$x_{11} = 0.64289$

3.60 $\lambda_1 = 2.46056$; $\lambda_2 = 8.43988$

$\lambda_{3,4} = 4.55457 \pm 0.62948\, i$

3.61 $[1, \quad 2.97988, \quad 1.64534, \quad -0.96639]^T$

3.62 $\lambda_{\text{dominante}} = 3$; $\mathbf{e} = [1, \quad 1, \quad 0]^T$

CAPÍTULO 4

4.1 $\mathbf{x} = [3, \quad 2, \quad 1, \quad 2, \quad 4, \quad 6]^T$

4.2 $\mathbf{x} = [2, \quad 4, \quad 1, \quad 1]^T$

4.3 $x^0 = 0.8$; $y^0 = 0.5$; $\bar{x} \approx 0.7718$; $\bar{y} \approx 0.4197$

4.4 $g_1(x,y) = \sqrt{37 - y}$, $g_2(x,y) = \sqrt{x - 5}$

$x > 5$ y $6 < y < 37$

4.5 *a*) $\bar{x} = 6$; $\bar{y} = 1$

b) $\bar{x} \approx 6.17107$; $\bar{y} \approx -1.08216$

4.6 *a*) $[0, 0]^T$; $[8000, \quad 4000]^T$

b) $\mathbf{x} \approx [0.529164, \quad 0.399996, \quad 0.100006]^T$

c) $\mathbf{x} = [1, \quad 1, \quad 1]^T$

d) $\mathbf{x} \approx [0.14966, \quad 0.000599, \quad 0.42364]^T$

4.9 *a*) $\mathbf{x} = [0, \quad 0.1, \quad 1]^T$

b) $\mathbf{x} \approx [-0.110949, \quad 0.4110824]^T$

4.11 $\mathbf{x} \approx [6.95, \quad 2.5, \quad -0.15]^T$

4.12 $\bar{C}_{A1} \approx 0.53292$; $\bar{C}_{A2} \approx 0.42435$; $\bar{C}_{A3} \approx 0.32879$

4.15 $\mathbf{x} \approx [0.61089, \quad 0.37899, \quad 0.24919, \quad 0.15622, \quad 0.07728]^T$

4.17 $\mathbf{x} \approx [1.4695, \quad 0., \quad -0.22777]^T$

4.18 $T = 57.85488935\, e^{(-0.10941272\, t)} + 33.4992038$

4.24 $t_{opt} = -1.15$

4.26 *a*) $\bar{x} \approx -2.13147$; $\bar{y} \approx 0.97941$; $\bar{z} \approx -1.36122$

b) No tiene solución.

c) $\mathbf{x} \approx [4.35734, \quad 1.66657, \quad -3.46610]^T$

4.31 *a*) $z_{min} = -3$ en $x = -7.85396$, $y = -1.33128$ E-6

b) $z_{min} = 0$ en $x_1 = 0$, $x_2 = 0$, $x_3 = 0$

CAPÍTULO 5

5.1 *a*) 1.1303 *b*) 1.2597 *c*) 1.2034 *d*) 16

5.2 205.82

5.4 $x\,(2) = 5.8$

5.8 $J_0\,(0.8) = 0.8463$

5.14 $p = 2.59$

5.15 $v = 67.8$

5.18 $a_0 = 99600 \quad a_1 = -1209.166667$
$a_2 = 5.375 \quad a_3 = -0.00833333$

5.19 $C_B\,(0.82) = 1.12$

5.23 $R_2\,(10) \approx 98$

5.28 $f\,(1;3,0.13) = 0.295$, con un polinomio de segundo grado.

5.29 $r = 10.12223 + 0.027975\, T$

5.30 $P = 481.03743\, v^{-1.06533}$

5.31 $a = 0.24033$

5.32 $z = 7.993487 \times 10^{10}$; $E = 19999.73634$

5.35 $n = 3$

5.36 $\tau = 0.92893$

5.37 $a = 1.78752$ $b = 0.0006533$ $c = 1.84624 \times 10^{-5}$

5.39 $a_0 = 161.33646$ $a_1 = 32.96875$ $a_2 = -0.0855$

CAPÍTULO 6

6.5 *a*) 20.9 kg/min. *b*) 30097 kg. *c*) 0.38153 *d*) 114.863 kg.

6.7 1.64711

6.8 81792338.66 con Simpson 1/3 y N = 100

6.12 *a*) 1.71125 *b*) 0.56343 *c*) 0.40546 *d*) 1.29584

6.13 *a*) $I_1^{(3)} = 1.26613$ *b*) $I_0^{(2)} = 0.07921$

c) K = 1: $I_1^{(2)} = 1.04417$; K = 2: $I_1^{(2)} = 0.87105$

6.15 *a*) $I_0^{(3)} = 0.006303$ *b*) $I_0^{(3)} = 0.946083$

6.17 0.84338 o bien 84.338 %

6.18 3.70387 con Simpson 1/3

6.19 analíticamente = $\text{sen}^{-1}(1) = 1.570796327$

6.20 *a*) 0 *b*) 0.25 *c*) 1/3

6.23 50403593.58 con N = 2, 50021079.17 con N = 3

6.24 1.21484

6.26 2.61198 con N = 2, 2.61945 con N = 3

6.27 *a*) –0.57722 *c*) 6 *d*) 1/3

6.28 *a*) 0.02 *b*) 0.22532 *c*) 0.08428

6.30 con N = 10 y M = 10 *a*) 1.47627 *b*) 1.47623 *c*)0.35593

con N = 2 y M = 2 *d*) 0.25

6.32 *a*) 6.93463 con N =10 y M =10 *b*) 0.33424 con N =20 y M =20

c) 0.83333 con N =2 y M =2 *d*) 4.38911 con N =10 y M =10

6.35 $\bar{x}$ = 0.53802 $\bar{y}$ = 0.52466

6.36 *a*) 14951.02 con N =20 y M = 20 *b*) 0.11267 con N =10 y M =10

6.39 – 0.014121 con P = 2, 8 y 15

6.40 $f''(3.7)$ = 0.02503 con N = 2

6.45 v = 96.62 m/s

6.46 E(0.95) = 122.831 E(0.96) = 77.704 E(0.97) = 2.54

E(0.98) = 2.88 E(0.99) = 3.04 E(1.00) = 3.10

6.47 0.6133 para t =10 0.6 para t =35 0.3466 para t =60

6.48 *a*) – 0.0006725 *b*) 0.01218 *c*) – 0.00038 *d*) 0.008833

CAPÍTULO 7

7.1 tiempo ≈ 432 s.

7.2 gasto ≈ 0.049 m^3 /s

7.3 tiempo ≈ 8000 s con h = 25 s y Euler simple

7.4 gasto = 7.73 l/s tiempo ≈ 32 horas

7.6

C_{A0}	0.5	1.0	1.5	1.0	2.0
C_{B0}	1.0	1.5	2.0	1.0	0.5
t (min)	7	6	5	18.5	3

7.7

h	0.1	0.5	1.0	0.05
I_1 (3)	2.047	1.992	5.280	2.052
I_2 (3)	0.658	0.590	2.736	0.665

7.11 *a*) 8.8 s *b*) 220 m *c*) 95.56 m

7.12 tiempo ≈ 30 min

7.13 C_1 = 49.00 C_2 = 44.57 C_3 = 41.84

7.14 C_1 = 46.31 C_2 = 44.58 C_3 = 42.38 con h =1 min y RK-4

7.15 $(C_{A1}, C_{A2}) =$ *a*) (0.67,0.62) *b*) (0.68,0.86) *c*) (0.73,0.61)
d) (0.813,0.62) *e*) (0.655,0.63)

7.16

t	200	400	600	800	1000
C_A	5.0000	4.2347	3.0407	1.7833	1.6097
T	300.0	310.6	322.3	333.2	333.8

7.17 $C_A = 4.33$, $T = 307.31$ a $t = 1200$ para $T_J = 310$
$C_A = 0.448$, $T = 346.38$ a $t = 1200$ para $T_J = 350$
$C_A = 0.531$, $T = 344.28$ a $t = 1200$ para $T_J = 340$

Cuando la temperatura T alcanza un valor mayor que T_J, se lleva a cabo violentamente la reacción, por lo que se acostumbra enfriar el reactor cuanto T está muy cerca de T_J o la rebasa.

7.18

a)	x	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
	y	.004	.014	.029	.048	.069	.093	.118	.143	.169	.195
b)	x	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
	y	.004	.016	.034	.057	.083	.111	.141	.172	.203	.234

7.19

x	0 0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
y	0.006	.023	.050	.084	.123	.168	.216	.266	.317	.369

7.21 T (5 días) = 66.82 con h = 12 horas y RK-4

7.22 57 g, usando RK-4 como inicializador y h = 1 día

7.23 ciclo = 26 unidades de tiempo

7.24 $T_2 = 106.7$ con $h = 0.25$

7.25 *a*) $y(1) = 1$ *b*) $y(2) = 3.38629$
c) $y(2) = 10.04277$ *d*) $y(.5) = 2$

7.28 *a*) $y(2) = 0.13534$ *b*) $y(2.5) = 5.25193$ *c*) $y(1) = 0.87628$
d) $y(-1) = 1.35914$ *e*) $y(1.5) = 8.33311$

731 $y(1) = 0.36788$ $z(1) = -0.36788$

7.32 $y(1) = 0.19876$

7.36 con RK-4 *a*) y (1) = – 0.35 $z(1) = 2.58$

b) y (2) = 1.97 $z(2) = 1.62$
c) y (3) = 34.04 $u(3) = 37.37$ $v(3) = 40.74$

7.37 $N_A = 0.02383$ $N_B = 0.12319$ $N_C = 0.85298$
con h = 10 min y RK-4

CAPÍTULO 8

8.1 *a*1) Elíptica *a*2) parabólica *a*3) hiperbólica

b) Elíptica en $x < 0$ o en $y < 0$

Parabólica en $x = 0$ o en $y = 0$

Hiperbólica en $x > 0, y > 0$ o en $x < 0, y < 0$

c) Parabólica en el plano x-y

d) Elíptica en el plano x-y

e) Hiperbólica en el plano x-y

8.2 Elíptica en $y > 0$; $-\infty < x < \infty$
Parabólica en $y = 0$; $-\infty < x < \infty$
Hiperbólica en $y < 0$; $-\infty < x < \infty$

8.5
$$\frac{T_{i,j,k+1} - T_{i,j,k}}{\Delta t} = \alpha \frac{T_{i+i,j,k} - 2T_{i,j,k} + T_{i-1,j,k}}{\Delta x^2}$$

$$+ \alpha \frac{T_{i,j+1,k} - 2T_{i,j,k} + T_{i,j-1,k}}{\Delta y^2}$$

8.6 El algoritmo matemático para resolver este PVF por el método explícito es

$$T_{i,j,k+1} = T_{i,j,k} + \alpha \frac{\Delta t}{\Delta x^2} [T_{i+1,j,k} - 2T_{i,j,k} + T_{i-1,j,k}]$$

$$+ \alpha \frac{\Delta t}{\Delta y^2} [T_{i,j+1,k} - 2T_{i,j,k} + T_{i,j-1,k}]$$

Con $\Delta x = 0.02$, $\Delta y = 0.02$, $\Delta t = 0.01$, se anotan algunos resultados

Tiempo = 0.00

y						
0.20	56.67	35.00	35.00	35.00	35.00	40.00
0.16	60.00	20.00	20.00	20.00	20.00	35.00
0.12	60.00	20.00	20.00	20.00	20.00	35.00
0.08	60.00	20.00	20.00	20.00	20.00	35.00
0.04	60.00	20.00	20.00	20.00	20.00	35.00
0.00	73.33	60.00	60.00	60.00	60.00	56.67
	0.00	0.02	0.04	0.06	0.08	0.10

x

Tiempo = 0.01

y						
0.20	75.00	50.00	50.00	50.00	50.00	50.00
0.16	100.00	30.00	20.00	20.00	23.75	50.00
0.12	100.00	30.00	20.00	20.00	23.75	50.00
0.08	100.00	30.00	20.00	20.00	23.75	50.00
0.04	100.00	30.00	20.00	20.00	23.75	50.00
0.00	100.00	100.00	100.00	100.00	100.00	75.00
	0.00	0.02	0.04	0.06	0.08	0.10 x

Tiempo = 0.10

y						
0.20	75.00	50.00	50.00	50.00	50.00	50.00
0.16	100.00	74.25	57.06	48.73	47.43	50.00
0.12	100.00	75.02	56.16	46.59	45.73	50.00
0.08	100.00	76.44	58.38	48.81	47.15	50.00
0.04	100.00	83.99	71.21	62.88	57.16	50.00
0.00	100.00	100.00	100.00	100.00	100.00	75.00
	0.00	0.02	0.04	0.06	0.08	0.10 x

Tiempo = 0.50

y						
0.20	75.00	50.00	50.00	50.00	50.00	50.00
0.16	100.00	82.89	70.91	62.44	55.83	50.00
0.12	100.00	88.61	77.96	68.16	58.95	50.00
0.08	100.00	90.88	81.56	71.76	61.22	50.00
0.04	100.00	94.06	87.39	78.92	67.00	50.00
0.00	100.00	100.00	100.00	100.00	100.00	75.00
	0.00	0.02	0.04	0.06	0.08	0.10 x

Tiempo = 1.00

y						
0.20	75.00	50.00	50.00	50.00	50.00	50.00
0.16	100.00	82.94	71.00	62.52	55.88	50.00
0.12	100.00	88.69	78.10	68.30	59.03	50.00
0.08	100.00	90.97	81.70	71.90	61.30	50.00
0.04	100.00	94.12	87.48	79.00	67.06	50.00
0.00	100.00	100.00	100.00	100.00	100.00	75.00
	0.00	0.02	0.04	0.06	0.08	0.10 x

8.9 Con $\Delta x = 2$ cm, $\Delta t = 0.5$ min, $\lambda = 0.1275$ y el método de Crank-Nicholson, se anotan algunos resultados

	x (cm)					
t (min)	0.0	4.0	8.0	12.0	16.0	20.0
0.0	1.0100	0.0200	0.0200	0.0200	0.0200	0.01
1.0	2.0000	0.0655	0.0203	0.0200	0.0195	0.00
5.0	2.0000	0.4481	0.0574	0.0213	0.0157	0.00
10.0	2.0000	0.7631	0.1815	0.0397	0.0143	0.00

8.10 Con $\Delta x = 2$ cm, $\Delta t = 0.5$ min, $\lambda = 0.1275$ y el método de Crank-Nicholson, se anotan algunos resultados.

	x (cm)					
t (min)	0.0	4.0	8.0	12.0	16.0	20.0
0.0	1.0100	0.0200	0.0200	0.0200	0.0200	0.51
1.0	2.0000	0.0655	0.0203	0.0202	0.0425	1.00
5.0	2.0000	0.4481	0.0582	0.0402	0.2319	1.00
10.0	2.0000	0.7640	0.1923	0.1214	0.3896	1.00

8.11 Con el uso del método implícito con $\Delta x = 0.25$, $\Delta t = 0.01$ y $\lambda = 0.01$, se anotan algunos resultados

	x (pies)				
t (hrs)	0.0	0.25	0.50	0.75	1.00
0.0	100.00	4.95	9.59	13.63	50.00
0.1	100.00	63.47	42.86	40.67	50.00
0.5	100.00	86.87	74.10	61.87	50.00
1.0	100.00	87.49	75.00	62.49	50.00

8.12 Con $\Delta x = 0.1$, $\Delta t = 0.24$, $\lambda = 0.4166666$, se anotan algunos resultados para $x < 0.5$, ya que la distribución de temperaturas es simétrica.

	x (pies)					
t (min)	0.0	0.1	0.2	0.3	0.4	0.5
0	60.00	20.00	20.00	20.00	20.00	20.00
2	100.00	75.07	54.02	39.09	30.62	27.93
6	100.00	88.02	77.22	68.65	63.16	61.27
8	100.00	91.36	83.57	77.38	73.41	72.05
12	100.00	95.50	91.44	88.23	86.16	85.45
14	20.00	44.44	64.04	76.62	82.93	84.74
20	20.00	27.96	35.15	40.85	44.51	45.77
40	20.00	20.30	20.58	20.80	20.94	20.99
60	20.00	20.01	20.02	20.03	20.04	20.04

8.14 Con $\Delta x = 0.05$, $\Delta t = 0.001$, $\lambda = 1.76984127$, $\beta = 28.57188572$, se anotan algunos resultados

	x (pies)					
t (hr)	0.0	0.2	0.4	0.6	0.8	1.0
0.000	140.00	78.00	76.00	74.00	72.00	69.00
0.001	200.00	80.71	75.82	73.83	71.84	68.00
0.006	200.00	118.48	83.00	73.83	70.80	68.00
0.020	200.00	140.07	103.90	83.94	73.76	68.00
0.100	200.00	146.54	113.79	93.12	79.09	68.00

8.15

$$\frac{\partial T}{\partial t} = \frac{k}{\rho Cp}\frac{\partial^2 T}{\partial x^2} - \frac{k}{\rho Cp\,(1-x)}\frac{\partial T}{\partial x} - \frac{2\sqrt{1.0625}\;a\;h}{0.25\,(1-x)\,\rho Cp}\,(T-68)$$

C.I. $T(x,0) = 68\ ^\circ F$

C.F.1 $T(0, t) = 200\ ^\circ F$

C.F.2 $T(1, t) = 68\ ^\circ F$

ÍNDICE ANALÍTICO

Esta obra se terminó de imprimir en diciembre de 1999
en los talleres de Impresos Naucalpan, S.A. de C.V.
San Andrés Atoto No. 12, Col. San Esteban
C.P. 53550, Naucalpan, Edo. de México