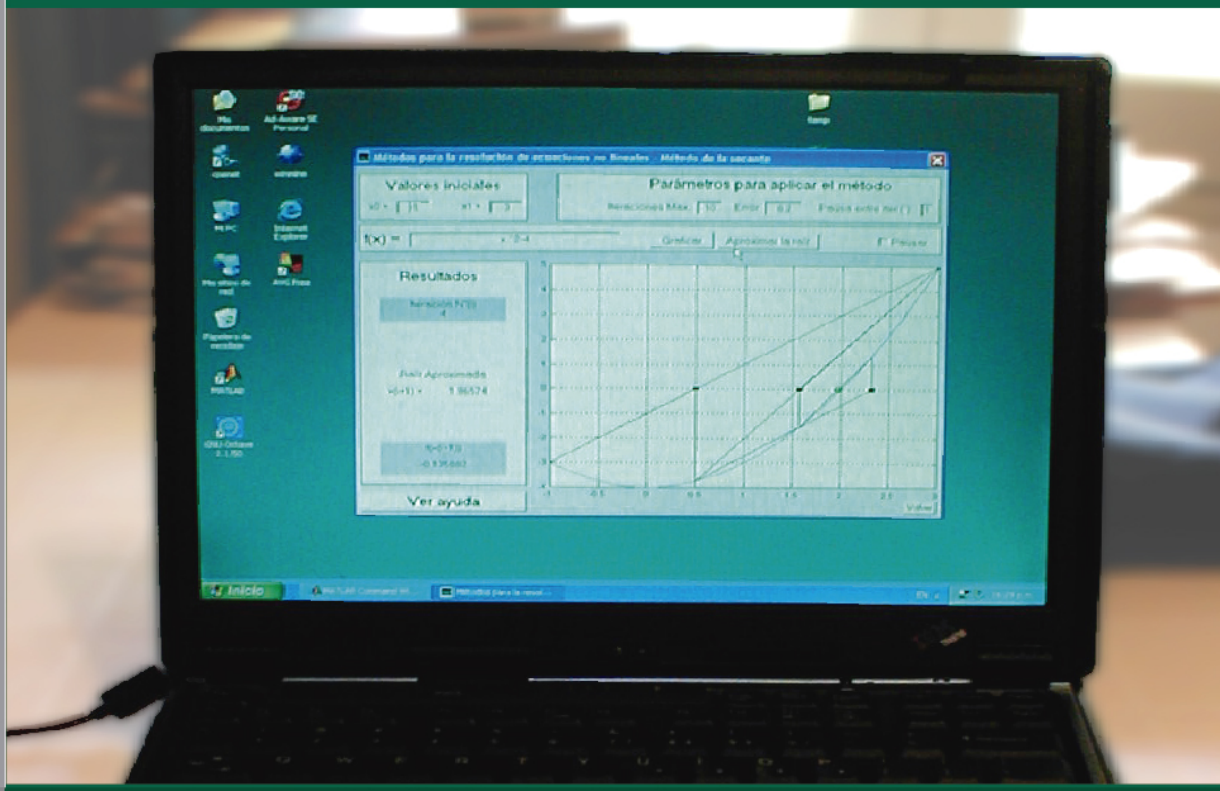
LIBRO DE TEXTO PARA **ESTUDIANTES UNIVERSITARIOS**

Cálculo Numérico



María E. ASCHERI
Rubén A. PIZARRO

EdUNLPam

2007

[2007] LIBRO DE TEXTO PARA ESTUDIANTES UNIVERSITARIOS

Cálculo Numérico

María Eva ASCHERI
Rubén Adrián PIZZARRO

Pizarro, Rubén
Cálculo numérico / Rubén Pizarro y María Eva Ascheri - 1a ed. - Santa Rosa : Univ. Nacional de La Pampa, 2008.
380 p. ; 18x25 cm.

ISBN 978-950-863-100-8

1. Ciencias Económicas. I. Aschieri, María Eva II. Título
CDD 330

Fecha de catalogación: 26/02/2008

LIBRO DE TEXTO PARA ESTUDIANTES UNIVERSITARIOS

Cálculo Numérico

María Eva Ascheri y Rubén Adrián Pizzaro.

Marzo de 2008, Santa Rosa, La Pampa

Coordinación de Diseño y Diagramación: Gabriela HERNÁNDEZ (DCV-EdUNLPam).

Impreso en Argentina
ISBN: 978-950-863-100-8
Cumplido con lo que marca la ley 11.723
EdUNLPam - Año 2008
Cnel. Gil 353 PB - CP L6300DUG
SANTA ROSA - La Pampa - Argentina

Índice

Prólogo

Este libro surge a partir de apuntes que servían de guías para dictar nuestras clases de Cálculo Numérico a los estudiantes de segundo y tercer año de Matemática, Física e Ingeniería de la Facultad de Ciencias Exactas y Naturales de la Universidad Nacional de La Pampa. Estos apuntes se fueron consolidando desde su origen hasta dar lugar a este texto. Proporciona una introducción a temas del Análisis Numérico adecuada para aquellos estudiantes que estén familiarizados, fundamentalmente, con el Álgebra Lineal y el Análisis Matemático, y que hayan realizado un curso de Programación estructurada. Su contenido puede ser desarrollado en un curso cuatrimestral. Se utiliza para sentar las bases de cursos avanzados sobre ciertos temas de Análisis Numérico, como el Álgebra Lineal Numérica y la Teoría de Aproximación.

A lo largo de todo este material presentamos diversos métodos numéricos que permiten encontrar soluciones aproximadas a problemas complejos, utilizando sólo las operaciones más simples de aritmética. Los métodos que estudiamos permiten simplificar los procedimientos matemáticos de manera que podamos ayudarnos con una calculadora o una computadora. Entre los métodos que desarrollamos se encuentran los referidos a: Solución de Ecuaciones no Lineales (capítulo 2), Solución de Sistemas de Ecuaciones (capítulo 3), Interpolación y Aproximación Polinomial, Aproximación por Mínimos Cuadrados (capítulo 4) y Derivación e Integración Numérica (capítulo 5). Debido a que aun en la resolución de problemas sencillos se requiere de una cierta cantidad de cálculos, resulta, entonces, imprescindible conocer con qué precisión estamos realizando estos cómputos. Por ello, previo al desarrollo de los distintos métodos numéricos, nos referimos con especial énfasis a la teoría relativa a Errores y Sistemas Numéricos (capítulo 1).

Los métodos numéricos son muy útiles e interesantes para estudiantes de diversa procedencia. Por ello es que hemos tratado de que el contenido de este libro sea fácil de leer para el estudiante, con un enfoque orientado hacia las aplicaciones, presentando numerosos ejemplos ilustrativos, resolviéndo-

los detalladamente con distintos procedimientos que permitan comparar los métodos, pudiendo, así, dar respuesta a por qué funcionan y cuáles son sus limitaciones. Adicionalmente, al finalizar cada uno de los cinco capítulos que conforman este texto se enuncian una serie de ejercicios propuestos entre los que se incluyen situaciones problemáticas de aplicación de diversas áreas de la ingeniería, así como de las ciencias físicas y biológicas. Hacemos una deducción matemática y presentamos un análisis del error de aquellos métodos que utilizan resultados elementales. Se presentan tablas y gráficas que permiten visualizar e interpretar cómo trabajan cada uno de los métodos analizados y las aproximaciones numéricas obtenidas. Cada capítulo comienza con un resumen del contenido teórico a tratar. Concluido el tratamiento de los contenidos teóricos incluidos en los capítulos 2, 3 (al finalizar el tema: Solución de Sistemas de Ecuaciones Lineales), 4 y 5, se presentan un cuadro en donde se comparan las principales características de los métodos estudiados y el pseudocódigo de uno de estos métodos. Dicha estructura hace que el usuario no se sienta obligado a emplear un lenguaje de programación determinado, sino a utilizar el que conozca o el que juzgue más conveniente. Estamos convencidos de que los estudiantes aprenden y entienden mejor los métodos numéricos a través de la observación de cómo se desarrollan los algoritmos a partir de la teoría matemática, y luego intentando por sí mismos formular y someter a prueba los programas de computadora por ellos elaborados. Es claro que estos programas no contendrán todos los procedimientos de verificación que poseen los paquetes matemáticos ya existentes. Sin embargo, muchas veces se presenta un problema muy sencillo que no está contemplado por el programa más elaborado, y con mínimos recursos de programación y de análisis numérico puede resolverse. Esto no impide que a los estudiantes les aconsejemos utilizar programas probados, como los que pueden encontrar en las bibliotecas de programas.

Incluimos un apéndice que contiene una introducción a la sintaxis del lenguaje Octave, numerosos ejemplos elaborados, algunos programas completos correspondientes a los métodos numéricos estudiados y una lista de ejercicios propuestos. Creemos que debe ser el capítulo inicial para aquellos estudiantes que no estén familiarizados con dicho lenguaje. Incorporamos este software, que es libre y de código abierto, debido a que es un lenguaje de programación similar y compatible con su contraparte comercial MATLAB, un paquete de programas de uso casi estándar en muchas ramas de la Ingeniería y de la Matemática Aplicada, y con una tendencia de inserción mayor en el futuro.

Capítulo 1

1

Errores. Sistemas Numéricos.

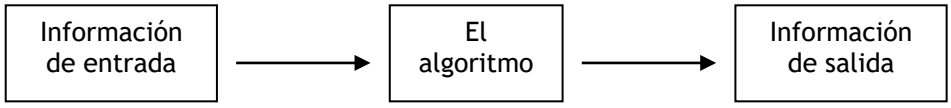
1

1.1. El algoritmo y la presencia de error.

El *Análisis Numérico* tiene que ver con el desarrollo y evaluación de métodos para calcular los resultados numéricos requeridos a partir de datos numéricos. Esto sitúa al Análisis Numérico como parte del moderno *procesamiento de información.*

Ejemplo 1. Reservación automática de asientos en líneas aéreas, impresión automática de cheques y cuentas de teléfono, cálculo automático de los promedios del mercado de cambios, evaluación de ciertos registros médicos (como los encefalogramas), entre otros.

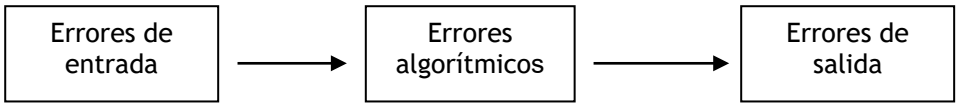
Los datos constituyen *la información de entrada*, los resultados *la información de salida* y el método de cálculo se conoce como *el algoritmo*, con vista a ser tratado en una computadora. Los ingredientes esenciales de un problema de Análisis Numérico se pueden resumir en un diagrama de flujo:



La descripción que acabamos de escoger está orientada definitivamente hacia las aplicaciones, pues enfoca nuestros esfuerzos hacia la búsqueda de algoritmos. Con frecuencia encontraremos que hay varios algoritmos disponibles para producir la información de salida que se requiere y deberemos escoger entre ellos. Hay varias razones para preferir un algoritmo en lugar de otro, pero dos criterios obvios son: la rapidez y la exactitud. La rapidez es, desde luego, una ventaja; con prioridad de otras condiciones, el método rápido tiene seguramente la preferencia. La obtención de la exactitud consumirá mucha de nuestra energía y pondrá en evidencia un segundo aspecto importante del Análisis Numérico: *la presencia de error*. Rara vez la información de entrada es exacta, pues proviene ordinariamente de algún tipo de mecanismo de medida. Las propiedades físicas medibles y todas las características calculadas en base a ellas están representadas en forma de cifras, es decir, números que son "aproximados" al verdadero valor (más adelante se verá en detalle el

concepto de "números aproximados"). Las constantes matemáticas como π, e, así como los resultados de operaciones matemáticas, tales como logaritmos, raíces, potencias de números, funciones trigonométricas o simplemente fracciones en sistema decimal, son también números para los cuales, en general, no se puede dar una representación mediante un número finito de cifras. Al vernos obligados a representar un número mediante una cantidad limitada de cifras decimales estamos ocasionando un error. Conviene hacer notar que la aproximación de los números con los que operamos, en parte está vinculada con el sistema de numeración que utilizamos (en general, decimal y binario). Por ejemplo, la fracción 2/7 con esta notación es exacta, pero en decimal es aproximada pues solamente podemos dar un número finito de las infinitas cifras que constituyen el número. El grado de aproximación se determina por la cantidad de cifras decimales que se utilizan al escribir el número.

Por otra parte, el algoritmo de cálculo introduce también error. La información de salida contiene, por consiguiente, errores de ambas fuentes como se muestra en un segundo diagrama de flujo. Un algoritmo que minimice el crecimiento de error merece, desde luego, una consideración seria.



Aunque nuestro punto de vista del Análisis Numérico está orientado hacia las aplicaciones, tendremos que entendernos naturalmente con la teoría básica, que será importante para nosotros, porque contribuye a la búsqueda de mejores algoritmos. La palabra algoritmo se usó primitivamente para designar procedimientos que terminasen después de un número finito de pasos. Los algoritmos finitos son, fundamentalmente, adecuados para la solución de problemas en Álgebra. Lo más probable es que los dos ejemplos siguientes sean familiares:

1. El algoritmo de Euclides para encontrar el MCD de dos números enteros positivos.
2. El algoritmo de Gauss para resolver un sistema de m ecuaciones lineales con n incógnitas.

Los problemas que se presentan en Análisis, sin embargo, usualmente no se pueden resolver en un número finito de pasos. Al contrario de lo que ocurre con las recetas de un libro de cocina, los algoritmos diseñados para la resolución de problemas del Análisis Matemático consisten, necesariamente, en una sucesión infinita (aunque numerable) de operaciones. Solamente pueden efectuarse un número finito de éstas, desde luego, en cualquier aplicación práctica. La idea es, sin embargo, que la precisión aumenta con el

número de pasos efectuados. Si se realiza un número suficiente de pasos, la precisión puede hacerse todo lo sensible que se quiera. Las mayores dificultades del Análisis Numérico se deben a la falta de precisión y esto surge, como hemos dicho, debido a que las operaciones sólo pueden efectuarse en un número finito, puesto que el tratamiento por una computadora de un problema numérico es finito, ya que la computadora posee una función finita, mientras que como sabemos, una función es un ente matemático definido sobre un espacio continuo. La información que tiene un matemático puro cuando trabaja con una función es infinita. Esto nos lleva a dar énfasis, dentro de los problemas del Análisis Numérico, al problema de aproximación que llega hasta los problemas de aproximación del Análisis Funcional.

1.2. Problemas numéricos y algoritmos.

Definición 1. Un *problema numérico* es una descripción clara y no ambigua de la relación funcional entre los datos de entrada (variables independientes del problema) y los datos de salida (resultados requeridos). Los datos de entrada y salida consisten en un número finito de cantidades reales (dado que un número complejo es un par de números reales, los datos de entrada y salida de complejos están incluidos en esta definición).

Los datos de entrada y salida son, por lo tanto, representables por vectores finitos - dimensionales y la relación funcional puede ser expresada en forma implícita o explícita.

Definición 2. Un *algoritmo* para un problema numérico dado es una descripción completa de operaciones bien definidas que transforma un vector de datos de entrada en un vector de datos de salida. Por operaciones entenderemos aquí las operaciones aritméticas y lógicas que una computadora puede ejecutar, junto con referencias a algoritmos previamente definidos.

Para un problema numérico dado, uno puede considerar muchos algoritmos diferentes. Estos pueden dar respuestas aproximadas, las cuales tienen una exactitud variable.

Ejemplo 2. Se quiere determinar la raíz real más grande de la ecuación

$$x^3 + a_2x^2 + a_1x + a_0 = 0$$

con coeficientes reales a_0, a_1, a_2.

Este es un problema numérico. El vector de datos de entrada es $\vec{a} = (a_0, a_1, a_2)$; el vector de datos de salida es la raíz $\vec{x}$; esta es una relación funcional definida implícitamente sobre los datos de entrada (ecuación). Un algoritmo para este problema puede basarse en la solución exacta de Cardano para una ecuación cúbica. La solución de Cardano usa raíces cuádricas y raíces cúbicas; así, uno necesita suponer que los algoritmos para la computación de estas funciones han sido especificados previamente.

A menudo, uno comienza la construcción de un algoritmo para un problema dado dividiendo el problema en subproblemas, de forma tal que los datos de salida de un subproblema sean los datos de entrada del próximo subproblema. Por lo tanto, distinguir entre problema y algoritmo no es siempre fácil de hacer. El punto esencial es que en la formulación de un problema, uno sólo está interesado en el estado inicial y en el estado final. En un algoritmo, sin embargo, uno definirá claramente cada paso a lo largo del camino desde el principio al final.

1.3. Fórmulas recursivas. Regla de Hörner.

Una de las partes más importantes e interesantes en la preparación de un problema para una computadora, es encontrar una descripción recursiva de la tarea. Algunas veces, una enorme cantidad de cálculos pueden ser descriptos por un pequeño conjunto de fórmulas recursivas.

Un ejemplo computacional común es la evaluación de un polinomio en un punto z dado, digamos

$$p(z) = a_0z^3 + a_1z^2 + a_2z + a_3.$$

Esto puede ser reformulado como

$$p(z) = ((a_0z + a_1)z + a_2)z + a_3.$$

Esta forma de escribir $p(z)$ es ventajosa, pues aquí sólo se hacen tres multiplicaciones mientras que originalmente hacíamos ocho, y esto se debe tener en cuenta en el momento de minimizar el efecto de los errores.

Para computación el siguiente esquema, regla de Hörner, ilustra el algoritmo indicado para la reformulación anterior

a_0	a_1	a_2	a_3
	zb_0	zb_1	zb_2
b_0	b_1	b_2	b_3

$$p(z) = b_3 \text{ (resultado).}$$

Ejemplo 3. Calculemos $p(8)$, donde, $p(x) = 2x^3 + x + 7$.

Debemos entonces desarrollar el siguiente esquema

2	0	1	7
	16	128	1032
2	16	129	1039

Por lo tanto, $p(8) = 1039$.

La Regla de Hörner para evaluar un polinomio de grado n

$$p(x) = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n$$

en un punto z, se escribe por la fórmula recursiva

$b_0 = a_0$
$b_i = a_i + z\, b_{i-1} \quad (i = 1, 2, \ldots, n)$
$b_n = p(z)$

Si los b_i no son de interés, en muchos lenguajes de programación el algoritmo puede ser descrito sin subindicar los b_i, tales como en la Figura 1 y (el símbolo := se lee como "dar el valor de")

$b := a[0]$;
para $i := 1$, hasta n, con paso 1, hacer
$b := a[i] + z*b$;

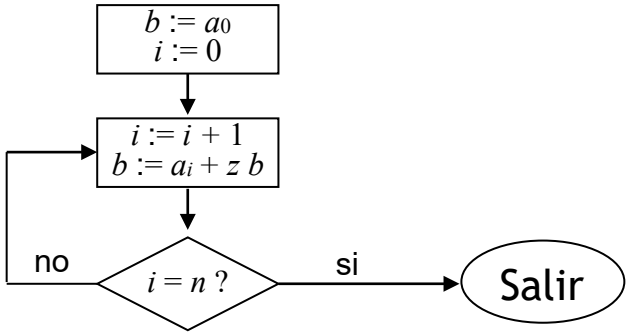


Figura 1

Ejemplo 4. Definamos un algoritmo como sigue: escójase z_0 arbitrariamente en el intervalo (0, 2) y calcúlese z_1, z_2, ... por la siguiente relación de recurrencia

$$z_{k+1} = \frac{1}{2 - z_k}.$$

Este algoritmo no está bien definido ya que la fórmula no tiene sentido cuando $z_k = 2$ que es el caso, por ejemplo, si $z_0 = \frac{5}{4}$.

Este algoritmo puede convertirse en uno bien definido por una proposición como la siguiente: "siempre que $z_k = 2$ hágase $z_k = 13$".

1.4. Ambigüedad, convergencia y estabilidad de los algoritmos.

La búsqueda de algoritmos debe tenerse en cuenta por varios factores que influyen de manera altamente negativa en la obtención de los resultados deseados. Estos factores son: *ambigüedad*, *convergencia* y *estabilidad*.

Veamos estos conceptos en detalle.

Ambigüedad. Un algoritmo debe estar bien definido, es decir, no debe presentar ambigüedades acerca de la operación que debe seguir o una cualquiera de él. Deben preverse y evitarse causas de interrupción, tales como división por cero, raíces cuadradas de números negativos en *R*, evaluación de funciones en puntos donde no estén definidas, entre otras, y deben contemplarse todas las posibilidades que se puedan presentar en función de los datos del problema y la o las respuestas deseadas.

Convergencia. Una vez que un algoritmo está propiamente formulado, deseamos conocer exactamente cuáles son las condiciones bajo las que el algoritmo produce la solución del problema considerado. Si como es el caso más común, el algoritmo da como resultado la construcción de una sucesión de números, deseamos conocer las condiciones para que esta sucesión sea convergente. El practicante del arte de la computación se encuentra frecuentemente inclinado a juzgar el comportamiento de un algoritmo de un modo excesivamente pragmático: el algoritmo se ha probado en cierto número de ejemplos y ha trabajado satisfactoriamente en un 95% de los casos. Los matemáticos tienden a considerar este tipo de investigación científica (no obstante que es el método de investigación tipo en disciplinas tan vitales como la medicina y la biología) de un modo muy poco satisfactorio. Lo cierto es que siempre es de desear que la base de nuestro juicio sobre el comportamiento de cierto algoritmo sea la evidencia lógica en lugar de la empírica. Tales proposiciones lógicamente

demostrables se llaman en matemática *teoremas*. Como veremos, en muchos casos se pueden enunciar teoremas sobre la convergencia de algoritmos. Una vez resuelto el problema de la convergencia, pueden plantearse otras cuestiones sobre el comportamiento de un algoritmo. Se puede querer saber, por ejemplo, cuál es su velocidad de convergencia. En este caso, se puede hacer un estudio más profundo del algoritmo con el fin de optimizarlo y acelerar la convergencia del mismo. La convergencia teórica de un algoritmo no garantiza que éste sea prácticamente útil. Debe satisfacer un requerimiento más.

Estabilidad numérica. En la Escuela Media aprendimos como expresar el resultado de multiplicar dos decimales de seis dígitos (que, en general, tiene doce dígitos) en forma aproximada en términos de un decimal de seis dígitos, por un proceso conocido como *redondeo*. Si tenemos que efectuar varias multiplicaciones consecutivas, el redondeo se convierte en una necesidad práctica, pues es imposible manejar un número siempre creciente de cifras decimales. En vista de que los errores individuales debido al redondeo son pequeños, usualmente presumimos que la exactitud del resultado final no queda seriamente afectada por los errores individuales de redondeo. Las computadoras digitales electrónicas modernas trabajan también con un número limitado de cifras decimales. El número de operaciones aritméticas que pueden realizar por unidad de tiempo, sin embargo, es aproximadamente un millón de veces mayor que el realizado por la computación manual. Aunque los errores individuales de redondeo sean pequeños, su efecto acumulativo puede, en vista del gran número de operaciones aritméticas efectuadas, crecer rápidamente e invalidar completamente el resultado final. Para que sea confiable, un algoritmo debe permanecer inmune a la acumulación de los errores de redondeo. A esta inmunidad se le llama *estabilidad numérica*. En lugar de intentar establecer patrones absolutos de estabilidad, la meta de una teoría de estabilidad numérica debe ser la de predecir de un modo cuantitativo la extensión de la influencia de la acumulación de los errores de redondeo, si se aventuran ciertas hipótesis sobre los errores de redondeo individuales. Tal teoría será más descriptiva que categórica. Teóricamente, predecirá el resultado de un experimento numérico de modo en gran parte análogo a como una teoría física predice el resultado de experimentos físicos.

Ejemplo 5. Calculemos para $n = 0, 1,..., 10,$ $\quad y_n = \int_0^1 \frac{x^n}{4x+1}\,dx$.

Usamos la fórmula de recursión

$$y_n + 4y_{n+1} = \frac{1}{n+1}$$

que se obtiene desde

$$y_n + 4y_{n+1} = \int_0^1 \frac{x^n}{4x+1}dx + \int_0^1 \frac{4x^{n+1}}{4x+1}dx =$$

$$= \int_0^1 \frac{x^n(1+4x)}{4x+1}dx = \int_0^1 x^n dx = \frac{x^{n+1}}{n+1}\Big|_0^1 = \frac{1}{n+1}.$$

Usamos 6 decimales en los cálculos.

Algoritmo I. Calculamos

$$y_0 = \int_0^1 \frac{dx}{4x+1} = \frac{1}{4}ln(4x+1)\Big|_0^1 = \frac{1}{4}(ln5 - ln1) = \frac{1}{4}ln5 = 0.402360$$

y luego, usamos

$$y_{n+1} = \frac{1}{4}\left[\frac{1}{n+1} - y_n\right] \quad \text{para} \quad n = 0,\ 1,...,\ 9$$

o bien

$$y_{n+1} = \frac{1}{4(n+1)} - \frac{y_n}{4} \quad \text{para} \quad n = 0,\ 1,...,\ 9.$$

Así,

$$y_1 = \frac{1}{4} - \frac{y_0}{4} = 0.149410$$

$$y_2 = \frac{1}{8} - \frac{y_1}{4} = 0.087648$$

$$y_3 = \frac{1}{12} - \frac{y_2}{4} = 0.061421$$

$$y_4 = \frac{1}{16} - \frac{y_3}{4} = 0.047145$$

$$y_5 = \frac{1}{20} - \frac{y_4}{4} = 0.038214$$
$$y_6 = \frac{1}{24} - \frac{y_5}{4} = 0.032114$$
$$y_7 = \frac{1}{28} - \frac{y_6}{4} = 0.027686$$
$$y_8 = \frac{1}{32} - \frac{y_7}{4} = 0.024328$$
$$y_9 = \frac{1}{36} - \frac{y_8}{4} = 0.021696$$
$$y_{10} = \frac{1}{40} - \frac{y_9}{4} = 0.019576 \qquad \text{(correcto !)}$$

Este algoritmo trabaja bien, pues el error se divide por -4 en cada paso.

Algoritmo II. Usamos la fórmula de recursión en dirección contraria; esto es

$$y_n = \frac{1}{n+1} - 4y_{n+1}.$$

Pero ahora necesitamos un valor de partida. Podemos ver directamente desde la definición que y_n decrece cuando n crece. En efecto,

$$0 < y_n = \int_0^1 \frac{x^n}{4x+1}dx \le \int_0^1 x^n dx = \frac{1}{n+1}, \ \text{de donde}, \ 0 < y_n \le \frac{1}{n+1}$$

y como

$$\frac{1}{n+1} \xrightarrow[n\to\infty]{} 0, \ \text{entonces}$$
$$y_n \xrightarrow[n\to\infty]{} 0, \ \text{de donde}, \quad y_n < y_{n-1}, \forall\, n \in N.$$

Uno puede también conjeturar que y_n decrece lentamente cuando n es grande. Así, podemos ensayar

$$y_{12} \approx y_{11} \ \text{ y entonces,}$$
$$y_{11} + 4y_{11} \approx \frac{1}{12}, \ \text{de donde}, \ y_{11} \approx \frac{1}{60} = 0.016667\,.$$

$$y_{10} = \frac{1}{11} - 4y_{11} = 0.024241$$

$$y_9 = \frac{1}{10} - 4y_{10} = 3.036\text{x}10^{-3} \qquad (y_9 < y_{10}\ !)$$

$$y_8 = \frac{1}{9} - 4y_9 = 0.098967$$

$y_7 = \frac{1}{8} - 4y_8 = -0.270868$, obviamente absurdo que $y_7 < 0$!

La razón de este resultado absurdo es que el error de redondeo ε en y_{12}, cuya magnitud puede ser tan grande como $5\text{x}10^{-7}$, es multiplicado por -4 en el cálculo de y_{11} el cual tiene entonces un error de -4ε. Este error produce un error en y_{10} de 16ε, en y_9 de -64ε, en y_8 de 256ε. Así, el error en y_7 es -1024ε, valor que puede ser tan grande como $-1024 \times 5 \times 10^{-7} = -5.12 \times 10^{-4}$.

Si usáramos más decimales en los cálculos, el resultado absurdo aparecerá en etapas subsiguientes. Este algoritmo es un ejemplo de un fenómeno desagradable llamado *inestabilidad numérica.*

Algoritmo III. El mismo algoritmo II excepto que uno toma como valor inicial $y_{12} = 0$. Obtenemos, entonces $y_{11} = 0.083333$, $y_{10} = -0.242423$!

Ejemplo 6. Calculemos para $n = 0, 1, \ldots, 8$, $y_n = \int_0^1 \frac{x^n}{x+5}\, dx$.

Usamos la fórmula de recursión

$$y_n + 5y_{n-1} = \frac{1}{n}, \qquad n = 1, 2, \ldots, 8$$

que se obtiene desde

$$y_n + 5y_{n-1} = \int_0^1 \frac{x^n + 5x^{n-1}}{x+5}\, dx = \int_0^1 \frac{x^{n-1}(x+5)}{x+5}\, dx = \int_0^1 x^{n-1}\, dx = \frac{1}{n}.$$

Usamos tres decimales en los cálculos.

Algoritmo I. Calculamos

$$y_0 = \int_0^1 \frac{dx}{x+5} = ln(x+5) \Big|_0^1 = ln\,6 - ln\,5 = 0.182$$

$y_1 = 1 - 5y_0 = 1 - 0.910 = 0.009$

$y_2 = \frac{1}{2} - 5y_1 = 0.050$

$y_3 = \frac{1}{3} - 5y_2 = 0.083,\ y_3 > y_2$!

$y_4 = \frac{1}{4} - 5y_3 = -0.165$, obviamente absurdo que $y_4 < 0$!

La razón para el resultado absurdo es que el error de redondeo ε en y_0, cuya magnitud puede ser tan alta como $5x10^{-4}$, es multiplicado por -5 en el cálculo de y_1 el cual tiene entonces un error de -5ε. Este error produce un error en y_2 de 25ε, en y_3 de -125ε. Así, el error en y_4 es 625ε, valor que puede ser tan grande como $625 \times 5 \times 10^{-4} = 0.3125$.

Si usáramos más decimales en los cálculos el resultado absurdo aparecerá en etapas subsiguientes. Este algoritmo es inestable numéricamente. Podemos evitar esta inestabilidad eligiendo otro algoritmo.

Algoritmo II. Usamos la fórmula de recursión en otra dirección

$$y_{n-1} = \frac{1}{5n} - \frac{y_n}{5}.$$

Ahora el error se dividirá por –5 en cada paso. Pero necesitamos un valor inicial. Podemos ver directamente desde la definición que y_n decrece cuando n crece. Uno puede también conjeturar que y_n decrece lentamente cuando n es grande.

Así, podemos ensayar $y_{10} \approx y_9$ y entonces

$$y_9 + 5y_9 \approx \frac{1}{10} \quad \text{ó} \quad y_9 \approx \frac{1}{60} = 0.017$$

$$y_8 = \frac{1}{45} - \frac{y_9}{5} = 0.019$$

$$y_7 = \frac{1}{40} - \frac{y_8}{5} = 0.021$$

$y_6 = 0.025$
$y_5 = 0.028$
$y_4 = 0.034$
$y_3 = 0.043$
$y_2 = 0.058$
$y_1 = 0.088$
$y_0 = 0.182$ (correcto!)

Algoritmo III. El mismo algoritmo II excepto que uno toma como valor inicial $y_{10}=0$. Obtenemos, entonces

$y_9 = 0.020$
$y_8 = 0.018$
$y_7 = 0.021$

y el resto de los y_n tienen los mismos valores que en el algoritmo II.

1.5. Errores absolutos y relativos.

Un *número aproximado a* es un número tal que difiere ligeramente de un número exacto *A* y se utiliza en los cálculos en lugar de este último. Si se sabe que $a < A$, se dice que *a es una aproximación por defecto* (más pequeña) *de A*; si $a > A$, se dice entonces que *a es una aproximación por exceso* (mayor) *de A*.

Ejemplo 7. Para $\sqrt{2}$ el número 1.41 es una aproximación por defecto, mientras que 1.42 lo es por exceso, ya que $1.41 < \sqrt{2} < 1.42$.

Si *a* es un valor aproximado del número *A* se escribe $a \approx A$.

Por *error* Δa de un número aproximado *a*, por lo general, quiere indicarse la diferencia entre el número exacto *A* y el número aproximado dado; esto es

$$\Delta a = A - a$$

(a veces la diferencia a - A se denomina error). Si $A > a$, entonces el error es positivo: $\Delta a > 0$; por el contrario, si $A < a$, el error entonces es negativo: $\Delta a < 0$. Para obtener el número exacto A, se añade el error Δa al número aproximado a:

$$A = a + \Delta a.$$

Por lo tanto, un número exacto puede considerarse como aproximado con error cero.

En muchos casos el signo del error es desconocido. Resulta entonces aconsejable usar el error absoluto del número aproximado:

$$\Delta = |\Delta a|.$$

Notación. $\Delta = \varepsilon$.

Definición 3. El *error absoluto* Δ de un número aproximado a es el valor absoluto de la diferencia entre el número exacto correspondiente y el número a

$$\Delta = | A - a |. \qquad (1)$$

Deben considerarse dos casos:

(i) el número A es conocido y entonces el error absoluto se determina en forma inmediata a partir de (1);

(ii) el número A es desconocido, caso más corriente y, por lo tanto, el error absoluto Δ no puede encontrarse a partir de la fórmula (1).

Resulta conveniente, entonces, introducir el error superior estimado, denominado cota del error absoluto, en lugar del error absoluto Δ teórico desconocido.

Definición 4. La *cota del error absoluto* (ó error absoluto límite ó error absoluto máximo) de un número aproximado es cualquier número no menor que el error absoluto de dicho número.

Por lo tanto, si Δ_a es la cota del error absoluto de un número aproximado a tomado en lugar del número exacto A, se tiene

$$\Delta = |A - a| \leq \Delta_a \underset{Not.}{=} \varepsilon_a. \qquad (2)$$

De aquí, se deduce que el número exacto A cae dentro del margen

$$a - \Delta_a \leq A \leq a + \Delta_a. \qquad (3)$$

Por lo tanto, $a - \Delta_a$ es una aproximación por defecto de A y $a + \Delta_a$ lo es por exceso.

Para mayor brevedad, puede escribirse

$$A = a \pm \Delta_a.$$

Ejemplo 8. Determinemos la cota del error absoluto del número $a = 3.14$ usado en lugar del número π.

Como tenemos la desigualdad $3.14 < \pi < 3.15$, se deduce que $|\pi - a| < 0.01$ y, por lo tanto, podemos tomar $\Delta_a = 0.01$.

Partiendo de que

$$3.14 < \pi < 3.142$$

se tiene una mejor aproximación: $\Delta_a = 0.002$.

Obsérvese que el concepto de cota del error absoluto tal como se ha indicado anteriormente es muy amplio, esto es, la cota del error absoluto de un número aproximado a se entiende es cualquiera de los infinitos números no negativos Δ_a que satisfagan la desigualdad (2).

Se deduce por lo tanto, lógicamente, que cualquier número que excede la cota del error absoluto de un número aproximado dado, puede denominarse cota del error absoluto de dicho número. Para fines prácticos es conveniente tomar el número más pequeño de Δ_a (en las circunstancias dadas) que satisfaga la desigualdad (2).

Cuando se escribe un número aproximado procedente de una medida es común dar su cota del error absoluto, como se muestra en el siguiente ejemplo.

Ejemplo 9. Si la longitud de un segmento es $l = 214$ cm con un error de 0.5 cm, se escribe $l = 214 \text{ cm} \pm 0.5 \text{ cm}$. En este caso la cota del error absoluto es $\Delta_l = 0.5$ cm y la magnitud exacta de la longitud del segmento cae dentro del margen

$$213.5 \text{ cm} \leq l \leq 214.5 \text{ cm}.$$

El error absoluto (o cota del error absoluto) no es suficiente para definir la exactitud de una medida o un cálculo, como se observa en el siguiente ejemplo.

Ejemplo 10. Supongamos que al medir las longitudes de dos varillas se tienen $l_1 = 100.8$ cm ± 0.1 cm y $l_2 = 5.2$ cm ± 0.1 cm. Independientemente del hecho de que los errores absolutos límites coincidan, la primera medida es mejor que la segunda.

Un punto esencial en la exactitud de las medidas es el error absoluto relativo a la unidad de longitud; esto es, la relación del error absoluto con respecto a la magnitud en cuya determinación se ha cometido el error. Este se denomina error relativo.

Definición 5. El *error relativo* δ de un número aproximado a es la relación entre el error absoluto Δ del número y el módulo (valor absoluto) del correspondiente número exacto A ($A \neq 0$). Por lo tanto,

$$\delta = \frac{\Delta}{|A|} \tag{4}$$

de donde,

$$\Delta = |A|\delta.$$

Como en el caso del error absoluto, introducimos la noción de cota del error relativo.

Definición 6. La *cota del error relativo* δ_a (ó error relativo límite ó error relativo máximo) de un número aproximado a dado es cualquier número no menor que el error relativo de dicho número.

Por definición, se tiene

$$\delta \leq \delta_a. \tag{5}$$

Esto es,

$$\frac{\Delta}{|A|} \leq \delta_a$$

de donde,

$$\Delta \leq |A|\delta_a.$$

Por lo tanto, como cota del error absoluto de un número aproximado a se puede escribir

$$\Delta_a = |A|\delta_a. \tag{6}$$

Como en situaciones prácticas $A \approx a$, en lugar de la fórmula (6) se utiliza frecuentemente

$$\Delta_a \approx |a|\delta_a. \qquad (6')$$

De esta fórmula, conociendo la cota del error relativo δ_a se obtienen los límites del número exacto. El hecho de que el número exacto caiga entre $a(1-\delta_a)$ y $a(1+\delta_a)$ (pues de (6'), $\Delta \leq |a|\delta_a$) se simboliza de la manera

$$A = a(1 \pm \delta_a). \qquad (7)$$

Supongamos que a es un número aproximado que toma el lugar de un número exacto A y supongamos que Δ_a es la cota del error absoluto de a. Hagamos $A > 0$, $a > 0$ y $\Delta_a < a$. Entonces

$$\delta = \frac{\Delta}{A} \leq \frac{\Delta_a}{a - \Delta_a}$$

(pues: $\Delta_a < a$, entonces $a - \Delta_a > 0$ y como $A \geq a - \Delta_a$, entonces $\frac{1}{A} \leq \frac{1}{a - \Delta_a}$).

Puede tomarse, por lo tanto, el número

$$\delta_a = \frac{\Delta_a}{a - \Delta_a} \qquad (8)$$

como cota del error relativo del número aproximado a.

Análogamente, se tiene

$$\Delta = A\delta \leq (a + \Delta)\delta_a$$

(pues: $A - a \leq |A - a| = \Delta$, entonces $A \leq a + \Delta$ y por (5), $\delta \leq \delta_a$)

de donde,

$$\Delta \leq \frac{a\delta_a}{1 - \delta_a}.$$

Puede tomarse, por lo tanto, el número

$$\Delta_a = \frac{a\delta_a}{1 - \delta_a} \qquad (9)$$

como cota del error absoluto del número aproximado a.

Si como sucede comúnmente, $\Delta_a << a$ y $\delta_a << 1$ (el símbolo << significa "mucho menor que") puede tomarse entonces, aproximadamente,

$$\delta_a \approx \frac{\Delta_a}{a} \tag{10}$$

y

$$\Delta_a \approx a\ \delta_a. \tag{11}$$

Ejemplo 11. El peso de 1 dm^3 de agua a 0°C viene dado por p = 999.847 gf ± 0.001 gf (gf = gramo fuerza). Determinemos la cota del error relativo del resultado del peso del agua.

Evidentemente, se tiene $\Delta_p = 0.001$gf y $p \geq 999.847$ gf - 0.001 gf, o sea, $p \geq 999.846$gf.

En consecuencia, de (8)

$$\delta_p = \frac{0.001}{999.846} = 1.000154\text{x}10^{-6} \approx 1.0002\text{x}10^{-6}.$$

Ejemplo 12. Al determinar la constante gaseosa del aire se obtuvo R = 29.25. Conociendo que la cota del error relativo de este valor es 0.001, encontremos los límites dentro de los cuales está R.

Tenemos $\delta_R = 0.001$ y, por lo tanto, $\Delta_R = R\delta_R = 0.02925 \approx 0.03$.

En consecuencia, de (3)

$$29.22 \leq R \leq 29.28.$$

Definición 7. Se llama *error relativo porcentual*, δ %, al producto 100δ y *cota del error relativo porcentual* (ó error relativo máximo porcentual ó error relativo límite porcentual), δ_a%, al producto $100\delta_a$.

Ejemplo 13. Si el error que se comete en la determinación del espesor de una madera terciada de 3 mm es de 1 mm, entonces resulta $\delta = 1/3$ y el error relativo porcentual $\delta\% = 33.333\ldots$, o sea es de aproximadamente el 33%.

Ejemplo 14. En 1862 el físico Foucault utilizando un espejo giratorio calculó en 298000 km por segundo la velocidad de la luz. Aceptando como exacta la velocidad de 299776 km por segundo determinada por Anderson mediante una célula Kerr, el error absoluto cometido por Foucault es de 1776 km por segundo (sobre la cuantía absoluta de este error basta decir que

un vehículo con esa velocidad daría la vuelta al mundo en 23 segundos). Sin embargo, a pesar del gran valor del error absoluto la determinación de Foucault es muy buena, pues el error relativo, 1776/299776, es aproximadamente 0.006 y el error relativo porcentual resulta ser de 0.6%.

Ejemplo 15. La determinación de la constante universal h realizada por Planck en 1913 dio el valor

$$h = 6.41 \times 10^{-27} \text{ erg.seg.}$$

El valor adoptado en 1948 por los físicos es

$$h = 6.623 \times 10^{-27} \text{ erg.seg.}$$

El error absoluto en este caso es

$$0.213 \times 10^{-27}$$

cantidad muy pequeña con respecto a la unidad ergio por segundo.

Pero el error relativo de la determinación de Planck es

$$\delta = \frac{0.213 \text{x} 10^{-27}}{6.623 \text{x} 10^{-27}} = 0.032.$$

El error relativo porcentual es de 3.2 %, que es 5 veces mayor que el error correspondiente a la determinación de la velocidad de la luz realizada por Foucault.

1.6 Fuentes básicas de errores

La “aproximación” es un concepto central en casi todos los usos de la matemática. A menudo debemos conformarnos con valores aproximados de las cantidades con las cuales trabajamos. Otro tipo de aproximación ocurre cuando ignoramos algunas cantidades que son pequeñas comparadas con otras. Frecuentemente, tales aproximaciones son necesarias para asegurar que el tratamiento matemático y numérico no se convierta en algo complicado, sin esperanzas. La computación numérica es susceptible a un cierto número de errores tales que algunas fuentes de errores no son controlables. En cambio, otras pueden ser disminuidas o anuladas, por

ejemplo, mediante una reformulación del problema o un cambio en la secuencia computacional.

Los errores pueden ser clasificados según su origen en:

(a) **Errores de entrada o inherentes.** Son errores en los valores numéricos con los que se va a operar.

1. De medición: datos medidos experimentalmente, los que pueden estar afectados de errores sistemáticos (debido a la imprecisión de los aparatos de medición) y errores accidentales (debido a la apreciación del observador y a otras causas).
2. De representación: el trabajar con aritmética finita obliga a operar con una cantidad fija y finita de cifras, lo que impide representar exactamente números tales como los irracionales (e, π, 1/3,…). Han de redondearse también, números finitos de muchos dígitos.

(b) **Errores de redondeo durante la computación.** Si el dispositivo de cálculo que uno está usando no puede manejar números que tengan más de S dígitos, entonces el producto exacto de dos números de S dígitos (que contiene $2S$ o $2S$ - 1 dígitos) no puede ser usado en los cálculos subsiguientes; el producto debe ser redondeado a S dígitos. El efecto de tales errores puede ser muy dañino en cálculos largos o en algún algoritmo que es numéricamente inestable.

(c) **Errores de truncamiento.** No son ni errores inherentes ni errores de redondeo. Generalmente provienen del hecho de que mientras una sucesión de infinitos pasos proporcionaría un resultado exacto, el proceso se trunca después de un número finito de pasos por razones prácticas.

1. De series infinitas.
2. Por discretización de operadores continuos como, por ejemplo, aproximar la derivada primera por el cociente incremental.
3. Por linealización de problemas no lineales.

(d) **Elección de un modelo matemático simplificado.** La descripción matemática del problema no se ajusta exactamente a la realidad, sino que constituye una idealización de la misma, por ejemplo, en el caso del péndulo ideal se desprecia la masa del cuerpo suspendido, o en cálculos de economía uno debe suponer que la tasa de interés es constante en un período de tiempo dado. En muchos tipos de problemas es ventajoso considerar que un cuerpo está constituido por una masa homogénea en lugar de ser armada por átomos. Los efectos de tal fuente de error son usualmente más difíciles de estimar que los tipos dados en (a), (b) y (c).

(e) **Errores humanos y errores de máquina.** En todo trabajo numérico deben esperarse que ocurran errores de transcripción, errores en la parte de cálculo manual y errores de mala interpretación. Se debe tener cuidado aun de tablas impresas, pues pueden tener errores.
Cuando usamos la computadora podemos esperar: errores en el programa, errores del operador y errores de máquina.
Errores que son puramente errores de máquina son responsables de una pequeña parte de resultados extraños producidos por la computadora año tras año. La mayoría de los errores dependen del factor humano.

1.7 Propiedades peculiares de los números almacenados con el arte de la computación.

Definición 8. Se llama *unidad numérica* de un número escrito en forma decimal al nombre del lugar del último dígito del número, con exclusión de los números enteros que terminan en uno o varios ceros. En este caso excepcional, la unidad numérica es el nombre del lugar del último dígito no nulo del número o el nombre del lugar de cualquiera de los ceros que están a la derecha del último dígito no nulo. (Esto depende si los ceros se conocen con exactitud).

Si en el contexto del proceso numérico no está bien claro cual es la unidad numérica del caso excepcional mencionado, debe ser claramente especificado antes de comenzar el proceso.

Ejemplo 16

3.14 centésimo || último dígito del número

0.0700 diezmilésimo || último dígito del número

67 unidad || último dígito del número

74.2 décimo || último dígito del número

67000 Caso excepcional: se debe aclarar si son unidades, decenas, centenas, unidades de mil.

Es conveniente extender este concepto de unidad numérica para considerarlo no sólo como nombre de un lugar en la representación decimal

de un número, sino una potencia de diez. Así, las unidades numéricas un diezmilésimo, un centésimo, una decena serán representadas por las potencias 10^{-4}, 10^{-2}, 10^{1}.

Ejemplo 17. Los números dados en el ejemplo 16 pueden expresarse como

314×10^{-2}
700×10^{-4}
67×10^{0}
742×10^{-1}
67×10^{3} ó 670×10^{2} ó 6700×10^{1} ó 67000×10^{0}.

La potencia de 10 indica en cada caso la unidad numérica.

Definición 9. Cualquier número n puede escribirse en *forma de unidad numérica*

$$n = n' 10^{u}$$

donde n' es un número entero y 10^{u} es la unidad numérica, de donde se desprende que u es también un número entero.

Si todos los dígitos de un número son ceros, por ejemplo, 0.00 pondremos $n' = 0$ y escribiremos 0×10^{0}.

Definición 10. Las *cifras ó dígitos significativos* de un número n cualquiera es el número de dígitos en n' cuando n está escrito en forma de unidad numérica.

Ejemplo 18. Consideremos los números dados en el ejemplo 16. Entonces,

$3.14 = 314 \times 10^{-2}$		tiene 3 cifras significativas
$0.0700 = 700 \times 10^{-4}$		tiene 3 cifras significativas
$67 = 67 \times 10^{0}$		tiene 2 cifras significativas
$74.2 = 742 \times 10^{-1}$		tiene 3 cifras significativas
67000 puede ser	67×10^{3}	tiene 2 cifras significativas
ó	670×10^{2}	tiene 3 cifras significativas
ó	6700×10	tiene 4 cifras significativas
ó	67000×10^{0}	tiene 5 cifras significativas

Omitiendo el caso excepcional de un entero que termine en uno o varios ceros, el número de cifras significativas de un número escrito en notación decimal es igual al número de sus dígitos, excluyendo todos los dígitos que preceden al primer dígito no nulo.

Las cifras significativas de un número reciben este nombre porque son las únicas que especifican el número de unidades numéricas.

Observación. En el número 0.002080, los 3 primeros ceros son dígitos no significativos ya que sirven para fijar la posición de la coma e indicar el valor posicional de los otros dígitos. Los otros dos ceros son significativos, ya que el primero cae entre los dígitos 2 y 8 y el segundo (tal como viene indicado en la notación) indica que se conserva el lugar decimal 10^{-6} en el número dado. Si el último dígito de 0.002080 no es significativo, entonces el número debe escribirse 0.00208.

Desde este punto de vista, los números 0.002080 y 0.00208 no son los mismos ya que el primero tiene 4 dígitos significativos y el último solamente tres.

Al escribir números grandes los ceros a la derecha pueden servir tanto para indicar los dígitos significativos como para fijar la posición de los otros dígitos. Esto puede ocasionar alguna confusión cuando los números se escriben en la forma ordinaria. Considérese, por ejemplo, el número 689000. No queda claro cuantos dígitos significativos hay, aun cuando pueda decirse que hay al menos tres. Para evitar esta ambigüedad es que se usa la notación en forma de unidad numérica (notación de potencias de 10) y se escribe el número 689000 en la forma $689\text{x}10^3$ si tiene tres dígitos significativos, ó $68900\text{x}10^1$ si tiene 5 dígitos significativos, y así siguiendo.

Cualquier número positivo *a* puede representarse en forma de decimal con un número finito o infinito de dígitos

$$a = \alpha_m 10^m + \alpha_{m-1} 10^{m-1} + \alpha_{m-2} 10^{m-2} + ... + \alpha_{m-n+1} 10^{m-n+1} + ... \qquad (12)$$

donde α_i son los dígitos del número *a* (α_i = 0, 1, 2,…, 9), $\alpha_m \neq 0$ es el dígito más significativo y *m* es un entero (la potencia más elevada de 10 en el número *a*).

Ejemplo 19

$$341.59\ldots = 3\text{x}10^2 + 4\text{x}10^1 + 1\text{x}10^0 + 5\text{x}10^{-1} + 9\text{x}10^{-2} + \ldots$$

Cada unidad ocupa una posición específica en el número *a* escrito en forma de potencias de 10, ecuación (12), y tiene un valor definido. La unidad que ocupa la primera posición es igual a 10^m, la de la segunda posición 10^{m-1}, la de la *n*-ésima posición 10^{m-n+1}, y así sucesivamente.

Los casos reales, por lo general, tratan con números aproximados en forma de decimales con un número finito de cifras

$$b = \beta_m 10^m + \beta_{m-1} 10^{m-1} + \beta_{m-2} 10^{m-2} + ... + \beta_{m-n+1} 10^{m-n+1} \ (\beta_m \neq 0). \quad (13)$$

Todos los dígitos decimales β_i (i = m, m-1, …, m-n+1) son los que hemos denominado significativos del número aproximado b; téngase en cuenta que algunos de ellos pueden ser iguales a cero (con excepción de β_m). En el sistema posicional decimal para representar el número b, a menudo han de introducirse ceros al comienzo o al final del número.

Ejemplo 20

$b = 7\text{x}10^{-3} + 0\text{x}10^{-4} + 1\text{x}10^{-5} + 0\text{x}10^{-6} = \underline{0.00}7010$
ó
$b = 2\text{x}10^{9} + 0\text{x}10^{8} + 0\text{x}10^{7} + 3\text{x}10^{6} + 0\text{x}10^{5} = 2003\underline{00000}$.

Los ceros subrayados no son significativos.

Introduzcamos la noción de dígitos exactos de un número aproximado.

Definición 11. Se dice que los n primeros dígitos significativos de un número aproximado son *exactos* si el error absoluto del número no excede de media unidad situada en el lugar n-ésimo, contados de izquierda a derecha.

Por lo tanto, si en lugar de un número exacto A se toma un número aproximado a, ecuación (12), se sabe que

$$\Delta = |A - a| \leq \frac{1}{2} \text{x } 10^{m-n+1}$$

entonces, por definición, los n primeros dígitos $\alpha_m, \alpha_{m-1}, ..., \alpha_{m-n+1}$ de dicho número son exactos.

Observemos que si a tiene n dígitos exactos, entonces como cota del error absoluto puede tomarse

$$\Delta_a = \frac{1}{2} \text{x } 10^{m-n+1}.$$

Ejemplo 21. Con respecto al número exacto A = 35.97 el número aproximado a = 36.00 es una aproximación con 3 dígitos exactos, ya que $\Delta = |A - a| = 0.03 < \frac{1}{2}$ x 10^{-1}.

Ejemplo 22. Con respecto a los números exactos

$$\frac{28}{8} \qquad \sqrt[3]{31} \qquad \pi \qquad log1386$$

las aproximaciones respectivas

3.1	3.14	3.142	3.1418	tienen, respectivamente,
2	3	4	5	dígitos exactos ya que

$$\left|\frac{28}{8} - 3.1\right| = 0.0111... < \frac{1}{2}\text{x}10^{-1}$$

$$\left|\sqrt[3]{31} - 3.14\right| = 1.3806\text{x}10^{-3} < \frac{1}{2}\text{x}10^{-2}$$

$$\left|\pi - 3.142\right| = 4.07346\,\text{x}10^{-4} < \frac{1}{2}\text{x}10^{-3}$$

$$\left|log1386 - 3.1418\right| = 3.68\text{x}10^{-5} < \frac{1}{2}\text{x}10^{-4}.$$

El termino "n dígitos exactos" no debe tomarse literalmente, esto es, no es necesariamente cierto que en un número aproximado dado a con n dígitos exactos, sus primeros n dígitos significativos coincidan con los correspondientes dígitos del número exacto A. Por ejemplo, el número aproximado a = 9.995 utilizado en lugar del número exacto A = 10 tiene tres dígitos exactos (pues: |10-9.995| = 0.005 ≤ $\frac{1}{2}$ x 10^{-2}), aun cuando difieran todos los dígitos de ambos números. No obstante, se encuentra en muchos casos que los dígitos exactos del número aproximado son los mismos que los dígitos correspondientes del número exacto.

Notación científica. Existe una notación similar a la forma de una unidad numérica, frecuentemente utilizada en la impresión de tablas y en la tabulación de datos de observación, llamada *notación estándar o forma científica.*

Un número *n* escrito en notación estándar se lo expresa

$$n = n'' \text{ x } 10^{v}$$

donde *n*'' tiene los mismos dígitos que *n*' en la forma de unidad numérica, pero a diferencia de ésta tiene un dígito no nulo a la izquierda del punto decimal.

Ejemplo 23

$22.75 = 2.275 \text{ x } 10^{1}$
$1.76 = 1.76 \text{ x } 10^{0}$
$0.0071 = 7.1 \text{ x } 10^{-3}$
$945 = 9.45 \text{ x } 10^{2}$
$0 = 0.00 \text{ x } 10^{0}$

Esta notación es particularmente útil cuando los números son excesivamente pequeños o grandes. Así, por ejemplo, la carga de un electrón es

$$e = 0.00000000000000000016 \text{ cb} = 1.6 \text{ x } 10^{-19}\text{cb}.$$

1.8 Redondeo de números

Considérese un número *a*, aproximado o exacto, escrito en sistema decimal. Con frecuencia se precisa redondear este número, es decir, reemplazarlo por un número a_1 que tenga un número menor de dígitos significativos. El número a_1 se elige de forma que el error de redondeo $|a_1 - a|$ sea mínimo.

1.8.1. Regla de redondeo

Para redondear un número a *n* dígitos significativos elimínense todos los dígitos a la derecha del dígito significativo de lugar *n* o reemplácelos por ceros, si estos ceros son necesarios para mantener un valor relativo.

Al ejecutar esta operación obsérvese lo siguiente:

1. Si el primero de los dígitos despreciados es menor de 5, déjense como están los dígitos restantes.
2. Si el primer dígito despreciado excede de 5, añádase uno al último dígito de los que quedan.

3. Si el primer dígito despreciado es exactamente 5, los usos varían. Algunos tratan este caso como el precedente; otros hacen lo siguiente:
4. Si el primer dígito despreciado es exactamente 5 y hay dígitos no nulos entre los despreciados, se añade al último dígito conservado una unidad.
5. Sin embargo, si el primer dígito despreciado es exactamente 5 y el resto de los dígitos despreciados son ceros, el último dígito conservado se deja tal como está si es par y se incrementa en una unidad si es impar (regla del dígito par).

En otras palabras, si al redondear un número se desprecia menos de media unidad del último dígito conservado, todos los dígitos retenidos permanecen inalterables; pero si la parte despreciada del número es superior a media unidad del último lugar retenido, el dígito de este lugar se incrementa en una unidad. En el caso excepcional en el que la parte despreciada sea exactamente igual a media unidad del último orden de la retenida, se hace uso de la regla del dígito par o bien se lo trata como en el caso anterior. El razonamiento que está detrás de la regla del dígito par es aparentemente plausible, pero en la práctica efectiva no nos interesa cual sistema se usa, todos son válidos.

Es evidente que al aplicar la regla de redondeo, el error de redondeo, que es la omisión del resto de cifras significativas, no excede de media unidad del orden del último dígito significativo conservado.

Ejemplo 24. Redondeando el número

$$\pi = 3.1415926535\ldots$$

para 5, 4 y 3 dígitos significativos, se tienen los números aproximados 3.1416, 3.142, 3.14 con errores absolutos inferiores a

$$\frac{1}{2} \text{ x } 10^{-4}, \quad \frac{1}{2} \text{x } 10^{-3}, \frac{1}{2} \text{x } 10^{-2}.$$

Ejemplo 25. Redondeando el número 1.2500 a dos dígitos significativos, se obtiene el número 1.2 con un error absoluto igual a $\frac{1}{2} \text{ x } 10^{-1} = 0.05$.

Ejemplo 26

	Redondeado a			
Número	**5 cifras signif.**	**4 cifras signif.**	**3 cifras signif.**	**2 cifras signif.**
32.0769	32.077	32.08	32.1	32
0.856025	0.85602	0.8560	0.856	0.86
123456	123460	123500	123000	120000
1234.56	1234.6	1235	1230	1200
0.999777	0.99978	0.9998	1.00	1.0
1.34996	1.3500	1.350	1.35	1.3

La exactitud de un número aproximado no depende del número de dígitos significativos sino del número de dígitos significativos exactos. Cuando un número aproximado contiene un exceso de dígitos significativos incorrectos, se acude al redondeo. Como guía puede utilizarse la siguiente regla práctica: cuando se ejecutan cálculos aproximados, el número de dígitos significativos de los resultados intermedios no debe exceder del número de dígitos exactos en más de una o dos unidades. El resultado final no debe contener más de un dígito significativo en exceso sobre el número de dígitos exactos. Esta regla produce un gran ahorro de tiempo (reduciendo el número de dígitos), sin poner en peligro la exactitud del cálculo. La retención de dígitos adicionales tiene el significado de que normalmente el error estimado de los resultados se hace relativo a la versión más desfavorable, y el error real puede ser apreciablemente menor que el teórico. Por lo tanto, en muchos casos algunos dígitos significativos considerados como inexactos realmente son exactos.

También se redondean números exactos que contienen demasiados o bien un número infinito de dígitos significativos, dependiendo de la exactitud general requerida del cálculo. Nótese que si un número exacto A se redondea a n dígitos significativos mediante la regla de redondeo, el número aproximado a resultante tendrá n dígitos exactos (en el sentido de la Definición 11).

A continuación se demostrará un teorema que relaciona la magnitud del error relativo de un número aproximado con el número de dígitos exactos en dicho número.

Teorema. Si un número aproximado positivo a tiene n dígitos exactos, el error relativo δ de este número no excede de $\left(\frac{1}{10}\right)^{n-1}$ dividido por el primer dígito significativo del número dado, o sea

$$\delta \leq \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1}$$

donde α_m es el primer dígito significativo del número *a*.

Demostración. Supongamos que el número

$$a = \alpha_m 10^m + \alpha_{m-1}10^{m-1} + \alpha_{m-2}10^{m-2} + \ldots + \alpha_{m-n+1}10^{m-n+1} + \ldots \ (\alpha_m \geq 1)$$

sea un valor aproximado del número exacto *A* y consideremos que tiene *n* dígitos exactos. Por definición, se tiene entonces

$$\Delta = | A - a | \leq \frac{1}{2} 10^{m-n+1}$$

de donde,

$$A \geq a - \frac{1}{2} 10^{m-n+1}.$$

Esta desigualdad se acrecienta aun más si el número *a* se reemplaza por otro, $\alpha_m 10^m$, definitivamente menor

$$A \geq \alpha_m 10^m - \frac{1}{2} 10^{m-n+1} = \frac{1}{2} 10^m \left(2\alpha_m - \frac{1}{10^{n-1}}\right). \qquad (14)$$

El lado derecho de (14) es mínimo para $n = 1$. Por lo tanto,

$$A \geq \frac{1}{2} 10^m \left(2\alpha_m - 1\right) \qquad (15)$$

o, ya que

$$\left(2\alpha_m - 1\right) = \alpha_m + (\alpha_m - 1) \geq \alpha_m$$

se deduce que

$$A \geq \frac{1}{2} 10^m \alpha_m.$$

En consecuencia,

$$\delta = \frac{\Delta}{A} \leq \frac{\frac{1}{2}10^{m-n+1}}{\frac{1}{2}10^m \alpha_m} = \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1}.$$

Por lo tanto,

$$\delta \leq \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1}. \tag{16}$$

Nota. La desigualdad (15) puede utilizarse para obtener un estimado más exacto del error relativo δ.

Corolario 1. Como cota del error relativo del número aproximado positivo a puede tomarse

$$\delta_a = \frac{1}{\alpha_m}\left(\frac{1}{10}\right)^{n-1} \tag{17}$$

donde α_m es el primer dígito significativo del número aproximado positivo a.

Corolario 2. Si el número aproximado positivo a tiene más de dos dígitos exactos, esto es $n > 2$, entonces en la práctica se tiene

$$\delta_a = \frac{1}{2\alpha_m}\left(\frac{1}{10}\right)^{n-1}. \tag{18}$$

Realmente, para $n > 2$ puede despreciarse $\frac{1}{10^{n-1}}$ en la desigualdad (14). Entonces,

$$A \geq \frac{1}{2}10^m 2\alpha_m = \alpha_m 10^{\,m}$$

de donde,

$$\delta = \frac{\Delta}{A} \leq \frac{\frac{1}{2}10^{m-n+1}}{\alpha_m 10^m} = \frac{1}{2\alpha_m}\left(\frac{1}{10}\right)^{n-1}.$$

Por consiguiente,

$$\delta_a = \frac{1}{2\alpha_m}\left(\frac{1}{10}\right)^{n-1}.$$

Ejemplo 27. ¿Cuántos dígitos deben tomarse en el cálculo de $\sqrt{20}$ de forma que el error relativo no exceda de 0.1%?

Como el primer dígito es 4 se tiene que $\alpha_m = 4$ y $\delta \leq 0.001$.

Tenemos

$$\frac{1}{4\text{x}10^{n-1}} \leq 0.001$$

y, por lo tanto,

$$10^{n-1} \geq 250$$

de donde,

$$n \geq 4.$$

Ejemplo 28. ¿Cuál es la cota del error relativo si se toma $a = 3.14$ en lugar del número π?

Puesto que $\alpha_m = 3$ y $n = 3$, entonces

$$\delta_a = \frac{1}{2\text{x}3}\left(\frac{1}{10}\right)^{3-1} = \frac{1}{600} = \frac{1}{6}\%.$$

1.9. Operaciones aritméticas. Deducción elemental de las reglas para el cálculo del error de operaciones (exactas) con números aproximados.

1.9.1. Error de una suma.

Supongamos que $a_3 = a_1 + a_2$ es una aproximación de $A_3 = A_1 + A_2$. (Por simplicidad, suponemos que los números son positivos).

Entonces, como $A_3 = A_1 + A_2$ se tiene que

$$a_3 + \Delta a_3 = a_1 + \Delta a_1 + a_2 + \Delta a_2 = (a_1 + a_2) + (\Delta a_1 + \Delta a_2)$$

de donde,

$$\Delta a_3 = \Delta a_1 + \Delta a_2$$

y, por lo tanto,

$$|\Delta a_3| \leq |\Delta a_1| + |\Delta a_2|. \qquad (19)$$

Por inducción, esta fórmula puede ser extendida a un número finito n de sumandos, o sea, si $a = a_1 + a_2 + \ldots + a_n$ es una aproximación de $A = A_1 + A_2 + \ldots + A_n$, entonces

$$|\Delta a| \leq |\Delta a_1| + |\Delta a_2| + \ldots + |\Delta a_n|. \qquad (20)$$

Se tiene así que "el error absoluto de una suma de varios números aproximados no excede de la suma de los errores absolutos de los números".

Como cota del error absoluto de una suma puede tomarse la suma de las cotas de los errores absolutos de cada uno de los términos

$$\Delta_a = \Delta_{a_1} + \Delta_{a_2} + \ldots + \Delta_{a_n} .$$

De aquí se deduce que la cota del error absoluto de una suma no puede ser menor que la cota del error absoluto del término con menor exactitud (en el sentido del error absoluto), es decir, del término que tenga el máximo error absoluto. En consecuencia, por muy grande que sea el grado de exactitud de los otros términos, no se puede con ellos aumentar la exactitud de la suma. Por esta razón, no tiene sentido retener dígitos en exceso en los términos más exactos.

Para el cálculo de la cota del error relativo de la suma de números aproximados usaremos

$$\delta_a = \frac{\Delta_a}{A} \tag{21}$$

en donde, $a = a_1 + a_2 + \ldots + a_n$ ($a_i > 0$, $i = 1, 2, \ldots, n$) es una aproximación a la magnitud exacta $A = A_1 + A_2 + \ldots + A_n$ ($A_i > 0$, $i = 1, 2, \ldots, n$) y

$$\Delta_a = \Delta_{a_1} + \Delta_{a_2} + \ldots + \Delta_{a_n} = \sum_{i=1}^{n} \Delta_{a_i} .$$

Como

$$\delta_{a_i} = \frac{\Delta_{a_i}}{A_i}, \qquad i = 1, 2, \ldots, n$$

se deduce que

$$\Delta_{a_i} = A_i \, \delta_{a_i} . \qquad i = 1, 2, \ldots, n \tag{22}$$

Sustituyendo esta expresión en (21), resulta

$$\delta_a = \frac{A_1 \delta_{a_1} + A_2 \delta_{a_2} + \ldots + A_n \delta_{a_n}}{A_1 + A_2 + \ldots + A_n} .$$

Hagamos $\overline{\delta}$ el mayor de los errores relativos δ_{a_i} ó $\delta_{a_i} \leq \overline{\delta}$. Entonces,

$$\delta_a \leq \frac{\overline{\delta}(A_1 + A_2 + ... + A_n)}{A_1 + A_2 + ... + A_n} = \overline{\delta}.$$

En consecuencia,

$$\delta_a \leq \overline{\delta} = máx\{\delta_{a_1}, \delta_{a_2}, ..., \delta_{a_n}\}. \quad (23)$$

Esto es, "si todos los términos vienen afectados por el mismo signo, la cota del error relativo de su suma no excede de la máxima cota del error relativo de cualquiera de ellos".

1.9.2 Error de una diferencia.

Supongamos que $a_3 = a_1 - a_2$ es una aproximación de $A_3 = A_2 - A_1$. (Suponemos que $a_1 > a_2 > 0$, $A_1 > A_2 > 0$). Entonces,

$$a_3 + \Delta a_3 = (a_1 + \Delta a_1) - (a_2 + \Delta a_2) = (a_1 - a_2) + (\Delta a_1 - \Delta a_2)$$

de donde,

$$\Delta a_3 = \Delta a_1 - \Delta a_2$$

y, por lo tanto,

$$|\Delta a_3| \leq |\Delta a_1| + |\Delta a_2|. \quad (24)$$

Así, "el error absoluto de una diferencia de dos magnitudes aproximadas no excede de la suma de los errores absolutos de dichas magnitudes".

Como cota del error absoluto de una diferencia puede tomarse la suma de las cotas de los errores absolutos del minuendo y del sustraendo

$$\Delta_{a_3} = \Delta_{a_1} + \Delta_{a_2}. \quad (25)$$

A partir de (24), (25) y (20) se puede generalizar el concepto al caso de una suma algebraica: "el error absoluto de una suma algebraica de varios números aproximados no excede de la suma de los errores absolutos de los números, y como cota del error absoluto de una suma algebraica puede tomarse la suma de las cotas de los errores absolutos de cada uno de los términos".

La cota del error relativo de la diferencia de dos números aproximados está dada por

$$\delta_{a_3} = \frac{\Delta_{a_1} + \Delta_{a_2}}{A_1 - A_2} \approx \frac{\Delta_{a_1} + \Delta_{a_2}}{a_1 - a_2} = \frac{\Delta_{a_3}}{a_1 - a_2}. \qquad (26)$$

Observación. Aumento del error al restar números muy próximos. Si los números aproximados a_1 y a_2 son números prácticamente iguales y tienen errores absolutos pequeños, el número A_3 es pequeño. De (26) se deduce que la cota del error relativo δ_{a_3}, en este caso, puede ser muy grande aun cuando los errores relativos máximos del minuendo y sustraendo (δ_{a_1} y δ_{a_2}) permanezcan pequeños. Esto conduce a una *pérdida de exactitud*.

Ejemplo 29. Calculemos la diferencia entre los dos números siguientes

$$a_1 = 47.132 \qquad \text{y} \qquad a_2 = 47.111$$

donde cada uno de los cuales tiene los 5 dígitos significativos exactos.

Efectuando la resta, se tiene

$$a_3 = 47.132 - 47.111 = 0.021.$$

La diferencia a_3 tiene 2 dígitos significativos únicamente, de los cuales el último es incierto ya que la cota del error absoluto de la diferencia es

$$\Delta_{a_3} = 0.0005 + 0.0005 = 0.001.$$

Las cotas de los errores relativos del minuendo, sustraendo y diferencia son, respectivamente

$$\delta_{a_1} = \frac{0.0005}{47.132} \approx 0.00001$$

$$\delta_{a_2} = \frac{0.0005}{47.111} \approx 0.00001$$

$$\delta_{a_3} = \frac{0.001}{0.021} \approx 0.05.$$

La cota del error relativo de la diferencia es aproximadamente 5000 veces mayor que las cotas de los errores relativos de los valores originales.

Por lo tanto, es deseable en cálculos aproximados transformar las expresiones en las cuales el cálculo de valores aproximados conduzca a la sustracción de números aproximadamente iguales.

Ejemplo 30. Hallemos la diferencia $\sqrt{2.01} - \sqrt{2}$ con 3 dígitos exactos.

Como $\sqrt{2.01} = 1.41774469\ldots$ y $\sqrt{2} = 1.41421356\ldots$ el resultado deseado es $0.00353 = 3.53 \times 10^{-3}$.

Este resultado puede obtenerse escribiendo

$$\sqrt{2.01} - \sqrt{2} = \frac{0.01}{\sqrt{2.01} + \sqrt{2}}$$

y hallar, entonces, las raíces $\sqrt{2.01}$ y $\sqrt{2}$ con 3 dígitos exactos, como puede verse

$$\frac{0.01}{1.42 + 1.41} = \frac{0.01}{2.83} = 10^{-2} \times 3.53 \times 10^{-1} = 3.53 \times 10^{-3}.$$

De lo anterior puede sacarse una regla práctica: en los cálculos aproximados evítese en lo posible la resta de dos números aproximadamente iguales; si es necesario restar tales números, tómese el minuendo y el sustraendo con un número suficiente de dígitos exactos adicionales (si es posible). Por ejemplo, si se desea la diferencia de dos números, a_1 y a_2, con n dígitos significativos exactos y se sabe que los m primeros dígitos significativos desaparecerán en la resta, debe comenzarse entonces con $m + n$ dígitos significativos en cada uno de los números (a_1 y a_2).

En estos dos últimos ejemplos se produce el fenómeno de la *cancelación catastrófica*. Más adelante volveremos sobre este tema.

1.9.3 Error de un producto.

Supongamos que $a_3 = a_1\, a_2$ es una aproximación de $A_3 = A_1\, A_2$. (Por simplicidad, consideramos que trabajamos con números positivos). Entonces,

$$a_3 + \Delta a_3 = (a_1 + \Delta a_1)\,(a_2 + \Delta a_2) = a_1\, a_2 + a_1\, \Delta a_2 + \Delta a_1\, a_2 + \Delta a_1\, \Delta a_2.$$

Despreciando el término $\Delta a_1\, \Delta a_2$ (pues es muy pequeño frente a los otros), se tiene

$$\Delta a_3 \approx a_1 \Delta a_2 + a_2 \Delta a_1.$$

Tomando valor absoluto en la expresión anterior, se tiene

$$| \Delta a_3 | \leq | a_1 \Delta a_2 | + | a_2 \Delta a_1 |. \qquad (27)$$

Dividiendo ambos miembros por $a_3 = a_1\, a_2$, se obtiene el error relativo del producto. En efecto,

$$\left|\frac{\Delta a_3}{a_3}\right| \leq \left|\frac{\Delta a_1}{a_1}\right| + \left|\frac{\Delta a_2}{a_2}\right|. \qquad (28)$$

Como a los efectos prácticos

$$\left|\frac{\Delta a_i}{a_i}\right| \approx \left|\frac{\Delta a_i}{A_i}\right| = \delta_i \qquad i = 1, 2, 3$$

entonces (28) puede escribirse

$$\delta_3 \leq \delta_1 + \delta_2$$

o poniendo $\delta = \delta_3$

$$\delta \leq \delta_1 + \delta_2 \qquad (29)$$

donde δ_1 y δ_2 son los errores relativos de los factores a_1, a_2, respectivamente y δ es el error relativo del producto.

Así, "el error relativo de un producto de dos números aproximados no excede de la suma de los errores relativos de los números".

Es evidente que (29) es también cierto si los factores a_i, para $i = 1, 2$ tienen signos diferentes. Además, este resultado puede generalizarse a un producto de n factores (por inducción). Así (suponemos que $a = a_1\, a_2 \ldots a_n$ es una aproximación de $A = A_1 A_2 \ldots A_n$)

$$\delta \leq \delta_1 + \delta_2 + \ldots + \delta_n \ . \qquad (30)$$

La cota del error relativo de un producto es igual a la suma de las cotas de los errores relativos de los factores; esto es,

$$\delta_a = \delta_{a_1} + \delta_{a_2} + \ldots + \delta_{a_n}. \qquad (31)$$

1.9.4. Error de un cociente. Supongamos que $a_3 = a_1/a_2$ es una aproximación de $A_3 = A_1/A_2$. Entonces,

$$a_3 + \Delta a_3 = \frac{a_1 + \Delta a_1}{a_2 + \Delta a_2} = \frac{a_1 + \Delta a_1}{a_2 . \left(1 + \frac{\Delta a_2}{a_2}\right)} = \left(\frac{a_1 + \Delta a_1}{a_2}\right)\left(1 + \frac{\Delta a_2}{a_2}\right)^{-1} = {}^{(*)\text{ ver Nota}}$$

$$= \left(\frac{a_1 + \Delta a_1}{a_2}\right)\left(1 - \frac{\Delta a_2}{a_2} + \text{términos que se desprecian pues son de orden superior al primero de } \Delta a_2\right) \approx$$

$$\approx \frac{a_1}{a_2} - \frac{a_1 \Delta a_2}{{a_2}^2} + \frac{\Delta a_1}{a_2} - \underbrace{\frac{\Delta a_1 \Delta a_2}{{a_2}^2}}_{\text{se desprecia}} \approx \frac{a_1}{a_2} + \frac{\Delta a_1}{a_2} - \frac{a_1 \Delta a_2}{{a_2}^2}.$$

Luego,

$$\Delta a_3 \approx \frac{\Delta a_1}{a_2} - \frac{a_1 \Delta a_2}{{a_2}^2}$$

y, por lo tanto,

$$|\Delta a_3| \leq \left|\frac{\Delta a_1}{a_2}\right| + \left|\frac{a_1 \Delta a_2}{{a_2}^2}\right|. \tag{32}$$

Dividiendo ambos miembros de (32) por el módulo de $a_3 = a_1/a_2$, se obtiene el error relativo del cociente

$$\left|\frac{\Delta a_3}{a_3}\right| \leq \left|\frac{\Delta a_1}{a_1}\right| + \left|\frac{\Delta a_2}{a_2}\right| \tag{33}$$

esto es,

$$\delta \leq \delta_1 + \delta_2 \tag{34}$$

puesto que a los efectos prácticos

$$\delta_i = \left|\frac{\Delta a_i}{A_i}\right| \approx \left|\frac{\Delta a_i}{a_i}\right| \quad (i = 1, 2) \qquad \text{y} \qquad \delta \underbrace{=}_{\text{Not.}} \delta_3 = \left|\frac{\Delta a_3}{A_3}\right| \approx \left|\frac{\Delta a_3}{a_3}\right|.$$

Luego, de (34), “el error relativo de un cociente no excede de la suma de los errores relativos del dividendo y del divisor”, y “la cota del error relativo de un cociente es igual a la suma de las cotas de los errores relativos del dividendo y del divisor”, o sea

$$\delta_a = \delta_{a_1} + \delta_{a_2}. \tag{35}$$

Notación. $a = a_3$.

Nota. Recordemos que el desarrollo de $f(x)$ alrededor de $x = a$ por medio de su serie de Taylor es

$$f(x) = f(a) + f'(a)(x - a) + f''(a)\frac{(x-a)^2}{2!} + \ldots + f^n(a)\frac{(x-a)^n}{n!} + \text{Error}$$

y que el desarrollo de Taylor de $f(x)$ alrededor de $x = 0$ recibe el nombre de serie de Maclaurin.

En (*) usamos este concepto para $f(x) = (1 + x)^m$, $m \in R$, alrededor de $x = 0$; esto es,

$$(1 + x)^m = 1 + m\,x + m(m - 1)\frac{x^2}{2!} + \ldots + m(m - 1)\ldots(m - n + 1)\frac{x^n}{n!} + \text{Error.}$$

1.9.5. Error de una potencia y de una raíz. Si $a = (a_1)^p$ es una aproximación de $A = (A_1)^p$ (p cualquier número natural), entonces

$$a + \Delta a = (a_1 + \Delta a_1)^p = a_1{}^p\left(1 + \frac{\Delta a_1}{a_1}\right)^p = a_1{}^p\Big(1 + p\,\frac{\Delta a_1}{a_1} + \underbrace{\ldots}_{\text{se desprecian}}\Big)$$

$$\approx a_1{}^p + p\,a_1{}^p\,\frac{\Delta a_1}{a_1}.$$

Luego,

$$\Delta a \approx p a_1{}^p\,\frac{\Delta a_1}{a_1} \quad \text{, de donde,} \quad |\Delta a| \approx p\left|a_1{}^p\,\frac{\Delta a_1}{a_1}\right|.$$

Tenemos entonces que el error relativo es

$$\left|\frac{\Delta a}{a}\right| \approx p\left|\frac{\Delta a_1}{a_1}\right|$$

y la cota del error relativo es

$$\delta_a = p\,\delta_{a_1}$$

o sea, "la cota del error relativo de la potencia p-ésima de un número es p veces la cota del error relativo de dicho número".

Si ahora consideramos $a = \sqrt[p]{a_1} = a_1^{\frac{1}{p}}$ como aproximación de $A = \sqrt[p]{A_1}$ se tiene, evidentemente, que

$$\left|\frac{\Delta a}{a}\right| \approx \frac{1}{p}\left|\frac{\Delta a_1}{a_1}\right|$$

y que

$$\delta_a = \frac{1}{p}\,\delta_{a_1}$$

esto es, "la cota del error relativo de una raíz p-ésima es $1/p$ veces la cota del error relativo del radicando".

1.10. Fórmula general para el cálculo de errores

El objetivo primordial del cálculo de errores es: dado los errores de un cierto conjunto de cantidades, determinar el error de una función dada de estas cantidades.

Supongamos una función diferenciable

$$y = f(x_1, x_2, \ldots, x_n)$$

y sean $|\Delta x_i|$ (i = 1, 2, …, n) los errores absolutos de los argumentos de la función. En un caso práctico, $|\Delta x_i|$ son magnitudes extraordinariamente pequeñas cuyos productos, cuadrados y potencias de orden superior pueden despreciarse.

Analizaremos primero el caso de una variable; esto es, $y = f(x)$.

El error de la función es, entonces

$$\Delta y = \Delta f = f(x + \Delta x) - f(x)$$

o bien

$$y + \Delta y = f(x + \Delta x) \tag{36}$$

esto es, el error cometido en el argumento x da lugar al error en la función y que llamaremos Δy. Desarrollando a $f(x + \Delta x)$ por serie de Taylor, se tiene

$$f(x + \Delta x) \approx f(x) + f'(x)\,\Delta x \tag{37}$$

Reemplazando (37) en (36), se obtiene

$$y + \Delta y \approx f(x) + f'(x)\,\Delta x$$

de donde,

$$\Delta y \approx f'(x)\,\Delta x.$$

Luego,

$$|\Delta y| \approx |f'(x)|\,|\Delta x|. \tag{38}$$

Denotando con Δ_x la cota del error absoluto del argumento x y con Δ_y la cota del error absoluto de la función y, se tiene

$$\Delta_y = |f'(x)|\Delta_x. \tag{39}$$

Dividiendo ambos lados de (38) por $|y|$ tendremos un valor aproximado del error relativo de la función y, que denotamos con δ

$$\delta \approx \frac{|f'(x)|}{|f(x)|}|\Delta x| = \left| x\frac{f'(x)}{f(x)} \right| \delta_1$$

es decir (hemos llamado δ_1 al valor aproximado del error relativo del argumento x)

$$\delta \approx \left| x\frac{f'(x)}{f(x)} \right| \delta_1. \tag{40}$$

Denotando con δ_y la cota del error relativo de la función y, y con δ_x la cota del error relativo del argumento x, se tiene

$$\delta_y = \left| x\frac{f'(x)}{f(x)} \right| \delta_x \tag{41}$$

Veamos ahora el caso para una función de varias variables

$$y = f(x_1, x_2, \ldots, x_n) = f(\vec{x}).$$

Entonces,

$$y + \Delta y = f(x_1 + \Delta x_1, x_2 + \Delta x_2, \ldots, x_n + \Delta x_n).$$

Usando el desarrollo en series de Taylor para una función de n dimensiones, se obtiene

$$y + \Delta y \approx f(x_1, x_2, \ldots, x_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\Delta x_i$$

de donde,

$$\Delta y \approx \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \Delta x_i .$$

Luego,

$$|\Delta y| \leq \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| |\Delta x_i| . \tag{42}$$

Denotando con Δ_{x_i} (i = 1, 2, …, n) las cotas de los errores absolutos de los argumentos x_i y con Δ_y la cota del error absoluto de la función y, se tiene

$$\Delta_y = \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| \Delta_{x_i} . \tag{43}$$

Dividiendo ambos lados de la desigualdad (42) por el valor absoluto de la función y tendremos un valor aproximado del error relativo de la función y, que denotamos con δ

$$\delta \leq \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| \left| \frac{1}{f(\vec{x})} \right| |\Delta x_i| = \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| \left| \frac{x_i}{f(\vec{x})} \right| \delta_i$$

(denotando con δ_i un valor aproximado del error relativo del argumento x_i, para i = 1, 2, …, n), es decir

$$\delta \leq \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| \left| \frac{x_i}{f(\vec{x})} \right| \delta_i . \tag{44}$$

Denotando con δ_y la cota del error relativo de la función y, y con δ_{x_i} la cota del error relativo del argumento x_i, para i = 1, 2, …, n, se tiene

$$\delta_y = \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| \left| \frac{x_i}{f(\vec{x})} \right| \delta_{x_i} \tag{45}$$

Observación sobre $\left| \frac{\partial f}{\partial x_i} \right|$**.** Puesto que en la práctica debemos contentarnos con trabajar con números aproximados (y por ello en la mayoría de los casos calculamos las cotas del error absoluto y del error relativo), es común adoptar cotas superiores de estas derivadas. Para ello, se podrá someter a tales derivadas a todas las modificaciones que se quiera a

condición de que éstas modificaciones tengan por efecto mayorar la expresión, pero jamás disminuirla. Estas modificaciones consistirán principalmente en cambiar los signos de algunos términos o también en suprimir algunos términos. No se puede enunciar sobre este asunto ninguna regla general, es necesario guiarse por la experiencia y sobre todo el buen sentido.

Una vez calculadas las derivadas parciales, se evalúan convenientemente en los números aproximados.

Ejemplo 31. Si $x_1 = 2.31 \pm 0.02$ y $x_2 = 1.42 \pm 0.03$, ¿cuál es la cota del error absoluto para $x_1 - x_2$?

Como $x_1 - x_2 = (x'_1 - x'_2) \pm (\Delta_{x'_1} + \Delta_{x'_2})$, entonces

$$x'_1 - x'_2 = 0.89 \quad \text{y} \quad \Delta_{x'_1} + \Delta_{x'_2} = 0.05.$$

Luego,

$$x_1 - x_2 = 0.89 \pm 0.05.$$

Ejemplo 32. Hallemos las cotas del error absoluto y del error relativo del volumen de una esfera $V = \frac{1}{6}\pi d^3$, si el diámetro es $d = 3.7$ cm con un error de 0.05 cm y $\pi \approx 3.14$.

Considerando a π y d como magnitudes variables, calculamos las derivadas parciales correspondientes.

$$\frac{\partial V}{\partial \pi} = \frac{1}{6} d^3 = 8.442166\ldots < 8.45$$

$$\frac{\partial V}{\partial d} = \frac{1}{2}\pi d^2 = 21.4933\ldots < 21.5.$$

Luego, usando la fórmula (42), obtenemos

$$|\Delta V| \leq 8.45 \text{ x } 0.0016 + 21.5 \text{ x } 0.05 = 1.08852 \text{ cm}^3 \approx 1.1 \text{ cm}^3.$$

Por lo tanto, la cota del error absoluto del volumen es

$$\Delta_V = 1.1 \text{ cm}^3$$

y así,

$$V = \frac{1}{6}\pi d^3 \approx 26.5 \text{ cm}^3 \pm 1.1 \text{ cm}^3.$$

De aquí que la cota del error relativo del volumen sea

$$\delta_V = \frac{1.1}{26.5} = 0.041509\ldots \approx 4.2\ \%.$$

Observemos que como en π conocíamos el decimal siguiente a 4 en 3.14, entonces tomamos como cota del error absoluto al tomar el valor aproximado 3.14 en lugar de π a 0.0016. Eventualmente, también se podría haber tomado como cota el valor de $.5\text{x}10^{-2}$ obteniendo un Δ_V mayor, pero como hemos dicho anteriormente, siempre trataremos de tomar el más chico (siempre que esto sea posible) de los infinitos que exceden a $|\Delta V|$.

Ejemplo 33. Calculemos la siguiente suma alternada

$$1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\frac{1}{9}-\frac{1}{11}+\frac{1}{13}-\frac{1}{15}+\frac{1}{17}-\frac{1}{19}$$

tomando cinco cifras decimales.

Disponiendo en dos columnas los términos positivos y negativos (con la finalidad de que no se produzca la cancelación de cifras significativas)

1.00000		0.33333		
0.20000		0.14286		
0.11111		0.09091		
0.07692		0.06667		
0.05882		0.05263		
1.44685	-	0.68640	=	0.76045.

Hay, en esta suma, dos términos exactos (el 1° y el 3°) y en los ocho restantes se comete un error que no excede a $5\text{x}10^{-6}$ puesto que se ha "redondeado" cada sumando. En la suma habrá pues un error de $8\text{x}5\text{x}10^{-6} = 4\text{x}10^{-5}$.

Si consideramos 4 decimales en el resultado cometemos un error de $5\text{x}10^{-5}$ que sumado al anterior alcanza a ser de $9\text{x}10^{-5} < 1\text{x}10^{-4}$. Por lo tanto, la suma alternada de los recíprocos de los 10 primeros números impares es

$$0.7604 \pm 0.0001.$$

Ejemplo 34. Se han medido con errores de 1 cm los lados de un rectángulo obteniéndose las medidas: $x = 2.36$ m, $y = 3.78$ m. Calculemos el error que se comete en la determinación del área A.

Siendo $A = xy$, de acuerdo a la fórmula (42), resulta

$$| \Delta A | \leq y\,|\Delta x| + x\,|\Delta y| \leq 3.78 \text{ x } 0.01 + 2.36 \text{ x } 0.01 = 0.0614 \text{ m}^2$$

de donde,

$$| \Delta A | \leq 0.0614 \text{ m}^2.$$

El cálculo directo nos da $A \approx 2.36$ m x 3.78 m = 8.9208 m^2, pero siendo el error de 0.0614 sólo se puede asegurar que el área A ésta comprendida entre 8.8594 m^2 y 8.9822 m^2.

Así, si redondeamos 8.9208 a un decimal obtenemos un error inferior a $3\text{x}10^{-2}$ que sumado al anterior nos da $0.0914 < 0.1$.

Luego, podemos tomar

$$8.9 \text{ m}^2 \pm 0.1 \text{ m}^2$$

o bien 8.9 m$^2 \pm$ 0.2 m^2, si consideramos que el error de redondeo es inferior a $5\text{x}10^{-2}$ que sumado a 0.0614 da $0.1114 < 0.2$.

Ejemplo 35. Calculemos $y = x_1^{\,2} - x_2$ para $x_1 = 1.03 \pm 0.01$, $x_2 = 0.45 \pm 0.01$.

Como $\left|\dfrac{\partial y}{\partial x_1}\right| = | 2x_1 | \leq 2.1$, $\left|\dfrac{\partial y}{\partial x_2}\right| = | -1 | = 1$, entonces

$$| \Delta y | \leq 2.1 \text{ x } 0.01 + 1 \text{ x } 0.01 = 0.031$$

de donde,

$$| \Delta y | \leq 0.031.$$

El cálculo directo nos da $y \approx 0.6109$, y considerando 3 decimales en este resultado obtenemos un nuevo error que no excede de $5\text{x}10^{-4}$ que sumado al anterior alcanza a ser de $0.0315 < 0.032$.

Luego, encontramos que

$$y = 0.611 \pm 0.032.$$

Si redondeamos el resultado a 2 decimales, entonces

$$y = 0.61 \pm 0.04.$$

Ejemplo 36. El tiempo de una oscilación simple de un péndulo ideal de longitud l es

$$t = \pi\sqrt{\frac{l}{g}}$$

siendo g la aceleración de la gravedad. Supuesto que se ha medido $l = 1.50$ m con un error de 0.005, calculemos el tiempo t y la aproximación lograda. Adoptamos el valor normal de $g = 9.81\text{m/seg}^2$ con un error de 0.01 y de $\pi = 3.14$ con un error de 0.002.

Siendo

$$\frac{\partial t}{\partial \pi} = \sqrt{\frac{l}{g}} = 0.3910... < 0.4$$

$$\frac{\partial t}{\partial l} = \frac{\pi}{2l}\sqrt{\frac{l}{g}} = 0.40927... < 0.41$$

$$\frac{\partial t}{\partial g} = -\frac{\pi}{2g}\sqrt{\frac{l}{g}}, \text{ de donde, } \left|\frac{\partial t}{\partial g}\right| = 0.06258... < 0.063$$

entonces

$$|\Delta t| \leq 0.4 \text{ x } 0.002 + 0.41 \text{ x } 0.005 + 0.063 \text{ x } 0.01 = 3.48 \text{ x } 10^{-3} \text{ seg}$$

de donde,

$$|\Delta t| \leq 3.48\text{x}10^{-3} \text{ seg.}$$

El cálculo directo nos da $t \approx 1.227837162\ldots$seg, y considerando tres decimales en este resultado obtenemos un nuevo error menor que $5\text{x}10^{-4}$ que sumado al anterior alcanza a ser de $0.00398 < 0.004$.

Luego, encontramos que

$$t = 1.228 \text{ seg} \pm 0.004 \text{ seg.}$$

Si redondeamos el resultado a 2 decimales, entonces

$$t = 1.23 \text{ seg} \pm 0.01 \text{ seg.}$$

Ejemplo 37. Calculemos la cota del error absoluto en la siguiente expresión en la cual se supone que los valores están redondeados (son exactos con n cifras significativas). Redondearemos el resultado a un número de cifras significativas.

$$0.124 - 3.5791 + 8.61312$$

Sea $f(a_1, a_2, a_3) = a_1 - a_2 + a_3$, donde, $a_1 = 0.124$, $a_2 = 3.5791$, $a_3 = 8.61312$, la aproximación a un valor exacto que designamos con $f(A_1, A_2, A_3) = A_1 - A_2 + A_3$ (y que desconocemos).

Como $\left|\frac{\partial f}{\partial a_i}\right| = 1$, para $i = 1, 2, 3$ y $\Delta_{a_1} = 5 \text{ x } 10^{-4}$, $\Delta_{a_2} = 5 \text{ x } 10^{-5}$, $\Delta_{a_3} = 5 \text{ x } 10^{-6}$,

Entonces

$$|\Delta f| \leq 5 \text{ x } 10^{-4} + 5 \text{ x } 10^{-5} + 5 \text{ x } 10^{-6} = 5.55 \text{ x } 10^{-4}$$

de donde

$$|\Delta f| \leq 5.6 \text{ x } 10^{-4}.$$

Del hecho que $f(a_1, a_2, a_3) = 5.15802$, sigue que el valor exacto de la suma cae dentro del rango: 5.15802 ± 0.00056, es decir, $5.15746 \leq f(A_1, A_2, A_3) \leq 5.15858$, y considerando tres decimales en el resultado obtenemos un nuevo error que no excede de $5\text{x}10^{-4}$ que sumado al anterior es inferior a $2\text{x}10^{-3}$. Luego, encontramos que

$$f(A_1, A_2, A_3) = 5.158 \pm 0.002.$$

Observemos que obtendríamos una mejor aproximación teniendo en cuenta que en realidad el error de redondeo que cometemos al tomar 3 decimales en el resultado es inferior a $1\text{x}10^{-4}$ que sumado a $5.6\text{x}10^{-4}$ nos da $6.6\text{x}10^{-4} < 1\text{x}10^{-3}$ y tendríamos, entonces

$$f(A_1, A_2, A_3) = 5.158 \pm 0.001.$$

Si redondeamos a 2 decimales, entonces

$$f(A_1, A_2, A_3) = 5.16 \pm 0.01.$$

Observación. Si se quiere calcular $f(\vec{x})$ para un $\vec{x} = \overrightarrow{x_0}$ estamos obligados, por lo general, a reemplazar $\overrightarrow{x_0}$ por $\overrightarrow{x'_0}$ con lo cual cometemos un error que proviene de esa sustitución. Este error, comúnmente llamado *error sistemático*, no excluye la posibilidad de que se cometan también *errores provenientes del cálculo*. Los errores sistemáticos preexisten a los de cálculo y no tienen nada que ver con ellos. Con la fórmula general para el cálculo de errores uno obtiene un límite superior del error sistemático. Si luego debemos redondear el resultado obtenido para $f\left(\overrightarrow{x'_0}\right)$, cometemos un error de cálculo que es independiente al ya calculado y que sumado al anterior nos dará el error total en el cálculo de la expresión $f\left(\vec{x}\right)$.

Ejemplo 38. Calculemos el valor de la función $f(x) = \dfrac{x+4}{x^3+1}$ para $x = \pi \approx 3.14$.

Como conocemos el decimal siguiente en π sabemos que $3.14 < \pi < 3.1416$, de donde, $\Delta_\pi = 0.0016$.

Debemos calcular $f'(x)$:

$$f'(x) = \frac{-2x^3 - 12x^2 + 1}{(x^3+1)^2}.$$

Como no podemos pensar en obtener el máximo exacto del valor absoluto de ésta expresión, entonces estudiaremos sus variaciones en el intervalo [3.14, 3.1416]. Tendremos allí una cota superior tomando el máximo del valor absoluto del numerador y el mínimo del valor absoluto del denominador. Aquí es fácil obtener el máximo del numerador exactamente, pero es más cómodo y más rápido de obtener mayorando de la siguiente forma: como x es positivo, se tiene evidentemente que

$$|f'(x)| = \left|\frac{-2x^3 - 12x^2 + 1}{(x^3+1)^2}\right| \le \frac{2x^3 + 12x^2 + 1}{(x^3+1)^2} \le \frac{2(3.1416)^3 + 12(3.1416)^2 + 1}{\left[(3.14)^3 + 1\right]^2} = 0.1776... < 0.18.$$

Luego,

$$|\Delta f| \le |f'(x)|\,\Delta\pi < 0.18 \text{ x } 0.0016 = 2.88 \text{ x } 10^{-4}$$

y, por lo tanto, obtenemos como límite superior del error sistemático a 2.88 x 10^{-4}. Esto significa que cuando reemplazamos el número exacto $\frac{\pi + 4}{\pi^3 + 1}$ por el número aproximado $\frac{3.14 + 4}{(3.14)^3 + 1}$(*), cometemos por este hecho un error del orden de 2.88x10^{-4}.

Los errores que cometemos en el cálculo efectivo de (*) se suman al anterior. Por lo tanto, como $\frac{3.14 + 4}{(3.14)^3 + 1} = \frac{7.14}{31.959144} = 0.2234102...$, si tomamos 0.223 cometemos un error de cálculo inferior a 5 x 10^{-4}, de donde, el error absoluto total será inferior a

$$2.88 \text{ x } 10^{-4} + 5 \text{ x } 10^{-4} = 7.88 \text{ x } 10^{-4} < 1 \text{ x } 10^{-3}.$$

Luego, 0.223 es exacto en menos de 10^{-3} (por exceso o por defecto) como una aproximación del número a calcular.

Ejemplo 39. Calculemos $f(x_1, x_2, x_3) = \frac{7x_3 - x_1 x_2}{x_1^2 + x_2}$ para $x_1 = \pi \approx 3.1$, $x_2 = \sqrt{3} \approx 1.7$, $x_3 = \sqrt{2} \approx 1.4$.

Analizaremos separadamente los errores sistemáticos y de cálculo.

Como $3.1 < \pi < 3.15$, $1.7 < \sqrt{3} < 1.74$, $1.4 < \sqrt{2} < 1.42$, entonces $\Delta_\pi = 0.05$, $\Delta_{\sqrt{3}} = 0.04$, $\Delta_{\sqrt{2}} = 0.02$.

Calculemos primero las derivadas parciales.

$$\frac{\partial f}{\partial x_1} = -\frac{x_2^2 + 14x_1x_3 - x_1^2x_2}{\left(x_1^2 + x_2\right)^2}, \text{ de aquí, } \left|\frac{\partial f}{\partial x_1}\right| \le \frac{x_2^2 + 14x_1x_3 + x_1^2x_2}{\left(x_1^2 + x_2\right)^2} < \frac{82.91475}{127.9161} = 0.648...$$

de donde,

$$\left|\frac{\partial f}{\partial x_1}\right| < 0.7$$

$$\frac{\partial f}{\partial x_2} = -\frac{x_1^3 + 7x_3}{\left(x_1^2 + x_2\right)^2}, \text{ de aquí, } \left|\frac{\partial f}{\partial x_2}\right| = \frac{x_1^3 + 7x_3}{\left(x_1^2 + x_2\right)^2} < \frac{41.195875}{127.9161} = 0.322053...$$

de donde,

$$\left|\frac{\partial f}{\partial x_2}\right| < 0.4$$

$$\frac{\partial f}{\partial x_3} = \frac{7}{x_1^{\,2} + x_2}, \quad \text{de aquí,} \quad \left|\frac{\partial f}{\partial x_3}\right| < \frac{7}{11.31} = 0.618921...$$

de donde,

$$\left|\frac{\partial f}{\partial x_3}\right| < 0.7$$

Por lo tanto, obtenemos como límite superior del error sistemático a

$$|\Delta f| \leq 0.7 \text{ x } 0.05 + 0.4 \text{ x } 0.04 + 0.7 \text{ x } 0.02 = 0.065$$

esto es,

$$|\Delta f| \leq 0.065.$$

Esto significa que cuando reemplazamos el número exacto $\frac{7\sqrt{2} - \pi\sqrt{3}}{\pi^2 + \sqrt{3}}$ por el aproximado $\frac{7\text{x}1.4 - 3.1\text{x}1.7}{3.1^2 + 1.7}$ (*), cometemos por este hecho un error del orden de 0.065.

Los errores que cometemos en el cálculo efectivo de (*) se suman al anterior. Por lo tanto, como $\frac{7\text{x}1.4 - 3.1\text{x}1.7}{3.1^2 + 1.7}$ = 0.400530503, si tomamos 0.4 cometemos un error de cálculo inferior a $5\text{x}10^{-2}$, de donde, el error absoluto total será inferior a

$$0.065 + 0.05 = 0.115 < 0.2.$$

Luego, 0.4 es exacto en menos de 0.2 (por exceso o por defecto) como una aproximación del número a calcular. Esto se puede escribir también como sigue

$$f(x_1, x_2, x_3) = 0.4 \pm 0.2.$$

Observemos que si tomamos como error de cálculo a $1\text{x}10^{-2}$, entonces el error absoluto será inferior a $0.075 < 0.1$ y de aquí podemos escribir que

$$f(x_1, x_2, x_3) = 0.4 \pm 0.1.$$

Observaciones

(I) Como casos particulares de la fórmula general para el cálculo de errores de una función cuyos argumentos son números aproximados, se pueden hallar las fórmulas antes obtenidas para los errores de las operaciones elementales con números aproximados. En efecto,

1. Si $y = x_1 + x_2 + \ldots + x_n$ (suponemos $x_i > 0$, para $i = 1, 2, \ldots, n$), entonces $\frac{\partial y}{\partial x_i} = 1$, para $i = 1, 2, \ldots, n$. Resulta así, de (42)

$$| \Delta y | \leq \sum_{i=1}^{n} \left| \Delta x_i \right|$$

de (43)

$$\Delta_y = \sum_{i=1}^{n} \Delta_{x_i}$$

de (44)

$$\delta \leq \sum_{i=1}^{n} \left| \frac{x_i}{f(\vec{x})} \right| \delta_i = \sum_{i=1}^{n} \frac{x_i}{y} \delta_i$$

de (45)

$$\delta_y = \sum_{i=1}^{n} \frac{x_i}{y} \delta_{x_i} .$$

2. Si $y = \frac{x}{z}$ (suponemos $x, z > 0$), entonces $\frac{\partial y}{\partial x} = \frac{1}{z}$, $\frac{\partial y}{\partial z} = \frac{-x}{z^2}$. Luego, resulta de (42)

$$\left| \Delta y \right| \leq \left| \frac{\Delta x}{z} \right| + \left| \frac{x \Delta z}{z^2} \right| = \frac{\left| \Delta x \right|}{z} + \frac{x}{z^2} \left| \Delta z \right|$$

de (43)

$$\Delta_y = \frac{1}{z} \Delta_x + \frac{x}{z^2} \Delta_z$$

de (44)

$$\delta \leq \frac{1}{z} x \frac{z}{x} \delta_1 + \frac{x}{z^2} z \frac{z}{x} \delta_2$$

es decir,

$$\delta \leq \delta_1 + \delta_2$$

donde, con δ_1 denotamos el error relativo de x y con δ_2 denotamos el error relativo de z, y de (45)

$$\delta_y = \delta_x + \delta_z .$$

3. Si $y = x^p,\ p \in N$, entonces, $y' = px^{p-1}$, de donde, resulta de (38)

$$|\Delta y| \approx p|x^{p-1}|\ |\Delta x|$$

y entonces, de (39)

$$\Delta y = p|x^{p-1}|\Delta x.$$

De (40)

$$\delta \approx p\delta_1$$

donde, con δ_1 denotamos el error relativo de x, y de (41)

$$\delta_y = p\delta_x.$$

4. Si $y = \sqrt[p]{x} = x^{\frac{1}{p}}, p \in N$, entonces $y' = \frac{1}{p} x^{\frac{1}{p}-1}$, de donde, resulta de (38)

$$|\Delta y| \approx \frac{1}{p}\left|\frac{\sqrt[p]{x}}{x}\right||\Delta x|$$

y entonces, de (39)

$$\Delta_y = \frac{1}{p}\left|\frac{\sqrt[p]{x}}{x}\right|\Delta_x .$$

De (40)

$$\delta \approx \frac{1}{p}\delta_1$$

donde, con δ_1 denotamos el error relativo de x, y de (41)

$$\delta_y = \frac{1}{p}\delta_x .$$

Para las operaciones elementales restantes el análisis es similar.

(II) La metodología utilizada para determinar el error del número que se obtiene como resultado cuando se efectúan operaciones con números aproximados cuyos errores se conocen, constituye *el problema directo del cálculo de errores*.

Un método muy general es el de la *acotación*, recomendable sobre todo si se dispone de una máquina de calcular.

Ejemplo 40. Si se desea calcular el tiempo de oscilación de un péndulo ideal de longitud l, al aplicar la fórmula $t = \pi\sqrt{\frac{l}{g}}$ debe efectuarse una serie de operaciones con números aproximados; algunos como l y g que se han medido experimentalmente y otro como π del cual sólo pueden usarse un número finito de las infinitas cifras decimales de su valor exacto. Supongamos que sea $l = 1.50$ m con un error menor que 0.005 m y $g = 9.81$ m/seg^2 con un error menor que 0.005 m/seg^2.

Se tendrán las siguientes desigualdades

$$3.141 < \pi < 3.142, \qquad 1.495 < l < 1.505, \qquad 9.805 < g < 9.815.$$

Entonces, valdrán las desigualdades

$$3.141\sqrt{\frac{1.495}{9.815}} < t < 3.142\sqrt{\frac{1.505}{9.805}}$$

Obsérvese que las cotas de los valores que aparecen en el numerador se invierten con respecto a las que figuran en el denominador.

Así, resulta

$$1.226 < t < 1.232.$$

Tomando el promedio: $t = 1.229$ seg, se tiene el período determinado con un error menor de 0.003 seg o "redondeado" el resultado: $t = 1.23$ seg con un error menor que el centésimo de segundo.

Con este procedimiento podría evitarse el dar las reglas particulares para cada operación con números aproximados, pero este procedimiento muy seguro presenta las siguientes desventajas:

1. Exige cálculos dobles, uno para cada miembro de la desigualdad.

2. No permite ver la cuantía del error de modo que no puede saberse con qué aproximación deben tomarse los datos que aparecen en la fórmula para que el resultado tenga una aproximación prefijada.

1.11. Problema inverso del cálculo de errores

También de importancia práctica es el problema inverso: si una magnitud $y = f(x_1, x_2, \ldots, x_n)$ que depende de varias variables sujetas a errores debe ser calculada con un error absoluto total prefijado, ¿con qué aproximación deben medirse o computarse los valores $x_1, x_2, \ldots, x_n$ y cómo deben efectuarse los cálculos?

Se podría decir también: ¿cuáles deben ser los errores absolutos de los argumentos de una función así como también en las distintas operaciones a calcular, para que el error absoluto total de la función no exceda de una cierta cantidad límite dada?

La solución más simple al problema inverso viene dada por el denominado *principio de igualdad de efectos*. Supongamos dada la magnitud límite permitida en el cálculo de la expresión $y = f(x_1, x_2, \ldots, x_n)$, que notaremos Δ_l. Sabemos que en dicho cálculo tendremos los errores sistemáticos y de cálculo que notaremos con α. Entonces, se deberá cumplir que

$$\sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| \Delta_{x_i} + \alpha \leq \Delta_l . \tag{46}$$

Los números $\left| \frac{\partial f}{\partial x_i} \right|$, para $i = 1, 2, \ldots, n$ serán calculados como ya lo hicimos en el caso anterior, ya que se podrán determinar intervalos simples en los cuales los argumentos de la función estarán comprendidos. Surge así que las cantidades $\left| \frac{\partial f}{\partial x_i} \right|$, para $i = 1, 2, \ldots, n$ en (46) se conocerán. Además, si x_1 es un dato experimental, Δ_{x_1} se conocerá. En otros casos puramente numéricos (por ejemplo, $x_1 \approx \pi$), también se podrá disponer de Δ_{x_1}. Esto también vale para $x_2, \ldots, x_n$.

Finalmente, se sabe que se puede disponer de α, pues se pueden hacer siempre las operaciones con tanta precisión como se desee, al menos teóricamente.

Luego, será necesario elegir α así como también Δ_{x_1}, Δ_{x_2},…, Δ_{x_n} de forma tal que se satisfaga (46). En la práctica, es ventajoso disponer para ambas clases de errores partes iguales, es decir, tomar

$$\left|\frac{\partial f}{\partial x_1}\right|\Delta_{x_1}+\left|\frac{\partial f}{\partial x_2}\right|\Delta_{x_2}+\ldots+\left|\frac{\partial f}{\partial x_n}\right|\Delta_{x_n}\le\frac{1}{2}\Delta_l$$

lo que resta constituirá el límite superior tolerable para los errores de cálculo; o sea

$$\alpha\le\frac{1}{2}\Delta_l$$

pero entendiéndose bien que se puede adoptar cualquier otra distribución de los errores; lo esencial es respetar la condición (46).

Ejemplo 41. Calculemos $f(x_1, x_2, x_3)=\dfrac{7x_3-x_1x_2}{x_1^{\,2}+x_2}$, siendo $x_1=\pi$, $x_2=\sqrt{3}$, $x_3=\sqrt{2}$, con un error absoluto total que no exceda a 0.01.

Se trata de saber cuántos decimales guardaremos en $\pi, \sqrt{3}$ y $\sqrt{2}$ para ejecutar los cálculos.

Tomemos como intervalos de errores

$$3.1<\pi<3.15, \qquad 1.7<\sqrt{3}<1.74, \qquad 1.4<\sqrt{2}<1.42.$$

Según vimos en el ejemplo 39, podemos considerar

$$\left|\frac{\partial f}{\partial x_1}\right|\approx 0.7, \qquad \left|\frac{\partial f}{\partial x_2}\right|\approx 0.4, \qquad \left|\frac{\partial f}{\partial x_3}\right|\approx 0.7.$$

Se deberá, entonces, satisfacer la condición esencial

$$0.7\,\Delta_{x_1}+0.4\,\Delta_{x_2}+0.7\,\Delta_{x_3}+\alpha\le 0.01$$

o bien

$$70\,\Delta_{x_1}+40\,\Delta_{x_2}+70\,\Delta_{x_3}+100\,\alpha\le 1.$$

Si tomamos la parte proveniente del error sistemático menor o igual a 1/2, podemos dar la misma importancia a cada término; esto es

$$70\,\Delta_{x_1} \le 1/6 \qquad \text{implica} \qquad \Delta_{x_1} \le 0.00238\ldots < 3\text{x}10^{-3}$$
$$40\,\Delta_{x_2} \le 1/6 \qquad \text{implica} \qquad \Delta_{x_2} \le 0.00416\ldots < 5\text{x}10^{-3}$$
$$70\,\Delta_{x_3} \le 1/6 \qquad \text{implica} \qquad \Delta_{x_3} \le 0.00238\ldots < 3\text{x}10^{-3}.$$

Se ve así que es necesario tomar los números con tres decimales y puesto que conocemos el decimal siguiente, podemos sin ninguna dificultad corregir la tercera cifra si fuera necesario, lo que hará que nuestros errores sean realmente inferiores a las evaluaciones hechas arriba.

Así, tomamos

$$\pi \approx 3.142,\ \sqrt{3} \approx 1.732,\ \sqrt{2} \approx 1.414,\ \text{con}\ \ \Delta_\pi = 5 \text{ x } 10^{-4},\ \Delta_{\sqrt{3}} = 5 \text{ x } 10^{-4},$$
$$\Delta_{\sqrt{2}} = 5 \text{ x } 10^{-4}$$

Es decir,

$$70\Delta_{x_1} + 40\Delta_{x_2} + 70\Delta_{x_3} = (70 + 40 + 70) \text{ x } 5 \text{ x } 10^{-4} = 18 \text{ x } 5 \text{ x } 10^{-3} = 9 \text{ x } 10^{-2}$$

(y, por lo tanto, el error sistemático es igual a $\frac{9\text{x}10^{-2}}{100} = 0.9 \text{ x } 10^{-3}$).

Es necesario efectuar los cálculos de manera tal que el error de cálculo, es decir α, verifique la condición

$$100\,\alpha \ \le\ 1 - 0.09 = 0.91$$

de donde,

$$\alpha \le\ 0.0091 = 9.1\text{x}10^{-3}.$$

Calculemos

$$f(3.142,\ 1.732,\ 1.414) = \frac{4.456056}{11.604164} = 0.3840049\ldots$$

El único error de cálculo es el que se comete al detener la división. Deteniéndola luego del segundo decimal vemos que el error es inferior a $5\text{x}10^{-3}$, número que es inferior a $9.1\text{x}10^{-3}$. Por consiguiente, podemos tomar como aproximación a $f(x_1, x_2, x_3)$ el valor 0.38.

Recapitulando, vemos que el error total que nos han pedido no sólo no excede a 0.01 sino que es inferior a

$$0.9 \times 10^{-3} + 5 \times 10^{-3} = 5.9 \times 10^{-3} < 6 \times 10^{-3}$$

es decir, los límites del valor exacto son

$$0.374 < f(x_1, x_2, x_3) < 0.386.$$

1.12. Representación de la información. Punto fijo y punto flotante.

Los matemáticos al diseñar un método para resolver un problema suponen que todos los cálculos se hacen dentro del sistema de los números reales. Esta suposición simplifica en gran medida el análisis matemático de los problemas. Sin embargo, cuando realmente se va a computar la solución se debe hacerlo sin el sistema de los números reales, y esto es así, pues R es infinito y cualquier conjunto de números representables en una computadora es necesariamente finito. La ausencia de esta propiedad en una computadora es la fuente de los errores de redondeo.

Para ser más precisos, necesitamos discutir el tipo de sistema numérico finito usado por una computadora. En la memoria de una computadora, cada número es almacenado en una posición que consta de un signo + ó - más un número finito de dígitos. Un problema que debemos enfrentar es cómo usar estos dígitos para representar números.

Una forma es asignar un número fijo de ellos para la parte fraccionaria. Este sistema numérico se llama *sistema de punto fijo*.

La otra forma, que es la usada por la mayoría de las computadoras, es el *sistema de punto flotante* en el cual se trabaja con un número constante de dígitos.

El sistema de punto fijo (es decir, el conjunto de números representables en este sistema) está caracterizado por tres parámetros:

β = base del sistema numérico
t = número de dígitos (ó longitud de palabra, que depende de la máquina)
f = número de dígitos de la parte fraccionaria.

Designaremos a este sistema numérico con $P(\beta, t, f)$ y con $fix(x)$ a los números que pertenecen a $P(\beta, t, f)$.

Como el punto decimal está fijo, quedará determinada una cantidad máxima de cifras tanto en la parte entera como en la parte fraccionaria de cada número. En este sistema a los exponentes se les permite tomar un único valor.

Los números enteros son, generalmente, un ejemplo en punto fijo ya que la mayor parte de las computadoras trabajan para este tipo de datos con dicho sistema.

Ejemplo 42. Sea P(10, 4, 1) y $x = 865.54$. Entonces, $fix(x) = 865.5$.

Ejemplo 43. Sea P(10, 4, 2) y $x = 95.523$. Entonces, $fix(x) = 95.52$.

Ejemplo 44. Sea P(2, 4, 2) y representemos $x_1 = 3.75$ y $x_2 = 1.25$. Entonces,

$$\begin{array}{r|l} 3 & 2 \\ \hline 1 & 1 \end{array}$$ de aquí, $(3)_{10} = (11)_2$

$$\begin{array}{r|l} 0 & 75 \\ \hline 1 & 50 \\ 1 & 00 \end{array}$$ de aquí, $(0.75)_{10} = (0.11)_2$.

$(1)_{10} = (1)_2$

$$\begin{array}{r|l} 0 & 25 \\ \hline 0 & 50 \\ 0 & 00 \end{array}$$ de aquí, $(0.25)_{10} = (0.01)_2$.

Luego, $fix(x_1) = 11.11$ y $fix(x_2) = 01.01$.

Ejemplo 45. Sea P(10, 5, 3) y $x = .95523$. Entonces, $fix(x) = 00.955$.

Antes de definir el sistema de punto flotante, recordemos que cualquier número entero en el sistema decimal usual y corriente puede expresarse como un polinomio de base 10 con coeficientes enteros entre 0 y 9:

$$(a_n a_{n-1} \ldots a_1 a_0) = a_n 10^n + a_{n-1} 10^{n-1} + \ldots + a_1 10 + a_0.$$

Ejemplo 46.

$$9474 = 9\text{x}10^3 + 4\text{x}10^2 + 7\text{x}10^1 + 4\text{x}10^0 = 9474.$$

Si el número es una fracción decimal las potencias de 10 serán negativas.

Ejemplo 47.

$$0.8125 = 8\text{x}10^{-1} + 1\text{x}10^{-2} + 2\text{x}10^{-3} + 5\text{x}10^{-4} = 0.8 + 0.01 + 0.002 + 0.0005 = 0.8125.$$

En general, un número en el sistema de base 10 puede expresarse

$$d_n x10^n + d_{n-1} x10^{n-1} + \ ... \ + d_1 x10^1 + d_0 x10^0 + d_{-1} x10^{-1} + ... + d_{-m} x10^{-m}$$

donde los dígitos d_n, d_{n-1}, ..., d_1, d_0, d_{-1}, ..., d_{-m} son coeficientes enteros comprendidos entre 0 y 9.

Ejemplo 48

$31025.67085 = 3x10^4 + 1x10^3 + 0x10^2 + 2x10^1 + 5x10^0 + 6x10^{-1} + 7x10^{-2} + 0x10^{-3} +$
$+ 8x10^{-4} + 5x10^{-5} = 30000 + 1000 + 0 + 20 + 5 + 0.6 + 0.07 + 0.000 + 0.0008 + 0.00005 =$
$= 31025.67085.$

No existe ninguna razón esencial para usar la base 10. Por razones técnicas las computadoras representan los números internamente en forma binaria, es decir, en base 2. Por consiguiente, debe haber un procedimiento de conversión de decimal a binario cuando ingresamos información y de binario a decimal cuando queremos interpretar los resultados de salida.

Si en lugar de usar base 10 usamos una base β cualquiera, por ejemplo 2, 3, 8, 16, ..., podemos generalizar y escribir

$$d_n x\beta^n + d_{n-1} x\beta^{n-1} + ... + d_1 x\beta^1 + d_0 x\beta^0 + d_{-1} x\beta^{-1} + ... + d_{-m} x\beta^{-m}.$$

En particular, cualquier número binario puede expresarse

$$d_n x2^n + d_{n-1} x2^{n-1} + ... + d_1 x2^1 + d_0 x2^0 + d_{-1} x2^{-1} + ... + d_{-m} x2^{-m}$$

donde los coeficientes d_n, d_{n-1}, ..., d_1, d_0, d_{-1},..., d_{-m} son en este caso específico ó bien 0 ó bien 1.

En un sistema de base 3 el número podrá escribirse

$$d_n x3^n + d_{n-1} x3^{n-1} + ... + d_1 x3^1 + d_0 x3^0 + d_{-1} x3^{-1} + ... + d_{-m} x3^{-m}$$

donde los coeficientes d_n, d_{n-1}, ..., d_1, d_0, d_{-1},..., d_{-m} son ó bien 0 ó bien 1 ó bien 2.

Ahora sí veamos los cuatro parámetros que caracterizan a los sistemas numéricos de punto flotante:

β = base del sistema numérico
t = número de dígitos (o longitud de palabra)
L, U = rango del exponente (números enteros que dependen de la computadora).

A este sistema lo designamos con F(β, t, L, U) y con $fl(x)$ a los números que pertenecen a F(β, t, L, U).

Cualquier número $fl(x) \in$ F(β, t, L, U) tiene la forma

$$fl(x) = \pm\left(\frac{d_1}{\beta^1} + \frac{d_2}{\beta^2} + \ldots + \frac{d_t}{\beta^t}\right) \text{x}\, \beta^e = \pm(.d_1 d_2 \ldots d_t)_\beta \text{x}\beta^e$$

donde los enteros $d_1, d_2, \ldots, d_t$ satisfacen

$$0 \leq d_i \leq \beta\text{-}1$$

y el entero e satisface

$$L \leq e \leq U.$$

Si para cada $fl(x)$ en F no nulo $d_1 \neq 0$, se dice que el sistema de números de punto flotante F está *normalizado*. Esto es

$$1 \leq d_1 \leq \beta\text{-}1.$$

Se exceptúa el 0 que siempre pertenece al sistema y se escribe

$$0 = +.\underbrace{00\ldots0}_{t} \text{x}\beta^L .$$

El entero e se llama *exponente*. El número $(.d_1 d_2 \ldots d_t)_\beta$ se llama *fracción o mantisa*.

Observación. La representación del cero en punto flotante no está estandarizada y escapa a la regla de que internamente todo número debe normalizarse. En punto fijo nos vemos en la imposibilidad de normalizar los números, pues a los exponentes se les permite tomar un único valor.

Ejemplo 49. Sea F(10, 3, -1, 2) y $x = .86554$. Entonces, $fl(x) = +.866\text{x}10^0$.

Ejemplo 50. Sea F(10, 4, -2, 3) y $x = 987.34$. Entonces, $fl(x) = +.9873\text{x}10^3$.

Ejemplo 51. Sea F(2, 8, -7, 8) y x = 3.625. Entonces, $fl(x) = +.11101000\text{x}2^2$.

En efecto, $(3)_{10} = (11)_2$ y $(.625)_{10} = (.101)_2$, pues

0	625
1	250
0	500
1	000

Luego, $(3.625)_{10} = (11.101)_2 = (.11101)_2 \times (2^2)_{10}$.

1.13. Comparación entre los sistemas de computación de punto fijo y de punto flotante.

Cuando la computadora desarrolla aritmética sobre números en el sistema de punto fijo tales como la suma y el producto, éstas deben dar de tal manera que caigan dentro del sistema de punto fijo considerado; esto es, no deben salirse del rango preestablecido. Esta es una de las más serias desventajas de la aritmética en punto fijo. Observamos que en muchas computadoras es posible conservar todos los dígitos de los 2*t* dígitos en un producto, almacenando el producto en dos longitudes de palabras de *t* dígitos cada una. En este sistema vemos que aún si los datos de un problema calculatorio están dentro del mismo, usualmente no podemos estar seguros de que todos los datos intermedios caigan dentro de él. Sin embargo, con frecuencia es posible obtener una estimación a priori (posiblemente muy cruda) del tamaño de los resultados intermedios. Se puede entonces intentar reformular el problema (por ejemplo, mediante la introducción de nuevas unidades de medida) de tal forma que los datos al igual que los resultados intermedios y finales, queden dentro del sistema. A esta reformulación se la conoce como *modulación* (scaling). La frecuente necesidad de modular es una desventaja más de la aritmética en punto fijo. Aunque en punto flotante también existe un rango fijo de representación, su amplitud es tal que permite descartar en casi todos los problemas de orden práctico las consideraciones efectuadas para punto fijo. Los números con que trabajan las máquinas de aritmética de punto flotante son muy grandes con un error relativo pequeño y por ello, en general, no es necesaria la modulación y la posibilidad de salirse del sistema es mínima. En definitiva, la utilidad del sistema de punto fijo se restringe casi por completo a aquellos problemas en los que siempre se trabaja con una cantidad de decimales preestablecida y nunca se excede un cierto número de cifras enteras (por ejemplo, precios de artículos de un supermercado, liquidación de sueldos en un comercio tipo, ya que estos datos estarán comprendidos en un rango relativamente chico que puede ser bien delimitado a priori).

Por estas razones, la aritmética de punto flotante es hoy en día casi universalmente usada, aunque la adición y la sustracción en punto flotante

son sustancialmente más lentas que las correspondientes en punto fijo (pues es necesario igualar las partes exponenciales de los números). Las operaciones de multiplicación y de división, en general, requieren el mismo tiempo en ambos sistemas.

1.14. Aritmética de simple y doble precisión.

La aritmética desarrollada tanto en punto fijo como en punto flotante sobre números con *t* dígitos se llama *aritmética de simple precisión*.

Hemos ya mencionado que pueden obtenerse productos cuyas longitudes sean de 2*t* dígitos. Para usar tales productos en cálculos posteriores debe ser posible desarrollar una aritmética sobre estos números de doble longitud de palabra. Siempre es posible programar a la computadora para desarrollar las operaciones ordinarias de la aritmética sobre estos números de doble longitud de palabra. Tal aritmética se llama *aritmética de doble precisión*. En las computadoras de longitud de palabra variable no es significativa la distinción entre simple y doble precisión, puesto que los números pueden tener virtualmente cualquier número de dígitos y la computadora está construida para desarrollar aritmética sobre estos números de longitud variable.

Seguidamente desarrollaremos algunos ejemplos en los dos sistemas de numeración en donde se supone que los datos de entrada son exactos (sin errores), pero las operaciones aritméticas elementales (ó combinación de ellas) introducen error. Para minimizar este tipo de errores es común usar en las computadoras acumuladores de doble longitud o precisión (en punto fijo a veces se habla de longitud, ya que incrementarla implica aumentar considerablemente el rango de representación de la computadora, mientras que en punto flotante suele ser común hablar de precisión, ya que aumentarla no modifica el rango de representación de la computadora sino que lo "refina", es decir, distingue mejor entre dos puntos cualesquiera que pertenecen al mismo).

Salvo en algoritmos muy simples, se presentan aquí dos inconvenientes:

1. Se pierden propiedades aritméticas elementales, como la asociatividad.
2. Su cálculo es muy engorroso.

Pero, ¿por qué las operaciones aritméticas elementales introducen error? Esto se debe porque, a menudo, los resultados de operaciones

aritméticas no pertenecen al rango de representación. Este motivo ya anticipa que al trabajar con una computadora no se podrán reproducir exactamente dichas operaciones. Más bien deberán utilizarse operaciones discretas que aproximen lo mejor posible a las exactas. El acumulador de doble longitud o precisión permite guardar TODAS las cifras de la operación, obteniendo como resultado computado el representante del resultado exacto. Se sigue que casi todos los problemas que abordaremos por medio de una computadora se pueden clasificar como problemas físicos (hardware).

Ejemplo 52. Se desarrollan usando $t = 5$, $\beta = 10$ y acumuladores de doble longitud o precisión.

(i) En punto fijo: P(10, 5, 4)

1. $u = 0.3124$, $v = 0.0837$, $y = 9.9325$.
   $fix(u + v) = fix(0.3961) = 0.3961$
   $fix(u - v) = fix(0.2287) = 0.2287$
   $fix(u + y) = fix(10.2449)$ = OVERFLOW!! (sobrecarga).

2. $u = 0.7540$, $v = 0.0003$, $y = 9.9325$, $x = 8.6629$.
   $fix(u \cdot v) = fix(0.0002262) = 0.0002$
   $fix(x \cdot y) = fix(86.044254\ldots)$ = OVERFLOW!!
   $fix(v \cdot v) = fix(0.00000009)$ = UNDERFLOW!! (subcarga).

3. $u = 9.9999$, $v = 0.0001$, $y = 4.0694$.
   $fix(u/y) = fix(2.4573401\ldots) = 2.4573$
   $fix(y/v) = fix(40694)$ = OVERFLOW!!
   $fix(v/u) = fix(0.00001\ldots)$ = UNDERFLOW!!

(ii) En punto flotante (normalizado): F(10, 5, -10, 10)

1. $u = 0.89023 \times 10^{0}$, $v = 0.94537 \times 10^{-7}$.

Para sumar (restar) u y v ambos deben expresarse con el exponente mayor aunque ello requiera más de t dígitos para la mantisa (de v, en este caso particular). Se tiene

$$\begin{array}{l} 0.8902300000 \times 10^{0} \\ + \\ 0.0000000945 \times 10^{0} \\ \hline 0.8902300945 \times 10^{0} \end{array}$$

entonces $fl(u + v) = +.89023 \times 10^{0} = u$.

En la práctica no se realiza la suma; directamente se asigna u a $fl(u + v)$.

2. $u = 0.10943\text{x}10^{2}$, $v = 0.92281\text{x}10^{2}$, $y = 0.83504\text{x}10^{-3}$.
   $fl(u + v) = ?$ y $fl(v + y) = ?$

$$\begin{array}{rr} 0.1094300000 \times 10^{2} & 0.9228100000 \times 10^{2} \\ + & \\ \underline{0.9228100000 \times 10^{2}} & \underline{0.0000083504 \times 10^{2}} \\ 1.0322400000 \times 10^{2} & 0.9228183504 \times 10^{2} \end{array}$$

Luego,

$fl(u + v) = +.10322 \times 10^{3}$ (aumentamos en 1 el exponente del resultado para normalizar).
$fl(v + y) = +.92282 \times 10^{2}$.

3. $u = 0.86253 \times 10^{3}$, $v = 0.86249 \text{x}10^{3}$, $fl(u - v) = ?$

$$\begin{array}{r} 0.8625300000 \times 10^{3} \\ - \\ \underline{0.8624900000 \times 10^{3}} \\ 0.0000400000 \times 10^{3} \end{array}$$

Luego,

$fl(u - v) = +.40000 \times 10^{-1}$ (Cancelación).

4. $u = 0.12654 \times 10^{2}$, $v = 0.35861 \times 10^{1}$, $fl(u \,.\, v) = ?$
   $fl(u \,.\, v) = fl(0.045378509 \times 10^{3}) = +.45379 \times 10^{2}$.

5. $u = 0.54786 \times 10^{7}$, $v = 0.80584 \times 10^{3}$, $y = 0.14796\text{x}10^{-1}$.
   $fl(u / v) = ?$ y $fl(u / y) = ?$

   $fl(u / v) = fl(0.679862\ldots \text{x}\ 10^{5}) = +.67986 \times 10^{5}$
   $fl(u / y) = fl(3.7027575\ldots \text{x}\ 10^{8}) = +.37028 \times 10^{9}$.

Como podemos observar, el único error en este tipo de operaciones es el de representación.

En la práctica, cualquier función se obtiene como composición de operaciones aritméticas. Sin embargo, el orden en que se efectúen estas

operaciones puede alterar el resultado final. Esto que a primera vista resulta sorprendente, tiene su explicación en que la restricción de la aritmética elemental a un número finito de cifras limita en gran medida la validez de propiedades matemáticas esenciales como la asociatividad.

Ejemplo 53. Comparemos procedimientos diferentes para calcular $\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^3$ que matemáticamente son equivalentes, pero como veremos no numéricamente.

1. Racionalizando el denominador:

$$\frac{\sqrt{2}-1}{\sqrt{2}+1}=\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right).\left(\frac{\sqrt{2}-1}{\sqrt{2}-1}\right)=\frac{\left(\sqrt{2}-1\right)^2}{2-1}=\left(\sqrt{2}-1\right)^2, \text{entonces } \left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^3=\left(\sqrt{2}-1\right)^6$$

.

2. Racionalizando el denominador y desarrollando el binomio $\left(\sqrt{2}-1\right)^2$:

$\left(\sqrt{2}-1\right)^2=2-2\sqrt{2}+1=3-2\sqrt{2}$, entonces $\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^3=\left(3-2\sqrt{2}\right)^3$.

3. Hacer lo mismo que en 2 y luego desarrollar el trinomio $\left(3-2\sqrt{2}\right)^3$:

$\left(3-2\sqrt{2}\right)^3=3^3-3.3^2.2\sqrt{2}+3.3.\left(2\sqrt{2}\right)^2-\left(2\sqrt{2}\right)^3=99-70\sqrt{2}$, entonces $\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^3=99\text{-}70\sqrt{2}$.

Tenemos así cuatro igualdades

$$\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^3=\left(\sqrt{2}-1\right)^6=\left(3-2\sqrt{2}\right)^3=99\text{-}70\sqrt{2} \qquad (*)$$

Operando numéricamente, obtendremos resultados diferentes con cada una de ellas.

Por ejemplo, utilizando para $\sqrt{2}$ la aproximación $\frac{7}{5}$:

- $\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^3 \approx \left(\frac{\frac{7}{5}-1}{\frac{7}{5}+1}\right)^3 = \left(\frac{1}{6}\right)^3 = \frac{1}{216} = 0.004630$
- $\left(\sqrt{2}-1\right)^6 \approx \left(\frac{7}{5}-1\right)^6 = \left(\frac{2}{5}\right)^6 = \frac{64}{15625} = 0.004096$
- $\left(3-2\sqrt{2}\right)^3 \approx \left(3-2\frac{7}{5}\right)^3 = \frac{1}{125} = 0.008000$
- $(\ 99\text{-}70\sqrt{2}\ ) \approx \left(99-70\frac{7}{5}\right) = 1$

Utilizando para $\sqrt{2}$ la aproximación $\frac{17}{12}$:

- $\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^3 \approx 0.005125$
- $\left(\sqrt{2}-1\right)^6 \approx 0.005233$
- $\left(3-2\sqrt{2}\right)^3 \approx 0.004630$
- $\left(99-70\sqrt{2}\right) \approx$ -0.1667

En ambos casos redondeamos los resultados a 4 cifras significativas.

Todos los valores debieron aproximarse al valor exacto: 0.0050506339, en el segundo grupo con mejor aproximación que en el primero.

Las igualdades (*) predecían resultados idénticos, pero esto no fue así.

Si además expresamos

$$99\text{-}70\sqrt{2} = \sqrt{99^2} - \sqrt{70^2 2} = \sqrt{9801} - \sqrt{9800}$$

y tomando $\sqrt{9800}$ con 4, 6, 8 y 10 cifras significativas, se tiene

$\sqrt{9800}$	$\sqrt{9801} - \sqrt{9800}$
98.99	0.01
98.9949	0.0051
98.994949	0.005051
98.99494936	0.00505064

En este ejemplo se está produciendo el fenómeno de la cancelación.

Como hemos podido comprobar, la secuencia de operaciones afecta el resultado.

Conclusión. La asociatividad, en general, no vale para la adición en punto flotante.

Ejemplo 54. Consideremos la adición en punto flotante usando 7 decimales para la mantisa.

$a = 0.1234567 \times 10^0, \qquad b = 0.4711325 \times 10^4, \qquad c = -b.$

Como ya se dijo, la computadora resuelve el problema de la colocación del punto decimal desplazando hacia la derecha al de menor exponente tantos lugares como indique la diferencia entre los exponentes. El siguiente esquema indica cómo la adición en punto flotante es realizada.

$fl(b + c) = +.0000000 \times 10^0 \qquad fl(a + fl(b + c)) = a = +.12345467 \times 10^0.$

a	=	0.0000123	4567×10^4
+			
b	=	0.4711325	$\times 10^4$
$fl(a + b)$	=	0.4711448	$\times 10^4$
c	=	-0.4711325	$\times 10^4$

$fl(fl(a + b) + c) = fl(0.0000123 \times 10^4) = +.1230000 \times 10^0 \neq fl(a + fl(b + c)).$

Ejemplo 55. Consideremos la adición en punto flotante usando $t = 8$.

$a = 0.23371258 \times 10^{-4}, \qquad b = 0.33678429 \times 10^2, \qquad c = -0.33677811 \times 10^2.$

a	=	0.00000023	371258×10^2
+			
b	=	0.33678429	$\times 10^2$
$fl(a + b)$	=	0.33678452	$\times 10^2$
c	=	-0.33677811	$\times 10^2$

$fl(fl(a + b) + c) = fl(0.00000641 \times 10^2) = +.64100000 \times 10^{-3}.$ (I)

$$\begin{array}{rl} & b = 0.33678429 \times 10^{2} \\ + & \\ & c = -0.33677811 \times 10^{2} \\ \hline \end{array}$$

$fl(b + c) = fl(0.00000618 \times 10^{2}) = +.61800000 \times 10^{-3}$

$a = 0.023371258 \times 10^{-3}$.

Para calcular $fl(a + fl(b + c))$ escribimos

$$\begin{array}{rl|l} & a = & 0.02337125 & 8 \times 10^{-3} \\ + & & & \\ & fl(b + c) = & 0.61800000 & \times 10^{-3} \\ \hline a + fl(b + c) = & & 0.64137125 & 8 \times 10^{-3} \end{array}$$

$fl(a + fl(b + c)) = fl(0.641371258 \times 10^{-3}) = +.64137126 \times 10^{-3}$. (II)

Luego, de (I) y (II), resulta que

$$fl(fl(a + b) + c) \neq fl(a + fl(b + c)).$$

1.15. Control del error de redondeo. Algunos consejos prácticos.

Una de las tareas del analista numérico es dar una respuesta a un problema mediante un previo "análisis del error de redondeo". Este es un problema altamente teórico que no seguiremos aquí. En lugar de ello adoptaremos un criterio más pragmático, que es el de intentar minimizar el error en cada operación en vista a que esto produzca el menor error a ser propagado y haga el resultado final tan exacto como sea posible. Hay varias formas en las que el error de redondeo en cada operación o conjunto de operaciones puede minimizarse. Estas formas pertenecen a tres categorías: hardware, software y cuidadosa programación.

Ilustraremos esto con un ejemplo de cada tipo usando el sistema numérico F(10, 4, -50, 50) con *chopping* (chopping o truncado es lo mismo que redondeo a t dígitos en el cual los dígitos que siguen a d_t son siempre eliminados).

1. Relativo al hardware.

Ejemplo 56. Supongamos que queremos sustraer 0.5678 de 12.34. Antes de la sustracción, la representación de los números en la máquina debe ser ajustada para alinear los puntos decimales. En este procedimiento alguno de los dígitos menos significativos se perderán. La previsión de "guardar un dígito" (un dígito extra de la parte fraccionaria del número) en la unidad aritmética de la computadora puede prevenir de pérdidas indebidas de exactitud.

Para ilustrar haremos dos sustracciones, guardando un dígito extra en un caso y sin guardar un dígito en otro.

Sin guardar dígito	Guarda un dígito
$.1234 \times 10^2$	$.12340 \times 10^2$
-	-
$.0056 \times 10^2$	$.00567 \times 10^2$
$.1178 \times 10^2$	$.11773 \times 10^2$

El resultado cuando se guarda un dígito es más próximo al resultado exacto: 12.34 - 0.5678 = 11.7722. La tachadura del 3 indica que este dígito es choppeado cuando se almacena el resultado. A primera vista parece sin importancia querer argumentar sobre una diferencia de sólo una unidad en el último dígito de un número. Sin embargo, en computaciones de gran escala donde se involucran millones de operaciones aritméticas hay una acumulación potencial significativa de los errores de redondeo. En consecuencia, es importante asegurarse que el resultado de cada operación individual sea tan seguro como sea posible. Por esta razón, la previsión de "guardar un dígito" en la unidad aritmética es, generalmente, considerada como esencial en una computadora diseñada para la computación científica (aunque reduce la velocidad de procesamiento e incrementa, en general, el trabajo del programador).

2. Relativo al software. Una expresión muy frecuente en el cálculo científico es la combinación de operaciones

$$a + b\,c$$

conocida como *flop* (contracción abreviada de floating - point - operation). Debido al orden de procedencia de las operaciones, la multiplicación se efectúa primero. Esto produce un resultado en "doble longitud", esto es, en $2t$-1 o $2t$ dígitos. Normalmente, en una computadora ordinaria esto podría ser choppeado a t dígitos antes de hacer la suma. Sin embargo, si la adición se hace antes de choppear se puede asegurar una mejor exactitud.

Ejemplo 57. Si $a = 0.1945$, $b = 12.34$ y $c = 5.678$, entonces

Simple precisión	Doble precisión
$b\ c\ \ .7006 \times 10^2$	$.7006652 \times 10^2$
+	+
$a\ \ .0019 \times 10^2$	$.001945 \times 10^2$
$.7025 \times 10^2$	$.7026102 \times 10^2$

Puesto que los flops son frecuentes en los cálculos científicos, muchos compiladores están diseñados para reconocerlos dentro de una sentencia aritmética y asegurar el apropiado código de máquina para efectuar la adición usando el producto en doble longitud. Esto es entonces un aspecto del software para ayudar a minimizar el monto del error de redondeo originado.

3. Relativo a la programación. Cada uno de los ejemplos precedentes ilustra un método para mejorar la exactitud. Además, como cada una de estas facilidades pueden estar incorporadas en el sistema de la computadora, uno no tiene por qué preocuparse de ellas más allá de un simple chequeo inicial para asegurarse que estas facilidades están realmente dadas. El usuario debería sin embargo estar enterado de los posibles métodos para reducir el error y que pueden ser incorporados en el programa.

Veamos algunos ejemplos ilustrativos.

Ejemplo 58. Supongamos que queremos calcular el punto medio c del intervalo $[a, b]$. Podemos elegir dos fórmulas:

$$\text{(i) } c = \frac{a+b}{2} \qquad \text{ó} \qquad \text{(ii) } c = a + \frac{b-a}{2}.$$

La primera es más módica para calcular pues sólo exige una adición y una división.

Sin embargo, en cuanto a la exactitud, (i) no es necesariamente mejor.

Supongamos, por ejemplo, que $a = 3.483$ y $b = 8.765$. Entonces, las respectivas fórmulas dan

$$\text{(i)} \quad c = \frac{a+b}{2} = \frac{.1224\text{x}10^2}{2} = .6120\text{x}10^1.$$

$$\text{(ii)} \quad c = a + \frac{b-a}{2} = .3483\text{x}10^1 + \frac{.5282\text{x}10^1}{2} = .6124\text{x}10^1.$$

Puesto que el resultado correcto es 6.124, la segunda fórmula es claramente mejor.

Por otra parte, supongamos ahora que a = -3.483 y b = 8.765. Entonces, los resultados son

(i) $c = \dfrac{.5282\text{x}10^{1}}{2} = .2641\text{x}10^{1}.$

(ii) $c = -.3483\text{x}10^{1} + \dfrac{.1224\text{x}10^{2}}{2} = .2637\text{x}10^{1}.$

El valor correcto es 2.641 y, por lo tanto, la primera fórmula es mejor. De estos resultados concluimos que para obtener la mejor exactitud se debería usar (i) ó (ii) según si a y b difieren de signo. En consecuencia, disponemos de una tercera fórmula que en base a la exactitud es claramente mejor, a saber

(iii) Si $sgn(a) \neq sgn(b)$ entonces
$c = (a + b)/2$
si no
$c = a + (b - a)/2$

Este ejemplo, así como otros vistos al principio (fórmulas recursivas, por ejemplo), ilustran que el cuidado al elegir la fórmula correcta puede mejorar el resultado obtenido de un programa. Recordemos que analizamos algoritmos que eran inestables numéricamente, y que aunque se usaran más decimales de exactitud durante todo el cálculo, los resultados absurdos llegarán en etapas ulteriores. Evitamos esta inestabilidad numérica mediante la elección apropiada de la fórmula.

También debemos evitar, en la medida que sea posible, restar dos números próximos debido a que se pueden perder dígitos significativos (aunque usemos el acumulador de doble longitud, como vimos en el ejemplo 52 (ii) 3 del párrafo 1.14 en punto flotante, donde $u = 0.86253 \text{ x } 10^{3}$, $v = 0.86249 \text{ x } 10^{3}$ y $fl(u - v) = +.40000 \text{ x } 10^{-1}$, que pertenece al sistema considerado pero los cuatro últimos dígitos de $fl(u - v)$ no son significativos pues provienen de agregar a ambos sumandos cinco dígitos iguales a cero para poder usar el acumulador de doble longitud).

A este fenómeno se lo denomina *cancelación catastrófica* que ocurre, como hemos visto, cuando dos números del mismo signo y aproximadamente la misma magnitud son sustraídos.

A menudo, una gran cancelación puede evitarse rescribiendo la expresión. Ya hemos visto ejemplos donde se producía este fenómeno y

como podíamos superarlo, pero para ilustrar una vez más este problema veamos el siguiente ejemplo.

Ejemplo 59. Consideremos la evaluación de las raíces de la ecuación cuadrática

$$ax^2 + bx + c = 0, \qquad a \neq 0.$$

Las raíces vienen dadas por las fórmulas

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \qquad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

Supongamos $a = 1$, $b = -320$, $c = 16$ y que las operaciones se hacen en un sistema de punto flotante F(10, 4, -50, 50) con chopping.

Así, $a = .1000\text{x}10^1$, $b = -.3200 \text{ x } 10^3$, $c = .1600 \text{ x } 10^2$.

Luego,

$$x_1 = \frac{.3200\text{x}10^3 + \sqrt{.1024\text{x}10^6 - .6400\text{x}10^2}}{.2000\text{x}10^1} = \frac{.3200\text{x}10^3 + .319\underline{8}\text{x}10^3}{.2000\text{x}10^1} = \frac{.639\underline{8}\text{x}10^3}{.2000\text{x}10^1} = .319\underline{9}\text{x}10^3$$

$$x_2 = \frac{.3200\text{x}10^3 - \sqrt{.1024\text{x}10^6 - .6400\text{x}10^2}}{.2000\text{x}10^1} = \frac{.3200\text{x}10^3 - .319\underline{8}\text{x}10^3}{.2000\text{x}10^1} = \frac{.\underline{2}000\text{x}10^0}{.2000\text{x}10^1} = .\underline{1}000\text{x}10^0$$

El guión debajo del número indica la posición del primer dígito incorrecto como consecuencia del error de redondeo. Las raíces correctas son

$$x_1 = 319.949921, \qquad x_2 = 0.05000785.$$

En consecuencia, hemos obtenido un muy buen valor para x_1, pero uno muy malo para x_2 (el error relativo es casi del 100 %). Es fácil describir qué ocurre con el cálculo para x_2. Los dos números 320.0 y 319.8 en el numerador son casi de la misma magnitud. En consecuencia, cuando se restan uno del otro los dígitos más significativos se cancelan dejando los menos significativos para determinar el resultado. Pero el último dígito de 319.8 contiene algún error de redondeo. Por consiguiente, el resultado de la operación de sustracción es completamente sin importancia debido a que su dígito más significativo está afectado de error. Es por ello que podemos afirmar que estamos en presencia de la cancelación, que en este caso podemos evitarla considerando el par de fórmulas alternativas para las raíces

$$x_1 = \frac{-b - sgn(b)\sqrt{b^2 - 4ac}}{2a}, \qquad x_2 = \frac{c}{ax_1}.$$

Así, para nuestro ejemplo, calculamos x_1 como antes y

$$x_2 = \frac{.1600\text{x}10^2}{.3199\text{x}10^3} = .500\underline{1}\text{x}10^{-1}.$$

Hemos evitado así la cancelación eliminando la sustracción de dos números del mismo signo y próximos, obteniendo mejores resultados.

Como hemos visto aquí, la aritmética de la computación presenta serias deficiencias. Veamos, por último, otra situación que suele ocurrir a menudo.

Ejemplo 60. Puesto que la función exponencial e^x aparece con mucha frecuencia en la computación científica, es importante estar en condiciones de poder evaluarla en forma exacta y eficiente para cualquier número x en el sistema de punto flotante. Hay, en efecto, muchos métodos que nos dan un algoritmo para resolver tal problema (y muchos sistemas de computación incluyen algún algoritmo).

Recordemos que e^x puede representarse por la serie infinita

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!} + \ldots$$

que converge para cualquier número real (o complejo) x.

Suponemos que en nuestra computadora no existe un algoritmo para recalcular e^x cada vez que sea necesario.

En consecuencia, un método posible sobre el cual se puede basar un algoritmo es truncar la serie en algún punto y evaluar la serie resultante. El punto en el cual truncar dependerá de la magnitud de x y de la exactitud deseada.

Hay tres tipos de errores que ocurren en este problema. Dos de ellos son los errores debido a la representación y al cálculo (los cuales se han descrito y visto numerosos ejemplos). El tercero surge debido a que es imposible considerar todos los términos de la serie: error de truncamiento.

Supongamos que deseamos calcular el valor de $e^{-5.5}$ en el sistema numérico F(10, 5, -50, 50) con redondeo. Obtenemos, entonces

$$
\begin{array}{rr}
e^{-5.5} \approx & 1.0000 \\
& -5.5000 \\
& 15.125 \\
& -27.730 \\
& 38.129 \\
& -41.942 \\
& 38.447 \\
& -30.208 \\
& 20.768 \\
& -12.692 \\
& 6.9805 \\
& -3.4902 \\
& 1.5997 \\
& \vdots \\
\hline
& 0.0038363
\end{array}
$$

La suma es terminada después de 25 términos, pues los términos subsiguientes no afectan a la suma. En consecuencia, tenemos un valor tan seguro como posible usando este método y un sistema numérico particular. Ahora bien, ¿hemos obtenido una respuesta satisfactoria?

En realidad

$$e^{-5.5} = 0.00408677$$

y, por lo tanto, en la respuesta obtenida anteriormente no hay dígitos significativos. Es obvio de la lista de términos que la cancelación ha estado de nuevo presente. Es importante darse cuenta que esta gran cancelación que se produce no es la causa del error en la respuesta, sino que simplemente aumenta el error ya presente en los términos. Aunque es siempre posible tener más dígitos significativos en una computación (o sea, en un cálculo hecho con computadora), esto es costoso en tiempo de ejecución y almacenaje y a menudo requiere técnicas especiales de programación. Considerar dígitos extras, por ejemplo, hacer los cómputos en F(10, *t*, -50, 50), $t > 5$, puede ser una solución, pero no la más práctica. Para este problema, un método mejor es computar $e^{5.5}$ y luego tomar la recíproca; esto es

$$e^{-5.5} = \frac{1}{e^{5.5}} = \frac{1}{1 + 5.5 + 15.125 + \ldots} = \frac{1}{244.71} = 0.0040865$$

con nuestra aritmética de cinco decimales. No hay así ninguna cancelación y se obtiene un buen resultado.

La pérdida de dígitos significativos debido a grandes sumas parciales comparadas con el resultado final se llama *smearing*.

EJERCICIOS PROPUESTOS .....

1. i) Establecer la unidad numérica de cada uno de los siguientes números y escribirlos en forma de unidad numérica:

a) 74.24	d) 0.35	g) -1863.000
b) 13258	e) 0.005	h) - 0.00743
c) 8200.02	f) 1200	i) 0.050070

ii) Encontrar el número de cifras significativas de cada uno de los números dados en el apartado i).

iii) Expresar cada uno de los números dados en el apartado i) en notación estándar, en punto flotante y en punto flotante normalizado.

2. Redondear a 3 cifras significativas cada uno de los siguientes números:

a) 2.243	d) -13485	g) -108243.1
b) 6.789	e) 0.0895302	h) 0.02475
c) 28294	f) 0.278	i) 35.9957

3. Redondear a 4 cifras significativas cada uno de los siguientes números exactos:

a) $\frac{22}{7}$	d) $cos\ 2°$	g) $\sqrt[3]{0.00000685}$
b) π	e) $\sqrt{0.00809}$	h) 10 !
c) $\frac{100000}{3}$	f) $sen\ 25'$	i) $\sqrt{3}$

4. Cada uno de los siguientes pares de números está formado de un número y de una aproximación. Encontrar para cada par, el error absoluto, el error relativo y el error relativo porcentual, siempre que sea posible, o en su defecto, las cotas de dichos errores:

a) $\sqrt{5}$, 2.24	d) 190^2, 36000
b) e, 2.718	e) 28294, 28290
c) $\frac{17}{64}$, 0.27	f) $\sqrt[3]{19700}$, 27

5. Supóngase que se tiene que medir la longitud de un puente y de un remache obteniéndose 9999 cm y 9 cm, respectivamente. Si los valores verdaderos son 10000 cm y 10 cm, calcular:

a) El error absoluto.
b) El error relativo porcentual.
c) ¿Qué se puede concluir a partir de los resultados obtenidos en los apartados anteriores?

6. Dar los límites entre los que se encuentra la cantidad exacta *x*, sabiendo que las aproximaciones redondeadas *x*' de las mismas son las siguientes:

a) 2.467
b) 5.4387
c) 0.002178
d) 13

7. Como aproximación de:

a) $\pi = 3.141592...$ se toma el valor 3.14
b) $\sqrt{2} = 1.414213...$ se toma el valor 1.413
c) 1000000 se toma el valor 999996

¿Cuáles son las cifras significativas exactas en cada caso?

8. Hallar las cifras significativas exactas de la cantidad aproximada *c* = 52.135 que posee una cota del error relativo $\delta_c = 0.1 \times 10^{-4}$.

9. Calcular las cotas del error absoluto y del error relativo de cada uno de los siguientes números redondeados:

a) 23.655
b) 0.005
c) 38
d) 23.490
e) 4500.0
f) 0.2100

10. A una cinta métrica defectuosa le falta el primer centímetro. Después de medir una longitud con la misma se obtienen 15 cm. Determinar la verdadera longitud de la magnitud medida, el error absoluto de la medición, el relativo y el relativo porcentual.

11. Se anota una pesada de 2.5 kg y se supone que existe un error en el instrumento de medida de 0.05 kg ¿Cuál será el error absoluto de medición y el intervalo en el que se encuentra el verdadero valor?

12. Usando seis cifras significativas con redondeo, comparar los resultados de calcular:

a) f (500) y g (500), siendo f (x) = $x\left(\sqrt{x+1}-\sqrt{x}\right)$ y g (x) = $\dfrac{x}{\sqrt{x+1}+\sqrt{x}}$

b) f (0.01) y P (0.01), siendo f (x) = $\dfrac{e^x-1-x}{x^2}$ y P (x) = $\dfrac{1}{2}+\dfrac{x}{6}+\dfrac{x^2}{24}$. (La función P (x) es el polinomio de Taylor de grado 2 de f (x) alrededor de x = 0).

13. Sean $P(x) = x^3 - 3x^2 + 3x - 1$ y $Q(x) = ((x - 3)x + 3)x - 1$, donde esta última expresión es obtenida utilizando el esquema de Hörner. Usando tres cifras significativas con redondeo calcular $P(2.19)$ y $Q(2.19)$. Comparar estas aproximaciones con los valores exactos $P(2.19) = Q(2.19) = 1.685159$.

14. Sabiendo que $\int_0^{1/2} e^{x^2} dx = 0.544987104 = I$, debemos determinar el número de cifras significativas exactas de la aproximación obtenida al reemplazar el integrando $f(x) = e^{x^2}$ por la serie de Taylor truncada $P_8(x) = 1+x^2+\dfrac{x^4}{2!}+\dfrac{x^6}{3!}+\dfrac{x^8}{4!}$.

15. La fórmula de *sen x* es

$$sen\ x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$$

Calcular *sen* 2° con redondeo a 6 decimales (6 decimales significativos exactos), considerando 2° ≈ 0.034906585 radianes. ¿Cuántos términos son suficientes considerar en la sumatoria? ¿Por qué? Comparar con el resultado que proporciona su calculadora.

16. Obtener una expresión para el error absoluto límite de $y = f(x)$ en términos de x y del error absoluto límite de x, y una expresión para el error relativo límite de $y = f(x)$ en términos de x y del error relativo límite de x, para cada una de las siguiente funciones:

a) $y = sen\ x$ b) $y = log\ x$ c) $y = a^x$, $a > 0$

17. Determinar la cota del error absoluto de:

a) $y = ln\ 1.377$ b) $y = e^{1.377}$

sabiendo que el número 1.377 está redondeado; en consecuencia, redondear los resultados a un número de cifras significativas.

18. Determinar las cotas del error absoluto y relativo cometido en la evaluación del polinomio para el valor x = 1.5. Tanto los coeficientes como el valor de x están redondeados. Dar los límites entre los que se encuentra el valor exacto p.

$$p(x) = 2.3\ x + 5.7$$

19. Calcular la cota del error absoluto de cada una de las siguientes expresiones, en las cuales los valores se sobreentienden que están redondeados. En consecuencia, redondear los resultados a un número de cifras significativas:

a) 1.761 + 17.32 - 5.82 c) $\dfrac{0.125 \text{ x } 3.5791}{8.6131}$

b) 3.14 x 1.31^2 d) $\sqrt{2.1243 - (3.5791/8.6131)}$

20. Calcular el límite superior del error sistemático y del error de cálculo que se comete al evaluar las funciones:

a) $f(x) = \dfrac{x\text{-}1}{x+4}$ para el valor $x = \sqrt{2}$, usando la aproximación 1.4 para $\sqrt{2}$.

b) $f(x_1, x_2) = \dfrac{2x_1 + 4}{3x_2}$ para los valores $x_1 = \sqrt{2}$, $x_2 = \pi$ usando las aproximaciones 1.4 para $\sqrt{2}$ y 3.1 para π.

21. Calcular la expresión 1.314 x π con un error absoluto límite de $\varepsilon_l = 0.0005$

22. ¿Con cuántos decimales deben ser determinados los valores de b, A y B si la expresión $a = b\dfrac{sen\ A}{sen\ B}$ quiere evaluarse para los valores $b = 42.3625$ cm, $A = 29°33'$ y $B = 53°12'$ con un error absoluto límite de 0.005?

23. Se da la expresión

$$E = 4x^2\sqrt{y} - sen(2y)$$

que quiere evaluarse para los valores aproximados $\bar{x} = 15.2$, $\bar{y} = 57°10'$. Se trata de saber con cuántos decimales deberían ser determinados los valores de $\bar{x}$ e $\bar{y}$ para que la cota del error absoluto de E sea $\Delta_E = 0.005$.

24. Ejecutar los dos cálculos siguientes en un sistema numérico de punto flotante con $t = 2$ y con redondeo. Decida cuál es la mejor respuesta. ¿Puede sacar alguna conclusión acerca de la mejor manera o secuencia para sumar los números en una computadora?

i) $((1.0 + 0.5) + 0.14) + 0.042$
ii) $1.0 + (0.5 + (0.14 + 0.042))$

25. a) Calcular los errores absolutos y relativos causados por el redondeo al efectuar, manejando únicamente cuatro dígitos para las mantisas, la operación $x = 723.4 + 0.08261$.
b) Realizar lo mismo que en al apartado a), pero empleando una técnica de corte o truncamiento al operar con las mantisas.

26. Imaginemos un sistema de punto flotante normalizado en el que las mantisas tienen sólo tres dígitos binarios y los exponentes son -1, 0, ó 1. ¿Cuáles serían estos números?

Capítulo 2

2

Solución de ecuaciones no lineales.

2

2.1. Introducción.

Desde hace años se aprendió a usar la fórmula cuadrática

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{1}$$

para resolver

$$f(x) = ax^2 + bx + c = 0 \tag{2}$$

A los valores calculados con la ecuación (1) se les llama *raíces* de la ecuación (2). Estos representan los valores de x que hacen la ecuación (2) igual a cero. Por lo tanto, se puede definir la raíz de una ecuación como el valor de x que hace $f(x) = 0$. Por esta razón, algunas veces a las raíces se las conoce como *ceros* de la ecuación.

Aunque la fórmula cuadrática es útil para resolver la ecuación (2), hay muchas funciones diferentes que no se pueden resolver de manera tan fácil. En estos casos, los métodos numéricos que se describirán proporcionan métodos eficientes para obtener la respuesta. Antes del advenimiento de las computadoras digitales, había una serie de métodos para encontrar las raíces de ecuaciones algebraicas o trascendentales. Para algunos casos, las raíces se podían obtener con *métodos directos* como se hace con la ecuación (2). Aunque había ecuaciones como ésta que se podían resolver directamente, había muchas otras que no. Por ejemplo, hasta una función aparentemente simple tal como $f(x) = e^{-x} - x$, no se puede resolver analíticamente. En estos casos, la única alternativa es una *técnica de solución aproximada*. Un método de solución aproximada es el de graficar la función y determinar donde cruza al eje x (Figura 1). Este punto que representa el valor de x para el cual $f(x) = 0$ es la raíz.

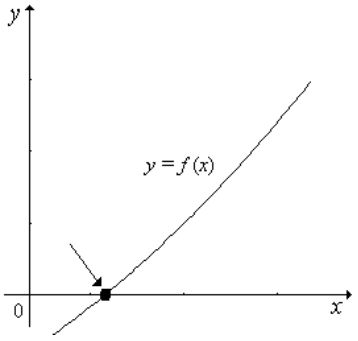


Figura 1

Aunque los *métodos gráficos* son útiles en la obtención de estimaciones aproximadas de las raíces, están limitados por la carencia de precisión. Una aproximación alternativa es usar la *técnica de prueba y error*. Esta "técnica" consiste en escoger un valor de x y evaluar si $f(x)$ es cero. Si no es así (como sucederá en la mayoría de los casos), se hace otra conjetura y se evalúa nuevamente $f(x)$ para determinar si el nuevo valor da una mejor estimación de la raíz. El proceso se repite hasta que se obtenga un valor que genere una $f(x)$ cercana a cero.

Estos métodos fortuitos, obviamente son ineficientes e inadecuados para las exigencias en la práctica de la ingeniería, estadística aplicada, entre otras disciplinas científicas. Las técnicas que emplearemos representan alternativas que no sólo aproximan, sino que emplean estrategias sistemáticas para encaminarse a la raíz verdadera. Además, se adaptan idealmente a la implementación en computadoras personales. Tal como se muestra en las páginas siguientes, la combinación de estos métodos sistemáticos con la computadora hacen de la solución de la mayor parte de los problemas sobre raíces de ecuaciones una tarea simple y eficiente.

En la mayor parte de las áreas estudiadas en Análisis Numérico, en general, existen algunos prerrequisitos de fundamentos matemáticos necesarios para conocer a fondo el tema. Por ejemplo, los conceptos de estimación de errores y la expansión en serie de Taylor tienen importancia directa en el análisis de raíces. Adicionalmente, antes se mencionaron los términos de ecuaciones "algebraicas" y "trascendentales". Puede resultar útil definir formalmente estos términos y discutir como se relacionan con esta parte.

Por definición, una *función* dada por $y = f(x)$ es *algebraica* si se puede expresar de la siguiente manera

$$f_n y^n + f_{n-1} y^{n-1} + \ldots + f_1 y^1 + f_0 = 0 \qquad (3)$$

donde las f_i, $i = 0, 1, ..., n$, son polinomios en x. Los polinomios son un caso simple de funciones algebraicas que se representan, generalmente, como

$$f(x) = a_0 + a_1x + ... + a_nx^n \tag{4}$$

donde las a_i, $i = 0, 1, ..., n$, son constantes. Algunos ejemplos simples son

$$f(x) = 1 - 2.37x + 7.5\,x^3 \tag{5}$$

y

$$f(x) = 5x^2 - x^3 + 7x^6 \tag{6}$$

Una *función trascendental* es una que no es algebraica. Incluye funciones trigonométricas, exponenciales, logarítmicas y otras menos familiares. Algunos ejemplos son

$$f(x) = e^{-x} - x \tag{7}$$

$$f(x) = sen\ x \tag{8}$$

$$f(x) = ln\ x^2 - 1 \tag{9}$$

Las raíces de las ecuaciones pueden ser reales o complejas. Un ejemplo simple de raíces complejas es el caso para el cual el término $b^2 - 4ac$ de la ecuación (1) es negativo. Por ejemplo, dado el polinomio de segundo orden

$$f(x) = 4x^2 - 16x + 17$$

la ecuación (1) se puede usar para determinar que las raíces son

$$x = \frac{16 \pm \sqrt{(-16)^2 - 4(4)(17)}}{2(4)} = \frac{16 \pm \sqrt{-16}}{8}$$

Por lo tanto, una raíz es

$$x = 2 + \frac{1}{2}\ i$$

y la otra es

$$x = 2 - \frac{1}{2}\ i$$

donde, $i = \sqrt{-1}$.

Aunque hay algunos casos donde las raíces complejas de las funciones no polinomiales son de interés, esta situación es menos común que para polinomios. Por lo tanto, los métodos estándar para encontrar raíces, en general, caen en dos áreas de problemas parecidos en principio, pero fundamentalmente diferentes:

1. *La determinación de raíces reales de ecuaciones algebraicas y trascendentales*. Estas técnicas se diseñaron para determinar el valor de una raíz simple de acuerdo a un conocimiento previo de su posición aproximada.
2. *La determinación de todas las raíces reales y complejas de un polinomio*. Estos métodos se diseñaron específicamente para polinomios. Determinar sistemáticamente todas las raíces del polinomio en lugar de simplemente una, dada una posición aproximada.

Los métodos diseñados expresamente para polinomios no se analizarán, ya que van más allá del alcance de este curso que está enfocado al área del primer caso. Sin embargo, como se verá más adelante, algunas de las técnicas diseñadas en el área del primer caso se pueden utilizar para la determinación de raíces complejas de ecuaciones algebraicas.

Antes de proceder con los métodos numéricos para determinar raíces de ecuaciones, será útil dar algunas orientaciones sobre la organización de los temas que se abordarán en este Capítulo.

Se desarrollarán *los métodos que usan intervalos* para encontrar raíces. Estos métodos empiezan con suposiciones que encierran o contienen a la raíz y reducen sistemáticamente el ancho del intervalo. Se cubren dos métodos: el de *bisección* y el de la *regla falsa* (ó *regula falsi*). Los métodos gráficos proporcionan conocimiento visual de las técnicas. Se desarrollarán formulaciones especiales para ayudar a determinar cuánto esfuerzo computacional se requiere para estimar la raíz hasta un nivel de precisión previamente especificado.

Se desarrollarán, además, los *métodos abiertos*. Estos métodos también involucran iteraciones sistemáticas de prueba y error, pero no requieren que la suposición inicial encierre a la raíz. Se descubrirá que estos métodos, en general, son más eficientes computacionalmente que los métodos que usan intervalos, pero no siempre trabajan. Se cubren los métodos de *iteración de punto fijo*, de *Newton* y de la *secante*. Los métodos gráficos proporcionan conocimiento en los casos donde los métodos abiertos no funcionan.

Los ejemplos que se analizarán se emplearán para ilustrar las ventajas y desventajas de cada uno de los métodos, y para proporcionar conocimiento sobre las aplicaciones de las técnicas en la práctica profesional.

Haremos una observación sobre la definición matemática de la raíz de una ecuación, ya que nuestro estudio está orientado hacia la búsqueda efectiva de las raíces a través de la computación. Por definición, s es una raíz de $f(x) = 0$ si $f(s) = 0$. Sin embargo, en las aplicaciones numéricas debemos entender que la expresión $f(s) = 0$ generalmente no se satisface exactamente, debido muchas veces a errores de redondeo o bien a la capacidad limitada de los equipos de computación. Por consiguiente, decimos que s es una raíz de la ecuación $f(x) = 0$ si satisface la condición $|f(s)| \leq \varepsilon$, donde ε es un número real positivo pequeño (cota de error admisible). En consecuencia, en un programa de computación se debe preguntar si $|f(s)| \leq \varepsilon$ luego de preguntar si $f(s) = 0$, pues puede ocurrir que esta última condición no se verifique nunca.

2.2. Métodos iterativos. Métodos abiertos y métodos que usan intervalos.

Los *métodos iterativos* son aquellos en los cuales se repite un determinado proceso hasta que se obtiene un resultado. Estos métodos dan lugar a una sucesión de valores que idealmente converge a la solución del problema.

Sobre raíces de ecuaciones, se estudian distintos métodos iterativos. Se analizan aquellos que aprovechan el hecho que una función, típicamente, cambia de signo en la vecindad de una raíz. A estas técnicas se les llaman *métodos que usan intervalos*, porque se necesita de dos valores iniciales para la raíz. Como su nombre lo indica, estos valores deben "encerrar" o estar uno de cada lado de la raíz. Los métodos particulares descriptos sobre este punto, emplean diferentes estrategias para reducir sistemáticamente el tamaño del intervalo y así, converger a la respuesta correcta.

Como preámbulo de estas técnicas, se discutirán los *métodos gráficos* para graficar funciones y sus raíces. Además de la utilidad de los métodos gráficos para determinar valores iniciales, también son útiles para visualizar las propiedades de las funciones y el comportamiento de los métodos numéricos.

Como ya se dijo anteriormente, en los métodos que usan intervalos la raíz se encuentra dentro del mismo, dado por un límite inferior y otro superior. La aplicación repetida de estos métodos siempre genera aproximaciones más y más cercanas a la raíz. A tales métodos se los conoce como *convergentes*, ya que se acercan progresivamente a la raíz a medida que crece el número de iteraciones.

En contraste con éstos, los *métodos abiertos* se basan en fórmulas que requieren de un solo valor de x o de un par de ellos, pero no necesariamente encierran a la raíz. Como tales, algunas veces *divergen* o se alejan de la raíz a medida que crece el número de iteraciones. Sin embargo, cuando los métodos abiertos convergen, en general, lo hacen más rápido que aquellos que usan intervalos.

2.2.1 Método gráfico.

Un método simple para obtener una aproximación a la raíz de la ecuación $f(x) = 0$ consiste en graficar la función y observar en dónde cruza al eje x. Este punto, que representa el valor de x para el cual $f(x) = 0$, proporciona una aproximación inicial de la raíz.

Ejemplo 1. Utilicemos una gráfica para obtener una raíz aproximada de $f(x) = e^{-x} - x$.

Se calculan los siguientes valores

x	f (x)
0.0	1.000
0.2	0.619
0.4	0.270
0.6	- 0.051
0.8	- 0.351
1.0	- 0.632

que se muestran en Figura 2.

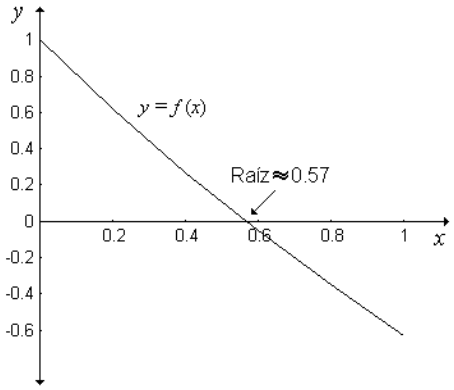


Figura 2

La curva resultante cruza al eje x entre 0.5 y 0.6. Mirando a la gráfica se obtiene una aproximada estimación de la raíz de 0.57, que se acerca a la raíz exacta de 0.56714328... y que se debe determinar utilizando métodos numéricos. La validez de la estimación visual se puede verificar sustituyendo su valor en la ecuación original, para obtener

$$f(0.57) = e^{-0.57} - 0.57 = -0.0045$$

que se acerca bastante a cero.

Las técnicas gráficas tienen un valor práctico limitado ya que no son precisas. Sin embargo, los métodos gráficos se pueden usar para obtener aproximaciones de la raíz. Estas aproximaciones se pueden emplear como valores iniciales para los métodos numéricos que analizaremos en este Capítulo.

Las interpretaciones geométricas, además de proporcionar aproximaciones iniciales de la raíz, son herramientas importantes en la asimilación de las propiedades de las funciones, previendo las fallas de los métodos numéricos.

En la siguiente sección veremos cómo realizar la separación de las raíces reales de una ecuación cualquiera, es decir, determinar intervalos en los cuales exista una sola raíz.

2.2.2 Separación de las raíces.

Cuando se trata de calcular las raíces reales de una ecuación $f(x) = 0$ conviene tener presente una serie de resultados del Análisis Matemático que facilitan el conocimiento de la ubicación de esas raíces.

Consideremos un intervalo (a, b) en el cual la función está definida y es continua. Sea AB el arco que la representa en coordenadas cartesianas. Admitiremos:

I. Si $f(x)$ tiene signos distintos en dos puntos de abscisas a y b, se anula por lo menos una vez en el intervalo (a, b) y, en general, un número impar de veces (Figura 3).

II. Si $f(x)$ tiene igual signo en dos puntos de abscisas a y b, o bien no se anula en el intervalo (a, b) o bien se anula un número par de veces. En la Figura 4, entre los puntos $x = c$, $x = d$ la función no se anula, mientras que en el intervalo (a, b) se anula cuatro veces.

III. Si $f(x)$ es constantemente creciente (o constantemente decreciente) en un intervalo (a, b), es decir, si $f'(x)$ tiene un signo determinado y $sgn\, f(a) \neq sgn\, f(b)$ hay una sola raíz s de la ecuación $f(x) = 0$, mientras que si $sgn\, f(a) = sgn\, f(b)$ con seguridad no hay ninguna raíz. Estos cuatro casos se han resumido en la Figura 5.

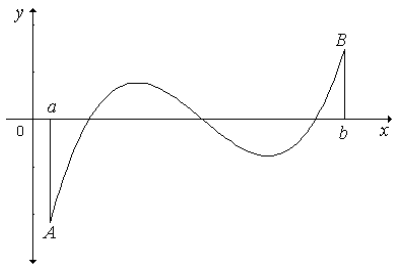


Figura 3

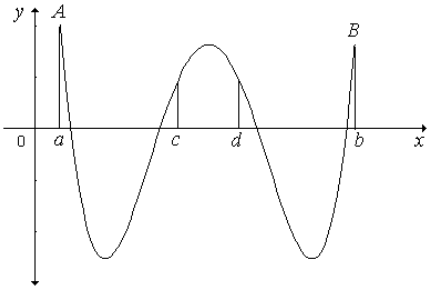


Figura 4

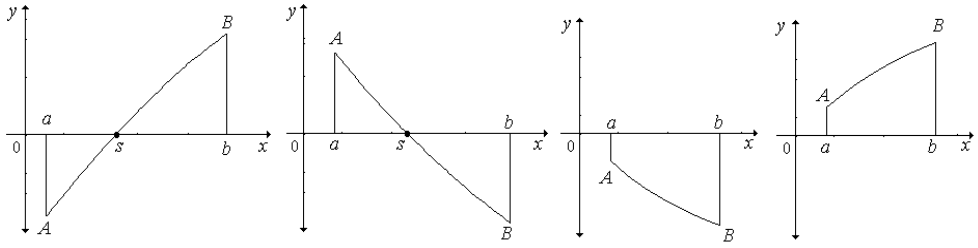


Figura 5

Todo lo expuesto anteriormente se resume en un teorema de Análisis Matemático que permite efectuar la separación de la raíz s en el intervalo (a, b), es decir, asegurar cuándo en el intervalo (a, b) hay una y solamente una raíz real de la ecuación $f(x) = 0$.

Teorema de Bolzano. Si una función continua $f(x)$ asume valores de signos opuestos en los extremos de un intervalo (a, b), es decir, $f(a) f(b) < 0$, entonces el intervalo contendrá al menos una raíz de la ecuación $f(x) = 0$.

La raíz será definitivamente única si la derivada $f'(x)$ existe y mantiene el signo en el intervalo (a, b); esto es, si $f'(x) > 0$ (ó $f'(x) < 0$), para $a < x < b$.

La amplitud del intervalo que contiene a s puede ir disminuyéndose, como se verá en los ejemplos siguientes.

Ejemplo 2. Sea $f(x) = x^3 - x^2 + x - 1 = 0$

Como $f'(x) = 3x^2 - 2x + 1$, resulta entonces que $f'(x)$ es estrictamente positiva $\forall x \in R$. Además, $f(-\infty) = -\infty$ y $f(+\infty) = +\infty$. Por consiguiente, $f(x)$ siempre es creciente y sólo tendrá una raíz real. Por ser $f(0) < 0$ y $f(+\infty) = +\infty$ la única raíz real es positiva. Más precisamente, por ser $f(2) = 5 > 0$ la raíz real está en el intervalo (0, 2).

Ejemplo 3. Sea $f(x) = x^3 - 3x - 1 = 0$

Como $f'(x) = 3x^2 - 3 = 3(x^2 - 1)$, resulta entonces que se anula en $x = -1$ y $x = 1$. Por ser $f(-1) = 1 > 0$ y $f(1) = -3 < 0$ hay una sola raíz en el intervalo (-1, 1).

Puesto que $f(-\infty) = -\infty$ y $f(-1) > 0$ hay otra raíz en el intervalo $(-\infty, -1)$ y, como $f(1) < 0$ y $f(+\infty) = +\infty$ hay otra raíz en el intervalo $(1, +\infty)$. Por lo tanto, las tres raíces reales de esta ecuación se encuentran en los intervalos $(-\infty, -1)$, (-1, 1), $(1, +\infty)$. Más precisamente (disminuyendo la amplitud de estos intervalos), como $f(-2) < 0$, $f(-1) > 0$, $f(0) < 0$, $f(1) < 0$, $f(2) > 0$, entonces las tres raíces reales se encuentran en los intervalos (-2, -1), (-1, 0), (1, 2). En la Figura 6 se muestra el comportamiento de esta función.

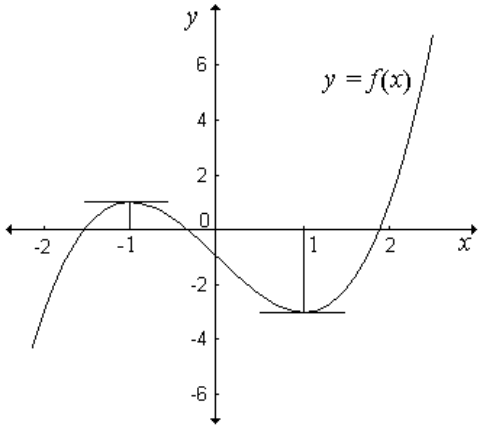


Figura 6

Vemos que si existe una derivada continua $f'(x)$ tal que $f'(x) = 0$, o sea que se pueden calcular fácilmente sus raíces, entonces el proceso de separación de raíces consiste en contar únicamente los signos de $f(x)$ para los ceros de la derivada y en los puntos extremos $x = a$ y $x = b$.

Ejemplo 4. Sea $f(x) = x + e^x = 0$

Como $f'(x) = 1 + e^x > 0$, $\forall x \in R$ y $f(-\infty) = -\infty$, $f(+\infty) = +\infty$, se deduce que $f(x)$ tiene una única raíz real. Por ser $f(-\infty) = -\infty$ y $f(0) > 0$ la única raíz real es negativa. Más precisamente, por ser $f(-1) < 0$ y $f(0) > 0$ la raíz real está en el intervalo (-1, 0).

Aunque las generalizaciones dadas en (I) y (II) son usualmente verdaderas, existen casos en que no se cumplen. Por ejemplo, las raíces

múltiples, esto es, funciones tangenciales al eje x (Figura 7 (a)). También, si las funciones son discontinuas (Figura 7 (b)) pueden no cumplir estos principios.

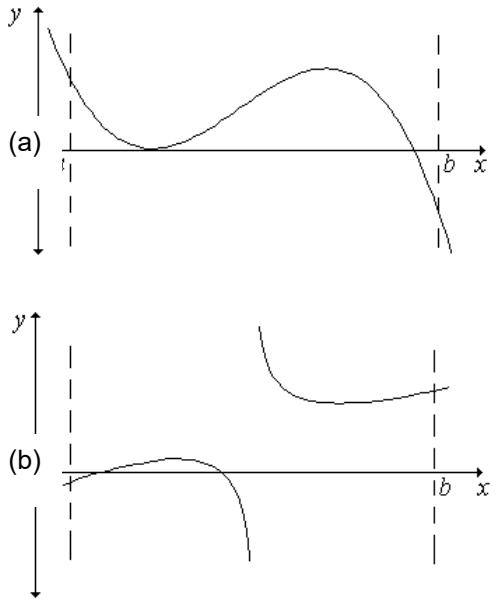


Figura 7

Un ejemplo de una función que tiene una raíz múltiple es la ecuación cúbica $f(x) = (x - 2)(x - 2)(x - 4) = 0$. Nótese que $x = 2$ anula dos veces a esta ecuación, de ahí que a este valor de x se le conozca como raíz múltiple. (En el párrafo 2.3 se presentan técnicas que están diseñadas expresamente para localizar raíces múltiples).

La existencia de casos del tipo mostrado en la Figura 7 dificulta el desarrollo de algoritmos generales que garanticen la localización de todas las raíces en un intervalo.

Sin embargo, usar los métodos numéricos en conjunción con los esquemas gráficos resulta de gran utilidad en la solución de problemas de muchas raíces que se presentan frecuentemente en el área de Ingeniería y de Matemática Aplicada.

2.2.3. Iteración de punto fijo (ó iteración escalar ó iteración funcional) (método abierto).

Este es un método abierto y como tal emplea una fórmula que predice una aproximación a la raíz. Tal fórmula se puede desarrollar para la iteración de punto fijo, rescribiendo la ecuación $f(x) = 0$ de tal forma que x quede del lado izquierdo de la ecuación

$$x = g(x) \qquad (10)$$

Esta transformación se puede llevar a cabo mediante operaciones algebraicas o simplemente agregando x a cada lado de la ecuación original.

Por ejemplo,

$$x^2 - 2x + 3 = 0$$

se puede reordenar para obtener

$$x = \frac{x^2 + 3}{2}$$

mientras que $sen\ x = 0$ puede transformarse en la forma de la ecuación (10), sumándole x a ambos miembros para obtener

$$x = sen\ x + x$$

Las ecuaciones $f(x) = 0$ y $x = g(x)$ son equivalentes. Luego, toda solución de $x = g(x)$ es también solución de $f(x) = 0$.

La Figura 8 nos muestra dos métodos gráficos alternativos para determinar la raíz de $f(x) = 0$: en (a) la raíz se encuentra en el punto donde $f(x)$ cruza al eje x, y en (b) la raíz se obtiene a partir de la intersección de las funciones componentes $y = x$ e $y = g(x)$.

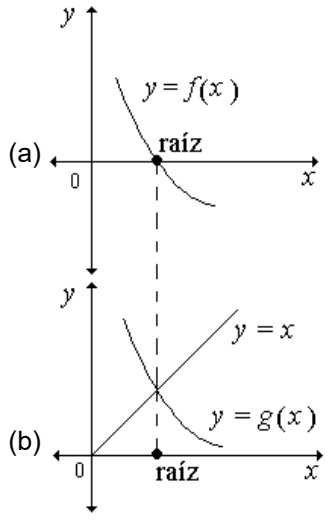


Figura 8

La utilidad de la ecuación (10) es que proporciona una fórmula para predecir un valor de x en función de x. De esta manera, dada una aproximación inicial a la raíz, x_{n-1}, la ecuación (10) se puede usar para obtener una nueva aproximación x_n expresada por la fórmula iterativa

$$x_n = g(x_{n-1}), \quad n = 1, 2, \ldots \tag{11}$$

Cada cálculo del tipo (11) se lo llama *una iteración* y a la función g se la llama *función generadora*. Para una función g, una solución de (10) es llamada *un punto fijo de* g (si la función g es interpretada como una aplicación de su dominio de definición en los reales, el punto s permanece fijo bajo esta aplicación).

En primer lugar, no podemos estar seguros de que este algoritmo esté bien definido. (Podría ser que g no estuviera definida en algún punto, $g(x_n)$).

Sea $D(g)$ el dominio de definición de la función g y $R(g)$ el rango de la función g. Para que la fórmula (11) tenga sentido para todo $n = 1, 2, \ldots$ y para toda elección de $x_0 \in D(g)$, es evidentemente necesario que

$$\mathrm{R}(g) \subseteq D(g) \tag{I}$$

Así, por ejemplo, si $D(g)$ es un intervalo finito cerrado, $D(g) = [a, b]$, entonces los valores de g deben satisfacer

$$a \leq g(x) \leq b, \text{ para } a \leq x \leq b$$

Expresado geométricamente (Figura 9), el gráfico de g debe yacer en la "ventana" del plano x, y descrito por: $a \leq x \leq b$, $a \leq y \leq b$.

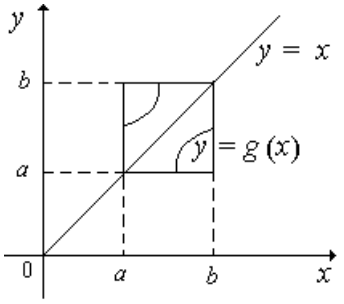


Figura 9

Considerando entonces esto, si $x_0 \in D(g) = [a, b]$ podemos decir que todos los elementos de la sucesión $x_1, x_2, \ldots$ están en $D(g)$. Si algún $x_n \in$

$D(g)$, con $n \geq 0$, entonces también $x_{n+1} = g(x_n) \in D(g)$, puesto que g tiene su valor en $D(g)$.

De aquí en adelante supondremos que el dominio de definición de g es un intervalo cerrado (pero no necesariamente acotado).

Suponiendo además que también vale (I), nos preguntamos:

(i) ¿g tiene un punto fijo?
(ii) ¿Es único este punto fijo?
(iii) ¿La sucesión iterativa generada por g converge para una elección arbitraria del valor inicial x_0?

Como se muestra en la Figura 9, el punto fijo puede no existir si g no es continua. Por ello postulamos

$$g \text{ continua en } D(g) \qquad \text{(II)}$$

Si $D(g)$ es un intervalo cerrado y acotado, $D(g) = [a, b]$, entonces seguro que g tiene punto fijo. (Por (I), $g(a) \geq a$ y $g(b) \leq b$).

Geométricamente vemos que si suponemos que g es continua, entonces la gráfica (Figura 10) muestra que la ecuación (10) tiene por lo menos una solución, pues la gráfica de $y = g(x)$ comienza en algún punto del segmento vertical que une los puntos (a, a) y (a, b) y termina en algún punto del segmento que une los puntos (b, a) y (b, b). Como la gráfica es continua debe intersecar a la recta $y = x$ ($a \leq x \leq b$) quizá en uno de sus extremos. Si s es la abscisa del punto de intersección, entonces $y = s$ e $y = g(s)$ en ese punto, es decir, el número s es una solución de (10).

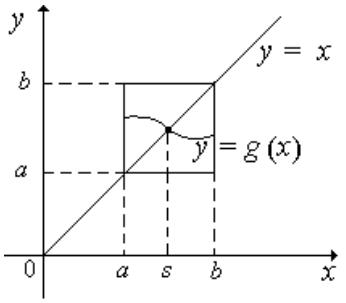


Figura 10

La anterior consideración intuitiva puede expresarse en términos puramente analíticos en la forma que sigue. Consideremos la función f definida por $f(x) = x - g(x)$, que es continua en el intervalo $[a, b]$. Entonces,

$[g(a) \geq a \quad \text{implica} \quad f(a) \leq 0]$ y $[g(b) \leq b \quad \text{implica} \quad f(b) \geq 0]$.

Así, al menos una de las tres condiciones es satisfecha:

(a) $f(a) = 0$
(b) $f(b) = 0$
(c) $f(a) < 0$ y $f(b) > 0$

De (a) y (b) es claro que $f(x)$ tiene una solución en $[a, b]$. De (c), como f es continua, entonces según el teorema del valor intermedio f toma todos los valores entre $f(a)$ y $f(b)$ en algún punto del intervalo $[a, b]$. Por lo tanto, debe tomar el valor cero en, digamos $x = s$, de donde, $0 = f(s) = s - g(s)$; o sea, $s = g(s)$. Luego, el número s es la solución buscada de (10).

El argumento anterior no funciona si $D(g)$ es un intervalo cerrado pero no acotado, por ejemplo, $D(g) = [0, +\infty)$ y $g(x) = x + 1$, que no tiene punto fijo (Figura 11).

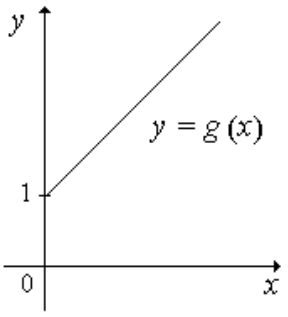


Figura 11

Por lo tanto, necesitamos un argumento que asegure que la curva $y = g(x)$ interseque a la recta $y = x$. Deseamos que la función g crezca menos fuertemente que la función x. Un camino simple para lograr esto es suponer:

g es diferenciable y la derivada satisface:

$|g'(x)| \leq \alpha$, $\forall x \in D(g)$, donde α es una constante que satisface: $\alpha < 1$. (III)

En este caso, g tiene un punto fijo aun si el intervalo de definición no es acotado (siempre suponemos que es cerrado). En efecto, sea $D(g) = [a, +\infty)$. Si $g(a) = a$, a es un punto fijo. Supongamos que $g(a) \neq a$. Entonces, por (I), $g(a) > a$. Para un $x > a$ arbitrario, se tiene

$$g(x) \leq |g(x)| = |g(a) + g(x) - g(a)| \leq |g(a)| + |g(x) - g(a)| \underset{\text{T.V.M.}}{=} |g(a)| + |g'(\xi)|\,|x - a| \leq$$

(ξ un punto comprendido entre a y x; T.V.M.: Teorema del Valor Medio)

$$\underset{\text{(III)}}{\leq} |g(a)| + \alpha\,|x - a| \leq |g(a)| + \alpha\,|x| + \alpha\,|a| = c + \alpha\,|x|$$

siendo $c = |g(a)| + \alpha\,|a| \geq 0$.

Debemos probar que $c + \alpha\,|x| < x$ (pues así $g(x) < x$). Para ello, basta tomar $x > \dfrac{c}{1-\alpha}$. En efecto, si $x > \dfrac{c}{1-\alpha}$, entonces $x > 0$.

Luego, $x - \alpha x > c$, de donde, $x > c + \alpha x = c + \alpha\,|x|$, es decir, $x > c + \alpha\,|x|$ y como $c + \alpha\,|x| \geq g(x)$, entonces $x > g(x)$. (*)

Además, notemos que tomando $x > \dfrac{c}{1-\alpha}$ sigue valiendo que $x > a$.

En efecto, $x > \dfrac{c}{1-\alpha} = \dfrac{|g(a)| + \alpha|a|}{1-\alpha} \geq \dfrac{g(a) + \alpha a}{1-\alpha} > \dfrac{a + \alpha a}{1-\alpha} > \dfrac{a(1+\alpha)}{1-\alpha} \geq a$, es decir, $x > a$.

Así, (de (*)) el gráfico de g corta a la recta $y = x$ en alguna parte entre los puntos a y x; esto es, g tiene un punto fijo (Figura 12).

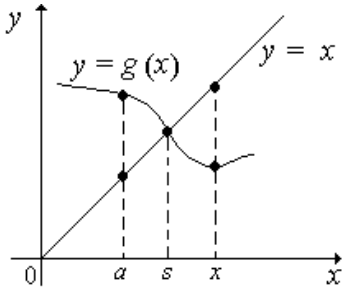


Figura 12

Los casos $D(g) = (-\infty, a]$ y $D(g) = (-\infty, +\infty)$ se tratan en forma análoga.

La condición (III) que nos permite responder afirmativamente la pregunta (i), es suficiente para dar respuestas positivas a las preguntas (ii) y (iii).

Para (ii). Suponemos que g tiene dos puntos fijos s_1 y s_2 distintos. Por definición, esto significa que

$$s_1 = g(s_1) \quad \text{y} \quad s_2 = g(s_2)$$

de donde,

$$s_1 - s_2 = g(s_1) - g(s_2)$$

Luego,

$$| s_1 - s_2 | = | g(s_1) - g(s_2) | \underset{\text{T.V.M.}}{=} |g'(\xi)| \, | s_1 - s_2 | \leq \alpha \, | s_1 - s_2 |$$

(ξ un punto comprendido entre s_1 y s_2). De aquí, $\alpha \geq 1$ contradiciendo el hecho que $\alpha < 1$. Por lo tanto, $s_1 = s_2$, como queríamos demostrar.

Para (iii). Sea s el único punto fijo de g y sea (x_n) la sucesión generada por g con un valor inicial x_0. Tenemos entonces para todo índice $n > 0$, usando (III), que

$$|x_n - s| = |g(x_{n-1}) - g(s)| \leq \alpha \, |x_{n-1} - s| \tag{12}$$

y así, por inducción

$$|x_n - s| \leq \alpha^n \, |x_0 - s| \tag{13}$$

En efecto, para $n = 1$ $|x_1 - s| \leq \alpha \, |x_0 - s|$ que se verifica por (12).

Supongamos cierta (13) para $n = k$, es decir

$$|x_k - s| \leq \alpha^k \, |x_0 - s|$$

y probémosla para $n = k + 1$.

$$|x_{k+1} - s| \underset{\text{por(12)}}{\leq} \alpha \, |x_k - s| \underset{\text{por hip.ind.}}{\leq} \alpha \, \alpha^k |x_0 - s| = \alpha^{k+1} |x_0 - s|$$

Es decir,

$$|x_{k+1} - s| \leq \alpha^{k+1} |x_0 - s|$$

como queríamos demostrar. Por lo tanto, (13) es válida.

En virtud que $0 \leq \alpha < 1$, entonces $\lim_{n\to\infty} \alpha^n = 0$, de donde, se sigue que $\lim_{n\to\infty} | x_n - s | = 0$, lo que significa lo mismo que $\lim_{n\to\infty} x_n = s$, como queríamos demostrar.

Es posible estimar además el error después de un número finito de pasos, es decir, *el error después de n iteraciones.*

Ningún proceso de cálculo puede proseguirse indefinidamente y en cualquier aplicación práctica, el algoritmo descrito según la fórmula (11) debe terminarse después de haberse calculado, digamos, el elemento x_n. Estamos interesados en encontrar una cota para $|x_n - s|$, es decir, para el error de x_n considerado como una aproximación de la solución s. Esta cota debe depender solamente de las cantidades que se conocen a priori y no debe depender de un conocimiento de la solución misma. (Es por esto que un resultado como el dado en (13) no satisface a nuestros propósitos).

Para establecer la cota necesitamos el siguiente resultado auxiliar:

$$\text{para } n = 0, 1, 2 \ldots, \qquad |x_{n+1} - x_n| \leq \alpha^n |x_1 - x_0| \tag{14}$$

Evidentemente, esto es cierto para $n = 0$. Suponiendo que (14) es válida para $n = k - 1$, donde k es un entero mayor que 0, tenemos

$$|x_{k+1} - x_k| = |g(x_k) - g(x_{k-1})| \leq \alpha |x_k - x_{k-1}| \underset{\text{por hip. ind.}}{\leq} \alpha\, \alpha^{k-1} |x_1 - x_0| = \alpha^k |x_1 - x_0|$$

estableciendo (14) para $n = k$. El que (14) sea válida para todos los enteros $n \geq 0$ se sigue del principio de inducción matemática.

Sea ahora $n \geq 0$ un entero fijo y sea $m > n$. Vamos a encontrar una cota para $x_m - x_n$. Escribiendo

$$x_m - x_n = (x_m - x_{m-1}) + (x_{m-1} - x_{m-2}) + \ldots + (x_{n+1} - x_n)$$

y aplicando la desigualdad triangular, obtenemos

$$|x_m - x_n| \leq |x_m - x_{m-1}| + |x_{m-1} - x_{m-2}| + \ldots + |x_{n+1} - x_n|$$

y usando (14) para estimar cada uno de los términos del segundo miembro, obtenemos

$$|x_m - x_n| \leq (\alpha^{m-1} + \alpha^{m-2} + \ldots + \alpha^n)\, |x_1 - x_0| = \alpha^n (1 + \alpha + \ldots + \alpha^{m-n-1}) |x_1 - x_0| \leq$$

$$\leq \alpha^n |x_1 - x_0| \underbrace{\sum_{i=0}^{\infty} \alpha^i}_{\text{serie geométrica de razón } \alpha < 1} = \alpha^n |x_1 - x_0| \frac{1}{1-\alpha} = \frac{\alpha^n}{1-\alpha} |x_1 - x_0|$$

Tenemos pues

$$|x_m - x_n| \leq \frac{\alpha^n}{1-\alpha} |x_1 - x_0|$$

En esta relación hagamos que $m \to \infty$ conservando fijo a n. Como $x_m \underset{m\to\infty}{\to} s$, obtenemos

$$|x_n - s| \leq \frac{\alpha^n}{1-\alpha} |x_1 - x_0| \qquad (15)$$

De (15) se ve claramente que la convergencia del proceso de iteración será más rápida a medida que sea más pequeño el número α.

Con los anteriores resultados hemos probado el siguiente teorema.

Teorema. Sea g una función real cuyo dominio de definición, $D(g)$, es un intervalo cerrado (pero no necesariamente acotado) y tal que g satisface las condiciones (I), (II) y (III). Entonces,

(a) g tiene un único punto fijo s.
(b) La sucesión iterativa (x_n) generada por g converge para cualquier elección del valor inicial $x_0 \in D(g)$, y su límite es s.
(c) Para $n = 0, 1, 2, \ldots$ el error de la aproximación n-ésima x_n definida por el algoritmo dado en (11) está acotado como sigue

$$|x_n - s| \leq \frac{\alpha^n}{1-\alpha} |x_1 - x_0|$$

Ejemplo 5. Sea la ecuación $e^{-x} - \ln x = 0$ con una raíz en el intervalo [1, 2]. Esta ecuación puede ser rescrita en las formas siguientes:

a) $x = x + e^{-x} - \ln x$
b) $\ln x = e^{-x}$, de donde, $x = e^{e^{-x}}$
c) $e^{-x} = \ln x$, de donde, $-x = \ln(\ln x)$, es decir, $x = -\ln(\ln x)$

Cualquiera de estas tres formas de escritura son del tipo (10). Usando (11) y tomando como valor inicial $x_0 = 1$, se obtienen los resultados que se detallan a continuación.

Caso a):
$x_1 = 1.3678794$, $x_2 = 1.3092641$, $x_3 = 1.3098176$,
$x_4 = 1.3097990$, $x_5 = 1.3097996$, $x_6 = 1.3097996$.

Caso b):
$x_1 = 1.4446679$, $x_2 = 1.2659520$, $x_3 = 1.3257399$, $x_4 = 1.3042216$,
$x_5 = 1.3117783$, $x_6 = 1.3091010$, $x_7 = 1.3100466$, $x_8 = 1.3097123$,
$x_9 = 1.3098304$, $x_{10} = 1.3097887$, $x_{11} = 1.3098034$, $x_{12} = 1.3097982$,
$x_{13} = 1.3098001$, $x_{14} = 1.3097994$, $x_{15} = 1.3097997$, $x_{16} = 1.3097996$,
$x_{17} = 1.3097996$.

Caso c):
$x_1 = -\infty$

Las formas de escritura a) y b) son convergentes al aplicar el algoritmo; se advierte en el caso a) una convergencia más rápida. En el caso c) el algoritmo diverge. Esto se debe a:

Caso a): $|g'(x)|_{x=1} = 0.36787944$
Caso b): $|g'(x)|_{x=1} = 0.531463618$
Caso c): $|g'(x)|_{x=1} = \infty$

Las condiciones (I) y (III) del teorema anterior aunque suficientes no son necesarias para la convergencia de la sucesión iterativa. No es difícil encontrar ecuaciones cuya convergencia esté asegurada aunque no se cumplan dichas condiciones, como se muestra en el siguiente ejemplo.

Ejemplo 6. La ecuación $x^3 - x^2 - x - 1 = 0$ tiene una raíz en el intervalo $[1, 2]$. Si escribimos $x^3 = x^2 + x + 1$, entonces $x = 1 + \frac{1}{x} + \frac{1}{x^2}$, donde $g(x) = 1 + \frac{1}{x} + \frac{1}{x^2}$. Luego, g es continua en $[1, 2]$, pero $g(1) = 3 \notin [1, 2]$ y además, $g'(x) = -\frac{1}{x^2} - \frac{2}{x^3}$, de donde, $|g'(x)|_{x=1} = 3 > 1$. Sin embargo, si se comienza la iteración con $x_0 = 1$, se obtiene

$x_1 = 3,\ x_2 = 1.4444444,\ x_3 = 2.1715976,\ x_4 = 1.67254193,\ x_5 = 1.9553676,$
$x_6 = 1.7729558,\ x_7 = 1.8821595,\ x_8 = 1.8735892,\ x_9 = 1.8554268,$
$x_{10} = 1.8294369, \ldots,\ x_{37} = 1.8392868,$

esto es, la sucesión iterativa

$$x_n = 1 + \frac{1}{x_{n-1}} + \frac{1}{x_{n-1}^2}, \quad n = 1, 2, \ldots$$

con $x_0 = 1$ converge a 1.8392868.

Observaciones

1. En las condiciones del Teorema, este método converge para cualquier valor inicial $x_0 \in D(g)$. Por esta razón es *autocorrector*, esto es, un error individual en los cálculos que no vaya por encima de los límites del $D(g)$ no afectará el resultado final, ya que un valor erróneo puede ser considerado como un nuevo valor inicial x_0. Únicamente se habrá trabajado más. La propiedad de autocorrección hace que el método de iteración de punto fijo sea uno de los más fiables. Naturalmente, los errores sistemáticos al aplicar el método pueden hacer que no se obtenga el resultado requerido.

2. Las aproximaciones pueden estimarse mediante otras fórmulas de aplicación en ciertos casos.

Pongamos

$$s - x_n = s - x_{n+1} + x_{n+1} - x_n$$

Resulta evidente que

$$|s - x_n| \leq |s - x_{n+1}| + |x_{n+1} - x_n|$$

de donde, teniendo en cuenta que $g(s) = s$ y $g(x_n) = x_{n+1}$, se tiene

$$|s - x_n| \leq |g(s) - g(x_n)| + |x_{n+1} - x_n| \underset{\text{T.V.M}}{=} |g'(\xi)|\,|s - x_n| + |x_{n+1} - x_n| \underset{\text{(III)}}{\leq} \alpha\,|s - x_n| + |x_{n+1} - x_n|$$

(ξ un punto comprendido entre s y x_n).

Luego,

$$|s - x_n| \leq \frac{1}{1-\alpha}\,|x_{n+1} - x_n| \leq \frac{\alpha}{1-\alpha}\,|x_n - x_{n-1}|$$

(En la última desigualdad de la expresión anterior usamos:

$|x_{n+1} - x_n| = |g(x_n) - g(x_{n-1})| = |g'(\xi')||x_n - x_{n-1}| \leq \alpha\,|x_n - x_{n-1}|$, ξ' un punto comprendido entre x_{n-1} y x_n).

De aquí,

$$|s - x_n| \leq \frac{\alpha}{1-\alpha}\,|x_n - x_{n-1}| \qquad (16)$$

de donde se deduce, en particular, que si $\alpha \leq \frac{1}{2}$, entonces ($1 - \alpha \geq \frac{1}{2}$, de donde, $\frac{\alpha}{1-\alpha} \leq \frac{1}{2}2 = 1$)

$$|s - x_n| \leq |x_n - x_{n-1}|$$

De aquí y de la desigualdad $|x_n - x_{n-1}| \leq \delta$ se deduce que

$$|s - x_n| \leq \delta$$

siendo δ la cota de error admisible para la raíz s.

La fórmula (16) nos permite estimar el error en el valor aproximado x_n a partir de la discrepancia entre dos aproximaciones sucesivas, x_{n-1} y x_n.

En todas las conclusiones anteriores hemos ignorado los errores de redondeo; se ha considerado que las aproximaciones sucesivas se han hallado exactas.

Veamos ahora un ejemplo de cómo utilizar la fórmula (16).

Ejemplo 7. Sabemos que la ecuación $e^x - 4x = 0$ tiene dos raíces, una en el intervalo [0, 1] y otra en el intervalo [2, 3]. Para calcular la raíz en el intervalo [2, 3] escribimos la ecuación en la forma $x = ln\,(4x)$, siendo $g(x) = ln\,(4x)$ que satisface las condiciones para que el algoritmo $x_n = g(x_{n-1})$ sea convergente, pues $ln\,(4x)$ es una función continua en el intervalo [2, 3]; además, $g(x) = ln\,(4x) \in [2, 3]$, $\forall\, x \in [2, 3]$ y $g'(x) = \frac{1}{x}$ es menor que 1 en el intervalo [2, 3].

Eligiendo $x_0 = 2$, se obtienen los valores

$x_1 = 2.0794415$, $x_2 = 2.1183937$, $x_3 = 2.1369525$, $x_4 = 2.1456751$,
$x_5 = 2.1497486$, $x_6 = 2.1516453$, $x_7 = 2.1525272$

Eligiendo para α el valor de la derivada en un punto

$$\xi = \frac{x_6 + x_7}{2} = 2.15208625, \text{ entonces } \alpha = 0.4646654.$$

Luego, se tiene aplicando (16)

$$|s - x_7| < \frac{0.4646654}{0.5353346} |\, 0.0008899 \,| = 0.00076548\,.$$

Por lo tanto, de aquí se infiere, ya que es obvio que $s > x_7$, que

$$s < x_7 + 0.00076548 = 2.1525272 + 0.00076548 = 2.1532927.$$

El valor exacto de esta raíz con 7 decimales es: 2.1532924.

Si se desea calcular la raíz que se encuentra en el intervalo [0, 1], la ecuación $e^x - 4x = 0$ se escribe $x = 1/4\ e^x$, que es continua en el intervalo [0, 1]; además, $g(x) = 1/4\ e^x \in [0, 1]$, $\forall\, x \in [0, 1]$ y $g'(x) = 1/4\ e^x$ se mantiene menor que 1 en el intervalo [0, 1]. Luego, consideramos el algoritmo $x_n = g\,(x_{n-1}) = 1/4\ e^{x_{n-1}}$, $n = 0, 1, \ldots$ Tomando $x_0 = 0$, se obtienen los valores

$x_1 = 0.25$, $x_2 = 0.3210064$, $x_3 = 0.3446286$, $x_4 = 0.3528664$,
$x_5 = 0.3557852$, $x_6 = 0.35682535$, $x_7 = 0.3571965$.

Eligiendo para α el valor de la derivada en un punto

$$\xi = \frac{x_6 + x_7}{2} = 0.3570109, \text{ entonces } \alpha = 0.3572629.$$

Aplicando (16), se tiene

$$|s - x_7| < 0.5558461 \text{ x } 0.0003712 = 0.0002063.$$

Luego,

$$s < 0.3571965 + 0.0002063 = 0.3574028.$$

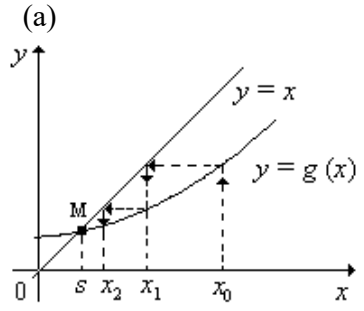
La raíz exacta con 7 decimales es: 0.3574030.

2.2.3.1. Interpretación geométrica.

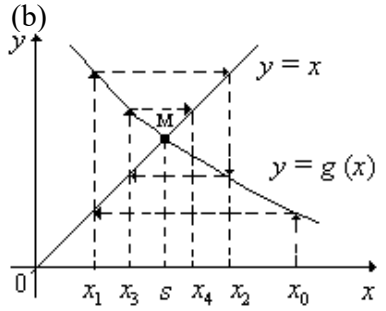
El método de iteración de punto fijo puede representarse geométricamente muy bien como vemos en la Figura 13, donde las curvas $y= x$ e $y = g(x)$ para distintas funciones $g(x)$ han sido representadas. Cada raíz real s de la ecuación $x = g(x)$ es la abscisa del punto de intersección M de la curva $y = g(x)$ con la línea recta $y = x$.

El valor inicial x_0 se usa para determinar el punto correspondiente a la curva $y = g(x)$, $(x_0, g(x_0))$. El punto (x_1, x_1) se encuentra moviéndose en forma horizontal hacia la recta $y = x$. Estos movimientos son equivalentes a la primera iteración del método de iteración de punto fijo: $x_1 = g(x_0)$. De esta manera, en la ecuación y en la gráfica se usa un valor inicial x_0 para obtener la aproximación x_1. La segunda iteración consiste en moverse al punto $(x_1, g(x_1))$ en dirección vertical, y después a (x_2, x_2) en dirección horizontal. Estos movimientos son equivalentes a la segunda iteración del método de iteración de punto fijo: $x_2 = g(x_1)$, y así sucesivamente se sigue con este proceso. Se ve directamente de este modo si el método converge o no. Como nos muestran los gráficos, nos aproximamos a la raíz buscada sólo en los casos (a) y (b), donde $|g'(x)| < 1$, o sea que el proceso de iteración de punto fijo converge. No obstante, si consideramos el caso en que $|g'(x)| > 1$, entonces el proceso de iteración de punto fijo puede ser divergente como lo muestran los casos (c) y (d) en donde las iteraciones divergen de la raíz.

A las gráficas (a) y (c) se las conoce como *patrones monótonos* o *en escalera*, mientras que a (b) y (d) se las conoce como *patrones oscilatorios* o *en espiral*.



$0 \le g'(x) < 1$



$-1 < g'(x) \le 0$

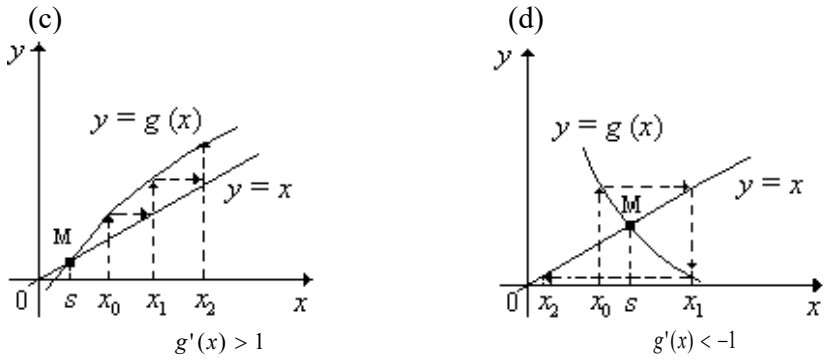


Figura 13

En general, este método converge hacia la raíz buscada y nos lleva al resultado final aun cuando en el transcurso del cálculo algunos errores puedan haber pasado inadvertidos. Verdad es que los errores pueden demorar la convergencia o pueden tener como consecuencia que la sucesión converja hacia otra raíz. Cuando se utiliza una calculadora existe un simple control de cálculo, que consiste en cerciorarse que en los sucesivos pasos de cálculo los valores correlativos de x_n tiendan en forma monótona a un determinado valor límite, como en el caso (a), o que se acerquen alternativamente de ambos lados al valor límite, como en (b).

Ejemplo 8. Resolvamos la ecuación cuadrática $x^2 - 5x + 2 = 0$ partiendo del valor inicial $x_0 = 0$.

Llevamos la ecuación cuadrática a la forma apropiada para un proceso de iteración de punto fijo, a saber

$$x = 0.2\, x^2 + 0.4$$

Designando con x_n a la n-ésima aproximación a una de las raíces, escribimos

$$x_{n+1} = 0.2\, x_n^2 + 0.4, \quad n = 0, 1, 2, \ldots$$

Introduciendo el valor $x_0 = 0$ en el segundo miembro de esta fórmula obtenemos $x_1 = 0.4$; este valor es nuevamente introducido en el segundo miembro lo que permite el cálculo de x_2, y así siguiendo. De esta manera, encontramos

$$x_2 = 0.432, \quad x_3 = 0.43732, \quad x_4 = 0.43825, \ldots$$

Se ve que x_n con n creciente se aproxima a una raíz s de la ecuación cuadrática dada.

Si por el contrario comenzamos con el valor $x_0 = 10$, entonces obtendremos la sucesión de valores

$$x_1 = 20.4, \quad x_2 = 83.6, \quad x_3 = 1398.2, \ldots$$

y encontramos que con n creciente también x_n crece sin límite.

De acuerdo a la fórmula iterativa usada es

$$g(x) = 0.2\,x^2 + 0.4 \quad \text{y} \quad g'(x) = 0.4x$$

La correspondiente ecuación cuadrática tiene dos raíces, $s \approx 0.44$ y $t \approx 4.6$, de modo que es

$$|g'(s)| = 0.176 < 1 \quad \text{y} \quad |g'(t)| = 1.84 > 1$$

De la teoría anterior, resulta que una sucesión iterativa x_n que parte de un valor x_0 cercano a s convergerá hacia la raíz s, mientras que para un valor x_0 cercano a t la sucesión iterativa probablemente divergirá.

Analizando geométricamente estos resultados (Figura 14), vemos que la sucesión iterativa se aproxima a la raíz más pequeña s cuando partimos de un valor inicial x_0 con $|x_0| < t$, pero que el método diverge si es $|x_0| > t$.

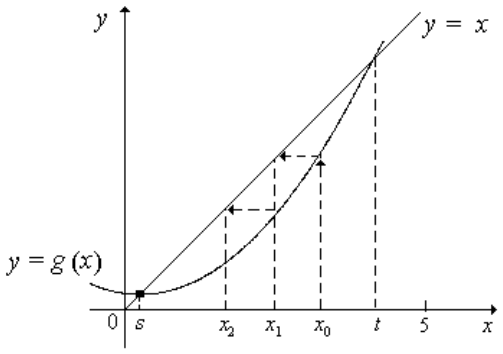


Figura 14

Como hemos dicho anteriormente, una ecuación $f(x) = 0$ puede naturalmente ser utilizada de múltiples maneras en la forma $x = g(x)$ y cada una de tales transformaciones lleva a un método iterativo. Los métodos, en general, sólo se distinguen por sus propiedades de convergencia. Por ejemplo, la transformación de $x^2 - 5x + 2 = 0$ a la forma $x = 5 - \dfrac{2}{x}$ nos proporciona la fórmula de iteración

$$x_{n+1} = 5 - \frac{2}{x_n}, \quad n = 0, 1, 2, \ldots$$

Ahora tenemos $g(x) = 5 - \frac{2}{x}$ y $g'(x) = \frac{2}{x^2}$, de modo que es

$$|g'(s)| = 10.33 > 1 \quad \text{y} \quad |g'(t)| = 0.0945 < 1.$$

Si elegimos la 0-ésima aproximación x_0 cerca de la raíz más grande t, entonces el método convergerá a t; por lo tanto, bajo ninguna condición el método nos proporcionará la raíz s, aun si partimos de un valor inicial cercano a s, como se muestra en la Figura 15.

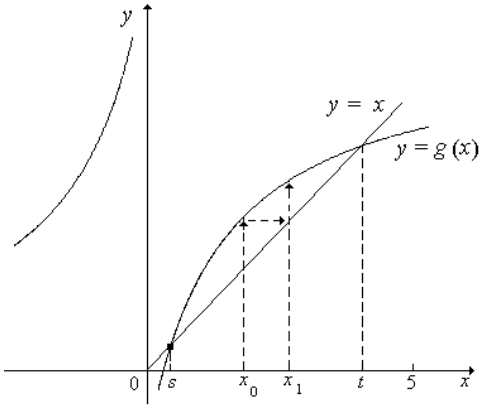


Figura 15

De la representación gráfica de esta situación, se desprende directamente que el método converge a la raíz más grande t para cualquier valor inicial arbitrario x_0.

Ejemplo 9. Consideremos la solución iterativa de la ecuación trivial $x = x^3$, representada en la Figura 16. El método converge hacia la raíz $s = 0$ si elegimos $|x_0| < 1$, pero es divergente si $|x_0| > 1$.

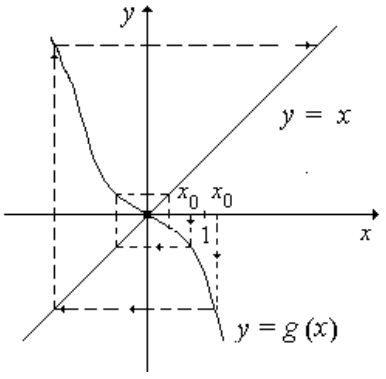


Figura 16

De las discusiones anteriores hemos aprendido que en el método de iteración de punto fijo los siguientes puntos son de considerar:

i) Para el cálculo completo de una determinada raíz de una ecuación $f(x) = 0$ se debe efectuar a modo de ensayo, bajo condiciones, un número de transformaciones de la ecuación a la forma $x = g(x)$ hasta haber encontrado finalmente una forma que lleve a la sucesión iterativa a converger hacia una de las raíces buscadas.

ii) Para la ejecución práctica de un método iterativo es deseable una convergencia lo más rápida posible, y esto se logra para este método cuando la cantidad $|g'(x)|$ sea lo más pequeña posible y el valor inicial x_0 esté lo más cerca posible de la raíz buscada (aunque esto último no es decisivo si $|g'(x)| \geq 1$).

iii) Para decidir si una sucesión iterativa converge o diverge se debe eventualmente no considerar algunos términos del comienzo de la sucesión, ya que una sucesión que por la consideración de sólo los primeros pares de términos aparece como divergente, en realidad puede ser convergente. Por consiguiente, para aplicar de una manera práctica el método de iteración de punto fijo tenemos que asegurarnos que se cumplan las condiciones suficientes de convergencia del proceso de iteración.

iv) En los casos que este método converja hacia la raíz buscada siempre nos proporcionará un resultado correcto, aun cuando en los cálculos intermedios algunos errores se hayan deslizado inadvertidamente.

v) El método de iteración de punto fijo corresponde al método de iteración general

$$x_{n+1} = g(x_n, x_{n-1}, \ldots, x_{n-m+1})$$

donde x_{n+1} estará determinada por los valores de la función $f(x)$ y los valores de la derivada de $f(x)$ en los m puntos $x_n, x_{n-1}, \ldots, x_{n-m+1}$ para el caso $m = 1$, pues en este caso tendríamos

$$x_{n+1} = g(x_n)$$

Es por ello que el método de iteración de punto fijo (y también el método de Newton, párrafo 2.2.7) es llamado método iterativo de un punto. Si $m = 2$, entonces se obtienen los métodos iterativos de dos puntos (por ejemplo, el método de la secante, párrafo 2.2.6), pues para computar x_{n+1} usamos información en dos valores previos.

vi) La exactitud de una raíz aproximada $\tilde{x}$ se estima en función de cómo satisfaga la ecuación dada $f(x) = 0$; es decir, si el número $|f(\tilde{x})|$ es

pequeño, o sea, $|f(\tilde{x})| \leq \varepsilon$, $\varepsilon > 0$ pequeño, prefijado, se considera entonces $\tilde{x}$ una buena aproximación a la raíz exacta s (salvo casos excepcionales, no encontraremos que $f(\tilde{x}) = 0$, es decir, $s = \tilde{x}$). Ahora bien, supongamos $f(x)$, α y β como los de la Figura 17 y un $\varepsilon > 0$ prefijado, también como se indica en dicho gráfico.

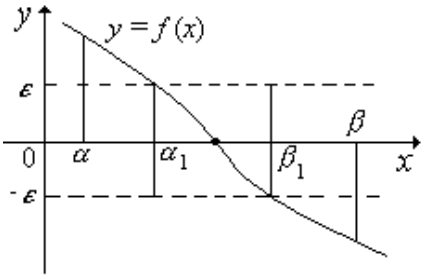


Figura 17

Entonces, todos los puntos $\lambda \in [\alpha_1, \beta_1]$ verifican que $|f(\lambda)| \leq \varepsilon$. Por lo tanto, "todos los puntos λ serían solución del problema". Pero entonces, esta forma de proceder para detener los cálculos porque hallamos la solución dentro de la exactitud prefijada (*criterio de stop*) no nos serviría. Hay que agregarle a esto que en situaciones prácticas no es siempre posible verificar todas las hipótesis del teorema antes visto. En tales casos, uno no tendría más opción que iterar indefinidamente. Lo peor que puede ocurrir es que la sucesión iterativa falle en su convergencia; en caso contrario, tal como hemos visto, el límite de la sucesión iterativa es un punto fijo de g, para una función g continua. Pero matemáticamente hablando, el punto fijo es obtenido tomando $n \to \infty$. Por lo tanto, no podemos permitir que el proceso de iteración se ejecute indefinidamente. Pero entonces, ¿cómo detendríamos el proceso iterativo? Una respuesta teórica está dada por la fórmula (16) que expresa que el error en la n-ésima iteración puede ser estimado en términos de la n-ésima y de la (n-1)-ésima aproximación. Pero en la práctica, una estimación de este tipo ignora el hecho de que α frecuentemente no es conocida (la estimación es arbitrariamente mala cuando $\alpha \to 1$) y además, ignora el hecho de que el sistema numérico de la computadora es discreto. Pero entonces, ¿qué deberíamos hacer?

Sea (x_n) una sucesión convergente matemáticamente con límite s. Denotando por x_n^* la representación de x_n en la máquina uno sospecha, debido a que el sistema de números de una computadora es finito, que la sucesión (x_n^*) finalmente se hace estacionaria, esto es que para algún índice n_0

$$x_n^* = x_{n_0}^*, \quad \forall\, n > n_0 \tag{I}$$

Uno entonces debería parar la iteración si

$$x_{n+1}^* = x_n^* \tag{II}$$

debido a que todos los valores iterados posteriores serían iguales a x_n^*. Sin embargo, aun en una máquina que redondea perfectamente, la condición (I) no necesariamente se verifica pues, por ejemplo, si el límite matemático s está exactamente entre dos números de máquina, los valores numéricos iterados pueden ciclar y la condición (II) nunca valdría. Así, en general, uno no puede más que usar un criterio de la forma

$$|x_n - x_{n-1}| \leq \delta \text{ , siendo } \delta > 0 \text{ pequeño, prefijado.}$$

Con esta última expresión podemos decir que si dos aproximaciones coinciden dentro de la exactitud especificada δ (por ejemplo, los m primeros decimales están estabilizados en esas aproximaciones), entonces se cumple que $s \approx x_n$ con la misma exactitud (esto es, el número aproximado x_n tiene m cifras exactas).

En el caso general, una coincidencia hasta de δ de dos aproximaciones sucesivas, x_{n-1} y x_n, no garantiza que los valores de x_n y la raíz exacta s coincidan con el mismo grado de exactitud, como se muestra en la Figura 18.

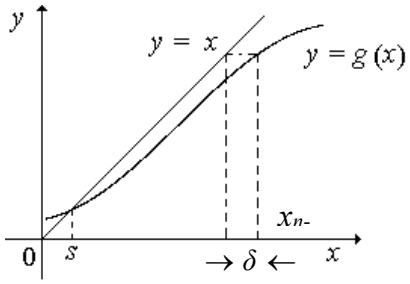


Figura 18

Lo que es más, resulta fácil demostrar que si $g'(x)$ está próxima a la unidad, entonces la cantidad $|s - x_n|$ puede ser grande aun cuando $|x_n - x_{n-1}|$ sea extremadamente pequeña.

Por lo expuesto, adoptaremos como *criterio de stop* o *convergencia* en situaciones prácticas el dado por las desigualdades siguientes

$$|x_n - x_{n-1}| \leq \delta, \quad \delta > 0 \text{ pequeño, fijo}$$

y (17)

$$|f(x_n)| \leq \varepsilon, \quad \varepsilon > 0 \text{ pequeño, fijo}$$

(por simplicidad es común tomar $\delta = \varepsilon$).

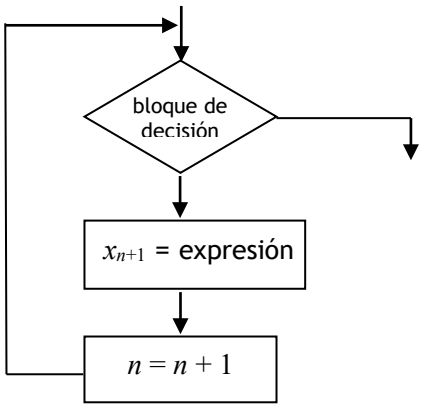
Por supuesto que este criterio de stop lo usaremos siempre y cuando no ocurra que $f(x_n) = 0$. Además, este criterio de stop es el que se adopta en todos los métodos iterativos que desarrollaremos en este Capítulo.

vii) Un punto final que debería tomarse en consideración es la posibilidad de que por una u otra razón la sucesión numérica (x_n) nunca satisfaga el criterio de stop o convergencia, independientemente o no de la máquina. Un programa bien hecho para un método iterativo cualquiera debería contar el número de iteraciones realizadas y el proceso debería detenerse si este número excede una tolerancia predeterminada $n_{máx}$.

viii) Independientemente del método iterativo que se use, el algoritmo numérico utilizado para el cálculo del problema $f(x) = 0$ debe contener los siguientes pasos:

1°) Inicio (ingreso de datos, incluidos valor o valores iniciales, ε, $n_{máx}$, etc.).
2°) $n = 0$ (n indica el paso de iteración).
3°) Test de convergencia (es cuando se toma la decisión de parar o no. Si se detiene, entonces ir a la etapa de salida del algoritmo; caso contrario, continuar con el paso siguiente).
4°) Obtención del valor x_{n+1}.
5°) Hacer $n = n + 1$ e ir al paso 3°) para completar el proceso iterativo.

En un diagrama



¿Qué ventaja tiene hacer al principio el test de convergencia? Ahorra tiempo, pues x_0 puede ser ya la solución buscada.

Notas

1. Lo expuesto en vi) (fórmula (17)), vii) y viii) es aplicable a todos los métodos iterativos analizados en este Capítulo.

2. Tanto el método de iteración de punto fijo como los métodos que se desarrollarán a continuación, están diseñados para determinar una raíz real simple de $f(x) = 0$ de acuerdo a un conocimiento previo de su posición aproximada. Otros casos se analizarán más adelante (párrafos 2.3 y 2.4).

2.2.4. Método de bisección (método que usa intervalos).

El *método de bisección*, conocido también como *de corte binario*, *de partición en dos subintervalos iguales* o *método de Bolzano*, es un método donde el intervalo se divide siempre en dos. Si la función cambia de signo en ese intervalo, se evalúa la función en el punto medio. La posición de la raíz se determina situándola en el punto medio del subintervalo dentro del cual ocurre un cambio de signo.

Más precisamente, sea la ecuación $f(x) = 0$, donde la función $f(x)$ es real y continua en el intervalo $[a_0, b_0]$ y $f(a_0) f(b_0) < 0$. Entonces, hay al menos una raíz real entre a_0 y b_0. Con el fin de localizarla más exactamente, se divide el intervalo $[a_0, b_0]$ por la mitad (Figura 19):

$$c_0 = (a_0 + b_0)/2.$$

Si $f(c_0) = 0$, entonces $s = c_0$ es una raíz de la ecuación $f(x) = 0$. Si $f(c_0) \neq 0$, elegimos dicha mitad $[a_0, c_0]$ ó $[c_0, b_0]$ en cuyos extremos la función $f(x)$ tiene signos opuestos. El nuevo intervalo reducido $[a_1, b_1]$ se divide por la mitad y se procede de igual forma, y así siguiendo.

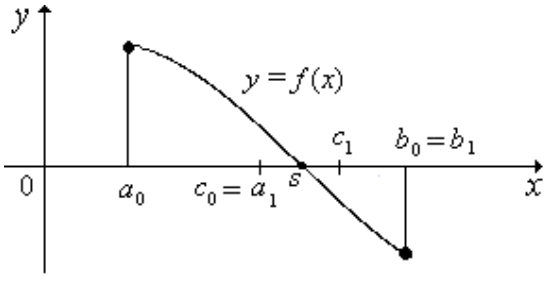


Figura 19

En cada paso, se toma el punto medio del intervalo como la aproximación más actualizada de la raíz. Finalmente, en una cierta etapa del proceso tendremos, o bien la raíz exacta de $f(x) = 0$ o una sucesión infinita de intervalos cada vez más reducidos $[a_1, b_1]$, $[a_2, b_2]$, ... , $[a_n, b_n]$, ... tal que para $n = 1, 2, ...$

$$f(a_n)\, f(b_n) < 0 \tag{18}$$

y

$$b_n - a_n = (b_0 - a_0) / 2^n \tag{19}$$

(tamaño del intervalo $[a_n, b_n]$ después de n pasos de la iteración).

En la Figura 20 se muestra la secuencia de los intervalos y puntos medios.

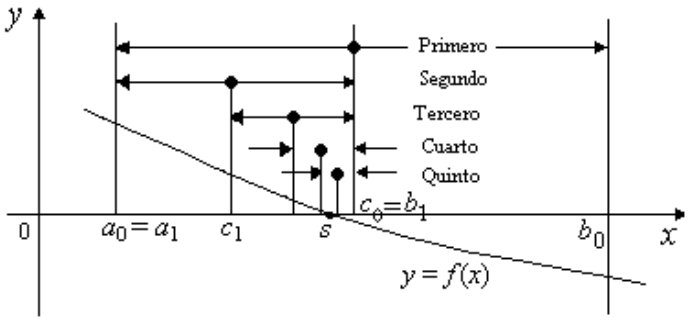


Figura 20

Como los puntos extremos de la izquierda $a_1, a_2, ... , a_n, ...$ forman una sucesión creciente y acotada y los de la derecha $b_1, b_2, ..., b_n, ...$ forman una sucesión decreciente y acotada, existe entonces para (19) un límite común, digamos s

$$s = \lim_{n \to \infty} a_n = \lim_{n \to \infty} b_n \tag{20}$$

(pues $\lim_{n \to \infty} (b_n - a_n) = 0$, de donde, $\lim_{n \to \infty} a_n = \lim_{n \to \infty} b_n$).

Pasando al límite en la desigualdad (18) para $n \to \infty$ tenemos, en virtud de la continuidad de la función $f(x)$, $[f(s)]^2 \leq 0$, de donde $f(s) = 0$, lo cual quiere decir que s es una raíz de la ecuación $f(x) = 0$. Es evidente que

$$0 \leq s - a_n \leq (b_0 - a_0) / 2^n \tag{21}$$

(porque $s - a_n \leq b_n - a_n$).

El tamaño del intervalo después de n pasos de la iteración es $(b_0 - a_0)/2^n$, donde el numerador es el tamaño del intervalo inicial (fórmula (19)).

Esto también representa el máximo error posible cuando la raíz se aproxima mediante el n-ésimo punto medio c_{n-1}.

Por lo tanto, si la tolerancia del error está dada por ε, *el número de pasos de iteración necesarios* es el menor entero positivo que satisface

$$(b_0 - a_0) / 2^n \leq \varepsilon$$

o en forma equivalente

$$n \geq \ln\left(\frac{b_0 - a_0}{\varepsilon}\right) / \ln 2 \qquad (22)$$

Ejemplo 10. Si queremos hallar una solución de $x^3 - 3x^2 + 2x = 0$ en el intervalo [1.5, 3] y con un error que no exceda de $\varepsilon = 1.0 \times 10^{-4}$, entonces la cantidad de ensayos necesarios para obtenerla será (fórmula (22))

$n \geq ln\ (1.5/\ 10^{-4}) / ln\ 2 = 9.6158055 / 0.69314718 = 13.8726$; esto es, $n \approx 14$.

Es decir, en aproximadamente 14 pasos se obtendrá lo requerido.

Observaciones.

1. Si $f'(x)$ existe y mantiene el signo dentro del intervalo $[a_0, b_0]$, entonces la raíz s será definitivamente única. Si las raíces simples de la ecuación $f(x) = 0$ no están separadas en el intervalo $[a_0, b_0]$, entonces del hecho que $f(a_0)\ f(b_0) < 0$ se satisface siempre que el intervalo tenga un número impar de raíces, tendremos que el método de bisección encontrará en este caso una de las raíces separadas en el intervalo dado. El método de bisección no puede encontrar una pareja de raíces dobles debido a que la función toca al eje x de manera tangencial en las raíces dobles (Figura 21), no cruza al eje x y, por lo tanto, no hay cambio de signos de la función en los extremos de un intervalo que contenga a la raíz.

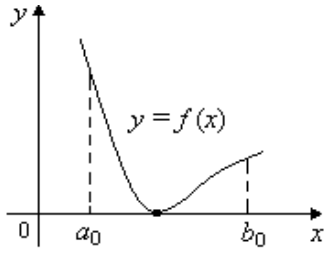


Figura 21

2. Una tarea importante que se debe realizar antes de aplicar el método de bisección es encontrar un intervalo que contenga a la raíz. La búsqueda de raíces se puede llevar a cabo, por ejemplo, listando una tabla de valores, graficando la función en la pantalla, etc.

3. El método de bisección encuentra una raíz de una ecuación si se sabe que la raíz existe en un intervalo dado, es decir que siempre converge para todas las funciones continuas, al igual que el método de la regula falsi (que veremos en el próximo párrafo), aunque la velocidad de convergencia de estos métodos es, por lo general, lenta (sólo un dígito binario es ganado en cada paso).

4. En resumen, si las condiciones se cumplen (que la función continua tenga signos contrarios en los extremos del intervalo $[a_0, b_0]$), entonces el método de bisección converge con seguridad y la convergencia se dice que es *global* (converge en todo el dominio). Es por ello que se suele usar como "arranque" de otros métodos, pues es muy conveniente para dar una idea rápida de la raíz de la ecuación dada, ya que el número de cálculos aumenta sustancialmente a medida que se desea mayor exactitud (convergencia lenta). Así, el último valor obtenido por este método se usa como aproximación inicial para otro método más rápido, en lo que respecta a la velocidad de convergencia.

5. Es un método muy didáctico y su algoritmo es sencillo. Es muy adecuado para computadoras. Se establece una rutina de cálculo de forma que la máquina halle el valor del segundo miembro de la ecuación $f(x) = 0$ para el punto medio de cada uno de los intervalos $[a_n, b_n]$, $n = 0, 1, 2, \ldots$ y elija la mitad apropiada. La iteración es terminada cuando la mitad del intervalo está dentro de una tolerancia dada ε y cuando el módulo del valor de la función en ese punto medio está dentro de esa tolerancia ε (a menos que el valor de la función se haga cero en ese punto medio).

Ejemplo 11. Se sabe que una raíz de $x^3 - 2x - 5 = 0$ está en el intervalo [2, 3]. Utilizando el método de bisección hallemos un valor aproximado a la raíz con una tolerancia de $\varepsilon = 1 \times 10^{-7}$.

Consideremos $f(x) = x^3 - 2x - 5$ que es continua en el intervalo [2, 3] y $f(2) = -1 < 0$, $f(3) = 16 > 0$, por lo tanto, debe anularse por lo menos una vez en el intervalo $[a_0, b_0] = [2, 3]$.

Como $f'(x) = 3x^2 - 2 > 0, \forall x \in [2, 3]$, entonces $f(x) = 0$ tiene una única raíz en el [2, 3].

El cálculo manual del método de bisección se puede llevar a cabo elaborando una tabla como se muestra a continuación. Cuando empieza la primera iteración, los valores de $a_0 = 2$, $b_0 = 3$ y el punto medio $c_0 = (2 + 3) / 2 = 2.5$ se escriben en la tabla, en el renglón $i = 1$. También se calculan $f(a_0)$, $f(b_0)$ y $f(c_0)$ y se escriben en el mismo renglón. Al examinar los signos de estos tres valores de f vemos que la raíz se localiza entre a_0 y c_0. Por lo tanto, a_0 y c_0 del paso $i = 1$ se convierten, respectivamente, en a_1 y b_1 para el paso $i = 2$. Así, $f(a_0)$ y $f(c_0)$ del paso $i = 1$ se copian a $f(a_1)$ y $f(b_1)$ para el paso $i = 2$. La c_1 para el paso $i = 2$ es $c_1 = (a_1 + b_1) / 2 = 2.25$ y se calcula $f(c_1)$ escribiendo su valor en la tabla. La iteración para el resto continúa de manera similar hasta que se alcanza la tolerancia. El último valor de c_{i-1} es la respuesta final.

Nro. de Iter. i	a_{i-1}	c_{i-1}	b_{i-1}	$f(a_{i-1})$	$f(c_{i-1})$	$f(b_{i-1})$	Cota del error $\lvert c_{i-1} - a_{i-1} \rvert$
1	2	2.5	3	-1	5.625	16	.5
2	2	2.25	2.5	-1	1.890625	5.625	.25
3	2	2.125	2.25	-1	0.3457031	1.890625	.125
4	2	2.0625	2.125	-1	-0.351384	0.3457031	.0625
5	2.0625	2.09375	2.125	-0.351384	$-8.9417^{(-3)}$	0.3457031	.03125
6	2.09375	2.109375	2.125	$-8.9417^{(-3)}$	0.1668358	0.3457031	.015625
7	2.09375	2.1015625	2.109375	$-8.9417^{(-3)}$	0.0785622	0.1668358	$7.8125^{(-3)}$
8	2.09375	2.09765625	2.1015625	$-8.9417^{(-3)}$	0.0347143	0.0785622	$3.90625^{(-3)}$
9	2.09375	2.095703125	2.09765625	$-8.9417^{(-3)}$	0.01286235	0.0347143	$1.9531^{(-3)}$
10	2.09375	2.094726562	2.095703125	$-8.9417^{(-3)}$	$1.954376^{(-3)}$	0.01286235	$9.7656^{(-4)}$
11	2.09375	2.094238281	2.094726562	$-8.9417^{(-3)}$	$-3.49516^{(-3)}$	$1.954316^{(-3)}$	$4.88281^{(-4)}$
12	2.094238281	2.094482421	2.094726562	$-3.49516^{(-3)}$	$-7.70842^{(-4)}$	$1.954316^{(-3)}$	$2.4414^{(-4)}$
13	2.094482421	2.094604491	2.094726562	$-7.70842^{(-4)}$	$5.91618^{(-4)}$	$1.954316^{(-3)}$	$1.2207^{(-4)}$
14	2.094482421	2.094543456	2.094604491	$-7.70842^{(-4)}$	$-8.4612^{(-5)}$	$5.41618^{(-4)}$	$6.1035^{(-5)}$
15	2.094543456	2.094573973	2.094604491	$-8.9612^{(-5)}$	$2.51054^{(-4)}$	$5.91618^{(-4)}$	$3.0517^{(-5)}$
16	2.094543456	2.094558714	2.094573973	$-8.9612^{(-5)}$	$8.0672^{(-5)}$	$2.51054^{(-4)}$	$1.5258^{(-5)}$
17	2.094543456	2.094551085	2.094558714	$-8.9612^{(-5)}$	$-4.47^{(-6)}$	$8.0672^{(-5)}$	$7.629^{(-6)}$
18	2.094551085	2.094554899	2.094558714	$-4.47^{(-6)}$	$3.8102^{(-5)}$	$8.0672^{(-5)}$	$3.814^{(-6)}$
19	2.094551085	2.094552992	2.094554899	$-4.47^{(-6)}$	$1.6816^{(-5)}$	$3.8102^{(-5)}$	$1.907^{(-6)}$
20	2.094551085	2.094552038	2.094552992	$-4.47^{(-6)}$	$6.224^{(-6)}$	$1.6816^{(-5)}$	$9.53^{(-7)}$
21	2.094551085	2.094551561	2.094552038	$-4.47^{(-6)}$	$8.78^{(-7)}$	$6.224^{(-6)}$	$4.76^{(-7)}$
22	2.094551085	2.094551323	2.094551561	$-4.47^{(-6)}$	$-1.746^{(-7)}$	$8.78^{(-7)}$	$2.38^{(-7)}$
23	2.094551323	2.094551442	2.094551561	$-1.746^{(-7)}$	$-4.84^{(-7)}$	$8.78^{(-7)}$	$1.19^{(-7)}$
24	2.094551442	2.094551501	2.094551561	$-4.84^{(-7)}$	$1.98^{(-7)}$	$8.78^{(-7)}$	$5.9^{(-8)}$
25	2.094551442	2.094551471	2.094551501	$-4.84^{(-7)}$	$-1.42^{(-7)}$	$1.98^{(-7)}$	$2.9^{(-8)}$
26	2.094551471	2.094551486	2.094551501	$-1.42^{(-7)}$	$2.8^{(-8)}$	$1.98^{(-7)}$	$1.5^{(-8)}$

Como $|f(c_{25})| = 2.8 \times 10^{-8} < 1 \times 10^{-7}$ y $|c_{25} - a_{25}| = 1.5 \times 10^{-8} < 1 \times 10^{-7}$, entonces la aproximación para la raíz s a 7 decimales es $x = 2.0945515$ que se obtuvo en 26 iteraciones.

Observemos que, evidentemente, $|c_{25} - a_{25}| = |c_{25} - b_{25}|$, pues cualquiera de las dos expresiones representan $|(b_{25} - a_{25}) / 2|$, es decir la mitad del intervalo $[a_{25}, b_{25}]$, o sea, la longitud del intervalo $[a_{26}, b_{26}]$ (pues, $|b_{26} - a_{26}| = |(b_{25} - a_{25}) / 2|$). También, como vemos en la tabla anterior $|c_{25} - a_{25}| = |c_{25} - c_{24}|$.

Además, si antes de efectuar los cálculos anteriores hubiésemos querido saber en forma aproximada la cantidad de iteraciones necesarias para obtener la aproximación de la raíz s con una tolerancia de $\varepsilon = 1 \times 10^{-7}$, entonces usaríamos la fórmula (22). En efecto, según (22) y los datos de este problema, $n \geq ln\,[1/1\text{x}10^{-7}] / ln\,2 = 23.25...$, o sea, $n \approx 24$. Según la tabla anterior, vemos que en las 24 iteraciones si bien $|c_{23} - a_{23}| = 5.8 \times 10^{-8} < \varepsilon$ aun no habíamos conseguido que $|f(c_{23})| < \varepsilon$. Ambas condiciones de stop se lograron en 26 iteraciones.

2.2.5. Método de la regula falsi (método que usa intervalos).

Aunque el método de bisección es una técnica perfectamente válida para determinar raíces, su enfoque es relativamente ineficiente. Una alternativa mejorada es la del *método de la regula falsi* (o *regla falsa*), que parte de las mismas suposiciones que el método de bisección pero que está basado en otra idea para aproximarse en forma más eficiente a la raíz. Un defecto del método de bisección es que al dividir el intervalo $[a_k, b_k]$, $k = 0, 1, ...$ por la mitad no se toma en consideración la magnitud de $f(a_k)$ y de $f(b_k)$. Por ejemplo, si $f(a_k)$ está mucho más cerca de cero que $f(b_k)$ es lógico que la raíz se encuentra mucho más cerca de a_k que de b_k, como se muestra en la Figura 22. Este es el principio en el cual se basa el método de la regula falsi.

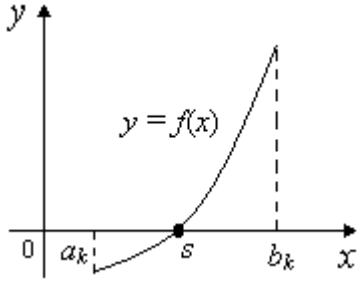


Figura 22

Este método alternativo aprovecha la idea de unir los puntos con una línea recta. La intersección de esta línea con el eje x proporciona una mejor estimación de la raíz, si la función está bien aproximada por la interpolación

lineal (recta que se ajusta a dos puntos extremos dados). Entonces, las raíces estimadas tendrán una buena precisión y en consecuencia, la iteración convergerá más rápido que cuando se utiliza el método de bisección. El hecho de reemplazar una curva por una línea recta da una "posición falsa" de la raíz; de aquí el nombre de *método de la regla falsa*, o en latín, *regula falsi*. También se conoce como *método de interpolación lineal* debido a que se aproxima la curva con una función lineal. Es análogo al método de bisección, puesto que el tamaño del intervalo que contiene a la raíz se reduce mediante iteración.

Como ya se mencionó, en este método se comienza con las mismas suposiciones que en el método de bisección para hallar una raíz s de la ecuación $f(x) = 0$ en un intervalo especificado $[a_0, b_0]$. Para una mejor definición, supongamos que $f(a_0) < 0$ y $f(b_0) > 0$ (Figura 23).

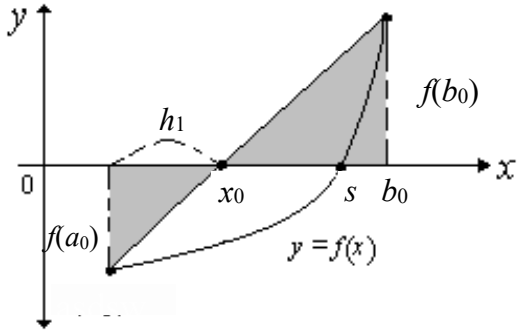


Figura 23

En lugar de dividir por la mitad el intervalo $[a_0, b_0]$ es más natural dividirlo en la relación $f(a_0):f(b_0)$. De esta forma, obtendremos un primer valor aproximado de la raíz

$$x_0 = a_0 + h_1 \tag{23}$$

siendo

$$h_1 = \frac{-f(a_0)}{-f(a_0)+f(b_0)}(b_0 - a_0) = -\frac{f(a_0)}{f(b_0)-f(a_0)}(b_0 - a_0) \tag{24}$$

pues (por triángulos semejantes)

$$\frac{-f(a_0)}{f(b_0)} = \frac{h_1}{b_0 - a_0 - h_1}$$

de donde,

$$-f(a_0)(b_0 - a_0) + f(a_0)h_1 = f(b_0)h_1$$

de aquí,

$$-f(a_0)(b_0-a_0)=[-f(a_0)+f(b_0)]h_1$$

y, por lo tanto,

$$h_1=\frac{-f(a_0)}{-f(a_0)+f(b_0)}(b_0-a_0).$$

Aplicando este procedimiento al intervalo $[a_0, x_0]$ ó $[x_0, b_0]$ en cuyos extremos la función $f(x)$ tenga signos opuestos tendremos una segunda aproximación x_1 de la raíz s; y así siguiendo. El tamaño del intervalo que contiene a la raíz se reduce en cada paso.

Geométricamente, el método de la regula falsi es equivalente a sustituir la curva $y = f(x)$ por una cuerda que pase por los puntos $P_0(a_0, f(a_0))$ y $P_1(b_0, f(b_0))$ (Figura 24).

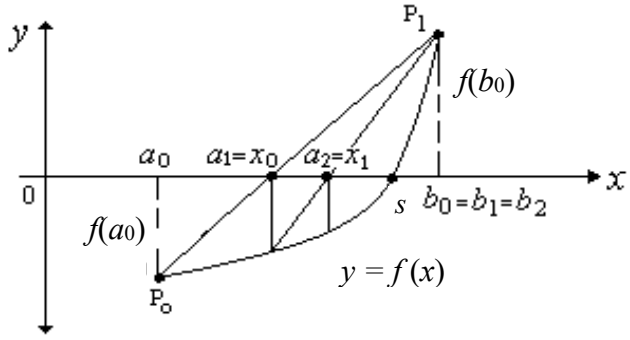


Figura 24

En efecto, la ecuación de la cuerda $P_0 P_1$ es

$$\frac{x-a_0}{b_0-a_0}=\frac{y-f(a_0)}{f(b_0)-f(a_0)}$$

De aquí, considerando $x = x_0$ e $y = 0$, tenemos

$$x_0=a_0-\frac{f(a_0)}{f(b_0)-f(b_0)}(b_0-a_0)$$

o bien

$$x_0=\frac{a_0f(b_0)-b_0f(a_0)}{f(b_0)-f(a_0)}$$

expresión totalmente equivalente a las fórmulas (23) y (24).

Luego,

$$x_n = \frac{a_n f(b_n) - b_n f(a_n)}{f(b_n) - f(a_n)}, \qquad n = 0, 1, 2, \ldots \tag{25}$$

donde a_n, b_n son los sucesivos valores que van tomando los extremos del intervalo, los cuales cumplen que

$$a_n \leq s \leq b_n, \quad n = 0, 1, 2, \ldots, \text{ siendo } s \text{ la raíz.}$$

Observaciones.

1. Si $f'(x)$ existe y mantiene el signo dentro del intervalo $[a_0, b_0]$, entonces la raíz será definitivamente única.

2. Vale la observación 3 dada en el método de bisección, ya que ambos métodos sólo necesitan que la función continua $f(x)$ cambie de signo, lo que asegura la convergencia global y además, que sean métodos siempre convergentes para todas las funciones continuas. Si no se considera que $f(x)$ tenga signos opuestos en los dos puntos usados para generar el siguiente punto, entonces se obtiene el método de la secante (que veremos en el párrafo siguiente), que puede no ser convergente, pero si lo es, converge más rápidamente que los métodos de bisección y de la regula falsi.

3. El método de la regula falsi es esencialmente igual al de bisección, excepto que éste se reemplaza por la interpolación lineal. Además, en el método de la regula falsi no podemos testear el número de iteraciones que se deben realizar.

4. En general, se espera que este método sea más rápido que el de bisección, pero no necesariamente esto ocurre debido a que un extremo puede permanecer fijo. La desventaja de este método es justamente cuando aparecen extremos fijos (Figura 24), en donde uno de los extremos de la sucesión de intervalos no se mueve del punto original, por lo que las aproximaciones a la raíz convergen a la raíz exacta solamente por un lado. Los extremos fijos no son deseables debido a que hacen más lenta la convergencia, en particular, cuando el intervalo inicial es muy grande o cuando la función se desvía de manera significativa de una línea recta en el intervalo, como se muestra en el siguiente ejemplo.

Ejemplo 12. Un caso donde el método de bisección es preferible al de la regula falsi es cuando se debe localizar la raíz de $f(x) = x^{10} - 1$ entre 0 y 1.3, debido a que esta curva viola una hipótesis sobre la cual se basa el método de la regula falsi: si $f(a_k)$ se encuentra mucho más cerca de cero que

$f(b_k)$, entonces la raíz se encuentra más cerca de a_k que de b_k. De acuerdo a la gráfica de esta función (Figura 25), la inversa es verdadera.

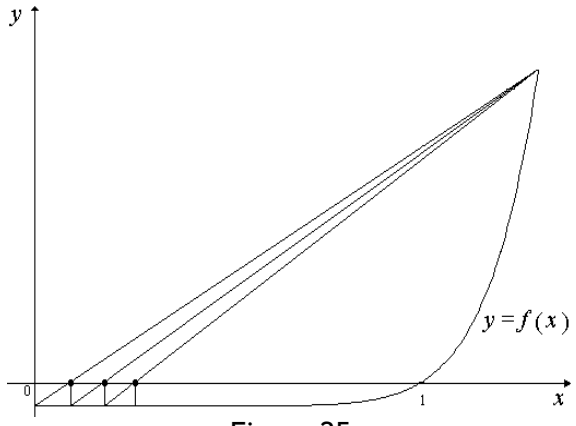


Figura 25

Este ejemplo ilustra que, en general, no es posible hacer generalizaciones con los métodos de obtención de raíces. Aunque un método como el de la regula falsi, en general, es superior al de bisección, hay invariablemente casos especiales que violan las conclusiones generales.

5. La Figura 26 muestra el comportamiento del método de la regula falsi para funciones continuas, no cóncavas.

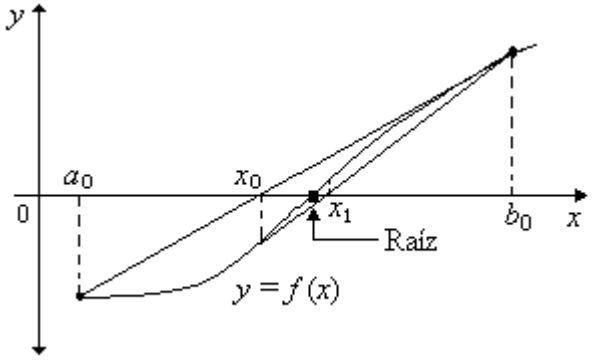


Figura 26

En este caso, la función iteración es de dos puntos.

Una condición suficiente para que este método sea un método de iteración de un punto es que $f(x)$ sea cóncava entre a_0 y b_0. Esto es, que $f''(x)$ sea continua y conserve el signo para $a_0 \leq x \leq b_0$.

Para interpretar geométricamente el resultado anterior, consideremos que la raíz s está separada y supongamos, por definición, que $f''(x) > 0$ para

$a_0 \le x \le b_0$ (el caso $f''(x) < 0$ se reduce al anterior si escribimos la ecuación de la forma $-f(x) = 0$). La curva $y = f(x)$ será cóncava desde arriba y, por lo tanto, estará localizada por debajo de su cuerda $P_0 P_1$. Son posibles dos casos:

(1) $f(a_0) > 0$ (Figura 27) (2) $f(a_0) < 0$ (Figura 28).

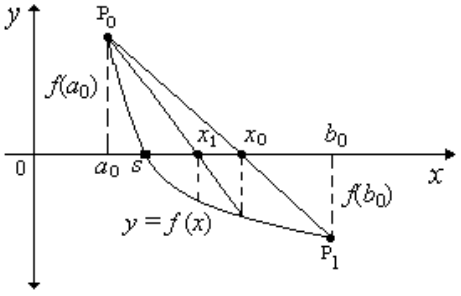


Figura 27

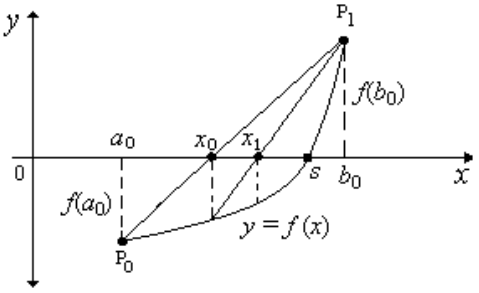


Figura 28

En el primer caso, el extremo a_0 está fijo y las aproximaciones sucesivas x_n de la raíz se encuentran a la derecha de la misma.

En el segundo caso, el extremo b_0 está fijo y las aproximaciones sucesivas x_n de la raíz se encuentran a la izquierda de la misma.

En ambos casos, cada aproximación sucesiva x_n está más próxima a la raíz s que la precedente x_{n-1}. Las aproximaciones sucesivas en el primer caso forman una sucesión decreciente y acotada y en el segundo una sucesión creciente y acotada, de donde se deduce que en ambos casos la sucesión (x_n) es convergente y converge a s.

6. El criterio de stop es el dado en la fórmula (17), analizando previamente si $f(x_n) = 0$.

Ejemplo 13. Hallemos una raíz de $f(x) = x^2 - 10 \log x - 3 = 0$ por medio del método de la regula falsi, la cual se sabe que se encuentra en el intervalo [2, 3] con una tolerancia de $\varepsilon = 5 \times 10^{-3}$.

Como $f(x) = x^2 - 10 \log x - 3$ es continua en el intervalo [2, 3] y $f(2) = -2.0103 < 0$, $f(3) = 1.2287875 > 0$, entonces f debe anularse por lo menos una vez en $[a_0, b_0] = [2, 3]$.

Como

$$f'(x) = 2x - \frac{10 \log e}{x} = \frac{2x^2 - 4.3429448}{x} > 0 \quad (*) \quad \forall\ x \in [2, 3]$$

entonces $f(x) = 0$ tiene una única raíz en el intervalo [2, 3].

Además,

$$f''(x) = 2 + \frac{10 \log e}{x^2} > 0, \qquad \forall\ x \in [2, 3]$$

de donde, la curva $y = f(x)$ es cóncava desde arriba. Además, es estrictamente creciente (por(*)) y como $f(2) < 0$, entonces $b_0 = 3$ será el extremo fijo y las aproximaciones x_n de la raíz s se encuentran a la izquierda de la misma.

Los cálculos se muestran en la tabla que se da a continuación. En el renglón de la primera iteración ($n = 1$) se escriben los valores de $a_0 = 2$, $b_0 = 3$ y los valores calculados de $f(a_0)$ y $f(b_0)$.

El valor de x_0 se halla mediante la fórmula (25) y, en consecuencia, se calcula $f(x_0)$. Estos dos números se escriben en el mismo renglón. Al examinar los signos de estos tres valores de f vemos que la raíz se localiza entre x_0 y b_0. Por lo tanto, los valores x_0, b_0, $f(x_0)$, $f(b_0)$ de la primera iteración se copian en a_1, b_1, $f(a_1)$ y $f(b_1)$ para $n = 2$, respectivamente. El valor de x_1 para $n = 2$ se encuentra mediante la ecuación (25) de la misma forma que en el paso $n = 1$ y se calcula $f(x_1)$ para $n = 2$. Al terminar el renglón $n = 2$, se examinan los signos de estos tres valores de f para localizar la raíz y se continúa de manera similar hasta que se alcanza la tolerancia. El último valor de x_{n-1} es la respuesta final.

Nro. de iter. n	a_{n-1}	x_{n-1}	b_{n-1}	$f(a_{n-1})$	$f(x_{n-1})$	$f(b_{n-1})$	Cota del error $\lvert x_n - x_{n-1} \rvert$
1	2	2.620637756	3	-2.0103	-0.316327654	1.2287875	
2	2.620637756	2.698303661	3	-0.316327654	-0.030065554	1.2287875	$7.7665905^{(-2)}$
3	2.698303661	2.705509161	3	-0.030065554	-0.00271028	1.2287875	$7.2055^{(-3)}$
4	2.705509161	2.706157276	3	-0.00271028	-0.000243098	1.2287875	$6.48115^{(-4)}$

Como $|f(x_3)| = 0.000243098 < 5 \times 10^{-3}$ y $|x_3 - x_2| = 6.48115 \times 10^{-4} < 5 \times 10^{-3}$, obtenemos, luego de 4 iteraciones que $x = 2.706$ con todos sus decimales exactos.

Ejemplo 14. Resolvamos la ecuación dada en el ejemplo 11 utilizando el método de la regula falsi.

Ya analizamos la unicidad de la solución de $f(x) = x^3 - 2x - 5 = 0$ en el intervalo [2, 3].

Además, como $f''(x) = 6x > 0$ en el intervalo [2, 3], entonces el extremo fijo es $b_0 = 3$ y las aproximaciones sucesivas se encuentran a la izquierda de la raíz.

La principal utilidad de disponer los cálculos en una tabla es para examinar los signos de los tres valores de *f*, cuando no hemos analizado si la gráfica de $y = f(x)$ es cóncava o no (por ejemplo, porque el cálculo de la derivada segunda de *f* resulta complicado).

Aquí ya hicimos ese análisis y, por lo tanto, daremos directamente los valores de las sucesivas aproximaciones a la raíz *s* (obtenidas según la fórmula (25)) y los valores de *f* en esas aproximaciones hasta alcanzar la tolerancia.

$x_0 = 2.058823529$, $f(x_0) = -0.390799958$
$x_1 = 2.081263662$, $f(x_1) = -0.147204024$
$x_2 = 2.08963921$, $f(x_2) = -0.05467652$
$x_3 = 2.092739575$, $f(x_3) = -0.02020285$
$x_4 = 2.093883707$, $f(x_4) = -7.450519 \times 10^{-3}$
$x_5 = 2.09430545$, $f(x_5) = -2.7457 \times 10^{-3}$
$x_6 = 2.094460845$, $f(x_6) = -1.01159 \times 10^{-3}$
$x_7 = 2.094518093$, $f(x_7) = -3.72686 \times 10^{-4}$
$x_8 = 2.094539183$, $f(x_8) = -1.37266 \times 10^{-4}$
$x_9 = 2.09454695$, $f(x_9) = -5.06 \times 10^{-5}$
$x_{10} = 2.094549813$, $f(x_{10}) = -1.8626 \times 10^{-5}$
$x_{11} = 2.094550866$, $f(x_{11}) = -6.932 \times 10^{-6}$
$x_{12} = 2.094551257$, $f(x_{12}) = -2.514 \times 10^{-6}$
$x_{13} = 2.094551399$, $f(x_{13}) = -8.98 \times 10^{-7}$
$x_{14} = 2.09455145$, $f(x_{14}) = -4 \times 10^{-7}$
$x_{15} = 2.094551472$, $f(x_{15}) = -1.44 \times 10^{-7}$
$x_{16} = 2.09455148$, $|f(x_{16})| = |-6 \times 10^{-8}| < \varepsilon = 1 \times 10^{-7}$

y además, $|x_{16} - x_{15}| = 8 \times 10^{-9} < \varepsilon$.

Luego, $x = 2.0945515$ obteniéndose el mismo resultado que con el método de bisección, pero ahora mucho más rápidamente (en 17 iteraciones).

2.2.6. Método de la secante (método abierto).

En este método no se utilizan intervalos cerrados donde sabemos que hay una raíz de la ecuación $f(x) = 0$ sino puntos. Como veremos, éste método es muy similar al de Newton y también está íntimamente ligado con el método de la regula falsi.

El método es como sigue: conjeturamos dos valores iniciales x_0, x_1 para x y consideramos la secante a $f(x)$ en $(x_0, f(x_0))$ y $(x_1, f(x_1))$. Luego tomamos como x_2, la abscisa del punto donde la recta secante corta al eje x.

Las aproximaciones sucesivas para la raíz en el *método de la secante* están dadas por la fórmula iterativa de dos puntos

$$x_{n+1} = \frac{x_{n-1} f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}, \quad n = 1, 2, \ldots \tag{26}$$

donde x_0 y x_1 son dos suposiciones iniciales para comenzar la iteración. En este método se toman siempre los dos últimos puntos obtenidos para calcular el siguiente valor. Notemos que el planteamiento requiere de dos puntos iniciales para x. Sin embargo, debido a que no se requiere que $f(x)$ cambie de signo entre estos valores, a este método no se le clasifica como aquellos que usan intervalos.

La fórmula (26) se obtiene teniendo en cuenta los triángulos semejantes $x_0 f(\overset{\Delta}{x_0}) x_2$, $x_1 f(\overset{\Delta}{x_1}) x_2$ (Figura 29).

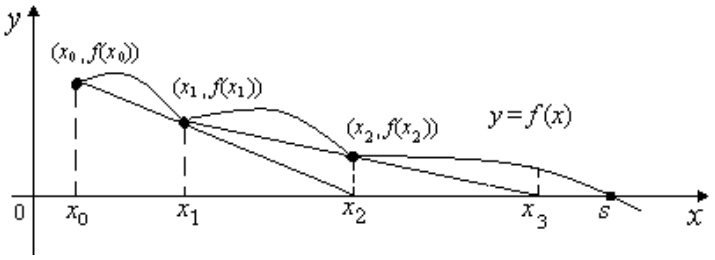


Figura 29

En efecto,

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

de donde,

$$f(x_0)x_2 - f(x_0)x_1 = f(x_1)x_2 - f(x_1)x_0$$

de aquí,

$$x_2[f(x_0) - f(x_1)] = f(x_0)x_1 - f(x_1)x_0$$

y, por lo tanto,

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

Luego,

$$x_{n+1} = \frac{x_{n-1} f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}, \quad n = 1, 2, \ldots$$

como queríamos demostrar.

Observaciones.

1. Este método es mucho más fácil para arrancar que los dos métodos anteriores; todo lo que necesitamos son dos puntos iniciales x_0 y x_1 (suficientemente próximos a la raíz s) y éstos no tienen que estar necesariamente en un intervalo cerrado en cuyos extremos la función cambie de signo. El precio que se paga por la no restricción de los signos opuestos de $f(x)$ en los extremos del intervalo, es que no se puede asegurar la convergencia global (suponiendo que converja); si converge, la convergencia es sólo local (esto es, en un entorno suficientemente pequeño de la solución) y en este caso converge más rápidamente que los anteriores. Pero el método de la secante puede converger a una raíz no deseada o puede no converger si la estimación inicial no es buena. Observemos que no se cumple que $x_{n-1} \leq s \leq x_n$.

2. Otra desventaja de este método parece ser la necesidad de usar aritmética de precisión doble para la iteración cerca de la convergencia, debido a que (26) involucraría entonces la diferencia de dos cantidades casi iguales y se podría producir la cancelación debido a los errores de redondeo si los x_{n-1} y x_n consecutivos son muy cercanos y $f(x_{n-1})\, f(x_n) > 0$. Sin embargo, si rescribimos (26)

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n), \quad n = 1, 2, \ldots \tag{27}$$

el segundo término puede considerarse como una corrección de x_n y como tal, requiere únicamente pocos dígitos significativos conforme se alcance la convergencia. Luego, es conveniente usar la expresión (27) para el método de la secante en lugar de la (26).

3. Notemos la similitud entre los métodos de la secante y de la regula falsi. Por ejemplo, las ecuaciones (26) y (27) del método de la secante y la (25) del método de la regula falsi son idénticas término a término. Ambas usan dos estimaciones iniciales para calcular una aproximación a la pendiente de la función, que se usa para proyectar hacia el eje x una nueva aproximación a la raíz. Sin embargo, existe una diferencia crítica entre ambos métodos y ésta estriba en la forma en que uno de los valores iniciales se reemplaza por la nueva aproximación. Recordemos que en el método de

la regula falsi, la última aproximación x_n de la raíz reemplaza a aquél valor donde la función tiene el mismo signo de $f(x_n)$. En consecuencia, las dos aproximaciones encierran a la raíz. Por lo tanto, en todos los casos prácticos el método siempre converge ya que la raíz se encuentra dentro del intervalo. En contraste, el método de la secante reemplaza los valores en una secuencia estricta, con el nuevo valor x_{n+1} se reemplaza a x_n y x_n reemplaza a x_{n-1}. Como resultado de esto, los dos valores pueden caer de un mismo lado de la raíz. En algunos casos, esto puede provocar la divergencia. En la Figura 30, se comparan los métodos de la regula falsi y de la secante. Las primeras iteraciones (a) y (b) de ambos métodos son idénticas (x_0 y x_2, respectivamente). Sin embargo, en las segundas (c) y (d) los puntos usados son diferentes (obtenemos, respectivamente, x_1 y x_3). En consecuencia, el método de la secante puede divergir, como lo muestra (d). En este caso, se prefiere regula falsi en lugar de secante.

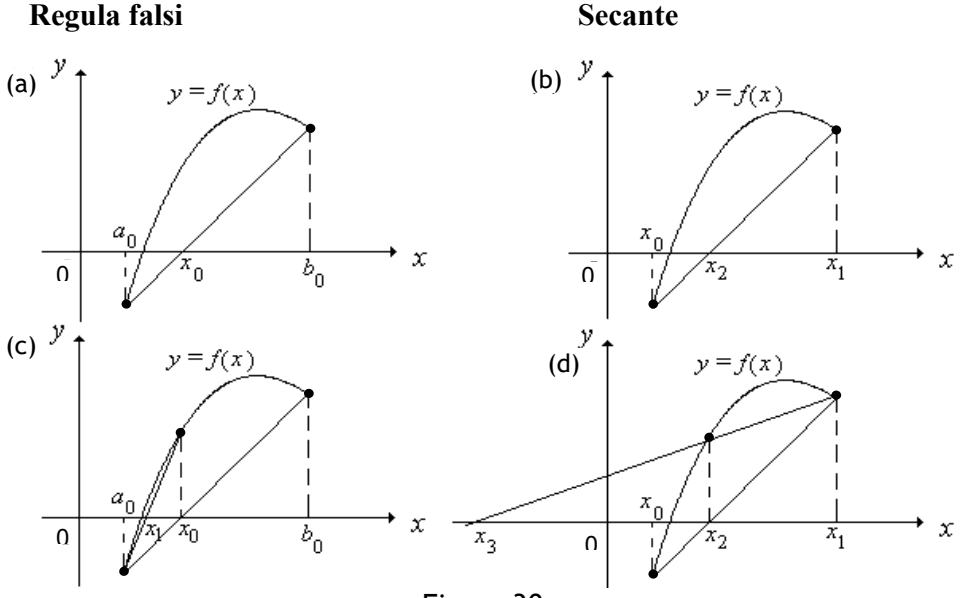


Figura 30

4. Daremos (sin demostración) las condiciones de convergencia para el método de la secante.

Supongamos que la función f está definida y es dos veces continuamente diferenciable en el intervalo $[a, b]$ y que se satisfacen las condiciones

i) $f(a)f(b) < 0$ (garantiza cuando menos un cruce al eje x).

ii) $f'(x), f''(x)$ conservan el signo para $a \leq x \leq b$ (la primera de las condiciones garantiza la unicidad de la solución y la segunda nos dice cómo es la gráfica, o sea si es cóncava desde arriba o desde abajo; esto es, f no tiene puntos de inflexión).

Entonces, partiendo de dos aproximaciones iniciales distintas $x_0, x_1 \in [a, b]$, la sucesión iterativa (x_n) generada por la fórmula (27) converge a la única solución s de la ecuación $f(x) = 0$.

La conclusión es que para un comienzo del método de la secante lo suficientemente bueno, dicho método converge a un cero simple de una función con una derivada segunda continua.

Ejemplo 15. Resolvamos la ecuación dada en el ejemplo 11 utilizando el método de la secante, tomando $x_0 = 2$ y $x_1 = 3$.

Obtenemos

$$\begin{aligned} x_2 &= 2.058823529 \quad , \quad f(x_2) = -0.390799958 \\ x_3 &= 2.081263662 \quad , \quad f(x_3) = -0.14720424 \\ x_4 &= 2.094824142 \quad , \quad f(x_4) = 3.043716 \times 10^{-3} \\ x_5 &= 2.094549433 \quad , \quad f(x_5) = -2.2866 \times 10^{-5} \\ x_6 &= 2.094551481 \quad , \quad f(x_6) = -6.2 \times 10^{-8} \\ x_7 &= 2.094551486 \quad , \quad f(x_7) = 2.8 \times 10^{-8} < \varepsilon \end{aligned}$$

y como $|x_7 - x_6| = 5 \times 10^{-9} < \varepsilon$ ($\varepsilon = 1 \times 10^{-7}$) se tiene que $x = 2.0945515$, obteniéndose el mismo resultado que en los casos anteriores, pero ahora mucho más rápidamente (sólo necesitamos 6 iteraciones para obtener la aproximación a la raíz s según la tolerancia especificada).

2.2.7. Método de Newton – Raphson (o método de las tangentes) (método abierto).

Como vimos, el método de la secante está basado en la idea de que localmente (esto es, en un entorno pequeño) podemos aproximar una función complicada con una función lineal y usamos esta idea para hallar una solución de $f(x) = 0$. Para aproximar una curva localmente según el método de la secante, se eligen dos puntos próximos sobre la curva y entonces aproximamos la curva con la secante que une esos puntos (Figura 31).

igu
ra
31

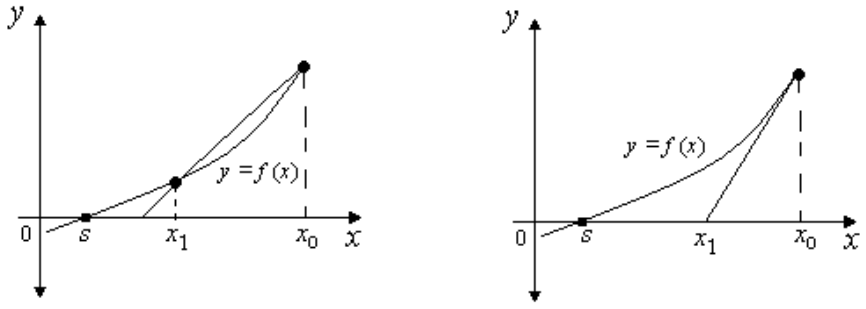


Figura 32

Pero también, otro camino para aproximar una curva localmente es el siguiente: supongamos que tenemos un valor aproximado x_0 de la raíz. Aproximamos, entonces, la curva con su tangente en el punto $(x_0, f(x_0))$. Llamemos x_1 la abscisa del punto de intersección entre el eje x y la tangente (Figura 32). Normalmente, x_1 debería ser mejor aproximación a la raíz que x_0 a menos que la aproximación inicial x_0 sea mala; entonces es posible que x_1 sea peor que x_0.

Una combinación de las ideas de iteración y aproximación lineal local nos da el método de Newton - Raphson. Tal vez, dentro de las fórmulas para localizar raíces, la fórmula de Newton - Raphson sea la más ampliamente usada.

Supongamos que la raíz s de la ecuación

$$f(x) = 0 \tag{28}$$

está separada en el intervalo $[a, b]$ y que $f'(x)$ y $f''(x)$ son continuas y conservan los signos para $a \leq x \leq b$. Habiendo hallado una aproximación n-ésima de la raíz, $x_n \approx s$ $(a \leq x_n \leq b)$, podemos mejorarla mediante el método de Newton - Raphson de la siguiente manera: sea

$$s = x_n + h_n \tag{29}$$

donde h_n es una cantidad pequeña. Aplicando la fórmula de Taylor, tendremos

$$0 = f(s) = f(x_n + h_n) \approx f(x_n) + h_n f'(x_n)$$

de donde,

$$h_n \approx -\frac{f(x_n)}{f'(x_n)}$$

Insertando esta corrección en la fórmula (29), tendremos la aproximación siguiente de la raíz

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)},\ n = 0,\ 1,\ 2,\ \ldots \qquad (30)$$

fórmula iterativa conocida con el nombre de m*étodo de Newton- Raphson* (o *método de las tangentes* o, simplemente, *método de Newton*).

En consecuencia, el método de Newton - Raphson es aquel que partiendo del valor x_0, que es una estimación inicial para la raíz s, se obtiene una sucesión x_1, x_2, ... a partir de la fórmula (30).

Geométricamente, como ya citamos anteriormente, el método de Newton es equivalente a sustituir un pequeño arco de la curva $y = f(x)$ por una tangente trazada por un punto de la curva. Supongamos, por definición, que $f''(x) > 0$ para $a \le x \le b$ y que $f(b) > 0$ (Figura 33).

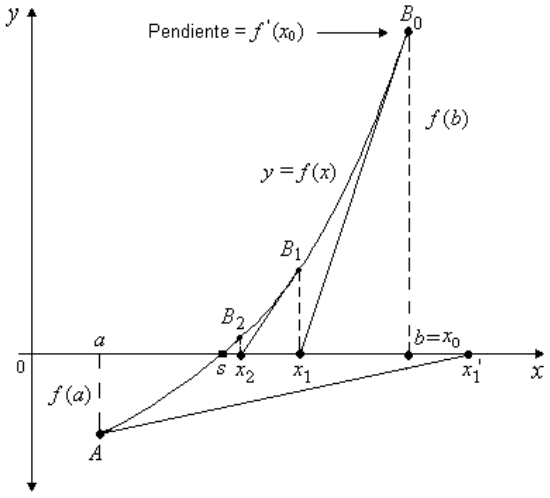


Figura 33

Tomamos, por ejemplo, $x_0 = b$ para el cual $f(x_0)\, f''(x_0) > 0$. Tracemos la tangente a la curva $y = f(x)$ en el punto $B_0(x_0, f(x_0))$. Como primera aproximación x_1 de la raíz s tomamos la abscisa del punto de intersección de esta tangente con el eje x. Tracemos nuevamente una tangente por el punto $B_1(x_1, f(x_1))$, cuya abscisa del punto de intersección con el eje x ofrece una segunda aproximación x_2 de la raíz s, y así sucesivamente. La ecuación de la recta tangente a la curva $y = f(x)$ en el punto $B_n(x_n, f(x_n))$, $n = 0, 1, 2, \ldots$ es la recta que pasa por B_n de pendiente $f'(x_n)$:

$$y - f(x_n) = f'(x_n)\,(x - x_n)$$

(la primera derivada en x_n es equivalente a la pendiente).

Haciendo $y = 0$, $x = x_{n+1}$, tendremos

$$-f(x_n) = f'(x_n)\,(x_{n+1} - x_n)$$

de donde,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)},\ n = 0,\ 1,\ 2,\ \ldots$$

es decir que obtenemos la fórmula (30).

Notemos que si en nuestro caso hacemos $x_0 = a$ y, por lo tanto, $f(x_0)\,f''(x_0) < 0$ y trazamos entonces la tangente a la curva $y = f(x)$ por el punto $A\ (a,\ f\ (a))$, tendremos que el punto x_1' (Figura 33) cae fuera del intervalo $[a,\ b]$; en otras palabras, el procedimiento de Newton no es práctico para este valor inicial. Por lo tanto, en el caso dado, una "buena" aproximación inicial x_0 es aquella para la cual resulta válida la desigualdad

$$f(x_0)\,f''(x_0) > 0. \qquad (31)$$

Demostraremos ahora que esta regla es general.

Teorema. Supongamos que la función f está definida y es dos veces continuamente diferenciable en el intervalo cerrado finito $[a,\ b]$ y que se satisfacen las condiciones

(i) $f(a)\,f(b) < 0$
(ii) $f'(x)$ y $f''(x)$ conservan el signo para $a \le x \le b$.

Entonces, prosiguiendo de la aproximación inicial $x_0 \in [a,\ b]$ la cual satisface (31), es posible utilizando el método de Newton – Raphson (fórmula (30)) calcular la raíz única s de la ecuación (28) con cualquier grado de exactitud.

Antes de pasar a la prueba del teorema daremos algunas explicaciones sobre las hipótesis del mismo. La condición (i) especifica que $f\ (a)$ y $f\ (b)$ deben tener signos distintos y, por lo tanto, que la ecuación $f\ (x) = 0$ tiene al menos una solución en el intervalo $[a,\ b]$. En virtud de la primera de las condiciones (ii), hay solamente una solución en el $[a,\ b]$ (pues esa condición nos dice que $f\ (x)$ es constantemente creciente ó constantemente decreciente); mientras que la segunda de las condiciones (ii) nos dice que la gráfica es cóncava desde arriba o cóncava desde abajo; y la desigualdad (31) nos dice que al aplicar el método de Newton uno debe guiarse por la

siguiente regla: para el punto inicial x_0 elíjase el final del intervalo $[a, b]$ asociado con una ordenada del mismo signo que el de $f''(x)$. En la Figura 34 están representados los cuatro casos bajo estas hipótesis.

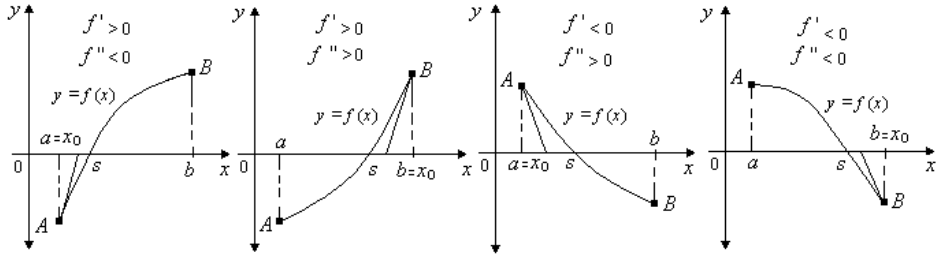


Figura 34

Es importante también observar que el procedimiento no sólo debe aplicarse cuando $f'(x)$ tenga un signo determinado en el intervalo considerado, sino que también se debe considerar que $f''(x)$ no se anule en ese intervalo, pues de lo contrario las tangentes en A y en B podrían ampliar el intervalo en vez de estrecharlo, como lo muestra la Figura 35. (Que $f''(x) = 0$ en el intervalo considerado quiere decir que hay un punto de inflexión en la vecindad de la raíz).

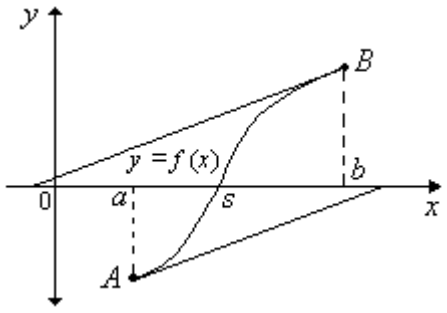


Figura 35

Demostración del teorema. Por ejemplo, supongamos que $f(a) < 0$, $f(b) > 0$, $f'(x) > 0, f''(x) > 0$ para $a \le x \le b$. Por la desigualdad (31) tenemos $f(x_0) > 0$ (podemos, por ejemplo, tomar $x_0 = b$).

Por inducción matemática demostraremos que todas las aproximaciones x_n satisfacen que $x_n > s$, $n = 0, 1, 2, \ldots$ y, por consiguiente, $f(x_n) > 0$ (pues $f(x_n) > f(s) = 0$). En efecto, ante todo $x_0 = b > s$.

Establezcamos ahora $x_n > s$ y consideremos

$$s = x_n + (s - x_n).$$

Desarrollando por la fórmula de Taylor de aproximación cuadrática en el entorno del punto $s = x_n$ tendremos

$$0 = f(s) = f(x_n) + f'(x_n)(s - x_n) + (1/2) f''(\zeta_n)(s - x_n)^2, \text{ donde, } s < \zeta_n < x_n$$

Como $f''(x) > 0$ tendremos, de la igualdad anterior

$$f(x_n) + f'(x_n)(s - x_n) < 0$$

y de aquí

$$s < x_n - \frac{f(x_n)}{f'(x_n)} = x_{n+1}$$

es decir

$$s < x_{n+1}$$

como queríamos demostrar.

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} < x_n$$

Tomando en consideración los signos de $f(x_n)$ y $f'(x_n)$ tenemos, de la fórmula (30)
esto es,

$$x_{n+1} < x_n \quad n = 0, 1, 2, ...$$

Es decir, las aproximaciones sucesivas $x_0, x_1, ..., x_n, ...$ forman una sucesión acotada y decreciente

$$a < s < ... < x_{n+1} < ... < x_2 < x_1 < x_0 = b$$

Por consiguiente, existe el límite $\bar{s} = \lim_{n \to \infty} x_n$. Pasando al límite en (30) para $n \to \infty$, se sigue de la continuidad de f y de f' que

$$\bar{s} = \bar{s} - \frac{f(\bar{s})}{f'(\bar{s})}$$

y de aquí, $f(\bar{s}) = 0$, de donde $s = \bar{s}$ (por unicidad de la solución de la ecuación $f(x) = 0$), completándose la demostración (es decir que la sucesión (x_n) generada por la fórmula iterativa de Newton - Raphson converge a la única raíz s de la ecuación $f(x) = 0$).

Nota 1. De la fórmula (30) está claro que cuanto mayor sea el valor numérico de la derivada, $f'(x)$, en la vecindad de la raíz, tanto menor será la

corrección que a de añadirse a la aproximación n-ésima para obtener la aproximación $(n + 1)$-ésima. El método de Newton es, por consiguiente, muy conveniente cuando la gráfica de la función es pendiente en la vecindad de la raíz dada; pero si el valor numérico de la derivada $f'(x)$ es pequeño cerca de ella, las correcciones serán entonces mayores y calcular la raíz mediante este procedimiento puede ser un proceso largo o a veces incluso imposible.

Resumiendo, no utilice el método de Newton para resolver una ecuación $f(x) = 0$ si la curva $y = f(x)$ es casi horizontal cerca del punto de intersección con el eje x.

Nota 2. Se debe tener en cuenta que en el caso general, una coincidencia hasta de ε de dos aproximaciones sucesivas x_{n-1} y x_n no garantiza que los valores de x_n y la raíz exacta s coincidan con el mismo grado de exactitud (Figura 36).

Además, tampoco suele ser suficiente con sólo analizar que $|f(x_n)| \leq \varepsilon$, $\varepsilon > 0$ pequeño dado (salvo raras ocasiones, sabemos que nunca ocurrirá que $f(x_n) = 0$), ya que puede suceder lo que mostramos en la Figura 37. Por todo lo expuesto, adoptaremos como condiciones de stop para el método de Newton (a menos que $f(x_n) = 0$) las siguientes

$$|x_n - x_{n-1}| \leq \varepsilon \qquad \text{y} \qquad |f(x_n)| \leq \varepsilon$$

para $\varepsilon > 0$ pequeño dado (que nos indica la exactitud deseada para la solución de la ecuación $f(x) = 0$).

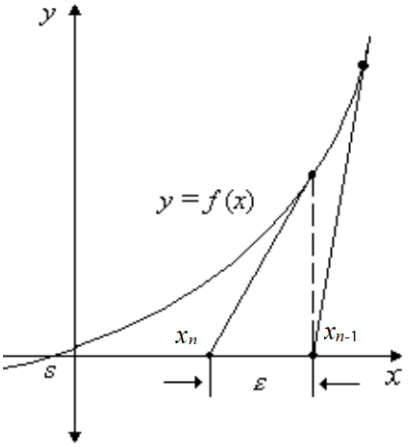


Figura 36

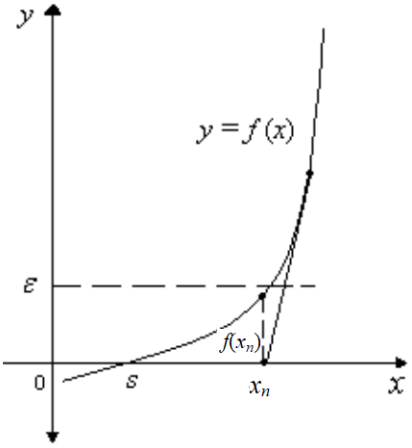


Figura 37

Nota 3. El teorema anterior nos da solamente una condición suficiente para la convergencia del método de Newton, pues puede ocurrir que alguna

de la hipótesis no se cumpla y, sin embargo, la sucesión iterativa obtenida por medio de este método sea igualmente convergente, como se muestra en el siguiente ejemplo.

Ejemplo 16. El propio Newton al aplicar su método a la ecuación $x^3 - 2x - 5 = 0$, que tiene una raíz en el intervalo [2, 2.1], partió del valor $x_0 = 2$ y para ese punto f y f'' tienen signos opuestos, a pesar de lo cual obtuvo correctamente la raíz (de la cual se han llegado a calcular las 101 primeras cifras decimales y cuyo valor es 2.09455148...).

En efecto, si $f(x) = x^3 - 2x - 5$, entonces $f(2) < 0$ y como $f'(x) = 3x^2 - 2$, entonces $f''(x) = 6x$, de donde, $f''(2) > 0$. Así, tomando $x_0 = 2$ en la fórmula iterativa de Newton

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \qquad n = 0,1,2, \ldots$$

se obtiene (con una calculadora con 10 dígitos)

$$x_1 = 2.1$$
$$x_2 = 2.094568122$$
$$x_3 = 2.09455148$$
$$x_4 = 2.09455148$$

Luego, podemos considerar $x = 2.0945515$ (resultado antes obtenido por los otros métodos, pero como podemos ver este método converge mucho más rápidamente que los otros; sólo se necesitaron 3 iteraciones).

Nota 4. El método de Newton es uno de los más utilizados por ser relativamente simple y de convergencia muy satisfactoria en la mayoría de los casos, además de que converge a la raíz mucho más rápidamente que los métodos anteriores. Si se conoce poco sobre la función $f(x)$ es necesario, sin embargo, que el valor de partida x_0 sea bastante cercano a la raíz para que la derivada f' no cambie de signo entre x_0 y s. Si en efecto, f' se anula sin que se anule también f, la tangente a la curva va a cortar al eje $0x$ en el infinito. Bastante cerca de este punto de tangencia horizontal, las tangentes cortan al eje a distancias considerables y se corre el riesgo de oscilar indefinidamente (Figura 38 (a)) o bien alcanzar una pendiente cercana a cero después de lo cual las aproximaciones se alejan del área de interés (Figura 38 (b)). Obviamente, una pendiente cero ($f'(x) = 0$) es un real desastre que causa una división por cero en la fórmula de Newton. Gráficamente (Figura 38 (c)), esto significa que la solución se dispara horizontalmente y jamás toca al eje x.

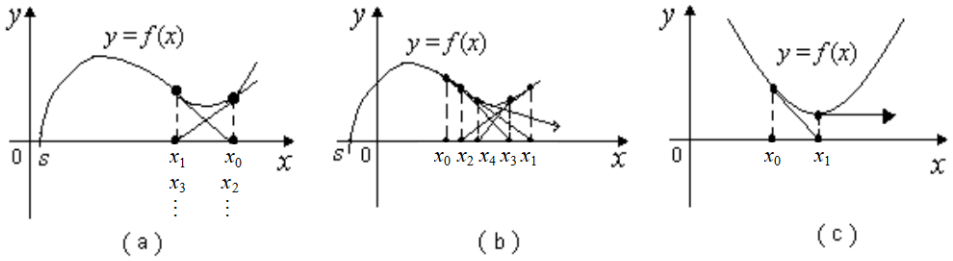


Figura 38

La única solución en estos casos es la de tener un valor inicial cercano a la raíz. Resulta por ello conveniente que este método sea precedido por algún otro método iterativo cuya convergencia sea global, para obtener una "buena" aproximación inicial x_0 a la raíz s de la ecuación $f(x) = 0$ con el fin de evitar la divergencia o convergencia lenta del método de Newton. Por ejemplo, los métodos de bisección y regula falsi son excelentes algoritmos cuando la información a priori sobre la localización de la raíz de $f(x) = 0$ es deficiente, pues en tal caso buscaríamos dos puntos en los que $f(x)$ tiene signos opuestos y entonces, aplicaríamos alguno de estos métodos que sabemos que convergen, aunque lentamente, para todas las funciones continuas, obteniendo así alguna aproximación inicial lo suficientemente cercana a la raíz s, a partir de la cual se puede aplicar el método de Newton para acelerar la convergencia. Pero si tuviéramos como dato un intervalo donde la función no cambia de signo, entonces el método de Newton puede ser precedido por el método de la secante, aunque como sabemos que si bien este método es una perfección substancial sobre el método de la regula falsi, se corre el riesgo de que puede no ser convergente, pues converge si las dos aproximaciones iniciales están suficientemente cerca de la raíz s; esto es, como sabemos, con este método solamente se da la convergencia local al igual que con el método de Newton (si es que convergen), mientras que los métodos de bisección y regula falsi son algoritmos que convergen globalmente.

Pero entonces, ¿por qué el método de la secante es una mejor elección en determinados casos que el método de Newton? Porque cada iteración del método de Newton requiere la evaluación no sólo de $f(x)$ sino que también de $f'(x)$. Aunque esto no es un inconveniente para los polinomios y para muchas otras funciones, existen algunas de éstas cuyas derivadas pueden ser extremadamente difíciles de evaluar. Por lo tanto, en estos casos, el método de la secante es más eficiente.

Nota 5. Método de Newton modificado ó método de von Mises

Si la derivada $f'(x)$ varía aunque ligeramente en el intervalo $[a, b]$, entonces en la fórmula (30) del método de Newton

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}\ , \qquad n = 0,\ 1,\ 2,\ ...$$

podemos poner $f'(x_n) \approx f'(x_0)$.

De aquí, para la raíz s de la ecuación $f(x) = 0$ tendremos las aproximaciones sucesivas

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}, \qquad n = 0,\ 1,\ 2,\ ... \tag{32}$$

fórmula iterativa que se conoce con el nombre *de método de Newton modificado ó método de von Mises* (debido a que *von Mises* introdujo esta modificación en el método de Newton - Raphson).

Geométricamente, este método significa que sustituimos las tangentes en los puntos $B_n(x_n, f(x_n))$, $n \in N$, por líneas rectas paralelas a la tangente a la curva $y = f(x)$ en el punto $B_0(x_0, f(x_0))$ (Figura 39).

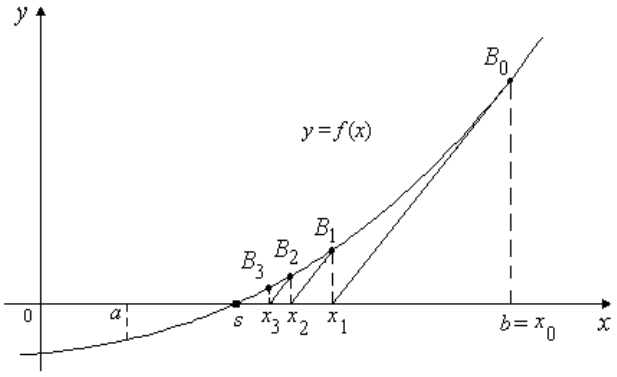


Figura 39

La fórmula (32) nos evita la necesidad de calcular los valores de la derivada $f'(x_n)$ cada vez y, por lo tanto, esta fórmula es muy útil si $f'(x_n)$ es complicada. Puede demostrarse que supuesta la constancia de los signos de las derivadas $f'(x)$ y $f''(x)$, las aproximaciones (32) presentan un proceso convergente.

Nota 6. Pasaje del método de Newton al método de la secante.

Según la fórmula iterativa de Newton

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}\ , \quad n = 0,\ 1,\ 2,\ ... \tag{33}$$

con

$$f'(x_n) = \lim_{\Delta \to 0} \frac{f(x_n + \Delta) - f(x_n)}{\Delta}.$$

Luego, tomando Δ suficientemente pequeño podemos poner

$$f'(x_n) \approx \frac{f(x_n + \Delta) - f(x_n)}{\Delta}$$

Elegimos Δ de modo que $x_n + \Delta = x_{n-1}$, de donde,

$$\Delta = x_{n-1} - x_n. \tag{34}$$

Así,

$$f'(x_n) \approx \frac{f(x_{n-1}) - f(x_n)}{\Delta} = \overline{f}(x_n) \tag{35}$$

y entonces, de (33) y (35), podemos tomar la sucesión

$$x_{n+1} = x_n - \frac{f(x_n)}{\overline{f}(x_n)} \tag{36}$$

método que recibe el nombre de *método de Newton discretizado* y que no es otro que el método de la secante. En efecto, usando en (36) las expresiones dadas en (34) y (35), resulta

$$x_{n+1} = x_n - \frac{f(x_n)}{\overline{f}(x_n)} = x_n - \frac{(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})} f(x_n) = \frac{x_{n-1} f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}$$

fórmula iterativa que corresponde al método de la secante. Luego, podemos considerar al método de la secante como una aproximación al método de Newton. Por ello es que el método de la secante es una alternativa natural para el método de Newton cuando existen dificultades para comenzar a resolver el problema por el proceso iterativo de Newton, como las que se mencionaron anteriormente.

2.3. Raíces múltiples.

Una raíz múltiple corresponde a un punto donde una función es tangencial al eje *x*. Por ejemplo, dos raíces repetidas resultan de

$$f(x) = (x - 3)\,(x - 1)\,(x - 1) \tag{37}$$

o multiplicando términos

$$f(x) = x^3 - 5x^2 + 7x - 3. \tag{38}$$

La ecuación tiene una raíz doble porque un valor de x anula dos términos de la ecuación (37). Gráficamente, esto significa que la curva toca tangencialmente al eje x en la raíz doble. Véase en la Figura 40 (a) en $x = 1$.

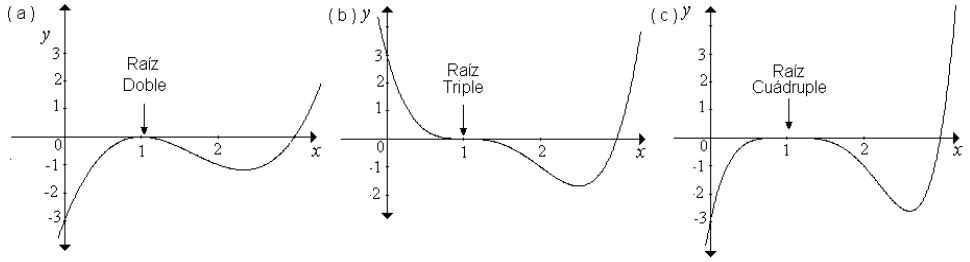


Figura 40

Notemos que la función toca al eje pero no lo cruza en la raíz.

Una raíz triple corresponde al caso en que un valor de x anula tres términos de la ecuación, como en

$$f(x) = (x-3)\,(x-1)\,(x-1)\,(x-1)$$

ó multiplicando

$$f(x) = x^4 - 6x^3 + 12x^2 - 10x + 3$$

Notemos que el gráfico (Figura 40 (b)) indica otra vez que la función es tangencial al eje x en la raíz, pero que en este caso sí cruza al eje. En general, la multiplicidad par de raíces no cruza el eje, mientras que la multiplicidad impar sí lo cruza. Por ejemplo, la raíz cuádruple en la Figura 40 (c) no cruza el eje.

Las raíces múltiples ofrecen ciertas dificultades a los métodos numéricos analizados:

1. El hecho de que la función no cambia de signo en una raíz de multiplicidad par impide el uso de los métodos confiables que usan intervalos. De esta manera, de los métodos incluidos en este Capítulo, los abiertos tienen la limitación de que pueden divergir.

2. Otro posible problema se relaciona con el hecho de que no sólo $f(x)$ se aproxima a cero. Estos problemas afectan a los métodos de Newton y de la secante, los que contienen derivadas o aproximaciones a ella en el denominador de sus respectivas fórmulas. Esto provocaría una división por cero cuando la solución se acerque a la raíz. Una forma simple de evitar estos problemas, que se ha demostrado teóricamente, se basa en el hecho de que si se verifica $f(x)$ contra cero dentro del programa, entonces los cálculos se pueden terminar antes de que $f'(x)$ llegue a cero.

3. Se puede demostrar que en el caso de una raíz simple de la ecuación $f(x) = 0$, el método de Newton converge cuadráticamente o es de orden 2 (el número de cifras decimales correctas se duplica en cada iteración) mientras que el método de la secante, para este mismo caso, tiene un orden de convergencia de 1.618034; pero ambos métodos convergen en forma lineal (orden de convergencia de 1) cuando hay raíces múltiples.

Se han propuesto algunas modificaciones para aliviar este problema. Ralston y Rabinowitz (1978) proponen que se haga un pequeño cambio en la formulación para que retome su convergencia cuadrática en el método de Newton, como

$$x_{n+1} = x_n - m\frac{f(x_n)}{f'(x_n)}, \qquad n = 0, 1, 2, \ldots \tag{39}$$

y en el método de la secante

$$x_{n+1} = x_n - m\frac{(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})} f(x_n), \qquad n = 1, 2, \ldots \tag{40}$$

en donde m es la multiplicidad de la raíz (esto es, $m = 2$ para una raíz doble, $m = 3$ para una raíz triple, y así sucesivamente). De hecho, puede resultar insatisfactorio porque presupone el conocimiento de la multiplicidad de las raíces.

Otra alternativa, también sugerida por Ralston y Rabinowitz (1978), es la de definir una nueva función $u(x)$ que es el cociente de $f(x)$ y de su derivada $f'(x)$; esto es,

$$u(x) = \frac{f(x)}{f'(x)} \tag{41}$$

Se puede demostrar que esta función tiene raíces simples en las mismas posiciones que la función original. Por lo tanto, la ecuación (41) se puede sustituir en la ecuación de la fórmula de Newton

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

y de esta forma desarrollar una alternativa del método de Newton

$$x_{n+1} = x_n - \frac{u(x_n)}{u'(x_n)}, \quad n = 0, 1, 2, \ldots \tag{42}$$

Se puede derivar la ecuación (41), obteniendo

$$u'(x) = \frac{f'(x)f'(x) - f(x)f''(x)}{[f'(x)]^2} \tag{43}$$

Se pueden sustituir las ecuaciones (41) y (43) en la ecuación (42), para obtener

$$x_{n+1} = x_n - \frac{f(x_n)/f'(x_n)}{[f'(x_n)f'(x_n) - f(x_n)f''(x_n)]/[f'(x_n)]^2} =$$

$$= x_n - \frac{f(x_n)f'(x_n)}{[f'(x_n)]^2 - f(x_n)f''(x_n)}$$

es decir,

$$x_{n+1} = x_n - \frac{f(x_n)f'(x_n)}{[f'(x_n)]^2 - f(x_n)f''(x_n)}, \qquad n = 0, 1, 2, \ldots \tag{44}$$

La ecuación (44) es la que se conoce como *método de Newton modificado para el cálculo de raíces múltiples*.

Análogamente, se puede desarrollar una versión modificada del método de la secante para raíces múltiples sustituyendo la ecuación (41) en la ecuación de la fórmula de la secante

$$x_{n+1} = x_n - \frac{(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})} f(x_n)$$

obteniendo

$$x_{n+1} = x_n - \frac{(x_n - x_{n-1})}{u(x_n) - u(x_{n-1})} u(x_n), \quad n = 1, 2, \ldots \tag{45}$$

Se puede sustituir la ecuación (41) en la ecuación (45), obteniendo

$$x_{n+1} = x_n - \frac{(x_n - x_{n-1})[f(x_n)/f'(x_n)]}{[f(x_n)/f'(x_n)] - [f(x_{n-1})/f'(x_{n-1})]} =$$

$$= x_n - \frac{(x_n - x_{n-1}) f(x_n)}{[f(x_n) f'(x_{n-1}) - f(x_{n-1}) f'(x_n)]/f'(x_{n-1})} =$$

$$= x_n - \frac{(x_n - x_{n-1}) f(x_n) f'(x_{n-1})}{f(x_n) f'(x_{n-1}) - f(x_{n-1}) f'(x_n)}$$

es decir,

$$x_{n+1} = x_n - \frac{(x_n - x_{n-1})}{f(x_n) f'(x_{n-1}) - f(x_{n-1}) f'(x_n)} f(x_n) f'(x_{n-1}), \; n = 1, 2, \ldots \tag{46}$$

La ecuación (46) se conoce con el nombre de *método de la secante modificado para el cálculo de raíces múltiples.*

Ejemplo 17. Usemos los dos métodos, el estándar y el modificado de Newton para evaluar la raíz múltiple de la ecuación (38), $f(x) = x^3 - 5x^2 + 7x - 3 = 0$ con un valor inicial de $x_0 = 0$ y una tolerancia de $\varepsilon = 1 \times 10^{-3}$.

Puesto que $f'(x) = 3x^2 - 10x + 7$, entonces el método de Newton estándar para este problema es

$$x_{n+1} = x_n - \frac{x_n^3 - 5x_n^2 + 7x_n - 3}{3x_n^2 - 10x_n + 7} \quad n = 0, 1, \ldots$$

que se puede resolver iterativamente, para obtener

$x_1 = 0.428571428$
$x_2 = 0.685714285$
$x_3 = 0.832865396$
$x_4 = 0.913329892$
$x_5 = 0.95578329$

$x_6 = 0.977655068$
$x_7 = 0.988766121$
$x_8 = 0.994367355$
$x_9 = 0.99717968$
$x_{10} = 0.998589039$
$x_{11} = 0.99929384, \qquad |f(x_{11})| = 9.96 \text{ x } 10^{-7} < \varepsilon$

y además $|x_{11} - x_{10}| = 7.048017 \text{ x } 10^{-4} < \varepsilon$, detenemos entonces el proceso iterativo.

Como ya se había anticipado, el método converge linealmente hasta el valor verdadero de 1.

Para el caso del método de Newton modificado la derivada segunda es $f''(x) = 6x - 10$ y la relación iterativa es (ecuación (44))

$$x_{n+1} = x_n - \frac{\left(x_n^{\,3} - 5x_n^{\,2} + 7x_n - 3\right)\left(3x_n^{\,2} - 10x_n + 7\right)}{\left(3x_n^{\,2} - 10x_n + 7\right)^2 - \left(x_n^{\,3} - 5x_n^{\,2} + 7x_n - 3\right)\left(6x_n - 10\right)} \quad n = 0, 1, \ldots$$

que se puede resolver iterativamente, para obtener

$x_1 = 1.105263157$
$x_2 = 1.003081512$
$x_3 = 1.000018937$
$x_4 = 1.000040715, \qquad |f(x_4)| = 5 \text{ x } 10^{-8} < \varepsilon$

y además $|x_4 - x_3| = 2.1778 \text{ x } 10^{-5} < \varepsilon$.

De esta forma, el método de Newton modificado para raíces múltiples converge más rápidamente (converge cuadráticamente) al valor verdadero de 1 que el método de Newton estándar o para raíces simples.

Se pueden usar ambos métodos para buscar la raíz simple $x = 3$. Usando un valor inicial de $x_0 = 4$, se obtienen los siguientes resultados:

n	Estándar	Modificado para raíces múltiples
1	3.4	2.636363637
2	3.1	2.820224703
3	3.008695653	2.961728234
4	3.000074596	2.998478679
5	3.00000004	2.999997721

De esta forma, ambos métodos son convergentes siendo el método estándar más eficiente.

El ejemplo anterior ilustra los factores de mayor importancia involucrados al escoger el método de Newton modificado para raíces múltiples. Aunque es preferible en raíces múltiples, algunas veces es menos eficiente y requiere más esfuerzo computacional que el método estándar para el caso de raíces simples.

Observemos en el ejemplo anterior que

para el estándar: $|x_5 - x_4| = 7.4556 \times 10^{-5} < \varepsilon$ y $|f(x_5)| = 8 \times 10^{-8} < \varepsilon$

para el modificado: $|x_5 - x_4| = 1.519042 \times 10^{-3} < \varepsilon$ y $|f(x_5)| \neq -9.59 \times 10^{-6} < \varepsilon$

es decir que deberíamos seguir iterando porque aun no hemos alcanzado la tolerancia especificada, mientras que con el estándar sí, y por ello es que resulta ser éste más eficiente para este caso (raíz simple) que el método de Newton modificado para raíces múltiples.

Ejemplo 18. Usemos los dos métodos, el estándar y el modificado de la secante para evaluar la raíz múltiple de la ecuación dada en el ejemplo 17, esto es, $f(x) = x^3 - 5x^2 + 7x - 3 = 0$ con valores iniciales de $x_0 = 0$ y $x_1 = 2$ y una tolerancia de $\varepsilon = 1 \times 10^{-3}$.

El método de la secante estándar para este problema es

$$x_{n+1} = x_n - \frac{(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})} f(x_n) = x_n - \frac{(x_n - x_{n-1})(x_n^3 - 5x_n^2 + 7x_n - 3)}{(x_n^3 - 5x_n^2 + 7x_n - 3) - (x_{n-1}^3 - 5x_{n-1}^2 + 7x_{n-1} - 3)}$$

y, a partir de esta fórmula obtenemos

$$x_2 = 2 - \frac{2}{(-1) - (-3)}(-1) = 2 + \frac{2}{2} = 2 + 1 = 3, \text{ esto es, } x_2 = 3$$

es decir que el método de la secante estándar converge en sólo una iteración pero a la raíz simple $x = 3$, o sea que directamente no detecta a la raíz múltiple $x = 1$ con los valores iniciales de $x_0 = 0$ y $x_1 = 2$. Por lo tanto, es convergente pero converge a una raíz que no es la que buscábamos.

Para el caso del método de la secante modificado para raíces múltiples, la relación iterativa es (ecuación (46))

$$x_{n+1} = x_n - \frac{(x_n - x_{n-1})(x_n^3 - 5x_n^2 + 7x_n - 3)(3x_{n-1}^2 - 10x_{n-1} + 7)}{(x_n^3 - 5x_n^2 + 7x_n - 3)(3x_{n-1}^2 - 10x_{n-1} + 7) - (x_{n-1}^3 - 5x_{n-1}^2 + 7x_{n-1} - 3)(3x_n^2 - 10x_n + 7)}$$

y a partir de la cual obtenemos

$$x_2 = 0.6$$
$$x_3 = 0.818181818$$
$$x_4 = 1.013793101$$
$$x_5 = 0.999422087$$
$$x_6 = 0.999999556$$
$$x_7 = 0.999618123$$
$$x_8 = 0.999713705$$

y como $|x_8 - x_7| = 9.55823 \times 10^{-5} < \varepsilon$ y $|f(x_8)| = 1.65 \times 10^{-7} < \varepsilon$ detenemos aquí el proceso iterativo (7 iteraciones). El método de la secante modificado para raíces múltiples converge al valor verdadero de 1.

2.4. Método de Newton y método de la secante para el caso de raíces complejas.

Ambos métodos se pueden emplear también para obtener raíces complejas aisladas de una ecuación algebraica $f(z) = 0$. Para buscar una raíz compleja la iteración se inicia, en ambos métodos, con un valor complejo utilizando las fórmulas

$$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)},\ n = 0, 1, 2, \ldots \qquad (47)$$

para el método de Newton y el método de la secante, respectivamente.

$$z_{n+1} = z_n - \frac{(z_n - z_{n-1})}{f(z_n) - f(z_{n-1})} f(z_n),\ n = 1, 2, \ldots \qquad (48)$$

Ejemplo19. Supongamos que tenemos como primera aproximación de una de las raíces complejas de la ecuación $z^3 - z - 1 = 0$ a $z_0 = -0.5 + 0.5i$.

Usemos la ecuación (47) para obtener las sucesivas aproximaciones z_n de una de las raíces complejas de $z^3 - z - 1 = 0$ con una tolerancia de $\varepsilon = 1 \times 10^{-2}$.

Evaluamos primero $f(z_0)$ y $f'(z_0)$ para obtener luego z_1.

$f(z) = z^3 - z - 1$, de donde, $f(z_0) = (-0.5 + 0.5i)^3 - (-0.5 + 0.5i) - 1 = -0.25 - 0.25i$

$f'(z) = 3z^2 - 1$, de donde, $f'(z_0) = 3(-0.5 + 0.5i)^2 - 1 = -1 - 1.5i$

$$z_1 = (-0.5 + 0.5i) - \frac{(-0.25 - 0.25i)}{(-1 - 1.5i)} = (-0.5 + 0.5i) - \frac{(0.25 + 0.25i)}{(1 + 1.5i)} =$$

$$= (-0.5 + 0.5i) - \frac{(0.25 + 0.25i)(1 - 1.5i)}{(1 + 1.5i)(1 - 1.5i)} = (-0.5 + 0.5i) - \frac{(0.625 - 0.125i)}{3.25} =$$

$$= \frac{-1.625 + 1.625i - 0.625 + 0.125i}{3.25} = -0.692307692 + 0.538461538i$$

Luego, $z_1 = -0.692307692 + 0.538461538i$

Entonces,

$$f(z_1) = (-0.692307692 + 0.538461538i)^3 - (-0.692307692 + 0.538461538i) - 1 =$$

$$= -0.33181611 + 0.774237594i + 0.6021848 - 0.15612198i + 0.692307692 -$$

$$-0.538461538i - 1 = -0.037323618 + 0.079654076i$$

esto es, $f(z_1) = -0.037323618 + 0.079654076i$

Además,

$$f'(z_1) = 3\,(-0.692307692 + 0.538461538i)^2 - 1 = 1.43786982 - 2.236686387i -$$

$$- 0.86982249 - 1 = - 0.43195267 - 2.236686387i$$

esto es, $f'(z_1) = - 0.43195267 - 2.236686387i$

Disponemos estos cálculos y los subsiguientes en la tabla que damos a continuación:

n	z_n	$f(z_n)$	$f'(z_n)$
0	-0.5 + 0.5i	-0.25 - 0.25i	-1 - 1.5i
1	-0.692307692 + 0.538461538i	-0.037323618 + 0.079654076i	-0.43195267 - 2.236686387i
2	-0.661082354 + 0.561178841i	-3.26287(-3) - 7.150449(-3)i	-0.6336754 - 2.25912592i
3	-0.662340395 + 0.562270273i	-3.24(-5) - 3.56782(-5)i	

En 3 iteraciones hemos obtenido la aproximación para la raíz compleja de $z^3 - z - 1 = 0$ según la tolerancia especificada, puesto que se satisface que $|z_3 - z_2| = 1.6654964 \times 10^{-3} < \varepsilon$ y $|f(z_3)| = 4.8194335 \times 10^{-5} < \varepsilon$.

Tomamos con 3 decimales a $z = -0.662 + 0.562i$ como aproximación para la raíz.

Ejemplo 20. Resolvamos ahora el mismo problema dado en el ejemplo 19 utilizando el método de la secante (48) con valores iniciales $z_0 = -0.5 + 0.5i$ y $z_1 = -0.7 + 0.7i$.

Evaluamos primero $f(z_0)$ y $f(z_1)$ para obtener luego z_2.

$f(z_0) = (-0.5 + 0.5i)^3 - (-0.5 + 0.5i) - 1 = -0.25 - 0.25i$

$f(z_1) = (-0.7+0.7i)^3 - (-0.7 + 0.7i) - 1 = -0.343 + 1.029i + 1.029 - 0.343i + 0.7 - 0.7i - 1 =$

$= 0.386 - 0.014i$

$$z_2 = (-0.7 + 0.7i) - \frac{(-0.2 + 0.2i)}{(0.386 - 0.014i) - (-0.25 - 0.25i)}(0.386 - 0.014i) =$$

$$= (-0.7 + 0.7i) + \frac{(1 - i)}{(15.9 + 5.9i)}(1.93 - 0.07i) =$$

$$= (-0.7 + 0.7i) + \frac{(1.86 - 2i)}{(15.9 + 5.9i)} = (-0.7 + 0.7i) + \frac{(17.774 - 42.774i)}{287.62} =$$

$= -0.638203184 + 0.551282942i$

Luego, $z_2 = -0.638203184 + 0.551282942\, i$

Entonces,

$f(z_2) = (-0.638203184 + 0.551282942i)^3 - (-0.638203184 + 0.551282942i) - 1 =$

$= -0.039864583 - 0.045206848i$

Disponemos estos valores y los subsiguientes en la tabla que damos a continuación:

n	z_n	$f(z_n)$
2	-0.638203184 + 0.551282942i	-0.039864583 - 0.045206848i
3	-0.659122914 + 0.560177184i	$-6.733131 \times 10^{-3} - 5.864732 \times 10^{-3}i$
4	-0.66247417 + 0.562261078i	$3.1613 \times 10^{-5} + 2.69084 \times 10^{-4}i$

En 3 iteraciones hemos obtenido la aproximación para la raíz compleja de $z^3 - z - 1 = 0$ según la tolerancia especificada, puesto que se satisfacen las relaciones: $|z_4 - z_3| = 3.9463313 \times 10^{-3} < \varepsilon$ y $|f(z_4)| = 2.7093464 \times 10^{-4} < \varepsilon$.

Tomamos con 3 decimales a $z = -0.662 + 0.562i$ como aproximación para la raíz.

2.5. Elementos de juicio.

El siguiente cuadro proporciona un resumen de los factores de mayor importancia que se emplean en la solución de ecuaciones algebraicas y trascendentales.

Comparación de las características de los métodos para encontrar raíces de ecuaciones algebraicas y trascendentales. Las comparaciones se basan en la experiencia general y no se toma en consideración el comportamiento de funciones especiales.

Método	Valores iniciales	Promedio relativo de convergencia	Estabilidad	Exactitud	Alcance de aplicación	Esfuerzo de programación	Comentarios
Directo	----	--------	--------	--------	Muy limitada	-------	--------
Gráfico	----	--------	--------	Baja	General	-------	Puede llevar más tiempo que el método numérico.
Bisección	2	Lenta	Siempre converge.	Buena	General	Fácil	Permite el conocimiento previo de pasos de iteración necesarios.
Regula falsi	2	Media	Siempre converge.	Buena	General	Fácil	----------
Iteración de punto fijo	1	Lenta	Puede no converger.	Buena	General	Fácil	----------
Newton	1	Rápida	Puede no converger.	Buena	Limitada si $f'(x) = 0$.	Fácil	Requiere la evaluación de $f'(x)$. Aplicable a raíces complejas.
Newton modificado para raíces múltiples	1	Rápida para raíces múltiples. Media para raíces simples.	Puede no converger.	Buena	Diseño específico para raíces múltiples.	Fácil	Requiere la evaluación de $f'(x)$ y $f''(x)$.
Secante	2	Entre media y rápida.	Puede no converger.	Buena	General	Fácil	Los valores iniciales no tienen que encerrar la raíz. Aplicable a raíces complejas.
Secante modificado para raíces múltiples	2	Entre media y rápida para raíces múltiples. Entre lenta y media para raíces simples.	Puede no converger.	Buena	Diseño específico para raíces múltiples.	Fácil	Requiere la evaluación de $f'(x)$.

2.6. Algoritmo de Newton - Raphson. Pseudocódigo.

En general, como vehículo para describir un algoritmo se usará un pseudocódigo. Este pseudocódigo especifica los datos de entrada a proporcionar y la forma de los resultados deseados. No todos los procedimientos numéricos darán resultados satisfactorios para datos de entrada escogidos arbitrariamente. Los pasos en los algoritmos se arreglan de tal manera que la dificultad de traducir el pseudocódigo en cualquier lenguaje de programación apropiado para aplicaciones científicas, sea mínima. Para ejemplificar daremos el correspondiente al método de Newton - Raphson.

Para encontrar una solución de $f(x) = 0$ dada una aproximación inicial x_0:
ENTRADA aproximación inicial x_0; tolerancia TOL; número máximo de iteraciones n_0.
SALIDA solución exacta x ó solución aproximada x ó mensaje de fracaso.
Paso 1 Tomar $n = 1$.
Paso 2 Mientras que $n \leq n_0$ seguir los Pasos 3 - 8.
Paso 3 Si $f'(x_0) = 0$ entonces
SALIDA ("División por cero. Antes de la división por cero se tenía como aproximación a la raíz: "; x_0; "en"; n - 1; "iteraciones"); (Procedimiento completado sin éxito).
PARAR.
Paso 4 Tomar $x = x_0 - f(x_0) / f'(x_0)$. (Calcular x_n).
Paso 5 Si $f(x) = 0$ entonces
SALIDA ("Raíz exacta: "; x; "en"; n; "iteraciones"); (Procedimiento completado satisfactoriamente).
PARAR.
Paso 6 Si $|x - x_0| \leq TOL$ y $|f(x)| \leq TOL$ entonces
SALIDA ("Raíz aproximada: "; x; "en"; n; "iteraciones"); (Procedimiento completado satisfactoriamente).
PARAR.
Paso 7 Tomar $n = n + 1$.
Paso 8 Tomar $x_0 = x$. (Redefinir x_0).
Paso 9 SALIDA ("El método fracasó después de n_0 iteraciones, n_0 = "; n_0;" y se tenía como solución aproximada: ";x); (Procedimiento completado sin éxito).
PARAR.

EJERCICIOS PROPUESTOS .....

1. Se desea resolver por el método iterativo de punto fijo la ecuación $x + ln\ x = 0$ cuya solución está próxima a $x_0 = 0.5 \in [0.1, 1]$ y se debe elegir entre las distintas fórmulas iterativas:

i) $x_{n+1} = -ln\ x_n$, $n = 0, 1, 2, ...$

ii) $x_{n+1} = e^{-x_n}$, $n = 0, 1, 2, ...$

iii) $x_{n+1} = \dfrac{(x_n + e^{-x_n})}{2}$, $n = 0, 1, 2, ...$

a) ¿Cuáles de las fórmulas **pueden** ser usadas?
b) ¿Cuál fórmula **debería** ser usada?
c) Dar una fórmula **mejor**.

2. a) Demostrar que la fórmula para determinar raíces cuadradas:

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{Q}{x_n} \right) \quad , \quad n = 0, 1, 2, ...$$

es un caso especial de la fórmula de Newton - Raphson con $f(x) = x^2 - Q$.
Aplicar la iteración de raíz cuadrada con $Q = 2, x_0 = 1.5 \in [1, 2]$.

b) Consideremos ahora como método iterativo:

$$x_{n+1} = \frac{Q}{x_n} \quad , \quad n = 0, 1, 2, \ \ldots$$

y tomemos $Q = 2, x_0 = 1.5$ ¿Es convergente a $\sqrt{2}$?

3. Probar que la función $g(x) = 2^{-x}$ tiene un único punto fijo en el intervalo [1/3, 1]. Luego utilizar el método de iteración de punto fijo para encontrar una aproximación con cuatro decimales redondeados y con una cota de error admisible de 5×10^{-4}.

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

sen x se mide en radianes). Con una cota de error admisible de 5 x 10^{-3} encontrar dicho valor, utilizando:

a) El método de bisección.
b) El método de la regula falsi.

5. La concentración de una bacteria contaminante en un lago decrece según la expresión:

$$c(t) = 80\, e^{-2t} + 20\, e^{-0.5t}$$

siendo t el tiempo en horas. Utilizando el método de Newton con un valor inicial de 0 y con una cota de error admisible de 5 x 10^{-3}, determinar el tiempo que se necesita para que el número de bacterias se reduzca a 7.

6. En el año 1225, Leonardo de Pisa estudió la ecuación polinómica $P(x) = x^3 + 2x^2 + 10x - 20 = 0$ y obtuvo $x = 1.368808107$. Nadie sabe qué método utilizó Leonardo para encontrar este valor, aunque fue un resultado notable en ese tiempo. Obtener este resultado y completar la siguiente tabla:

Método	Valor inicial o intervalo	Solución	Iteraciones
Punto fijo	1		
Bisección	[1, 2]		
Newton	1		

7. La ecuación $x - e^{-x} = 0$ tiene una raíz en el intervalo [0, 1].

a) Determinar tres fórmulas iterativas y analizar la convergencia de cada una de ellas en dicho intervalo.
b) Con una cota de error admisible de 1 x 10^{-5}, encontrar la raíz de la ecuación dada indicando el número de iteraciones necesarias para obtenerla, usando:

i) Una de las tres fórmulas obtenidas en el apartado a) y que a su criterio es la mejor (método iterativo de punto fijo).
ii) El método de la regula falsi.
iii) El método de Newton - Raphson modificado o método de von Mises tomando como aproximación inicial $x_0 = 0$. ¿Qué ocurre si se toma como aproximación inicial $x_0 = 1$?

8. Hallar una raíz aproximada en el intervalo [1, 2] de la ecuación $x^3 - x - 1 = 0$ utilizando en los cálculos cuatro decimales redondeados y con una cota de error admisible de 0.1×10^{-2}, mediante:

a) i) El método de bisección.
   ii) El método de la secante.
   iii) El método de Newton comenzando con $x_0 = 1$.
b) Hacer los programas correspondientes a estos tres métodos y verificar los resultados obtenidos en el apartado a).

9. Se sabe que la ecuación $x^3 - 9x^2 + 24x - 20 = 0$ tiene una raíz doble en el intervalo [1.5, 2.5]. Con una tolerancia de 1×10^{-3} obtenerla, usando:

a) Los dos métodos de la secante, el estándar y el modificado para el cálculo de raíces múltiples.
b) Los dos métodos de Newton, el estándar y el modificado para el cálculo de raíces múltiples tomando como aproximación inicial $x_0 = 1.5$.
c) Según los resultados obtenidos en los dos apartados anteriores, ¿cuál de los métodos es el más eficiente para resolver este problema?

10. Se sabe que una raíz compleja de la ecuación $x^2 - x + 4 = 0$ está próxima a $0.5 + 2i$. Con una tolerancia de 1×10^{-3} encontrar la raíz compleja de la ecuación dada, usando:

a) El método de la secante comenzando con $x_0 = 0.5 + 1.5\,i$, $x_1 = 0.5 + 2i$.
b) El método de Newton comenzando con $x_0 = 0.5 + 1.5\,i$.
c) Hacer el programa correspondiente al método de Newton para raíces complejas y verificar el resultado obtenido en el apartado b).

11. a) Usando los programas correspondientes a los métodos de bisección, de la secante y de Newton para calcular raíces reales de una ecuación no lineal y con una cota de error admisible de 1×10^{-5}, hallar las soluciones de las siguientes ecuaciones:

i) $ln\ x + 64 - e^x = 0$ iii) $sen\ x - ln\ x = 0$

ii) $ln\ x - (x-3)^2 + 2 = 0$ iv) $(x^2+1)e^{-x} - x = 0$

b) Usando el programa correspondiente al método de Newton para el cálculo de raíces complejas de una ecuación no lineal y con una cota de error admisible de 5×10^{-6}, encontrar:

i) Dos raíces complejas de la ecuación $2 - x + 2x^2 + x^4 = 0$ tomando como aproximaciones iniciales $-0.5 + 1.5i$ y $0.5 - 0.7i$, respectivamente.

ii) Una raíz compleja de la ecuación $4x^4 + 128x^3 - 95x^2 + 24x - 2 = 0$ tomando como aproximación inicial $0.2 + 0.8i$.

12. En una fábrica de chocolates se decidió envasar los bombones en un modelo de caja que sea un prisma de base x cm, de altura $(x - 2)$ cm, de profundidad $(x + 10)$ cm y cuyo volumen sea igual a 957 cm^3. Para poder armar esta caja se desea conocer las medidas de sus lados. Para ello:

a) Plantear la ecuación correspondiente, según los datos del problema.
b) Separar las raíces de esta ecuación realizando, primero manualmente y luego con la computadora, el gráfico de la función polinómica resultante.
c) En el apartado b) se localizaron las raíces de la ecuación polinómica. Utilizando estos datos y realizando 10 iteraciones del método de bisección, obtener la medida de los lados de este prisma.

13. Se quiere estudiar la variación de la temperatura promedio de la superficie de un planeta, sabiendo que recibe radiaciones de una estrella cuya temperatura aumenta, aunque muy lentamente. La temperatura promedio del planeta se puede calcular a través de la siguiente fórmula: $T(x) = x^3 - 6x^2 + 11x - 5$, con x en millones de años y T en ºC.

Se desea averiguar si en un período de 4 millones de años, la temperatura promedio puede llegar a tomar el valor de 0º C. Para ello:

a) Graficar la función $T(x)$ para observar dónde cruza al eje x.
b) Con los datos obtenidos en el apartado anterior y con una cota de error admisible de $\varepsilon = 0.0005$, emplear el método de Newton para obtener de manera aproximada la solución a este problema.

14. Con el objeto de ordenar sus CD's, César quiere construir una caja en forma de prisma sin tapa cuyo volumen sea 6293 cm^3. Para ello cuenta con un cartón duro rectangular de 60 cm de largo por 45 cm de ancho al que se le corta un cuadrado en cada esquina. Se debe:

a) Encontrar una expresión que sirva para calcular el volumen de la caja armada en función de la medida del lado de cada cuadrado cortado.
b) Graficar la función volumen para separar las raíces de la ecuación correspondiente.
c) Obtener la solución a este problema aplicando el método de Newton con una cota de error admisible de $\varepsilon = 0.0005$.

15. El crecimiento de dos poblaciones A y B está dado por las siguientes fórmulas:

$$P_A(t) = \frac{5}{2}t + 30 \qquad P_B(t) = t^3 - 12t^2 + 44t - 8$$

donde t es el tiempo de conteo expresado en semanas.

a) Realizar un gráfico para separar las raíces de la ecuación polinómica resultante.
b) Con los datos obtenidos en el apartado a) y utilizando los métodos de bisección y de Newton, obtener los puntos de intersección de ambas poblaciones con una cota de error admisible de $\varepsilon = 0.00005$.
c) Según los resultados obtenidos, expresar sus conclusiones.

16. En un Laboratorio comenzaron a las 6 horas a medir la temperatura, en grados, de una sustancia durante cierto día, y obtuvieron la fórmula $P(t) = 0.01t^3 - 0.36t^2 + 2.88t$, donde t es el tiempo medido en horas desde el inicio de la medición. Se desea saber:

i) ¿A qué hora la temperatura era sobre cero?
ii) ¿A qué hora la temperatura era bajo cero?

Para ello se debe:

a) Realizar un gráfico para separar las raíces de la ecuación polinómica resultante.

b) Según los datos obtenidos, aplicar los métodos iterativos de punto fijo y de Newton, con una cota de error admisible de $\varepsilon = 0.0001$, para obtener la información solicitada.
c) A partir de los resultados obtenidos, expresar sus conclusiones.

17. Una fábrica decide envasar 385.84 cm^3 de jugo natural en envases de forma cúbica (tetrabrik) y cilíndrica (latitas), considerando que la base del cilindro es un círculo de 6.4 cm de diámetro. La función resultante es $P(x) = x^3 - 14.16x^2 + 34.6852\,x - 105.15405$. Se quiere determinar la altura de ambos envases, sabiendo que el cubo tiene 4.72 cm de altura menos que el cilindro. Para ello:

a) Graficar la función $P(x)$ para obtener los intervalos en donde se encuentran las raíces reales aproximadas de la ecuación correspondiente.
b) Utilizando los datos obtenidos en el apartado anterior y mediante los métodos de bisección e iterativo de punto fijo, aproximar el valor de la raíz del polinomio con una cota de error admisible de $\varepsilon = 0.005$.
c) Según los resultados obtenidos, expresar sus conclusiones.

18. Sea $f(x) = tg\,x - 0.5x$

a) Graficar la función.
b) Calcular, con una cota de error admisible de 5×10^{-4}, la raíz positiva más pequeña de la ecuación $f(x) = 0$ y completar la siguiente tabla:

Método	Valor inicial o intervalo	Solución	Iteraciones
Punto fijo			
Bisección			
Newton			
Secante			
Regula falsi			

c) Según los resultados obtenidos, expresar sus conclusiones.

19. Consideremos el problema físico de hallar la porción de una esfera de radio r que queda sumergida al meter la esfera en agua (Figura 1). Supongamos que la esfera está construida de pino, que tiene una densidad de $p = 0.638$ gr/cm^3 y que su radio mide $r = 10$ cm.

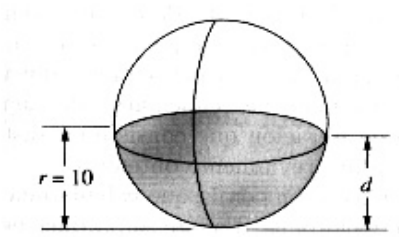


Figura 1

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

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

y la masa de la esfera es $M_e = 4\pi\, r^3 p/3$. Aplicando el principio de Arquímedes, $M_a = M_e$, se obtiene la siguiente ecuación que se debe resolver para d:

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

a) Realizar el gráfico de la función $y = d^3 - 3d^2 r + 4r^3 p$

b) Obtener, con una cota de error admisible de 5 x 10^{-4}, cuánto vale la profundidad a la que está sumergido el polo sur de la esfera y completar la siguiente tabla:

Método	Valor inicial o intervalo	Solución	Iteraciones
Punto fijo			
Bisección			
Newton			
Secante			
Regula falsi			

c) Según los resultados obtenidos, expresar sus conclusiones.

20. Los modelos de crecimiento en un grupo de microbios suponen que el promedio de cambio de la población (p) es proporcional a la población existente en un tiempo (t):

$$\frac{dp}{dt} = kp.$$

La población crece en un medio en el que existe alimento suficiente de manera que k no es una función de la concentración. Cuando el alimento no escasea, el crecimiento se limita sólo por el consumo de productos tóxicos o de espacio, si es que el tamaño de la población crece demasiado. Con el tiempo, estos factores retrasan la tasa de crecimiento de la población y la

detienen completamente cuando ésta alcanza una densidad máxima de $p_{máx}$. En este caso se modifica la ecuación anterior de la siguiente manera:

$$\frac{dp}{dt} = kp(p_{máx} - p)$$

en donde las unidades de k son litros por célula por día. Esta ecuación diferencial se puede integrar de forma analítica dando:

$$p(t) = \frac{p_{máx}}{1 + \left(\frac{p_{máx}}{p_0} - 1 \right) e^{-kp_{máx}t}} \qquad (1)$$

en donde $p(t=0) = p_0$. A esta ecuación se la conoce como "el modelo de crecimiento logístico".

Considérese el crecimiento de una población bacteriológica en un lago. El crecimiento se comporta como lo define la ecuación (1). La población es pequeña en la primavera del año en donde $t = 0$, $p(t=0) = 10$ células por litro. Es sabido que la población alcanza una densidad de 15000 células por litro cuando $t = 60$ días y que la tasa de crecimiento k es de 2×10^{-6} litros por célula por día. Se requiere calcular la densidad de la población bacterial cuando $t = 90$ días, con una cota de error admisible de 5×10^{-4}. Si su número excede de 40000 células por litro, entonces la calidad estándar del agua requiere la implementación de algún procedimiento para disminuirlas y proteger a las personas que se introduzcan al agua. Se debe:

a) Completar la siguiente tabla:

Método	Valor inicial o intervalo	Solución	Iteraciones
Bisección			
Newton			
Secante			
Regula falsi			

b) Según los resultados obtenidos, expresar sus conclusiones.

21. La ley de los gases ideales está dada por:

$$pV = nRT$$

en donde p es la presión absoluta, V es el volumen, n es el número de moles, R es la constante universal de los gases y T es la temperatura absoluta. Aunque esta ecuación la usan ampliamente los ingenieros y científicos, sólo es exacta sobre un rango limitado de presión y temperatura. Más aun, esa ecuación es más apropiada para algunos gases que para otros.

Una ecuación alternativa del estado de los gases está dada por:

$$\left(p + \frac{a}{v^2} \right)(v - b) = RT$$

a la que se conoce con el nombre de ecuación de van der Waals, $v = \frac{V}{n}$ es el volumen molal y a y b son constantes empíricas que dependen de un gas particular.

Un proyecto de ingeniería química requiere que se calcule exactamente el volumen molal (v) del bióxido de carbono y del oxígeno para combinaciones diferentes de la temperatura y de la presión, de tal forma que se pueda seleccionar una vasija apropiada que los contenga. Así mismo, es importante examinar qué tan bien se apega cada gas a la ley de los gases ideales, comparando los volúmenes molales calculados con las dos ecuaciones dadas anteriormente. Se proporcionan los siguientes datos:

R = *0.082054 l atm /(mol* K*)*

$$\left.\begin{array}{l} a = 3.592 \\ b = 0.04267 \end{array}\right\} \text{bióxido de carbono}$$

$$\left.\begin{array}{l} a = 1.360 \\ b = 0.03183 \end{array}\right\} \text{oxígeno}$$

Las presiones de interés en el diseño son de 1, 10 y 100 atm para combinaciones de la temperatura de 300, 500 y 700° K.

Los volúmenes molales de ambos gases se calculan con la ley de los gases ideales con $n = 1$.

Se debe entonces:

a) Realizar los cálculos del volumen molal a partir de la ecuación de van der Walls, usando el método de Newton y el de bisección con una tolerancia de $\varepsilon = 0.001$.

Repetir estos cálculos para todas las combinaciones de presión y temperatura y completar la siguiente tabla:

Temperatura *K*	**Presión atm**	**Volumen Molal (ley de los gases ideales) l/mol**	**Volumen Molal (van der Waals), por el método de Newton. Bióxido de carbono l/mol**	**Iteraciones**
300	1			
	10			
	100			
500	1			
	10			
	100			
700	1			
	10			
	100			

Volumen Molal (van der Waals), por el método de bisección. Bióxido de carbono l/mol	**Iteraciones**	**Volumen Molal (van der Waals), por el método de Newton. Oxígeno l/mol**	**Iteraciones**	**Volumen Molal (van der Waals), por el método de bisección. Oxígeno l/mol**	**Iteraciones**

b) Según los resultados obtenidos, expresar sus conclusiones.

……

Capítulo 3

3

Solución de sistemas de ecuaciones.

3

3.1. Introducción.

Hay sistemas de ecuaciones lineales (aquellos de orden m x n, $m \neq n$) que no se pueden resolver por la regla de Leibnitz-Cramer. Surge así un teorema del Álgebra Lineal que contempla todos los casos posibles de soluciones de ecuaciones algebraicas y que es el teorema de Rouché-Frobenius (cuya demostración puede encontrarse en los textos de Álgebra Lineal).

Repasemos primero el concepto de rango o característica de una matriz.

El *rango* o *característica de una matriz* es el orden del determinante no nulo de orden máximo que se puede extraer de esa matriz.

Teorema de Rouché-Frobenius. Dado un sistema de m ecuaciones lineales con n incógnitas, no homogéneo, se dice que es compatible (tiene solución común) si el rango p de la matriz del sistema es igual al rango p' de la matriz orlada (o ampliada). Si además $p = p' = n$ se dice que el sistema está determinado y la solución es única (la cual puede ser obtenida por la regla de Leibnitz-Cramer). Si $p = p' < n$, entonces existen infinitas soluciones; se tendrá una solución para cada valor de las $n - p$ incógnitas que no figuran en el sistema de p ecuaciones. Por otro lado, si $p \neq p'$, entonces el sistema es incompatible (no tiene solución común).

Si el sistema de ecuaciones lineales es homogéneo y $p = n$, entonces se dice que existe solamente la solución trivial (la nula) y si $p < n$, entonces existen infinitas soluciones y el grado de indeterminación será $n - p$.

En particular, para un sistema de n ecuaciones con n incógnitas no homogéneo se tiene:

a) Si el determinante de los coeficientes es distinto de cero se dice que el sistema está determinado y la solución es única.
b) Si el determinante de los coeficientes es nulo y su característica es p, pero los determinantes de orden $p + 1$ formados con los términos independientes son también nulos, es decir, si la matriz orlada tiene el mismo rango que p hay infinitas soluciones del sistema y el grado de indeterminación es $n - p$.
c) Si la matriz orlada tiene característica mayor que p el sistema es incompatible, es decir, no tiene solución.

Si el sistema es homogéneo y tiene tantas ecuaciones como incógnitas, para que $p < n$ debe ser nulo el determinante de los coeficientes. Es decir, la condición necesaria y suficiente para la existencia de soluciones de un sistema de n ecuaciones lineales homogéneas con n incógnitas es que el determinante de los coeficientes sea nulo.

Como se ve, se trata de una condición opuesta a la requerida en el sistema no homogéneo, es decir, de aquél cuyos términos independientes no son todos nulos.

Ejemplo 1. En el sistema de 3 ecuaciones con 2 incógnitas

$$\begin{cases} x + y = 3 \\ 2x - y = 0 \\ x + 2y = 5 \end{cases}$$

las matrices del sistema y la ampliada tienen rango $p = 2$. Como las incógnitas son también $2 = n$, entonces el sistema está determinado y la solución es única ($x = 1$, $y = 2$).

Ejemplo 2. En el sistema de 4 ecuaciones con 2 incógnitas

$$\begin{cases} 3x - 2y = 7 \\ 5x - y = 16 \\ 4x + 3y = 9 \\ 3x - 5y = 11 \end{cases}$$

el rango de la matriz del sistema es $p = 2$ y el rango de la matriz orlada es $p' = 3$. Como $p \neq p'$, entonces el sistema es incompatible, es decir, no tiene solución.

Ejemplo 3. Dado el sistema de 4 ecuaciones con 5 incógnitas

$$\begin{cases} x + y + z + t + u = 3 \\ 2x - y + z - 2t - 3u = -3 \\ x + 2y - z + t + 2u = 3 \\ 3x + 0y + 2z - t - 2u = 0 \end{cases}$$

la característica de la matriz de los coeficientes es $p = 3$ (la cuarta ecuación es la suma de las dos primeras y se puede eliminar del sistema); también es $p' = 3$ la característica de la matriz ampliada con los términos independientes.

Como $p = p' < n = 5$, entonces existen infinitas soluciones, una para cada valor de las 5 - 3 = 2 incógnitas (por ejemplo, t y u) que no figuran en el sistema de ecuaciones. Pasando al segundo miembro estas 2 incógnitas, t y u por ejemplo, resulta un sistema de 3 ecuaciones con 3 incógnitas:

$$\begin{cases} x + y + z = 3 - t - u & \text{(I)} \\ 2x - y + z = -3 + 2t + 3u & \text{(II)} \\ x + 2y - z = 3 - t - 2u & \text{(III)} \end{cases} \qquad (1)$$

que resuelto da

$$\begin{cases} x = \dfrac{1}{7}(-6 + 6u + 5t) \\ y = \dfrac{1}{7}(18 - 11u - 8t) \\ z = \dfrac{1}{7}(9 - 2u - 4t) \end{cases} \qquad (2)$$

Para cualquier par de valores u y t, las fórmulas (2) dan la terna de valores x, y, z correspondientes que resuelven el sistema. Se dice entonces que el sistema tiene a 2 como grado de indeterminación. (En general, el grado de indeterminación es $n - p$).

Resolvamos el sistema de 3 ecuaciones con 3 incógnitas (1), es decir, obtengamos (2).

$$\begin{array}{rl} \text{(I):} & x + y + z = 3 - t - u \\ - & \\ \text{(III):} & \underline{x + 2y - z = 3 - t - 2u} \\ & -y + 2z = u \end{array} \qquad \begin{array}{rl} \text{(I) x 2:} & 2x + 2y + 2z = 6 - 2t - 2u \\ - & \\ \text{(II):} & \underline{2x - y + z = -3 + 2t + 3u} \\ & 3y + z = 9 - 4t - 5u \end{array}$$

Multiplicando por 3 a $-y + 2z = u$, obtenemos $-3y + 6z = 3u$. Luego,

$$\begin{array}{rl} & -3y + 6z = 3u \\ + & \\ & \underline{3y + z = 9 - 4t - 5u} \\ & 7z = 9 - 4t - 2u \end{array}$$

de donde, $z = \dfrac{1}{7}(9 - 2u - 4t)$.

Por otro lado,

$-y + 2z = u$, de donde, $y = 2z - u = -u + \frac{1}{7}(18 - 4u - 8t)$, esto es, $y = \frac{1}{7}(18 - 11u - 8t)$.

Además, de (I) $x + y + z = 3 - t - u$, de donde, $x = 3 - t - u - \frac{1}{7}(18 - 11u - 8t) - \frac{1}{7}(9 - 2u - 4t)$, es decir, $x = \frac{1}{7}(-6 + 6u + 5t)$.

3.2. Sistemas de ecuaciones lineales.

Analizaremos distintos métodos para resolver un sistema de n ecuaciones con n incógnitas

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \qquad\qquad \vdots \\ \qquad\qquad \vdots \\ \qquad\qquad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n \end{cases} \tag{3}$$

que sintéticamente puede ser escrito como

$$\sum_{i=1}^{n} a_{ji}x_i = b_j \qquad (j = 1, 2, \ldots, n) \tag{3'}$$

o bien en forma matricial

$$Ax = B \tag{3"}$$

donde A es la matriz de los coeficientes, matriz cuadrada, x es la matriz de las incógnitas, matriz columna, y B es la matriz columna formada por los términos independientes. Es decir,

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ & \cdot & & \\ & \cdot & & \\ & \cdot & & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}, x = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}, \qquad B = \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{pmatrix}.$$

La teoría elemental establece que si el determinante de la matriz A es distinto de cero, det $(A) \neq 0$, entonces es posible resolver el sistema (3) por las fórmulas de Cramer obteniéndose una solución única. Los determinantes no son de interés práctico para la resolución de los sistemas lineales, ya que el cálculo de un determinante tiene la misma dificultad que resolver el sistema lineal.

Si nos propusiéramos resolver por la regla de Cramer un sistema de 10 ecuaciones con 10 incógnitas tendríamos que realizar más de 79.800.000 multiplicaciones. Luego, esta regla es evidentemente poco práctica de modo que nuestro objetivo será desarrollar algoritmos computacionales más eficientes que nos permitan resolver el sistema (3). El mismo sistema de 10 ecuaciones con 10 incógnitas resuelto con cualquiera de los métodos directos que veremos para la resolución de sistemas alcanzan a 400 multiplicaciones, aproximadamente.

Existen dos tipos de métodos numéricos para resolver sistemas de ecuaciones lineales: *métodos directos* y *métodos iterativos*.

Los *métodos directos* son aquellos en los cuales suponiendo que no existan errores de redondeo, se obtiene la solución exacta después de un número finito de operaciones aritméticas.

Los *métodos iterativos* comienzan con valores iniciales aproximados y aplicando un algoritmo oportunamente seleccionado, se obtienen aproximaciones sucesivas mejoradas. Los métodos iterativos varían con el algoritmo seleccionado y según la rapidez de su convergencia. Son muy simples y adecuados para ser usados en computadoras.

El proceso de decidir cuál método es conveniente usar es muy delicado. Si la matriz del sistema es de dimensión no muy grande (no mayor de 40 x 40), es densa (es decir, cuando sus términos no son nulos, ni próximos a ser nulos; en caso contrario, se dice que la matriz es rala o dispersa), y sus elementos están homogéneamente distribuidos en la matriz, entonces se opta por los métodos directos que darán una solución más exacta, pero con una aritmética más complicada. Si la matriz del sistema es rala o no, pero es de dimensión grande (100 x 100, por ejemplo), entonces se opta por los métodos iterativos. Pero también es complicado elegir dentro de

estos dos grupos el proceso que nos lleve a la solución, puesto que existen una gran cantidad de tales procesos. Comenzaremos con el estudio de algunos de los procedimientos que entran dentro del grupo de los métodos directos.

3.2.1. Método de eliminación de Gauss.

El proceso de eliminar una incógnita por vez del conjunto de ecuaciones (3) es tal vez el más simple, breve y conciso para resolver dicho conjunto de ecuaciones. Suponemos que $a_{11} \neq 0$, pues en caso contrario podemos intercambiar 2 filas o columnas y obtener un elemento distinto de cero en el extremo superior izquierdo. Consideramos la matriz $(A|B)$ que queremos triangularla. Debemos hacer cero el primer elemento de cada fila, a partir de la segunda fila. Para ello multiplicamos la primera ecuación por un factor apropiado de modo que al restarla de las demás filas obtengamos 0 en el primer elemento de cada una de ellas:

$$\text{Factor: } m_{i1} = \frac{a_{i1}}{a_{11}}, \quad i = 2, 3, \ldots, n.$$

multiplicador o coeficiente de eliminación

Luego de restar se elimina la variable x_1 y obtenemos otro sistema (el del primer paso) de la forma

$$\begin{aligned} a_{22}^{(1)} x_2 + a_{23}^{(1)} x_3 + \ldots + a_{2n}^{(1)} x_n &= b_2^{(1)} \\ a_{32}^{(1)} x_2 + a_{33}^{(1)} x_3 + \ldots + a_{3n}^{(1)} x_n &= b_3^{(1)} \\ &\vdots \\ a_{n2}^{(1)} x_2 + a_{n3}^{(1)} x_3 + \ldots + a_{nn}^{(1)} x_n &= b_n^{(1)} \end{aligned} \qquad (4)$$

es decir, transformamos el sistema original en otro de $(n\text{-}1)\text{x}(n\text{-}1)$, donde

$$\begin{aligned} a_{ij}^{(1)} &= a_{ij} - m_{i1}\, a_{1j} \qquad i, j = 2, 3, \ldots, n \\ b_i^{(1)} &= b_i - m_{i1}\, b_1 \end{aligned}$$

(el superíndice (1) indica el paso 1). Este es un sistema de $(n\text{-}1)$ ecuaciones en las $(n\text{-}1)$ incógnitas $x_2, x_3, \ldots, x_n$. Si $a_{22}^{(1)} \neq 0$ repetimos el procedimiento y podemos de manera análoga eliminar x_2 de las últimas $(n\text{-}2)$ ecuaciones del

sistema (4). Obtenemos así un sistema de (n-2) ecuaciones en las (n-2) incógnitas x_3, x_4, …, x_n cuyos coeficientes y términos independientes están calculados por

$$a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i2}\, a_{2j}^{(1)}$$

$$b_i^{(2)} = b_i^{(1)} - m_{i2}\, b_2^{(1)}, \quad \text{con } m_{i2} = \frac{a_{i2}^{(1)}}{a_{22}^{(1)}}, \quad i, j = 3, 4, \ldots, n.$$

Los elementos a_{11}, $a_{22}^{(1)}$, $a_{33}^{(2)}$, $a_{44}^{(3)}$, … que se van despreciando durante la eliminación se llaman *elementos pivotales*. Si todos ellos son no nulos podemos continuar la eliminación hasta que en (n-1) pasos llegamos a la ecuación final

$$a_{nn}^{(n-1)}\, x_n = b_n^{(n-1)}.$$

Seleccionamos ahora la primera fila (o ecuación) de cada paso y formamos el sistema triangular superior

$$\begin{aligned} a_{11}^{(0)}x_1 + a_{12}^{(0)}x_2 + \ldots + a_{1,n-1}^{(0)}x_{n-1} + a_{1n}^{(0)}x_n &= b_1^{(0)} \\ a_{22}^{(1)}\, x_2 + \ldots + a_{2,n-1}^{(1)}x_{n-1} + a_{2n}^{(1)}x_n &= b_2^{(1)} \\ \vdots \qquad\qquad & \\ a_{n-1,n-1}^{(n-2)}x_{n-1} + a_{n-1,n}^{(n-2)}x_n &= b_{n-1}^{(n-2)} \\ a_{nn}^{(n-1)}x_n &= b_n^{(n-1)} \end{aligned} \tag{5}$$

donde hemos introducido la notación $a_{ij}^{(0)}$, $b_i^{(0)}$ en lugar de a_{ij}, b_i para los coeficientes y términos independientes del sistema original.

Se ha reducido así el sistema de partida a un sistema triangular superior que se resuelve por sustitución regresiva (se calcula primero x_n, luego x_{n-1}, y así siguiendo hasta x_1); esto es

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}}, \qquad x_{n-1} = \frac{b_{n-1}^{(n-2)} - a_{n-1,n}^{(n-2)}x_n}{a_{n-1,n-1}^{(n-2)}}.$$

Luego,

$$x_i = \frac{b_i^{(i-1)} - \sum_{k=i+1}^{n} a_{ik}^{(i-1)}x_k}{a_{ii}^{(i-1)}}, \quad i = n, n-1, \ldots, 1. \tag{6}$$

Notemos que el segundo miembro de (3) es transformado exactamente en la misma forma que las columnas de A. En consecuencia, la descripción de la eliminación es simplificada si consideramos a la matriz columna B como la última columna de la matriz A, poniendo

$$a_{i,n+1}^{(k)} = b_i^{(k)}, \quad k = 0, 1, \ldots, n\text{-}1, \quad i = 1, 2, \ldots, n.$$

Luego, el algoritmo puede ser resumido como sigue: la eliminación es realizada en (n-1) pasos de modo que en el paso k los elementos $a_{ij}^{(k-1)}$, $i, j \geq k+1$ son transformados según las siguientes expresiones

$$m_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}, \qquad a_{ij}^{(k)} = a_{ij}^{(k-1)} - m_{ik}\, a_{kj}^{(k-1)}, \qquad k = 1, 2, \ldots, n\text{-}1$$

$$i = k+1, \ldots, n, \quad j = k+1, \ldots, n+1$$

$$a_{ij}^{(0)} = a_{ij}$$

y la solución se calcula mediante (7)

$$x_i = \frac{a_{i,n+1}^{(i-1)} - \sum_{j=i+1}^{n} a_{ij}^{(i-1)} x_j}{a_{ii}^{(i-1)}}, \qquad i = n, n\text{-}1, \ldots, 1.$$

Ejemplo 4. Apliquemos el método de eliminación de Gauss al sistema de ecuaciones lineales siguiente

$$\begin{cases} 2x_1 - 5x_2 + x_3 + 3x_4 = 7 \\ 4x_1 + x_2 + 2x_3 - 2x_4 = 4 \\ 3x_1 + 2x_2 - 3x_3 + x_4 = 2 \\ x_1 + 5x_2 + 3x_3 - 4x_4 = 4 \end{cases}$$

Primer paso: $k = 1$ ($i, j \geq k+1$, por lo tanto, $i = 2, 3, 4$, $j = 2, 3, 4, 5$).

De (7): $m_{21} = 2$
$m_{31} = 3/2$
$m_{41} = 1/2$
$a_{ij}^{(1)} = a_{ij} - m_{i1}\, a_{1j}$

y, en consecuencia, obtenemos

$$\begin{cases} 11x_2 + 0x_3 - 8x_4 = -10 \\ \frac{19}{2}x_2 - \frac{9}{2}x_3 - \frac{7}{2}x_4 = -\frac{17}{2} \\ \frac{15}{2}x_2 + \frac{5}{2}x_3 - \frac{11}{2}x_4 = \frac{1}{2} \end{cases}$$

Segundo paso: $k = 2$ $(i, j \geq k+1$, por lo tanto, $i = 3, 4, \quad j = 3, 4, 5)$.

De (7): $m_{32} = 19/22$
$m_{42} = 15/22$
$a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i2}\ a_{2j}^{(1)}$

y, en consecuencia, obtenemos

$$\begin{cases} -\frac{9}{2}x_3 + \frac{75}{22}x_4 = \frac{3}{22} \\ \frac{5}{2}x_3 - \frac{1}{22}x_4 = \frac{161}{22} \end{cases}$$

Tercer paso: $k = 3$ $(i, j \geq k+1$, por lo tanto, $i = 4, \quad j = 4, 5)$.

De (7): $m_{43} = -5/9$
$a_{ij}^{(3)} = a_{ij}^{(2)} - m_{i3}\ a_{3j}^{(2)}$

y, en consecuencia, obtenemos

$$\begin{cases} \frac{366}{198}x_4 = \frac{1464}{198}. \end{cases}$$

El sistema triangular superior obtenido por el método de eliminación de Gauss es

$$\begin{cases} 2x_1 - 5x_2 + x_3 + 3x_4 = 7 \\ 11x_2 + 0x_3 - 8x_4 = -10 \\ -\frac{9}{2}x_3 + \frac{75}{22}x_4 = \frac{3}{22} \\ \frac{366}{198}x_4 = \frac{1464}{198} \end{cases}$$

Efectuando sustitución regresiva (fórmulas (7)), obtenemos

$$x_4 = \frac{1464}{366} = 4$$

$$x_3 = \frac{\frac{3}{22} - \frac{75}{22}4}{-\frac{9}{2}} = \frac{-\frac{297}{22}}{-\frac{9}{2}} = 3$$

$$x_2 = \frac{-10 - 0(3) - (-8)4}{11} = \frac{22}{11} = 2$$

$$x_1 = \frac{7 - (-5)2 - 1(3) - 3(4)}{2} = \frac{2}{2} = 1.$$

Luego,

$$x_1 = 1, \quad x_2 = 2, \quad x_3 = 3, \quad x_4 = 4.$$

3.2.1.1. Esquema práctico del método de eliminación Gaussiana.

Los cálculos de este método pueden disponerse convenientemente en una tabla. En el cálculo manual conviene establecer mecanismos de control para evitar errores. Si en el sistema (3') cada incógnita x_i la reemplazamos por $\bar{x}_i - 1$, obtenemos un sistema de ecuaciones en las incógnitas $\bar{x}_i$ cuyos coeficientes son los mismos que los coeficientes del sistema (3') pero los términos independientes son

$$s_j = b_j + \sum_{i=1}^{n} a_{ji}, \quad j = 1, 2, \ldots, n. \tag{8}$$

En efecto,

$$\sum_{i=1}^{n} a_{ji}(\bar{x}_i - 1) = b_j, \quad j = 1, 2, \ldots, n$$

de donde,

$$\sum_{i=1}^{n} a_{ji}\bar{x}_i - \sum_{i=1}^{n} a_{ji} = b_j, \quad j = 1, 2, \ldots, n$$

es decir,

$$\sum_{i=1}^{n} a_{ji}\bar{x}_i = \sum_{i=1}^{n} a_{ji} + b_j = s_j, \quad j = 1, 2, \ldots, n.$$

De este modo resolvemos simultáneamente dos sistemas de ecuaciones que tienen los mismos coeficientes y cuyos términos independientes son b_j y s_j, respectivamente. Para formar los s_j se observa que en cada ecuación se suman los coeficientes de las incógnitas y el término independiente b_j. En cada etapa del cálculo de eliminación de las incógnitas, esta columna debe dar la suma de los coeficientes y término independiente de cada nueva ecuación (salvo errores de redondeo).

Ejemplo 5. Resolvamos el sistema dado en el ejemplo 4 usando el esquema práctico. Sea, entonces, el sistema

$$\begin{cases} 2x_1 - 5x_2 + x_3 + 3x_4 = 7 \\ 4x_1 + x_2 + 2x_3 - 2x_4 = 4 \\ 3x_1 + 2x_2 - 3x_3 + x_4 = 2 \\ x_1 + 5x_2 + 3x_3 - 4x_4 = 4 \end{cases}$$

Disponemos los cálculos en el siguiente esquema, donde S es la columna de suma y que nos sirve de control. A la derecha de cada fila se ha calculado el coeficiente de eliminación y se hace la operación: la fila en que está el coeficiente menos la primera fila de cada panel multiplicada por dicho coeficiente, lo que nos permite ir eliminando en cada etapa una incógnita (de acuerdo a las fórmulas (7)):

x_1	x_2	x_3	x_4	B	S	
2	-5	1	3	7	8	
4	1	2	-2	4	9	$m_{21} = a_{21}/a_{11} = 2$
3	2	-3	1	2	5	$m_{31} = a_{31}/a_{11} = 3/2$
1	5	3	-4	4	9	$m_{41} = a_{41}/a_{11} = 1/2$
	11	0	-8	-10	-7	
	19/2	-9/2	-7/2	-17/2	-7	$m_{32} = a_{32}/a_{22} = 19/22$
	15/2	5/2	-11/2	1/2	5	$m_{42} = a_{42}/a_{22} = 15/22$
		-9/2	75/22	3/22	-21/22	
		5/2	-1/22	161/22	215/22	$m_{43} = a_{43}/a_{33} = -5/9$
			366/198	1464/198	1830/198	

Hemos obtenido la matriz triangular superior

$$\left(\begin{array}{cccc:c:c} 2 & -5 & 1 & 3 & 7 & 8 \\ & 11 & 0 & -8 & -10 & -7 \\ & & -9/2 & 75/22 & 3/22 & -21/22 \\ & & & 366/198 & 1464/198 & 1830/198 \end{array}\right)$$

Debemos encontrar la solución de $Ax = B$ y también de $A\bar{x} = S$ que satisface que $\bar{x}_i = x_i + 1$, la cual proporciona un chequeo de la sustitución regresiva:

Solución **Verificación**

$$x_4 = \frac{1464}{366} = 4 \qquad \bar{x}_4 = \frac{1830}{366} = 5 \quad \text{y} \quad x_4 = \bar{x}_4 - 1 = 4$$

$$x_3 = \frac{\frac{3}{22} - 4\frac{75}{22}}{-\frac{9}{2}} = 3 \qquad \bar{x}_3 = \frac{-\frac{21}{22} - 5\frac{75}{22}}{-\frac{9}{2}} = 4 \quad \text{y} \quad x_3 = \bar{x}_3 - 1 = 3$$

$$x_2 = \frac{-10 + 4(8) - 0(3)}{11} = 2$$

$$\bar{x}_2 = 3 \qquad \text{y} \qquad x_2 = \bar{x}_2 - 1 = 2$$

$$x_1 = \frac{7 - 4(3) - 3(1) + 5(2)}{2} = 1$$

$$\bar{x}_1 = 2 \qquad \text{y} \qquad x_1 = \bar{x}_1 - 1 = 1.$$

Luego, la solución es $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$.

Ejemplo 6

x_1	x_2	x_3	x_4	B	S	
1	0.5	0.33333	0.25	0.58333	2.66666	
0.5	0.33333	0.25	0.2	0.21667	1.5	$m_{21} = 0.5$
0.33333	0.25	0.2	0.16667	0.11666	1.06666	$m_{31} = 0.33333$
0.25	0.2	0.16667	0.14286	0.07381	0.83334	$m_{41} = 0.25$
	0.08333	0.08333	0.075	-0.075	0.16667	
	0.08333	0.08889	0.08333	-0.07778	0.17778	$m_{32} = 1$
	0.075	0.08333	0.08036	-0.07202	0.16667	$m_{42} = 0.9003$
		0.00556	0.00833	-0.00278	0.01111	
		0.00833	0.01286	-0.00452	0.01667	$m_{43} = 1.4982$
			0.00038	-0.00035	0.00002	

Obtenemos la matriz triangular superior

$$\left(\begin{array}{cccc|c|c} 1 & 0.5 & 0.33333 & 0.25 & 0.58333 & 2.66666 \\ & 0.08333 & 0.08333 & 0.075 & -0.075 & 0.16667 \\ & & 0.00556 & 0.00833 & -0.00278 & 0.01111 \\ & & & 0.00038 & -0.00035 & 0.00002 \end{array}\right)$$

Solución **Verificación**

$x_4 = -0.92105$ $\quad \bar{x}_4 = 0.07895$ y $x_4 = \bar{x}_4 - 1 = -0.92105$

$x_3 = 0.87992$ $\quad \bar{x}_3 = 1.87992$ y $x_3 = \bar{x}_3 - 1 = 0.87992$

$x_2 = -0.95098$ $\quad \bar{x}_2 = 0.04914$ y $x_2 = \bar{x}_2 - 1 = -0.95086$

$x_1 = 0.99578$ $\quad \bar{x}_1 = 1.99573$ y $x_1 = \bar{x}_1 - 1 = 0.99573.$

Los cálculos han sido efectuados con 5 decimales. Nótese que en la sustitución regresiva el acuerdo con las sumas de control no es muy bueno; la razón de esto es que en este ejemplo los errores de redondeo perturban la solución. El error de la solución computada es de la misma magnitud que los desacuerdos con la columna de control.

3.2.1.2. Mejora de raíces.

Los valores aproximados de las raíces obtenidas mediante el método de eliminación de Gauss pueden mejorarse. Indicaremos el procedimiento, si las correcciones de las raíces son pequeñas en valor absoluto.

Supongamos que se ha hallado una solución aproximada x_0 para el sistema

$$Ax = B.$$

Estableciendo

$$x = x_0 + \delta$$

tenemos entonces para la corrección $\delta = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_n \end{pmatrix}$ de la raíz x_0

$$A\,(x_0 + \delta) = B$$

ó

$$A\delta = \varepsilon$$

donde $\varepsilon = B - A\,x_0$ es el resto de la solución aproximada x_0. De este modo, para hallar δ resulta necesario resolver un sistema lineal con la matriz primera A y un nuevo término constante ε

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Para ello, todo lo que se necesita es adjuntar al esquema principal de cálculo una columna ε de términos constantes y transformarla mediante las reglas generales. Las correcciones δ_1, …, δ_n se hallan de manera similar a como se calculan la solución y los valores $\bar{x}_i$ que se utilizan para verificar dicha solución.

Ejemplo 7. Resolvamos el siguiente sistema de ecuaciones mediante el método de eliminación de Gauss con 3 dígitos significativos:

$$\begin{cases} 6x_1 - x_2 - x_3 = 11.33 \\ -x_1 + 6x_2 - x_3 = 32 \\ -x_1 - x_2 + 6x_3 = 42 \end{cases}$$

Luego, utilizando los valores obtenidos como aproximaciones iniciales mejoraremos las raíces para 10^{-4}.

Ejecutamos todas las operaciones con 3 dígitos significativos. Los valores aproximados de las raíces son

$$x_1^{(0)} = 4.67, \quad x_2^{(0)} = 7.62, \quad x_3^{(0)} = 9.05.$$

Sustituyendo estos valores en el sistema dado, calculemos los restos apropiados (esto es, las diferencias entre los miembros de la izquierda y de la derecha de dicho sistema). Obtenemos

$$\varepsilon_1^{(0)} = -0.02, \quad \varepsilon_2^{(0)} = 0, \quad \varepsilon_3^{(0)} = -0.01.$$

x_1	x_2	x_3	B	S	Resto (ε)
6 -1 -1	-1 6 -1	-1 -1 6	11.33 32 42	15.33 36 46	-0.02 0 -0.01
	5.83 -1.17	-1.17 5.83	33.9 43.9	38.6 48.6	-0.00333 -0.0133
		5.60	50.7	56.3	-0.0140

Así,

$$\delta_3^{(0)} = \frac{-0.0140}{5.60} = -0.0025$$

$$\delta_2^{(0)} = \frac{-0.00333-(-0.0025)(-1.17)}{5.83} = -0.0011$$

$$\delta_1^{(0)} = \frac{-0.02-(-0.0025)(-1)-(-0.0011)(-1)}{6} = -0.0039$$

Luego, conseguimos los valores mejorados de las raíces

$$x_1 = 4.6661, \quad x_2 = 7.6189, \quad x_3 = 9.0475$$

siendo los restos iguales a

$$\varepsilon_1 = -2 \text{ x } 10^{-4}, \quad \varepsilon_2 = 2 \text{ x } 10^{-4}, \quad \varepsilon_3 = 0.$$

Observación. Se ha dicho que la regla de Cramer sólo es aplicable para sistemas pequeños porque requiere la evaluación de los determinantes, lo cual es pesado para conjuntos grandes de ecuaciones.

Sin embargo, el método de eliminación Gaussiana proporciona una forma simple de calcular el determinante de la matriz A, $\Delta = \det(A)$.

El método se basa en el hecho que el determinante de la matriz triangular se puede calcular simplemente con el producto de los elementos de su diagonal:

$$\Delta = a_{11}\, a_{22} \ldots a_{nn}.$$

La validez de esta fórmula se puede ilustrar en sistemas de 3 por 3

$$\Delta = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{vmatrix}$$

en donde el determinante se puede evaluar como sigue

$$\Delta = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ 0 & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} 0 & a_{23} \\ 0 & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} 0 & a_{22} \\ 0 & 0 \end{vmatrix} =$$

$$= a_{11}\, a_{22}\, a_{33} - a_{12}\,(0) + a_{13}\,(0) = a_{11}\, a_{22}\, a_{33}.$$

Recordemos que el paso de eliminación Gaussiana genera un sistema triangular superior. Ya que el valor del determinante se puede evaluar al final de este paso

$$\Delta = a_{11}\; a_{22}^{(1)}\; a_{33}^{(2)}\; \ldots\; a_{nn}^{(n-1)}$$

donde los superíndices indican que los elementos se han modificado durante el proceso de eliminación. Por lo tanto, se puede aprovechar el esfuerzo que se ha hecho al reducir el sistema a su forma triangular, y por añadidura obtener una aproximación al determinante. Hay una pequeña modificación en el planteamiento anterior cuando se deben intercambiar dos líneas paralelas (filas o columnas) para obtener el elemento pivotal. En este caso, el determinante cambia de signo cada vez que hay uno de estos intercambios. Una manera de representar esto es modificando la ecuación anterior por

$$\Delta = (-1)^p\, a_{11}\; a_{22}^{(1)}\; a_{33}^{(2)}\; \ldots\; a_{nn}^{(n-1)}$$

donde p representa el número de veces en que dos líneas paralelas son intercambiadas para obtener el elemento pivotal. Esta modificación se puede incorporar de forma simple en un programa: únicamente se toma en cuenta el número de veces que estos intercambios se han llevado a cabo durante los cálculos y después se usa la ecuación anterior para evaluar el determinante.

En el ejemplo 5:

$$\Delta = 2 \text{ x } 11 \text{ x } \left(\frac{-9}{2}\right) \text{x } \frac{366}{198} = \text{-}183$$

En el ejemplo 6:

$$\Delta = 1 \text{ x } 0.08333 \text{ x } 0.00556 \text{ x } 0.00038 \approx 0.00084.$$

En el ejemplo 7:

$$\Delta = 6 \text{ x } 5.83 \text{ x } 5.60 = 195.888 \approx 196.$$

Nota. Propiedades de los determinantes.

1. El cambio de filas por columnas no altera el valor del determinante.
2. El cambio de dos líneas (filas o columnas) paralelas cambia el signo del determinante, pero no su valor absoluto.
3. Si se multiplican todos los elementos de una línea por un mismo número el determinante queda multiplicado por ese número.
4. Si un determinante tiene dos líneas proporcionales (en particular, iguales) el determinante es nulo.
5. Si a los elementos de una línea se le suman los elementos de una línea paralela multiplicados cada uno por un mismo factor numérico (positivo o negativo) el determinante no varía.

3.2.1.3. Estrategias pivotales.

Cuando los coeficientes de las incógnitas en el sistema de ecuaciones y los coeficientes sucesivos en los sistemas derivados de orden menor no varían mucho en magnitud, el procedimiento de eliminación de Gauss da resultados altamente satisfactorios; pero cuando algún coeficiente de la diagonal principal, o sea alguno de los divisores $a_{ii}^{(i-1)}$ es muy pequeño comparado con $a_{ji}^{(i-1)}, j > i$, entonces se cometen errores de redondeo que se propagan de tal manera que los resultados obtenidos no sean tan satisfactorios. Para evitar esto así como también las divisiones por cero, surge lo que se llama *estrategia pivotal*. Analicemos este concepto con algunos ejemplos.

Ejemplo 8. Sea el sistema

$$\begin{cases} x_1 + x_2 + x_3 = 1 \\ x_1 + x_2 + 2x_3 = 2 \\ x_1 + 2x_2 + 2x_3 = 1 \end{cases} \qquad (9)$$

que tiene como solución $x_1 = -x_2 = x_3 = 1$.

Sin embargo, después del primer paso de eliminación, obtenemos

x_1	x_2	x_3	B	S
1	1	1	1	4
1	1	2	2	6
1	2	2	1	6
	0	1	1	2
	1	1	0	2

De modo que $a_{22}^{(1)} = 0$ y no podemos usar el proceso en la forma usual. El remedio es, obviamente, intercambiar segunda y tercera ecuaciones antes del próximo paso, obteniéndose

1	1	0	2
0	1	1	2
	1	1	2

Luego, $x_3=1$, $x_2 = -1$, $x_1 = 1$ (y la verificación: $\bar{x}_3 = 2$, $\bar{x}_2 = 0$, $\bar{x}_1 = 2$).

También podría haberse cambiado el orden de las columnas segunda y tercera; en este caso el orden de las incógnitas habría cambiado.

Consideremos el caso general en que en el paso k (k = 1, …, n-1) tengamos $a_{kk}^{(k-1)} = 0$. Entonces, algún otro elemento $a_{ik}^{(k-1)}$, $i = k$+1, …, n, de la columna k debe ser no nulo, pues de lo contrario las primeras k columnas resultarán linealmente dependientes y A será singular (det (A) = 0 y, por lo tanto, el sistema no está determinado).

Supongamos que $a_{rk}^{(k-1)} \neq 0$. Podemos entonces intercambiar las filas k y r, y proseguir con la eliminación. De esto se sigue que cualquier sistema de ecuaciones no singular puede ser reducido a la forma triangular por eliminación Gaussiana combinado con intercambios de filas.

Para asegurar la estabilidad numérica será necesario frecuentemente efectuar intercambio de filas, no sólo cuando un elemento pivotal es exactamente cero sino también cuando es cercano a cero.

Ejemplo 9. Supongamos que en el sistema de ecuaciones (9) del ejemplo 8 el elemento a_{22} es cambiado por 1.0001 y la eliminación Gaussiana la efectuamos sin intercambios:

x_1	x_2	x_3	B
1	1	1	1
1	1.0001	2	2
1	2	2	1
	0.0001	1	1
	1	1	0
		-9999	-10000

$$x_3 = 1.00010001$$
$$x_2 = -1.0001$$
$$x_1 = 0.99999999.$$

Si efectuamos la sustitución regresiva usando aritmética en punto flotante con 4 cifras significativas, obtenemos

$$x_3' = 1.000$$
$$x_2' = 0$$
$$x_1' = 0$$

mientras que la solución verdadera redondeada a 4 cifras decimales es

$$x_3 = 1.0001, \quad x_2 = -1.0001, \quad x_1 = 1.0000.$$

Si efectuamos el intercambio de segunda y tercera filas en el sistema original y usamos la misma precisión, obtenemos

x_1	x_2	x_3	B
1	1	1	1
1	2	2	1
1	1.0001	2	2
	1	1	0
	0.0001	1	1
		0.9999	1

$$x_3 = 1.00010001$$
$$x_2 = -1.00010001$$
$$x_1 = 1$$

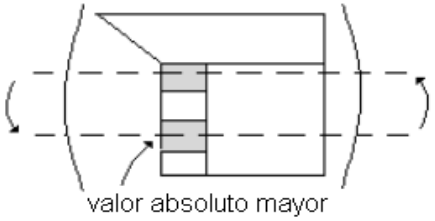
mientras que usando aritmética en punto flotante con 4 cifras significativas, obtenemos

$$x_3' = 1.000$$
$$x_2' = -1.000$$
$$x_1' = 1.000$$

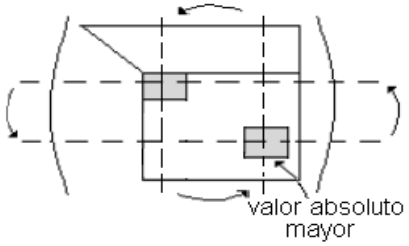
correcta a 3 decimales (la solución verdadera redondeada a 3 decimales es $x_3 = 1.000$, $x_2 = -1.000$, $x_1 = 1$).

Con el fin de prevenir posibles errores catastróficos como los ilustrados en los ejemplos 8 y 9, es generalmente necesario elegir el elemento pivotal en el paso k por una de las siguientes estrategias:

1. ***Pivotación parcial.*** Elíjase el índice r como el entero más pequeño para el cual
$|a_{rk}^{(k-1)}| = máx_i \{|a_{ik}^{(k-1)}|\} \quad (k \le i \le n)$
e intercámbiense las filas k y r.



2. ***Pivotación completa.*** Elíjanse r y s como los enteros más pequeños para los cuales
$|a_{rs}^{(k-1)}| = máx_{i,j} \{|a_{ij}^{(k-1)}|\} \quad (k \le i, j \le n)$
e intercámbiense las filas k y r, y las columnas k y s.



Así, en la pivotación completa seleccionamos como pivote en cada etapa el elemento de mayor valor absoluto en toda la parte relevante de la matriz.

En la práctica, la pivotación parcial es generalmente satisfactoria; la pivotación completa no es muy usada debido a la gran cantidad de trabajo que requiere. De aquí que la pivotación parcial se prefiera sobre la otra y es la más usada en las situaciones prácticas.

Hay dos importantes casos en los que la eliminación Gaussiana puede realizarse sin intercambios de filas o columnas:

a) *Matrices diagonalmente dominantes*: $|a_{ii}| \ge \sum_{\substack{j=1 \\ j \ne i}}^{n} |a_{ij}|, \quad i = 1, 2, ..., n$

b) *Matrices simétricas y definidas positivas*: $A^t = A$ y $\det(A_k) > 0$, $k = 1, 2, ..., n$, donde A_k es la matriz k x k formada por la intersección de las primeras k filas y columnas de A.

3.2.1.4. Descomposición *LU*.

Teorema de la descomposición de una matriz (Teorema *LU*). Sea A una matriz n x n y denotemos por A_k la matriz k x k formada por la intersección de las primeras k filas y columnas de A. Si $\det(A_k) \ne 0$ ($k = 1, 2,$

..., n -1), entonces existe una única matriz triangular inferior $L = (m_{ij})$, con $m_{ii} = 1$ ($i = 1, 2, \ldots, n$) y una única matriz triangular superior $U = (u_{ij})$ tal que $A = LU$.

No daremos aquí la demostración de este teorema (la que puede encontrarse en cualquier texto de Álgebra Lineal), pero veremos su aplicación. La formulación matricial involucrando la descomposición de A en el producto LU (descomposición LU) para obtener la solución del sistema $A\,x = B$, es totalmente equivalente a la eliminación Gaussiana para resolver el mismo sistema $A\,x = B$.

¿Por qué cuando la matriz A ha sido descompuesta en el producto LU podemos obtener la solución de $A\,x = B$? Pues si conocemos L y U, resolver $A\,x = B$ es lo mismo que resolver $LUx = B$ cuya solución se determina resolviendo primero el sistema $Ly = B$ y después resolviendo $Ux = y$. Por lo tanto, se debe encontrar la descomposición de A en LU cuya existencia está garantizada por el teorema anterior.

Se puede probar que las ecuaciones obtenidas durante la eliminación Gaussiana son equivalentes a la ecuación matricial $A = LU$, donde los elementos no nulos en L y U están dados por $(L)_{ik} = (m_{ik})$, $i \geq k$, $(U)_{kj} = \left(a_{kj}^{(k-1)}\right)$, $k \leq j$.

Concluimos que los elementos en L (matriz triangular inferior) son los multiplicadores obtenidos durante la eliminación Gaussiana y que la matriz U es la matriz triangular final (superior) obtenida por eliminación Gaussiana. Esto es

$$L=\begin{pmatrix} 1 & & & & & & \\ m_{21} & 1 & & & & & \\ m_{31} & m_{32} & 1 & & & & \\ \vdots & \vdots & \ddots & \ddots & & & \\ \vdots & \vdots & \vdots & & \ddots & & \\ \vdots & \vdots & \vdots & & & \ddots & \\ m_{n-1,1} & m_{n-1,2} & & \cdots & m_{n,n-1} & 1 & \\ m_{n1} & m_{n2} & m_{n3} & \cdots & & & 1 \end{pmatrix}, \quad U=\begin{pmatrix} u_{11} & u_{12} & \cdots & & u_{1n} \\ & u_{22} & \cdots & & u_{2n} \\ & & \ddots & & \vdots \\ & & & \ddots & \vdots \\ & & & & \vdots \\ & & & & u_{nn} \end{pmatrix}$$

donde,

$$m_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}, \qquad k = 1, 2, ..., n, \quad i = k, \ldots, n$$

$$u_{kj} = a_{kj}^{(k-1)}, \qquad k = 1, 2, ..., n, \quad j = k, \ldots, n.$$

Es decir que hay una equivalencia entre la formulación matricial $A = LU$ y la eliminación Gaussiana.

Notemos además que, como det $(L_k) = 1$ se tiene

$$\det(A_k) = a_{11}^{(0)}\ a_{22}^{(1)}\ \ldots\ a_{kk}^{(k-1)}, \quad k = 1, 2, ..., n$$

esto es, el producto de los k primeros pivotes. Entonces, la eliminación Gaussiana puede aplicarse (sin intercambios de filas o columnas) si y sólo si det $(A_k) \neq 0$, $k = 1, 2, ..., n$ -1, que es la condición en el teorema LU.

Es importante observar que si se hace pivotación parcial durante la eliminación, entonces estos intercambios de filas nos pueden dar L y U tales que $LU = A'$, donde A' es la matriz que resulta si los intercambios de filas son realizados en el mismo orden sobre la matriz inicial A.

Observemos que además para $k = n$ es

$$\det(A) = \det(A_n) = a_{11}^{(0)}\ a_{22}^{(1)}\ \ldots\ a_{nn}^{(n-1)}$$

(ver observación del párrafo 3.2.1.2).

Luego, el determinante de la matriz de los coeficientes es el producto de los pivotes, y esto es efectivamente así por las propiedades de los determinantes, ya que ninguna etapa del algoritmo de Gauss altera el valor del determinante de la matriz de los coeficientes, excepto el pivoteo parcial que modifica el signo del determinante si se ha efectuado un número impar de intercambios. Esto se representa, como hemos mencionado anteriormente, por

$$\det(A) = (-1)^p\ a_{11}^{(0)}\ a_{22}^{(1)}\ \ldots\ a_{nn}^{(n-1)}$$

donde p representa el número de veces que se han intercambiado las filas durante la eliminación.

Ejemplo 10. En el ejemplo 5

$$L = \begin{pmatrix} 1 & & & \\ 2 & 1 & & \\ 3/2 & 19/22 & 1 & \\ 1/2 & 15/22 & -5/9 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 2 & -5 & 1 & 3 \\ & 11 & 0 & -8 \\ & & -9/2 & 75/22 \\ & & & 366/198 \end{pmatrix}$$

y $LU = A$. (Esto es, la eliminación Gaussiana nos conduce a la descomposición de la matriz A en LU).

Ejemplo 11. En el ejemplo 6

$$L = \begin{pmatrix} 1 & & & \\ 0.5 & 1 & & \\ 0.33333 & 1 & 1 & \\ 0.25 & 0.90003 & 1.4982 & 1 \end{pmatrix}, \; U = \begin{pmatrix} 1 & 0.5 & 0.33333 & 0.25 \\ & 0.08333 & 0.08333 & 0.075 \\ & & 0.00556 & 0.00833 \\ & & & 0.00038 \end{pmatrix}$$

y $LU = A$.

Ejemplo 12. En el ejemplo 8, det (A) = -1 (hubo un intercambio de filas).

$$L = \begin{pmatrix} 1 & & \\ 1 & 1 & \\ 1 & 0 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 1 & 1 & 1 \\ & 1 & 1 \\ & & 1 \end{pmatrix} \text{ y } \quad LU = A' = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 1 & 2 \end{pmatrix}.$$

Si hacemos el mismo intercambio de filas en A (segunda por tercera), obtenemos

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 1 & 2 \end{pmatrix}$$

coincidiendo con A'.

Ejemplo 13. Aplicación del teorema LU.

Si para algún k, det (A_k) = 0, entonces puede no existir la descomposición LU. Un simple ejemplo de esto es la matriz no singular

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

Supongamos que $A = LU$, donde

$$\boldsymbol{LU} = \begin{pmatrix} m_{11} & 0 \\ m_{21} & m_{22} \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{pmatrix} = \begin{pmatrix} m_{11}u_{11} & m_{11}u_{12} \\ m_{21}u_{11} & m_{21}u_{12} + m_{22}u_{22} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

Entonces, $m_{11} = 0$ ó $u_{11} = 0$. Pero entonces, o la primera fila (si $m_{11} = 0$) o la primera columna (si $u_{11} = 0$) de LU también se anuló. Por lo tanto, $A \neq LU$, esto es, A no se puede descomponer en el producto LU.

Notemos que si intercambiamos las dos filas de A, entonces la matriz resultante es triangular y la descomposición LU existe trivialmente, pues

$$L = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad U = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \text{con } A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

En cualquier matriz no singular A, las filas pueden ser reordenadas de tal manera que exista la descomposición LU. Esto se sigue de la equivalencia entre la descomposición LU y la eliminación Gaussiana.

3.2.2. Método de Gauss – Jordan.

Este método es una variante del método de eliminación de Gauss. La principal diferencia consiste en que en el *método de Gauss - Jordan* cuando se elimina una incógnita, no sólo se elimina de las ecuaciones siguientes sino de todas las otras ecuaciones. De esta forma, el paso de eliminación genera una matriz diagonal en vez de una matriz triangular superior. Por consiguiente, no es necesario emplear la sustitución hacia atrás para obtener la solución. Las incógnitas se obtienen mediante una simple división.

El primer paso de eliminar la incógnita x_1 lo hacemos como antes, quedándonos el primer sistema derivado

$$\begin{aligned} a_{11}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \cdots + a_{1n}^{(1)}x_n &= a_{1,n+1}^{(1)} \\ a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \cdots + a_{2n}^{(1)}x_n &= a_{2,n+1}^{(1)} \\ \vdots & \\ a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \cdots + a_{nn}^{(1)}x_n &= a_{n,n+1}^{(1)} \end{aligned}$$

donde los nuevos coeficientes $a_{ij}^{(1)}$ están dados por

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j}, \quad i = 2, 3, ..., n, \quad j = 2, 3, ..., n+1$$

y llamamos (por una cuestión de notación)

$$a_{1j}^{(1)} = a_{1j}, \quad j = 2, 3, ..., n+1$$

El segundo paso de eliminar x_2 sigue la norma anterior, pero también eliminamos x_2 de la primera ecuación de modo que $a_{12}^{(1)} = 0$, quedándonos el segundo sistema derivado

$$\begin{aligned} a_{11}x_1 + \quad & a_{13}^{(2)}x_3 + \cdots + a_{1n}^{(2)}x_n = a_{1,n+1}^{(2)} \\ a_{22}^{(1)}x_2 + \; & a_{23}^{(2)}x_3 + \cdots + a_{2n}^{(2)}x_n = a_{2,n+1}^{(2)} \\ & \vdots \\ & a_{n3}^{(2)}x_3 + \cdots + a_{nn}^{(2)}x_n = a_{n,n+1}^{(2)} \end{aligned}$$

donde

$$a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{i2}^{(1)}}{a_{22}^{(1)}} a_{2j}^{(1)}, \quad i = 1, 3, 4, ..., n, \;\; j = 3, 4, ..., n+1$$

y llamamos (por una cuestión de notación)

$$a_{2j}^{(2)} = a_{2j}^{(1)}, \;\; j = 3, 4, ..., n+1$$

En forma similar, al eliminar x_3 de las últimas (n-3) ecuaciones también debe eliminarse de las dos primeras ecuaciones de modo que $a_{13}^{(2)} = 0,\ a_{23}^{(2)} = 0$. Nos queda así el tercer sistema derivado

$$\begin{aligned} a_{11}x_1 + \quad & \quad & a_{14}^{(3)}x_4 + \cdots + a_{1n}^{(3)}x_n = a_{1,n+1}^{(3)} \\ & a_{22}^{(1)}x_2 + \; & a_{24}^{(3)}x_4 + \cdots + a_{2n}^{(3)}x_n = a_{2,n+1}^{(3)} \\ & & a_{33}^{(2)}x_3 + a_{34}^{(3)}x_4 + \cdots + a_{3n}^{(3)}x_n = a_{3,n+1}^{(3)} \\ & & a_{44}^{(3)}x_4 + \cdots + a_{4n}^{(3)}x_n = a_{4,n+1}^{(3)} \\ & & \vdots \\ & & a_{n4}^{(3)}x_4 + \cdots + a_{nn}^{(3)}x_n = a_{n,n+1}^{(3)} \end{aligned}$$

donde

$$a_{ij}^{(3)} = a_{ij}^{(2)} - \frac{a_{i3}^{(2)}}{a_{33}^{(2)}} a_{3j}^{(2)}, \quad i = 1, 2, 4, ..., n, \;\; j = 4, 5, ..., n+1$$

y llamamos (por una cuestión de notación)

$$a_{3j}^{(3)} = a_{3j}^{(2)}, \quad j = 4, 5, ..., n+1$$

Continuando de esta forma de tal manera que en cada etapa todos los elementos de una columna excepto el elemento diagonal han sido reducidos a cero, obtenemos finalmente

$$\begin{aligned} a_{11}x_1 \qquad\qquad\qquad\qquad &= a_{1,n+1}^{(n)} \\ a_{22}^{(1)}x_2 \qquad\qquad\qquad &= a_{2,n+1}^{(n)} \\ a_{33}^{(2)}x_3 \qquad\qquad &= a_{3,n+1}^{(n)} \\ \ddots \qquad\qquad & \\ a_{nn}^{(n-1)}x_n &= a_{n,n+1}^{(n)} \end{aligned} \tag{10}$$

Se ha reducido así el sistema original a un sistema diagonal. En tal caso, la solución del sistema (10) se obtiene fácilmente por

$$x_i = \frac{a_{i,n+1}^{(n)}}{a_{ii}^{(i-1)}}, \quad i = 1, 2, 4, ..., n \quad (a_{ii}^{(0)} = a_{ii}) \tag{11}$$

Generalizando las distintas fórmulas que nos dan los coeficientes de los sistemas derivados en cada una de las etapas, obtenemos

$$\begin{cases} a_{ij}^{(k)} = a_{ij}^{(k-1)} - \dfrac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k-1)} & i = 1, 2, ..., k-1, k+1, ..., n, \quad j = k+1, ..., n+1 \\ a_{ij}^{(k)} = a_{ij}^{(k-1)} & i = k, \quad j = k+1, ..., n+1 \end{cases} \tag{12}$$

$(a_{ij}^{(0)} = a_{ij})$

Ejemplo 14. Apliquemos el método de Gauss - Jordan al sistema de ecuaciones lineales dado en el ejemplo 4, es decir al siguiente sistema

$$\begin{cases} 2x_1 - 5x_2 + x_3 + 3x_4 = 7 \\ 4x_1 + x_2 + 2x_3 - 2x_4 = 4 \\ 3x_1 + 2x_2 - 3x_3 + x_4 = 2 \\ x_1 + 5x_2 + 3x_3 - 4x_4 = 4 \end{cases}$$

Disponemos los cálculos en el siguiente esquema, donde *S* es la columna de suma que nos sirve de control. A la derecha de cada fila se ha colocado el coeficiente de eliminación y se hace la operación: la fila en que está el coeficiente menos la fila donde no hay coeficiente multiplicada por el coeficiente:

x_1	x_2	x_3	x_4	*B*	*S*	
2	-5	1	3	7	8	
4	1	2	-2	4	9	$a_{21}/\ a_{11} = 2$
3	2	-3	1	2	5	$a_{31}/\ a_{11} = 3/2$
1	5	3	4	4	9	$a_{41}/\ a_{11} = 1/2$
2	-5	1	3	7	8	$a_{12}/\ a_{22} = -5/11$
0	11	0	-8	-10	-7	
0	19/2	-9/2	-7/2	-17/2	-7	$a_{32}/\ a_{22} = 19/22$
0	15/2	5/2	-11/2	1/2	5	$a_{42}/\ a_{22} = 15/22$
2	0	1	-7/11	27/11	53/11	$a_{13}/\ a_{33} = -2/9$
0	11	0	-8	-10	-7	$a_{23}/\ a_{33} = 0$
0	0	-9/2	75/22	3/22	-21/22	
0	0	5/2	-1/22	161/22	215/22	$a_{43}/\ a_{33} = -5/9$
2	0	0	4/33	82/33	152/33	$a_{14}/\ a_{44} = 4/61$
0	11	0	-8	-10	-7	$a_{24}/\ a_{44} = -264/61$
0	0	-9/2	75/22	3/22	-21/22	
0	0	0	61/33	244/33	305/33	$a_{34}/\ a_{44} = 225/122$
2	0	0	0	66/33	132/33	
0	11	0	0	726/33	1089/33	
0	0	-9/2	0	-1647/122	-2196/122	
0	0	0	61/33	244/33	305/33	

En el último panel se tiene la matriz diagonal y de aquí se obtiene directamente

Solución	**Verificación**
$x_4 = 244/61 = 4$	$\overline{x_4} = 305/61 = 5$
$x_3 = 3$	$\overline{x_3} = 4$
$x_2 = 2$	$\overline{x_2} = 3$
$x_1 = 1$	$\overline{x_1} = 2$

Observaciones

1. A pesar de que a primera vista parecería que se prefiere la reducción de Gauss - Jordan sobre la eliminación Gaussiana ocurre todo lo contrario, pues en la primera se deben hacer muchas más operaciones que en la segunda.

2. Si designamos con $\Delta = \det(A)$, entonces $\Delta = a_{11}a_{22}^{(1)} \ldots a_{nn}^{(n-1)}$

En el ejemplo anterior $\Delta = 2 \text{x} 11 \text{x} \left(\frac{-9}{2} \right) \text{x} \frac{61}{33} = -183$

3. En caso de ser necesario, el método de Gauss - Jordan puede combinarse con pivoteo.

4. Es común hacer una pequeña modificación en el método de Gauss – Jordan, dando origen al *método de Gauss – Jordan normalizado*. Esta modificación consiste en lo siguiente: se normaliza cada vez la ecuación pivotal dividiendo por el pivote de modo que el coeficiente de la incógnita a eliminar sea 1, o sea, para la primera

$$1 \qquad \frac{a_{12}}{a_{11}} \qquad \frac{a_{13}}{a_{11}} \quad \ldots \quad \frac{a_{1n}}{a_{11}} \qquad \frac{a_{1,n+1}}{a_{11}}$$

El sistema se reduce a

$$\begin{matrix} 1 & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} & a_{1,n+1}^{(1)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} & a_{2,n+1}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3n}^{(1)} & a_{3,n+1}^{(1)} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & a_{n2}^{(1)} & a_{n3}^{(1)} & \cdots & a_{nn}^{(1)} & a_{n,n+1}^{(1)} \end{matrix}$$

donde

$$a_{1j}^{(1)} = a_{1j} / a_{11}, \quad j = 2, 3, ..., n+1$$

$$a_{ij}^{(1)} = a_{ij} - a_{i1}\, a_{1j}^{(1)} = a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j}, \quad i = 2, 3, ..., n, \quad j = 2, 3, ..., n+1$$

En el sistema derivado procedemos de igual forma que en el método de Gauss-Jordan, sólo que aquí se divide primero la segunda ecuación por el coeficiente $a_{22}^{(1)}$, obteniéndose

$$\begin{matrix} 1 & 0 & a_{13}^{(2)} & \cdots & a_{1n}^{(2)} & a_{1,n+1}^{(2)} \\ 0 & 1 & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} & a_{2,n+1}^{(2)} \\ 0 & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} & a_{3,n+1}^{(2)} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} & a_{n,n+1}^{(2)} \end{matrix}$$

donde

$$a_{2j}^{(2)} = a_{2j}^{(1)} / a_{22}^{(1)}, \quad j = 3, 4, ..., n+1$$

$$a_{ij}^{(2)} = a_{ij}^{(1)} - a_{i2}^{(1)}\ a_{2j}^{(2)} = a_{ij}^{(1)} - \frac{a_{i2}^{(1)}}{a_{22}^{(1)}} a_{2j}^{(1)}, \quad i = 1, 3, ..., n, \quad j = 3, 4, ..., n+1$$

Continuando de esta forma obtenemos finalmente

$$\begin{matrix} 1 & 0 & 0 & \cdots & 0 & a_{1,n+1}^{(n)} \\ 0 & 1 & 0 & \cdots & 0 & a_{2,n+1}^{(n)} \\ 0 & 0 & 1 & \cdots & 0 & a_{3,n+1}^{(n)} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & a_{n,n+1}^{(n)} \end{matrix} \qquad (13)$$

es decir que se genera una matriz identidad y, por lo tanto, la incógnita se obtiene como sigue

$$x_i = a_{i,n+1}^{(n)}, \quad i = 1, 2, ..., n \qquad (14)$$

Ejemplo 15. Resolvamos el sistema dado en el ejemplo 4 usando el método de Gauss-Jordan normalizado.

x_1	x_2	x_3	x_4	B	S	
2	-5	1	3	7	8	
4	1	2	-2	4	9	$a_{21}/\ a_{11}$ = 2
3	2	-3	1	2	5	$a_{31}/\ a_{11}$ = 3/2
1	5	3	-4	4	9	$a_{41}/\ a_{11}$ = 1/2
1	-5/2	1/2	3/2	7/2	4	$a_{12}/\ a_{22}$ = -5/22
0	11	0	-8	-10	-7	
0	19/2	-9/2	-7/2	-17/2	-7	$a_{32}/\ a_{22}$ = 19/22
0	15/2	5/2	-11/2	1/2	5	$a_{42}/\ a_{22}$ = 15/22
1	0	1/2	-7/22	27/22	53/22	$a_{13}/\ a_{33}$ = -1/9
0	1	0	-8/11	-10/11	-7/11	$a_{23}/\ a_{33}$ = 0
0	0	-9/2	75/22	3/22	-21/22	
0	0	5/2	-1/22	161/22	215/22	$a_{43}/\ a_{33}$ = -5/9
1	0	0	2/33	41/33	76/33	$a_{14}/\ a_{44}$ = 2/61
0	1	0	-8/11	-10/11	-7/11	$a_{24}/\ a_{44}$ = -24/61
0	0	1	-25/33	-1/33	7/33	$a_{34}/\ a_{44}$ = -25/61
0	0	0	61/33	244/33	305/33	
1	0	0	0	1	2	
0	1	0	0	2	3	
0	0	1	0	3	4	
0	0	0	1	4	5	

Luego,

Solución	**Verificación**
$x_4 = 4$	$\overline{x_4} = 5$
$x_3 = 3$	$\overline{x_3} = 4$
$x_2 = 2$	$\overline{x_2} = 3$
$x_1 = 1$	$\overline{x_1} = 2$

y $\Delta = 2 \text{ x } 11 \text{ x} \left(\frac{-9}{2} \right) \text{x} \frac{61}{33} = -183$.

3.2.2.1. Inversión de matrices.

Suponiendo que A es una matriz cuadrada no singular $n \text{x} n$, dado que la inversa de A, es decir, A^{-1} satisface la condición

$$A A^{-1} = I = A^{-1} A \tag{15}$$

nuestro problema será, de acuerdo a (15), determinar una matriz $n \text{x} n$, G, tal que

$$A G = I . \tag{16}$$

Si designamos con $G_1, G_2, \cdots, G_n$ los vectores formados por la primera, segunda, ..., n-ésima columna de G y, análogamente, si definimos los vectores unitarios $I_1, I_2, ..., I_n$ formados por la primera, segunda, ..., n-ésima columna de I, la ecuación matricial (16) puede reemplazarse por n sistemas de ecuaciones, todos con los mismos coeficientes de la matriz A

$$A G_r = I_r , \qquad r = 1, 2, ..., n. \tag{17}$$

Los n sistemas (17) tienen solución única dado que hemos supuesto que $\det(A) \neq 0$. Para encontrar la matriz G, o sea la inversa de la matriz A, aplicamos el método de eliminación de Gauss - Jordan normalizado en forma simultánea a los n sistemas de ecuaciones.

Ejemplo 16. Para ilustrar el procedimiento de Gauss – Jordan normalizado, nos proponemos encontrar la inversa de la matriz

$$A = \begin{pmatrix} 2 & -5 & 1 & 3 \\ 4 & 1 & 2 & -2 \\ 3 & 2 & -3 & 1 \\ 1 & 5 & 3 & -4 \end{pmatrix}. \tag{18}$$

Dado que encontrar la inversa de esta matriz es equivalente a resolver cuatro sistemas de ecuaciones lineales que tienen los mismos coeficientes y difieren los términos independientes, escribimos en detalle los cuatro sistemas (17) aplicados a la matriz (18):

$$\begin{array}{cccc:cccc|cc} & & & & & & & & S & \\ 2 & -5 & 1 & 3 & 1 & 0 & 0 & 0 & 2 & \\ 4 & 1 & 2 & -2 & 0 & 1 & 0 & 0 & 6 & 2 \\ 3 & 2 & -3 & 1 & 0 & 0 & 1 & 0 & 4 & 3/2 \\ 1 & 5 & 3 & -4 & 0 & 0 & 0 & 1 & 6 & 1/2 \end{array} \qquad (19)$$

Para controlar los cálculos en caso de hacerse en forma manual, hemos agregado la columna *S*, suma de los coeficientes y términos independientes.

Observemos que en (19) tenemos la matriz de coeficientes *A* aumentada con la matriz identidad. Se aplica entonces, el método de eliminación de Gauss - Jordan normalizado, exactamente en la misma forma que al resolver un conjunto de ecuaciones lineales para reducir la matriz *A* a la matriz identidad. Cuando se completa esta tarea, la mitad derecha de la matriz aumentada contiene la matriz inversa A^{-1} .

Usando $a_{11} = 2$ como pivote, el primer renglón se normaliza y se usa para eliminar a x_1 de los otros renglones:

$$\begin{array}{cccc:cccc|cc} & & & & & & & & S & \\ 1 & -5/2 & 1/2 & 3/2 & 1/2 & 0 & 0 & 0 & 1 & -5/22 \\ 0 & 11 & 0 & -8 & -2 & 1 & 0 & 0 & 2 & \\ 0 & 19/2 & -9/2 & -7/2 & -3/2 & 0 & 1 & 0 & 1 & 19/22 \\ 0 & 15/2 & 5/2 & -11/2 & -1/2 & 0 & 0 & 1 & 5 & 15/22 \end{array}$$

Enseguida se usa $a_{22} = 11$ como pivote y x_2 se elimina de la primera, tercera y cuarta fila. Nos queda

$$\begin{array}{cccc:cccc|cc} & & & & & & & & S & \\ 1 & 0 & 1/2 & -7/22 & 1/22 & 5/22 & 0 & 0 & 32/22 & -1/9 \\ 0 & 1 & 0 & -8/11 & -2/11 & 1/11 & 0 & 0 & 2/11 & 0 \\ 0 & 0 & -9/2 & 75/22 & 5/22 & -19/22 & 1 & 0 & -16/22 & \\ 0 & 0 & 5/2 & -1/22 & 19/22 & -15/22 & 0 & 1 & 80/22 & -5/9 \end{array}$$

Ahora se usa $a_{33} = -9/2$ como pivote y x_3 se elimina de la primera, segunda y cuarta fila:

								S	
1	0	0	$2/33$	$7/99$	$13/99$	$1/9$	0	$136/99$	$2/61$
0	1	0	$-8/11$	$-2/11$	$1/11$	0	0	$2/11$	$-24/61$
0	0	1	$-25/33$	$-5/99$	$19/99$	$-2/9$	0	$16/99$	$-25/61$
0	0	0	$61/33$	$98/99$	$-115/99$	$5/9$	1	$320/99$	

Finalmente se usa $a_{44} = 61/33$ como pivote y x_4 se elimina de las tres primeras filas:

							S
1	0	0	0	$7/183$	$31/183$	$17/183$	$-2/61$	$232/183$
0	1	0	0	$38/183$	$-67/183$	$40/183$	$24/61$	$266/183$
0	0	1	0	$65/183$	$-52/183$	$1/183$	$25/61$	$272/183$
0	0	0	1	$98/183$	$-115/183$	$55/183$	$33/61$	$320/183$

Por lo tanto, la inversa es

$$A^{-1} = \begin{pmatrix} 7/183 & 31/183 & 17/183 & -2/61 \\ 38/183 & -67/183 & 40/183 & 24/61 \\ 65/183 & -52/183 & 1/183 & 25/61 \\ 98/183 & -115/183 & 55/183 & 33/61 \end{pmatrix}. \qquad (20)$$

Cada una de las columnas de esta matriz es solución del sistema respectivo (19).

Se puede verificar que la matriz (20) es la matriz inversa de A multiplicando a la matriz A a la izquierda o a la derecha por A^{-1} , que como sabemos nos debe dar I, suponiendo que no existen errores de redondeo.

Aunque no utilizamos pivoteo en la explicación anterior, éste es necesario puesto que el esquema de inversión es, en esencia, un proceso de eliminación. Por fortuna, la matriz inversa no se ve afectada por un cambio en el orden secuencial de las ecuaciones, esto es, A^{-1} no se altera por el pivoteo. Ilustraremos esto con un ejemplo.

Ejemplo 17. Consideremos el siguiente sistema de ecuaciones lineales

$$\begin{cases} x_1 + x_2 + x_3 = 1 \\ x_1 + x_2 + 2x_3 = 2 \\ x_1 + 2x_2 + 2x_3 = 1 \end{cases}$$

(resuelto por eliminación de Gauss, sistema (9) dado en el ejemplo 8).

Llamemos

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \end{pmatrix}$$

la matriz cuyos elementos son los coeficientes del sistema. Debemos hallar A^{-1} según el procedimiento anterior.

$$\begin{array}{ccc:ccc|c} & & & & & & S \\ 1 & 1 & 1 & 1 & 0 & 0 & 4 \\ 1 & 1 & 2 & 0 & 1 & 0 & 5 \quad 1 \\ 1 & 2 & 2 & 0 & 0 & 1 & 6 \quad 1 \end{array}$$

$$\begin{array}{ccc:ccc|c} & & & & & & S \\ 1 & 1 & 1 & 1 & 0 & 0 & 4 \\ 0 & 0 & 1 & -1 & 1 & 0 & 1 \\ 0 & 1 & 1 & -1 & 0 & 1 & 2 \end{array}$$

Intercambiamos 2° y 3° filas.

$$\begin{array}{ccc:ccc|cc} & & & & & & S & \\ 1 & 1 & 1 & 1 & 0 & 0 & 4 & 1 \\ 0 & 1 & 1 & -1 & 0 & 1 & 2 & \\ 0 & 0 & 1 & -1 & 1 & 0 & 1 & 0 \end{array}$$

$$\begin{array}{ccc:ccc|cc} & & & & & & S & \\ 1 & 0 & 0 & 2 & 0 & -1 & 2 & 0 \\ 0 & 1 & 1 & -1 & 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & -1 & 1 & 0 & 1 & \end{array}$$

$$\begin{array}{ccc:ccc|c} & & & & & & S \\ 1 & 0 & 0 & 2 & 0 & -1 & 2 \\ 0 & 1 & 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 1 & -1 & 1 & 0 & 1 \end{array}$$

Luego,

$$A^{-1} = \begin{pmatrix} 2 & 0 & -1 \\ 0 & -1 & 1 \\ -1 & 1 & 0 \end{pmatrix}.$$

Para verificar, hacemos

$$A\ A^{-1} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \end{pmatrix} \begin{pmatrix} 2 & 0 & -1 \\ 0 & -1 & 1 \\ -1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = I$$

3.2.3. Esquemas compactos para la eliminación Gaussiana.

Para resolver un sistema lineal por eliminación Gaussiana son muchas las operaciones que debemos realizar, lo cual es muy pesado y da origen a muchos errores. Es posible, sin embargo, ordenar los cálculos de modo que los elementos de L y U se determinen directamente. Esto da origen al llamado *método de Gauss – Doolitle*, método que es equivalente a la eliminación Gaussiana. No desarrollaremos este método aquí, pero en cambio sí analizaremos otro método que es más usado en los programas para computadoras y que se obtiene haciendo una ligera modificación en el método de eliminación Gaussiana. Este es el llamado *método de Crout* (1941) (similar al método de Gauss - Doolitle, con la variante que se normaliza la matriz U de modo que $u_{kk} = 1$).

3.2.3.1. Método de Crout.

Supongamos que queremos hallar la solución de

$$Ax = B \tag{21}$$

donde A es la matriz de los coeficientes que es transformada en el producto de dos matrices L y U, donde ahora L es una matriz triangular inferior y U es una matriz triangular superior con unos en la diagonal principal, o sea, $u_{ii} = 1$ ($i = 1, 2, ..., n$).

Esto es, si

$$\begin{pmatrix} m_{11} & 0 & 0 & \cdots & 0 \\ m_{21} & m_{22} & 0 & \cdots & 0 \\ m_{31} & m_{32} & m_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ m_{n1} & m_{n2} & m_{n3} & \cdots & m_{nn} \end{pmatrix} \begin{pmatrix} 1 & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & 1 & u_{23} & \cdots & u_{2n} \\ 0 & 0 & 1 & \cdots & u_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & m_{n3} & \cdots & a_{nn} \end{pmatrix} \quad (22)$$

es decir que si $LU = A$, nuestro sistema se reduce a

$$LUx = B$$

y si llamamos

$$LH = B$$

resulta ser

$$Ux = H. \quad (23)$$

De modo que el sistema (21) queda reducido, según la ecuación (23), a

$$\begin{aligned} x_1 + u_{12}x_2 + u_{13}x_3 \cdots + u_{1n}x_n &= h_1 \\ x_2 + u_{23}x_3 \cdots + u_{2n}x_n &= h_2 \\ x_3 \cdots + u_{3n}x_n &= h_3 \\ \vdots \quad \vdots \quad \vdots & \\ x_n &= h_n \end{aligned}$$

Este sistema nos permite determinar las incógnitas a partir de x_n, luego x_{n-1}, x_{n-2}, ..., hasta x_1.

Para determinar los elementos m_{ij} y u_{ij}, se observa que si en la ecuación (22) multiplicamos las filas de L por la primera columna de U, obtenemos

$$\begin{aligned} m_{11} &= a_{11} \\ m_{21} &= a_{21} \\ m_{31} &= a_{31} \\ &\cdot \\ &\cdot \\ &\cdot \\ m_{n1} &= a_{n1} \end{aligned}$$

es decir, la primera columna de L es la primera columna de A.

Multiplicando ahora la primera fila de L por las columnas de U, obtenemos

$$\begin{aligned}
m_{11}u_{12} &= a_{12}\ , \text{ de donde,} \quad u_{12} = a_{12}/m_{11}\\
m_{11}u_{13} &= a_{13}\ , \text{ de donde,} \quad u_{13} = a_{13}/m_{11}\\
m_{11}u_{14} &= a_{14}\ , \text{ de donde,} \quad u_{14} = a_{14}/m_{11}\\
&\vdots\\
m_{11}u_{1n} &= a_{1n}\ , \text{ de donde,} \quad u_{1n} = a_{1n}/m_{11}
\end{aligned}$$

y determinamos así la primera fila de U.

En este método se obtienen de manera alternada una columna de L y una fila de U, de modo que para obtener la segunda columna de L se multiplican las filas de L por la segunda columna de U. Se obtiene, entonces

$$\begin{aligned}
m_{21}u_{12} + m_{22} &= a_{22}\ , \text{ de donde,} \quad m_{22} = a_{22} - m_{21}u_{12}\\
m_{31}u_{12} + m_{32} &= a_{32}\ , \text{ de donde,} \quad m_{32} = a_{32} - m_{31}u_{12}\\
&\vdots\\
m_{n1}u_{12} + m_{n2} &= a_{n2}\ , \text{ de donde,} \quad m_{n2} = a_{n2} - m_{n1}u_{12}
\end{aligned}$$

Para obtener la segunda fila de U multiplicamos la segunda fila de L por las columnas de U, y obtenemos

$$\begin{aligned}
m_{21}u_{13} + m_{22}u_{23} &= a_{23}\ , \text{ de donde,} \quad u_{23} = (a_{23} - m_{21}u_{13})/m_{22}\\
m_{21}u_{14} + m_{22}u_{24} &= a_{24}\ , \text{ de donde,} \quad u_{24} = (a_{24} - m_{21}u_{14})/m_{22}\\
&\vdots\\
m_{21}u_{1n} + m_{22}u_{2n} &= a_{2n}\ , \text{ de donde,} \quad u_{2n} = (a_{2n} - m_{21}u_{1n})/m_{22}
\end{aligned}$$

Las fórmulas generales para obtener los elementos de las matrices L y U pueden escribirse como sigue

$$\begin{cases}
m_{ij} = a_{ij} - \sum_{k=1}^{j-1} m_{ik}u_{kj}, & i \geq j\\
u_{ij} = (a_{ij} - \sum_{k=1}^{i-1} m_{ik}u_{kj})/m_{ii}, & i < j
\end{cases} \tag{24}$$

Conociendo L y siendo $LH = B$, podemos escribir

$$\begin{pmatrix} m_{11} & 0 & 0 & \cdots & 0 \\ m_{21} & m_{22} & 0 & \cdots & 0 \\ m_{31} & m_{32} & m_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ m_{n1} & m_{n2} & m_{n3} & \cdots & m_{nn} \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ \vdots \\ h_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{pmatrix}.$$

Luego,

$$m_{11} h_1 = b_1, \text{ de donde, } h_1 = b_1 / m_{11}$$
$$m_{21} h_1 + m_{22} h_2 = b_2, \text{ de donde, } h_2 = (b_2 - m_{21} h_1)/ m_{22}$$
$$m_{31} h_1 + m_{32} h_2 + m_{33} h_3 = b_3, \text{ de donde, } h_3 = (b_3 - m_{31} h_1 - m_{32} h_2)/ m_{33}$$

y, en general,

$$h_i = (b_i - \sum_{k=1}^{i-1} m_{ik} h_k) / m_{ii} \quad (i = 1, 2, \ldots, n). \tag{25}$$

Los elementos de la matriz columna H se calculan en forma simultánea cuando se calculan los elementos de la matriz u_{ij} dados por la ecuación (24), y siendo además de la ecuación (23) $Ux = H$, es decir

$$\begin{pmatrix} 1 & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & 1 & u_{23} & \cdots & u_{2n} \\ 0 & 0 & 1 & \cdots & u_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ \vdots \\ h_n \end{pmatrix}$$

entonces,

$$x_i = h_i - \sum_{k=i+1}^{n} u_{ik} x_k \quad (i = n, n-1, \ldots, 1). \tag{26}$$

Este método adquirió cierta popularidad, dado que permite economizar almacenamiento de memoria en la programación para computadoras. No es necesario almacenar los ceros de las matrices L y U y los unos de la matriz U. Los unos pueden omitirse de la diagonal principal y pueden almacenarse valores de u_{ij} de la matriz U donde aparecen ceros de la matriz L, y pueden almacenarse elementos m_{ij} de la matriz L donde hay ceros de la matriz U. Si uno examina las ecuaciones (24), se observa que los

elementos a_{ij} de la matriz A se usan sólo una vez y después no aparecen en las ecuaciones. Por lo tanto, el almacenamiento de los a_{ij} de la matriz A puede utilizarse como almacenamiento de las matrices L y U. En otros términos, si consideramos la matriz ampliada formada por los coeficientes a_{ij} y los términos independientes b_i, esta matriz después de las operaciones aritméticas indicadas tiene el aspecto siguiente

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} & b_2 \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} & b_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} & b_n \end{pmatrix} \rightarrow \begin{pmatrix} m_{11} & u_{12} & u_{13} & \cdots & u_{1n} & h_1 \\ m_{21} & m_{22} & u_{23} & \cdots & u_{2n} & h_2 \\ m_{31} & m_{32} & m_{33} & \cdots & u_{3n} & h_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ m_{n1} & m_{n2} & m_{n3} & \cdots & m_{nn} & h_n \end{pmatrix}. \quad (27)$$

En resumen, para obtener los elementos de la matriz (27) a partir de la matriz original aumentada con los términos independientes, deben tenerse presente las reglas siguientes:

a) Se comienza determinando los elementos de la primera columna $m_{i1} = a_{i1}$; luego los elementos de la primera fila a la derecha de la primera columna; a continuación se determinan los elementos de la segunda columna debajo de la primera fila; luego los elementos de la segunda fila a la derecha de la segunda columna; siguen los elementos de la tercera columna debajo de la segunda fila; luego los elementos de la tercera fila a la derecha de la tercera columna, y así siguiendo hasta que todos los elementos se han determinado.

b) Los elementos de la primera columna de la matriz (27) son los $m_{i1} = a_{i1}$ ($i = 1, 2, ..., n$) de la matriz original. Cada elemento de la primera fila de (27), con excepción del primero, se obtiene dividiendo el correspondiente elemento de la matriz original por el primer elemento; o sea

$$u_{1j} = a_{1j}/a_{11} = a_{1j}/m_{11}, \quad j = 2, 3, ..., n+1.$$

Hemos considerado en estas últimas expresiones que

$$a_{1,\,n+1} = b_1, \qquad u_{1,\,n+1} = h_1.$$

Observemos que, en general, en (27) podemos considerar

$$a_{i,\,n+1} = b_i, \qquad u_{i,\,n+1} = h_i, \qquad i = 1, 2, \ldots, n.$$

c) Cada elemento de la diagonal principal o debajo de ella es igual al correspondiente elemento de la matriz original menos la suma de los

productos de los elementos de la fila y los correspondientes de la columna previamente calculados

$$m_{ij} = a_{ij} - \sum_{k=1}^{j-1} m_{ik} u_{kj}, \qquad i \geq j \quad (i = j,\ j+1,\ \ldots,\ n).$$

d) Cada elemento a la derecha de la diagonal se calcula de acuerdo a c), dividiendo por el elemento de la diagonal

$$u_{ij} = \left(a_{ij} - \sum_{k=1}^{i-1} m_{ik} u_{kj} \right) / m_{ii}, \qquad i < j \quad (j = i+1,\ i+2,\ldots,\ n+1).$$

e) Como $A = LU$, entonces es claro que

$$\det(A) = \det(LU) = \det(L)\det(U) = m_{11} m_{22} \ldots m_{nn}$$

esto es, el determinante de la matriz A es el producto de los elementos diagonales de la matriz L.

Este método puede ser combinado con pivoteo, y en este caso si el número de intercambios es impar se modifica el signo del determinante.

Con el fin de simplificar lo anteriormente expuesto y a modo de ejemplo, hallemos la solución de un sistema de 3 ecuaciones con 3 incógnitas usando el método de Crout:

$$\begin{aligned} a_1 x &+ b_1 y &+ c_1 z &= d_1 \\ a_2 x &+ b_2 y &+ c_2 z &= d_2 \\ a_3 x &+ b_3 y &+ c_3 z &= d_3 \end{aligned} \qquad (28)$$

Escribamos la matriz del sistema ampliada más una columna de control:

$$\left(\begin{array}{ccc:c|c} a_1 & b_1 & c_1 & d_1 & e_1 \\ a_2 & b_2 & c_2 & d_2 & e_2 \\ a_3 & b_3 & c_3 & d_3 & e_3 \end{array} \right), \text{ donde, } \quad e_i = a_i + b_i + c_i + d_i \quad (i = 1, 2, 3).$$

Computamos ahora una matriz a partir de la anterior, que escribiremos

$$\left(\begin{array}{ccc:c|c} A_1 & B_1 & C_1 & D_1 & E_1 \\ A_2 & B_2 & C_2 & D_2 & E_2 \\ A_3 & B_3 & C_3 & D_3 & E_3 \end{array}\right)$$

donde,

$A_i = a_i \quad (i = 1, 2, 3)$ Primera columna.

$B_1 = \frac{b_1}{a_1}, C_1 = \frac{c_1}{a_1}, D_1 = \frac{d_1}{a_1}, E_1 = \frac{e_1}{a_1}$ Primera fila a la derecha de la primeracolumna.

$B_2 = b_2 - B_1A_2$ Remanente de la segunda columna.
$B_3 = b_3 - B_1A_3$

$C_2 = (c_2 - C_1A_2)/B_2$ Remanente de la segunda fila.
$D_2 = (d_2 - D_1A_2)/B_2$
$E_2 = (e_2 - E_1A_2)/B_2$

$C_3 = c_3 - C_2B_3 - C_1A_3$ Remanente de la tercera columna.

$D_3 = (d_3 - D_2B_3 - D_1A_3)/C_3$ Remanente de la tercera fila.
$E_3 = (e_3 - E_2B_3 - E_1A_3)/C_3$

La solución viene dada por

$$z = D_3, \qquad y = D_2 - C_2z, \qquad x = D_1 - C_1z - B_1y.$$

La quinta columna de los E_i (i = 1, 2, 3) sirve como control de los cálculos, en los que los E_i son iguales a 1 más la suma de los elementos de la misma fila a la derecha de la diagonal principal, es decir

$$\begin{aligned} E_1 &= 1 + B_1 + C_1 + D_1 \\ E_2 &= 1 + C_2 + D_2 \\ E_3 &= 1 + D_3 \end{aligned} \tag{29}$$

chequeo que debe realizarse después de completar cada fila. Se utilizan también para controlar la solución

$$\bar{z} = E_3, \quad \bar{y} = E_2 - C_2\bar{z}, \quad \bar{x} = E_1 - C_1\bar{z} - B_1\bar{y}$$

que excederán en una unidad a x, y, z, respectivamente.

Observemos además que

$$L = \begin{pmatrix} A_1 & 0 & 0 \\ A_2 & B_2 & 0 \\ A_3 & B_3 & C_3 \end{pmatrix}, \quad U = \begin{pmatrix} 1 & B_1 & C_1 \\ 0 & 1 & C_2 \\ 0 & 0 & 1 \end{pmatrix}.$$

Como los elementos diagonales de U son unos y los que están por debajo de dicha diagonal son ceros, entonces esto justifica la ecuación (29).

En el caso que los coeficientes de la matriz dada sean simétricos respecto de la diagonal principal, el trabajo para computar la matriz auxiliar se reduce casi a la mitad. Esto puede observarse, ya que si dividimos los elementos debajo de la diagonal principal de la matriz auxiliar por sus elementos diagonales, obtenemos los elementos simétricamente opuestos respecto de esta diagonal, o sea, los elementos a la derecha de dicha diagonal.

En efecto, si el sistema (28) fuera simétrico tendríamos

$$b_1 = a_2, \quad c_1 = a_3, \quad c_2 = b_3$$

de modo que la matriz ampliada puede escribirse

$$\left(\begin{array}{ccc|c} a_1 & a_2 & a_3 & d_1 \\ b_1 & b_2 & b_3 & d_2 \\ c_1 & c_2 & c_3 & d_3 \end{array}\right)$$

y en la matriz transformada tendríamos

$$A_1 = a_1, \quad A_2 = b_1, \quad A_3 = c_1$$

$$B_1 = \frac{b_1}{a_1} = \frac{A_2}{A_1}$$

$$C_1 = \frac{c_1}{a_1} = \frac{A_3}{A_1}$$

$$C_2 = \frac{c_2 - C_1 A_2}{B_2} = \frac{b_3 - B_1 A_3}{B_2} = \frac{B_3}{B_2}$$

(en la última ecuación hemos usado que: $C_1A_2=B_1A_3$, pues $C_1 = \frac{A_3}{A_1}$, de donde $C_1A_2 = \frac{A_3}{A_1}A_2$, o sea $C_1A_2 = \frac{A_2}{A_1}A_3$, y como $\frac{A_2}{A_1} = B_1$, entonces $C_1A_2 = B_1A_3$).

Esto reduce considerablemente los cálculos debido a que los elementos simétricamente correspondientes de la matriz que se computa pueden ser calculados con una división adicional.

Ejemplo 18. Resolvamos con Crout, el siguiente sistema de ecuaciones lineales

$$\begin{cases} x + y + z = 1 \\ 3x + y - 3z = 5 \\ x - 2y - 5z = 10 \end{cases}$$

Disponemos los datos y los cálculos en el siguiente esquema práctico:

x	y	z	B	S
1	1	1	1	4
3	1	-3	5	6
1	-2	-5	10	4
1	1	1	1	4
3	-2	3	-1	3
1	-3	3	2	3

Luego,

Solución	**Verificación**
$z = 2$	$\bar{z} = 3$
$y = -1 - 3x2 = -7$	$\bar{y} = 3 - 3x3 = -6$
$x = 1 - 1x2 - 1x(-7) = 6$	$\bar{x} = 4 - 1x3 - 1x(-6) = 7.$

Esto es,

$$x = 6, \quad y = -7, \quad z = 2.$$

Además,

$$\det(A) = 1 \text{ x } (-2) \text{ x } 3 = -6$$

y

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 3 & -2 & 0 \\ 1 & -3 & 3 \end{pmatrix}, \quad U = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix}.$$

Ejemplo 19. Resolvamos con Crout, el sistema dado en el ejemplo 4.

x_1	x_2	x_3	x_4	B	S
2	-5	1	3	7	8
4	1	2	-2	4	9
3	2	-3	1	2	5
1	5	3	-4	4	9
2	-5/2	1/2	3/2	7/2	4
4	11	0	-8/11	-10/11	-7/11
3	19/2	-9/2	-25/33	-1/33	7/33
1	15/2	5/2	61/33	4	5

Luego,

Solución	**Verificación**
$x_4 = 4$	$\overline{x_4} = 5$
$x_3 = 3$	$\overline{x_3} = 4$
$x_2 = 2$	$\overline{x_2} = 3$
$x_1 = 1$	$\overline{x_1} = 2$

y

$$\det(A) = 2 \text{ x } 11 \text{ x } (-9/2) \text{x } (61/33) = -183.$$

Ejemplo 20. Resolvamos con el método de Crout, el sistema simétrico que se muestra en el siguiente esquema

x_1	x_2	x_3	x_4	x_5	B	S
1	1	1	1	1	8	13
1	2	3	4	5	25	40
1	3	6	10	15	61	96
1	4	10	20	35	125	195
1	5	15	35	70	228	354
1	1	1	1	1	8	13
1	1	2	3	4	17	27
1	2	1	3	6	19	29
1	3	3	1	4	9	14
1	4	6	4	1	2	3

Luego,

Solución	Verificación
$x_5 = 2$	$\overline{x_5} = 3$
$x_4 = 1$	$\overline{x_4} = 2$
$x_3 = 4$	$\overline{x_3} = 5$
$x_2 = -2$	$\overline{x_2} = -1$
$x_1 = 3$	$\overline{x_1} = 4$

y $\det(A) = 1$.

3.2.3.1.1. Inversión de matrices.

Mediante el método de Crout podemos obtener la traspuesta de la matriz inversa de un conjunto dado de ecuaciones lineales. Para hacerlo, la matriz de coeficientes se aumenta con una matriz identidad. Posteriormente se aplica el método de Crout y cuando se completa esta tarea, resolvemos para $\vec{x}$ usando las constantes obtenidas en la mitad derecha (a su turno, según corresponda). La matriz resultante es la traspuesta de la matriz inversa del sistema original. Esta técnica se ilustra en los ejemplos siguientes.

Ejemplo 21. Determinemos la matriz inversa del sistema dado en el ejemplo 4, usando el método de Crout.

Aumentamos la matriz de los coeficientes con una matriz identidad y para controlar los cálculos en caso de hacerse en forma manual agregamos la columna *S*, suma de los elementos de cada renglón. Luego aplicamos el método de Crout.

x	y	z	e_1	e_2	e_3	S
1	1	1	1	0	0	4
3	1	-3	0	1	0	2
1	-2	-5	0	0	1	-5
1	1	1	1	0	0	4
3	-2	3	3/2	-1/2	0	5
1	-3	3	7/6	-1/2	1/3	2
11/6	-2	7/6				
-1/2	1	-1/2	$(A^{-1})^t$			
2/3	-1	1/3				

Luego,

$$A^{-1} = \begin{pmatrix} 11/6 & -1/2 & 2/3 \\ -2 & 1 & -1 \\ 7/6 & -1/2 & 1/3 \end{pmatrix}$$

y para verificar hacemos

$$A\,A^{-1} = \begin{pmatrix} 1 & 1 & 1 \\ 3 & 1 & -3 \\ 1 & -2 & -5 \end{pmatrix} \begin{pmatrix} 11/6 & -1/2 & 2/3 \\ -2 & 1 & -1 \\ 7/6 & -1/2 & 1/3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = I.$$

Observemos que estamos hallando simultáneamente las soluciones de 3 sistemas de ecuaciones lineales, a saber

$$As_1 = e_1$$
$$As_2 = e_2$$
$$As_3 = e_3$$

donde,

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 3 & 1 & -3 \\ 1 & -2 & -5 \end{pmatrix}, \quad s_i = \begin{pmatrix} x^{(i)} \\ y^{(i)} \\ z^{(i)} \end{pmatrix} \quad (i = 1, 2, 3; \text{ el supraíndice } (i) \text{ indica la fila } i \text{ de } (A^{-1})^t)$$

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Es decir que la solución de cada uno de estos sistemas es una matriz columna. Así, la solución del primer sistema es la traspuesta de (11/6 -2 7/6), la solución del segundo sistema es la traspuesta de (-1/2 1 -1/2) y por último, la solución del tercer sistema es la traspuesta de (2/3 -1 1/3). Esto es, la matriz resultante obtenida usando el método de Crout es la traspuesta de la matriz inversa del sistema original. Luego,

$$A^{-1} = \begin{pmatrix} 11/6 & -1/2 & 2/3 \\ -2 & 1 & -1 \\ 7/6 & -1/2 & 1/3 \end{pmatrix}.$$

Podemos hallar simultáneamente la solución del sistema $Ax = B$ y la inversa de la matriz A. Lo único que cambia en el esquema práctico es la columna S a la que además, en este caso, se le suman los términos independientes correspondientes. Considerando el ejemplo anterior, resulta

x	y	z	B	e_1	e_2	e_3	S
1	1	1	1	1	0	0	5
3	1	-3	5	0	1	0	7
1	-2	-5	10	0	0	1	5
1	1	1	1	1	0	0	5
3	-2	3	-1	3/2	-1/2	0	4
1	-3	3	2	7/6	-1/2	1/3	4
6	-7	2		Solución			
11/6	-2	7/6					
-1/2	1	-1/2		$(A^{-1})^t$			
2/3	-1	1/3					

Luego,

$$x = 6, \quad y = -7, \quad z = 2$$

y

$$A^{-1} = \begin{pmatrix} 11/6 & -1/2 & 2/3 \\ -2 & 1 & -1 \\ 7/6 & -1/2 & 1/3 \end{pmatrix}.$$

Ejemplo 22. Determinemos la matriz inversa de

$$A = \begin{pmatrix} 3 & -1 & 4 \\ 1 & 2 & -1 \\ 4 & -1 & 1 \end{pmatrix}$$

utilizando el método de Crout.

x	y	z	e_1	e_2	e_3	S
3	-1	4	1	0	0	7
1	2	-1	0	1	0	3
5	-1	1	0	0	1	6
3	-1/3	4/3	1/3	0	0	7/3
1	7/3	-1	-1/7	3/7	0	2/7
5	2/3	-5	11/35	2/35	-1/5	41/35
-1/35	6/35	11/35				
3/35	17/35	2/35	$(A^{-1})^t$			
1/5	-1/5	-1/5				

Luego,

$$A^{-1} = \begin{pmatrix} -1/35 & 3/35 & 1/5 \\ 6/35 & 17/35 & -1/5 \\ 11/35 & 2/35 & -1/5 \end{pmatrix}.$$

3.2.3.1.2. Mejoramiento en la exactitud de las soluciones.

Los valores obtenidos para las incógnitas del sistema lineal $Ax = B$ no son, por lo general, exactos y sus sustituciones en las ecuaciones originales puede que no balancee exactamente las ecuaciones.

Cuando desarrollamos el método de Gauss vimos que un mejoramiento en las soluciones puede obtenerse calculando las diferencias ($\varepsilon = B - Ax_0$, siendo x_0 una solución aproximada) y tratándolas como una nueva columna de términos constantes. Para ello, todo lo que se necesita es adjuntar al esquema principal de cálculo esta nueva columna de términos constantes y transformarla mediante las reglas generales del método de Crout (en este caso). Obtenemos así la nueva columna en la matriz auxiliar, que nos da las correcciones, digamos $\delta_1, \delta_2, ..., \delta_n$, que se hallan de forma similar a como se calculan la solución y los valores $\overline{x}_i$ que se utilizan para verificar la solución. Estas correcciones sumadas a los valores de las raíces aproximadas obtenidas previamente nos dan valores mejorados de las mismas, que reemplazados en las ecuaciones originales nos darán nuevamente los restos, pudiendo continuar con este procedimiento tantas veces como se desee.

Es esencial que los métodos vistos puedan ser combinados con pivotación parcial (la pivotación completa es muy difícil de realizar).

Si el sistema de ecuaciones dado es simétrico y es tal que la matriz de los coeficientes de las incógnitas es definida positiva, entonces surge otro método conocido con el nombre de *Cholesky* (o *método de la raíz cuadrada*). Este método es, dentro de los esquemas compactos muy atractivo, pues no es necesario ninguna técnica de pivotación.

Para matrices simétricas y definidas positivas tenemos la siguiente variante del teorema *LU*.

Teorema. Sea A una matriz simétrica y definida positiva. Entonces, existe una única matriz triangular superior S con elementos diagonales positivos tal que $A = S^t S$ (S^t es la matriz traspuesta de S).

3.2.3.2. Método de Cholesky.

Sea el sistema dado

$$Ax = B \tag{30}$$

donde A es una matriz simétrica y definida positiva.

Conociendo la matriz A, nos proponemos determinar la matriz triangular superior S

$$S = \begin{pmatrix} s_{11} & s_{12} & s_{13} & \cdots & s_{1n} \\ 0 & s_{22} & s_{23} & \cdots & s_{2n} \\ 0 & 0 & s_{33} & \cdots & s_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & s_{nn} \end{pmatrix}$$

tal que

$$S^t S = A \tag{31}$$

donde S^t es la matriz traspuesta de S, es decir, $S_{ij}^t = S_{ji}$. Entonces, reemplazando en (30), se tiene

$$S^t S\, x = B. \tag{32}$$

Esta última ecuación matricial puede expresarse mediante las dos ecuaciones siguientes

$$S^t K = B$$

donde,

$$Sx = K \quad (K \text{ es una matriz columna}).$$

Una vez determinada la matriz S mediante la ecuación (31), puede fácilmente determinarse K y conociendo K y S se determina la incógnita x. La ecuación (31) puede escribirse

$$\begin{pmatrix} s_{11} & 0 & 0 & \cdots & 0 \\ s_{12} & s_{22} & 0 & \cdots & 0 \\ s_{13} & s_{23} & s_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{1n} & s_{2n} & s_{3n} & \cdots & s_{nn} \end{pmatrix} \begin{pmatrix} s_{11} & s_{12} & s_{13} & \cdots & s_{1n} \\ 0 & s_{22} & s_{23} & \cdots & s_{2n} \\ 0 & 0 & s_{33} & \cdots & s_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & s_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{pmatrix}.$$

Multiplicando la primera fila de S^t por cada una de las columnas de *S*, obtenemos

$$\begin{aligned} s_{11}\, s_{11} &= a_{11} \\ s_{11}\, s_{12} &= a_{12} \\ s_{11}\, s_{13} &= a_{13} \\ &\vdots \\ s_{11}\, s_{1n} &= a_{1n} \end{aligned}$$

Luego,

$$s_{11} = \sqrt{a_{11}} \tag{33}$$

$$s_{1j} = a_{1j}/s_{11}, \quad j = 2, 3, \ldots, n$$

quedando determinada la primera fila de la matriz *S*.

Multiplicando la segunda fila de S^t por la segunda, tercera, ... , *n*-ésima columna de *S*, obtenemos

$$\begin{aligned} s_{12}\, s_{12} + s_{22}\, s_{22} &= a_{22} \\ s_{12}\, s_{13} + s_{22}\, s_{23} &= a_{23} \\ s_{12}\, s_{14} + s_{22}\, s_{24} &= a_{24} \\ &\vdots \\ s_{12}\, s_{1n} + s_{22}\, s_{2n} &= a_{2n}. \end{aligned}$$

De aquí, surge

$$\begin{aligned} s_{22} &= \sqrt{a_{22} - s_{12}s_{12}} \\ s_{23} &= (a_{23} - s_{12}\, s_{13})/s_{22} \\ s_{24} &= (a_{24} - s_{12}\, s_{14})/s_{22} \\ &\vdots \\ s_{2n} &= (a_{2n} - s_{12}s_{1n})/s_{22} \end{aligned} \tag{34}$$

esto es,

$$s_{22} = \sqrt{a_{22} - s_{12}^{\,2}}$$
$$s_{2j} = (a_{2j} - s_{12}\, s_{1j})/s_{22}, \quad j = 3, 4, \ldots, n$$

expresiones que nos dan la segunda fila de S.

En general, multiplicando la fila i de S^t por las columnas i, i+1, i+2, ..., n de S, obtenemos

$$\sum_{r=1}^{i-1} s_{ri}\, s_{ri} \;+\; s_{ii}\, s_{ii} \;=\; a_{ii}$$

$$\sum_{r=1}^{i-1} s_{ri}\, s_{rj} \;+\; s_{ii}\, s_{ij} \;=\; a_{ij}, \quad j > i$$

de aquí, obtenemos que en general se verifica

$$s_{ii} = \left(a_{ii} - \sum_{r=1}^{i-1} s_{ri}^2 \right)^{1/2} \tag{35}$$

$$s_{ij} = \frac{1}{s_{ii}} \left(a_{ij} - \sum_{r=1}^{i-1} s_{ri} s_{rj} \right), \quad j > i$$

Simultáneamente, mientras se van calculando los coeficientes s_{ij} mediante (35) podemos obtener la matriz columna K de la relación $S^t K = B$, pues se tiene

$$\begin{pmatrix} s_{11} & 0 & 0 & \cdots & 0 \\ s_{12} & s_{22} & 0 & \cdots & 0 \\ s_{13} & s_{23} & s_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{1n} & s_{2n} & s_{3n} & \cdots & s_{nn} \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \\ k_3 \\ \vdots \\ k_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{pmatrix}$$

de donde resulta que

$s_{11}\, k_1 = b_1$, es decir, $k_1 = b_1/s_{11}$

$s_{12}\, k_1 + s_{22}\, k_2 = b_2$, es decir, $k_2 = (b_2 - s_{12}\, k_1)/s_{22}$

$s_{13}\, k_1 + s_{23}\, k_2 + s_{33}\, k_3 = b_3$, es decir, $k_3 = (b_3 - s_{13}\, k_1 - s_{23}\, k_2)/s_{33}$

y, en general,

$$k_i = \frac{1}{s_{ii}}\left(b_i - \sum_{r=1}^{i-1} s_{ri} k_r \right), \quad i = 1, 2, \ldots, n. \tag{36}$$

Finalmente, de la ecuación $Sx = K$, obtenemos

$$\begin{pmatrix} s_{11} & s_{12} & s_{13} & \cdots & s_{1n} \\ 0 & s_{22} & s_{23} & \cdots & s_{2n} \\ 0 & 0 & s_{33} & \cdots & s_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & s_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} k_1 \\ k_2 \\ k_3 \\ \vdots \\ k_n \end{pmatrix}$$

esto es,

$$\begin{aligned} s_{nn}\, x_n &= k_n \\ s_{n\text{-}1,\, n\text{-}1}\; x_{n\text{-}1} + s_{n\text{-}1,\, n}\, x_n &= k_{n\text{-}1} \\ s_{n\text{-}2,\, n\text{-}2}\; x_{n\text{-}2} + s_{n\text{-}2,\, n\text{-}1}\; x_{n\text{-}1} + s_{n\text{-}2,\, n}\, x_n &= k_{n\text{-}2} \end{aligned}$$

y, en general,

$$s_{ii} x_i + \sum_{r=i+1}^{n} s_{ir} x_r = k_i, \quad i = n, \ldots, 2, 1.$$

Por consiguiente,

$$x_i = \frac{1}{s_{ii}}\left(k_i - \sum_{r=i+1}^{n} s_{ir} x_r \right), \quad i = n, n\text{-}1, \ldots, 1. \tag{37}$$

En el cálculo manual conviene establecer el control de suma como se hizo en los métodos anteriores y resolver simultáneamente los dos sistemas. Siendo el sistema simétrico, sólo hace falta escribir los coeficientes siguientes

a_{11}	a_{12}	a_{13}	...	a_{1n}	b_1	$\overline{b_1}$
	a_{22}	a_{23}	...	a_{2n}	b_2	$\overline{b_2}$
		a_{33}	...	a_{3n}	b_3	b_3
					.	.
					.	.
					.	.
a_{nn}					b_n	$\overline{b_n}$

donde el coeficiente $\overline{b_i}$ se obtiene sumando todos los coeficientes de la fila y la columna i, esto es

$$\overline{b_i} = \sum_{j=1}^{i-1} a_{ji} + \sum_{j=i}^{n} a_{ij} + b_i \ , \quad i = 1, 2, ..., n.$$

Del sistema dado, deducimos el sistema y las incógnitas siguientes

s_{11}	s_{12}	s_{13}	...	s_{1n}		k_1	$\overline{k_1}$
	s_{22}	s_{23}	...	s_{2n}		k_2	k_2
		s_{33}	...	s_{3n}		k_3	k_3
					.	.	.
					.	.	.
					.	.	.
				s_{nn}		k_n	k_n
x_1	x_2	x_3	...	x_n			
x_1	x_2	x_3	...	x_n			

Los términos independientes k_i del sistema transformando se obtienen en forma análoga a los s_{ij} para $j > i$, según se deduce de (36) en donde se llama $b_i = a_{i,\, n+1}$, y así, según (35), $k_i = s_{i,\, n+1}$, para $i = 1, 2, ..., n$.

Los términos k_i se obtienen en forma análoga a los k_i y s_{ij}, y deben ser tales que verifiquen la siguiente expresión

$$\overline{k_i} = k_i + \sum_{j=i}^{n} s_{ij} \ , \quad i = 1, 2, \ldots, n.$$

Observaciones

1. Se insiste en que el método de Cholesky puede aplicarse si la matriz de los coeficientes es simétrica y definida positiva.

Por ejemplo, el siguiente sistema simétrico

$$\begin{cases} 1.423x_1 + 2.316x_2 + 3.218x_3 = 8.553 \\ 2.316x_1 + 3.751x_2 + 1.244x_3 = 7.342 \\ 3.218x_1 + 1.244x_2 + 6.173x_3 = 13.349 \end{cases}$$

no puede resolverse por el método de Cholesky, porque si bien la matriz de los coeficientes es simétrica, no es definida positiva; por ejemplo,

$$\begin{vmatrix} 1.423 & 2.316 \\ 2.316 & 3.751 \end{vmatrix} < 0 \, .$$

2. Como $A = S^t S$, entonces

$$\Delta = \det(A) = \det(S^t)\det(S) =$$
$$= (s_{11}s_{22}\ldots s_{nn})(s_{11}s_{22}\ldots s_{nn}) = s_{11}^2 s_{22}^2 \ldots s_{nn}^2.$$

Ejemplo 23. Usando el método de Cholesky, resolvamos el siguiente sistema

$$\begin{cases} x_1 + 2x_2 + 3x_3 + 4x_4 = -2 \\ 2x_1 + 5x_2 + x_3 + 10x_4 = -12 \\ 3x_1 + x_2 + 35x_3 + 5x_4 = 32 \\ 4x_1 + 10x_2 + 5x_3 + 45x_4 = -46 \end{cases}$$

x_1	x_2	x_3	x_4	B	S
1	2	3	4	-2	8
	5	1	10	-12	6
		35	5	32	76
			45	-46	18
1	2	3	4	-2	8
	1	-5	2	-8	-10
		1	3	-2	2
			4	-4	0
1	-1	1	-1	x_i	
2	0	2	0	$\overline{x_i}$	

Luego,

$$x_1 = 1, \quad x_2 = -1, \quad x_3 = 1, \quad x_4 = -1$$

y

$$\Delta = 1^2 \text{ x } 1^2 \text{ x } 1^2 \text{ x } 4^2 = 16.$$

Ejemplo 24. Usando el método de Cholesky, resolvamos el sistema que se da a continuación

$$\begin{cases} x_1 + 0.42x_2 + 0.54x_3 + 0.66x_4 = 0.3 \\ 0.42x_1 + x_2 + 0.32x_3 + 0.44x_4 = 0.5 \\ 0.54x_1 + 0.32x_2 + x_3 + 0.22x_4 = 0.7 \\ 0.66x_1 + 0.44x_2 + 0.22x_3 + x_4 = 0.9 \end{cases}$$

x_1	x_2	x_3	x_4	B	S
1	0.42 1	0.54 0.32 1	0.66 0.44 0.22 1	0.3 0.5 0.7 0.9	2.92 2.68 2.78 3.22
1	0.42 0.90752	0.54 0.10270 0.83538	0.66 0.17939 -0.18533 0.70560	0.3 0.41211 0.59335 1.04597	2.92 1.60173 1.24339 1.75156
-1.25777 -0.25776	0.04349 1.04350	1.03914 2.03913	1.48238 2.48237	$\frac{x_i}{x_i}$	

Además,

$$\Delta = 1^2 \text{ x } 0.90752^2 \text{ x } 0.83538^2 \text{ x } 0.70560^2 = 0.28615.$$

Veamos ahora cómo obtener la solución de un sistema de ecuaciones lineales por iteración.

3.2.4. Métodos iterativos.

La solución de un sistema de ecuaciones lineales por iteración es, en cierto modo, análogo a los métodos iterativos que vimos para obtener raíces de una ecuación. Comenzando con valores iniciales de las incógnitas generamos una sucesión de valores de las mismas.

Primero analizaremos el *método de Jacobi* (o *método de iteración global*) y luego veremos el *método de Gauss-Seidel* (o *método de iteración individual*).

3.2.4.1. Método de Jacobi.

Dado un sistema lineal

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases} \tag{38}$$

introduciendo las matrices

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix}, \quad B = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{pmatrix}$$

podemos escribir el sistema (38) en forma abreviada como ecuación matricial

$$A\,x = B. \tag{38'}$$

Suponiendo que los coeficientes diagonales

$$a_{ii} \neq 0\,, \quad i = 1, 2, ..., n$$

resolveremos la primera ecuación de (38) para x_1, la segunda para x_2, y así sucesivamente. Tendremos entonces el siguiente sistema equivalente

$$\begin{aligned} x_1 &= \frac{b_1}{a_{11}} - \left(\frac{a_{12}}{a_{11}} x_2 + \frac{a_{13}}{a_{11}} x_3 + \cdots + \frac{a_{1n}}{a_{11}} x_n \right) \\ x_2 &= \frac{b_2}{a_{22}} - \left(\frac{a_{21}}{a_{22}} x_1 + \frac{a_{23}}{a_{22}} x_3 + \cdots + \frac{a_{2n}}{a_{22}} x_n \right) \\ &\vdots \\ x_n &= \frac{b_n}{a_{nn}} - \left(\frac{a_{n1}}{a_{nn}} x_1 + \frac{a_{n2}}{a_{nn}} x_2 + \cdots + \frac{a_{n,n-1}}{a_{nn}} x_{n-1} \right) \end{aligned} \tag{39}$$

Introduciendo las matrices

$$C = \begin{pmatrix} \frac{b_1}{a_{11}} \\ \frac{b_2}{a_{22}} \\ \vdots \\ \frac{b_n}{a_{nn}} \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}, \quad M = \begin{pmatrix} 0 & \frac{-a_{12}}{a_{11}} & \frac{-a_{13}}{a_{11}} & \cdots & \frac{-a_{1n}}{a_{11}} \\ \frac{-a_{21}}{a_{22}} & 0 & \frac{-a_{23}}{a_{22}} & \cdots & \frac{-a_{2n}}{a_{22}} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \frac{-a_{n1}}{a_{nn}} & \frac{-a_{n2}}{a_{nn}} & \frac{-a_{n3}}{a_{nn}} & \cdots & 0 \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & \cdots & m_{1n} \\ m_{21} & m_{22} & m_{23} & \cdots & m_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ m_{n1} & m_{n2} & m_{n3} & \cdots & m_{nn} \end{pmatrix}$$

podemos escribir la ecuación (39) en forma matricial

$$x = C + M\,x. \tag{39'}$$

Resolveremos el sistema (39) por el método de iteración de punto fijo. Para la aproximación de orden cero tomaremos la columna de constantes $x^{(0)} = C$.

A continuación construiremos en forma consecutiva las matrices de columnas

$$x^{(1)} = C + Mx^{(0)} \ \text{(primera aproximación)}$$
$$x^{(2)} = C + Mx^{(1)} \ \text{(segunda aproximación)}$$

y así sucesivamente.

Hablando en términos generales, cualquier aproximación de orden (k+1) se calcula a partir de la fórmula

$$x^{(k+1)} = C + M\,x^{(k)}, \quad k = 0, 1, \ldots \tag{40}$$

Si la sucesión de aproximaciones $x^{(0)}, x^{(1)}, \cdots, x^{(k)}, \cdots$ tiene límite x, esto es

$$x = \lim_{k\to\infty} x^{(k)}$$

entonces este límite es la solución del sistema (39). Pasando al límite en (40), tenemos

$$\lim_{k\to\infty} x^{(k+1)} = C + M \lim_{k\to\infty} x^{(k)}$$

ó

$$x = C + M\,x$$

lo cual es decir que el vector límite x es la solución del sistema (39') y, en consecuencia, del sistema (38).

Escribimos por completo las fórmulas de las aproximaciones:

$$\begin{cases} x_i^{(0)} = c_i \\ x_i^{(k+1)} = c_i + \sum_{j=1}^{n} m_{ij} x_j^{(k)} \\ (m_{ii} = 0;\ i = 1, 2, \cdots, n;\ k = 0, 1, 2, \cdots) \end{cases} \tag{40'}$$

El método dado por (40) o por (40') se denomina *método de Jacobi*. El proceso de iteración (40) converge rápidamente, esto es, el número necesario de aproximaciones para obtener las raíces de (38) con una exactitud dada es menor, si los elementos de la matriz M son pequeños en valor absoluto. En otras palabras, para el mayor éxito del proceso de Jacobi se requiere que los módulos de los coeficientes diagonales del sistema (38) sean grandes en comparación con los módulos de los coeficientes no diagonales de este sistema (los términos constantes no tienen importancia alguna).

Daremos, sin demostración, una condición suficiente para la convergencia del método de Jacobi.

Teorema. Si al menos es válida una de las dos condiciones siguientes en el sistema reducido (39)

(a) $\sum_{\substack{j=1 \\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| < 1, \ i = 1, 2, ..., n$ (por filas)

ó

(b) $\sum_{\substack{i=1 \\ i\neq j}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| < 1, \ j = 1, 2, ..., n$ (por columnas)

el método de Jacobi (40) converge a una solución única del sistema, independientemente de la aproximación inicial elegida.

Observaciones

1. Cuando se utiliza el método de Jacobi no es necesario tomar para la aproximación de orden cero la columna de términos constantes C. La convergencia del proceso de Jacobi depende únicamente de las propiedades de la matriz M. Es aconsejable que las componentes del vector inicial tomen los valores aproximados de las raíces del sistema halladas tras un cálculo previo. Esto es, el vector inicial $x^{(0)}$ puede elegirse (dentro de ciertos límites) arbitrariamente. Además, como todo proceso iterativo convergente, este método es *autocorrectivo*, es decir, un error de cálculo individual no afectará el resultado final ya que cualquier aproximación errónea puede considerarse como un nuevo vector inicial.

2. La *condición de stop* es aquella donde se analiza la desigualdad

$$\left|x^{(k+1)} - x^{(k)}\right| \leq \varepsilon \qquad (41)$$

para $\varepsilon > 0$ pequeño dado.

(Observar que la condición anterior se debe verificar componente a componente, es decir que si

$$\left|x_1^{(k+1)} - x_1^{(k)}\right| \le \varepsilon \quad \text{y} \quad \left|x_2^{(k+1)} - x_2^{(k)}\right| \le \varepsilon \quad \text{y} \cdots \text{y} \quad \left|x_n^{(k+1)} - x_n^{(k)}\right| \le \varepsilon$$

entonces se finaliza el proceso iterativo porque se ha obtenido la solución dentro de la exactitud deseada).

3. Como ya se citó anteriormente, el mayor éxito del método de Jacobi se produce si los elementos de la diagonal principal del sistema dado son *preponderantes*; en otras palabras, si los módulos de los coeficientes diagonales del sistema dado son grandes en comparación con los módulos de los coeficientes no diagonales. Es por esto que dado un sistema lineal, debemos analizar en primer lugar si se verifica la condición suficiente de convergencia, pues si no se verificara debemos tratar de escribir el sistema dado en forma tal que sobre la diagonal predominen los valores mayores. Luego de esto, debemos analizar nuevamente si el sistema resultante satisface la condición suficiente de convergencia.

Ejemplo 25. Supongamos que debemos resolver el siguiente sistema de ecuaciones

$$\begin{cases} 2x_1 + 12x_2 + x_3 - 4x_4 = 13 \\ 2x_1 + x_2 - 3x_3 + 9x_4 = 31 \\ 13x_1 + 5x_2 - 3x_3 + x_4 = 18 \\ 3x_1 - 4x_2 + 10x_3 + x_4 = 29 \end{cases}$$

Es inmediato que la condición (b) no se verifica ni tampoco la condición (a), de modo que rescribiremos el sistema dado de forma tal que sobre la diagonal predominen los valores mayores (hacemos pivotación parcial):

$$\begin{cases} 13x_1 + 5x_2 - 3x_3 + x_4 = 18 \\ 2x_1 + 12x_2 + x_3 - 4x_4 = 13 \\ 3x_1 - 4x_2 + 10x_3 + x_4 = 29 \\ 2x_1 + x_2 - 3x_3 + 9x_4 = 31 \end{cases}$$

Este sistema satisface la condición (b), pues se tiene

para $j = 1$ $$\left|\frac{a_{21}}{a_{22}}\right| + \left|\frac{a_{31}}{a_{33}}\right| + \left|\frac{a_{41}}{a_{44}}\right| = \left|\frac{2}{12}\right| + \left|\frac{3}{10}\right| + \left|\frac{2}{9}\right| < 1$$

para $j = 2$ $$\left|\frac{a_{12}}{a_{11}}\right| + \left|\frac{a_{32}}{a_{33}}\right| + \left|\frac{a_{42}}{a_{44}}\right| = \left|\frac{5}{13}\right| + \left|\frac{-4}{10}\right| + \left|\frac{1}{9}\right| < 1$$

para $j = 3$ $$\left|\frac{a_{13}}{a_{11}}\right| + \left|\frac{a_{23}}{a_{22}}\right| + \left|\frac{a_{43}}{a_{44}}\right| = \left|\frac{-3}{13}\right| + \left|\frac{1}{12}\right| + \left|\frac{-3}{9}\right| < 1$$

para $j = 4$ $$\left|\frac{a_{14}}{a_{11}}\right| + \left|\frac{a_{24}}{a_{22}}\right| + \left|\frac{a_{34}}{a_{33}}\right| = \left|\frac{1}{13}\right| + \left|\frac{-4}{12}\right| + \left|\frac{1}{10}\right| < 1$$

Luego de este análisis podemos aplicar el método de Jacobi con la seguridad de que va a ser convergente.

Cuando debemos hacer los cálculos manualmente conviene tener presente el siguiente esquema práctico (con columna de control) que explicaremos con un ejemplo.

3.2.4.1.1. Esquema práctico.

Ejemplo 26. Utilizando el método de Jacobi, resolvamos el siguiente sistema de ecuaciones lineales

$$\begin{cases} 13x_1 + x_2 + 5x_3 = 30 \\ 3x_1 - 14x_2 + x_3 = -22 \\ -4x_1 - 2x_2 + 17x_3 = 43 \end{cases}$$

que verifica la condición suficiente de convergencia (b).

		A		*B*		
	13	1	5	30		
	3	-14	1	-22		
	-4	-2	17	43		
		M		*C*		
	0	-1/13	-5/13	30/13		
	3/14	0	1/14	22/14		
	4/17	2/17	0	43/17		
S	1391/3094	126/3094	-969/3094	19828/3094	σ	$\sum$
$x^{(0)}$	30/13	22/14	43/17			
$x^{(1)}$	1.2141	2.2465	3.2572		6.7178	6.7178
$x^{(2)}$	0.8822	2.0641	3.0792		6.0255	6.0256
$x^{(3)}$	0.9647	1.9803	2.9797		5.9247	5.9247
$x^{(4)}$	1.0094	1.9909	2.9893		5.9896	5.9896
$\vdots$	$\vdots$	$\vdots$	$\vdots$			
$\bar{x}$	1	2	3			

Las matrices *M* y *C* son las obtenidas según (39). En *S* ponemos la suma de las columnas de las matrices *M* y *C*. Tomamos como aproximación inicial $x^{(0)}=C$; esto es

$$x_1^{(0)} = \frac{30}{13}, \quad x_2^{(0)} = \frac{22}{14}, \quad x_3^{(0)} = \frac{43}{17}.$$

Realizamos los cálculos con cuatro cifras decimales.

Para obtener las restantes aproximaciones utilizamos las fórmulas (40) ó (40'). Por ejemplo, para obtener el valor 1.2141, hacemos

$$1.2141 = \frac{30}{13} + 0\frac{30}{13} + \left(-\frac{1}{13}\right)\frac{22}{14} + \left(\frac{-5}{13}\right)\frac{43}{17} = 2.3077 - 0.0769\,\text{x}\,1.5714 - 0.3846\,\text{x}\,2.5294,$$

para obtener el valor 2.2465, hacemos

$$2.2465 = \frac{22}{14} + \frac{3}{14}\,\frac{30}{13} + 0\,\frac{22}{14} + \frac{1}{14}\,\frac{43}{17} = 1.5714 + 0.2143\,\text{x}\,2.3047 + 0.0714\,\text{x}\,2.5294,$$

para obtener el valor 0.8822, hacemos (usamos ahora los nuevos valores iterados $x^{(1)}$)

$$0.8822 = \frac{30}{13} + 0(1.2141) + \left(-\frac{1}{13}\right)2.2465 + \left(\frac{-5}{13}\right)3.2572 = 2.3077 - 0.0769 \text{ x } 2.2465 - 0.3846 \text{ x } 3.2572$$

y así siguiendo hasta obtener que $\left|x^{(k+1)} - x^{(k)}\right| \leq \varepsilon$, para $\varepsilon > 0$ prefijado (cota de error admisible).

Las dos columnas anexadas a la tabla a la derecha, σ y $\sum$, son utilizadas para controlar los cálculos cuando éstos son realizados manualmente. σ es igual a la suma de los elementos de cada fila.

Por ejemplo,

$$6.7178 = 1.2141 + 2.2465 + 3.2572.$$

Los valores de la columna σ deben coincidir con los valores de la columna $\sum$ que se obtienen realizando la siguiente operación: la suma de los elementos de la matriz columna *C*, más la suma de los productos obtenidos multiplicando cada elemento de *S* por el valor iterado anterior correspondiente. Por ejemplo,

$$6.7178 = \frac{19828}{3094} + \frac{1391}{3094}\frac{30}{13} + \frac{126}{3094}\frac{22}{14} + \left(-\frac{969}{3094}\right)\frac{43}{17} =$$
$$= 6.4085 + 0.4496 \text{ x } 2.3077 + 0.0407 \text{x } 1.5714 - 0.3132 \text{ x } 2.5294$$

$$6.0256 = \frac{19828}{3094} + \frac{1391}{3094}1.2141 + \frac{126}{3094}2.2465 + \left(-\frac{969}{3094}\right)3.2572 =$$
$$= 6.4085 + 0.4496\text{x}1.2141 + 0.0407\text{x}2.2465 - 0.3132\text{x}3.2572$$

y así siguiendo.

Debemos observar que $\sigma = \sum$ suponiendo que no hay errores de redondeo.

Este procedimiento se puede acelerar si a medida que se "corrige" el valor de cada incógnita se utilizan los valores mejorados logrados para aproximar las demás incógnitas, aunque sean de la misma etapa. Surge así el *método de Gauss - Seidel*.

3.2.4.2. Método de Gauss – Seidel.

Su fundamento estriba en que al calcular la aproximación (k+1) de la incógnita x_i se tienen en cuenta las (k+1) aproximaciones anteriores de las incógnitas $x_1, x_2, \ldots, x_{i-1}$.

Supongamos el sistema lineal reducido

$$x_i = c_i + \sum_{j=1}^{n} m_{ij} x_j, \quad i = 1, 2, ..., n \tag{42}$$

Se eligen arbitrariamente las aproximaciones iniciales de las raíces $x_1^{(0)}, x_2^{(0)}, \cdots, x_n^{(0)}$ tratando, por supuesto, de que correspondan de manera aproximada a las incógnitas deseadas: $x_1, x_2, \cdots, x_n$.

Suponiendo ahora que sean conocidas las k aproximaciones $x_i^{(k)}$ de las raíces, formemos de acuerdo con el procedimiento de Seidel las (k+1) aproximaciones de las raíces usando las siguientes fórmulas

$$\begin{aligned}
x_1^{(k+1)} &= c_1 + \sum_{j=1}^{n} m_{1j} x_j^{(k)} \\
x_2^{(k+1)} &= c_2 + m_{21} x_1^{(k+1)} + \sum_{j=2}^{n} m_{2j} x_j^{(k)} \\
&\vdots \qquad\qquad (k = 0, 1, ...) \\
x_i^{(k+1)} &= c_i + \sum_{j=1}^{i-1} m_{ij} x_j^{(k+1)} + \sum_{j=i}^{n} m_{ij} x_j^{(k)} \\
&\vdots \\
x_n^{(k+1)} &= c_n + \sum_{j=1}^{n-1} m_{nj} x_j^{(k+1)} + m_{nn} x_n^{(k)}
\end{aligned} \tag{43}$$

El teorema de convergencia dado anteriormente para el método de Jacobi permanece válido para la iteración por el método de Gauss - Seidel (fórmulas (43)).

En general, el método de Gauss - Seidel ofrece una mejor convergencia que el método de Jacobi y lo que es más, el proceso de Seidel puede incluso converger cuando el proceso de Jacobi diverge. Esto no siempre ocurre, por supuesto. Estos casos son posibles cuando el procedimiento de Gauss - Seidel converge mas lentamente que el de Jacobi.

Ocurre también a veces que el proceso de Jacobi converge cuando el proceso de Seidel diverge.

3.2.4.2.1. Esquema práctico.

Ejemplo 27. Utilizando el método de Gauss – Seidel, resolvamos el sistema dado en el ejemplo 26, que satisface la condición suficiente de convergencia (b).

	A			*B*		
	13	1	5	30		
	3	-14	1	-22		
	-4	-2	17	43		
	M			*C*		
	0	-1/13	-5/13	30/13		
	3/14	0	1/14	22/14		
	4/17	2/17	0	43/17		
S	1391/3094	126/3094	-969/3094	19828/3094	σ	$\sum$
$x^{(0)}$	30/13	22/14	43/17			
$x^{(1)}$	1.2141	2.0122	3.0517		6.2780	6.2440
$x^{(2)}$	0.9793	1.9992	2.9949		5.9734	5.9744
$x^{(3)}$	1.0021	2.0000	3.0004		6.0025	6.0024
$\vdots$	$\vdots$	$\vdots$	$\vdots$			
$\vec{x}$	1	2	3			

M, *C*, *S*, $x^{(0)}$ se obtienen y se disponen en el esquema práctico de la misma forma que en el método de Jacobi.

Según las fórmulas (43) con $x^{(0)}$, obtenemos

$$x_1^{(1)} = 1.2141 = \frac{30}{13} + 0\frac{30}{13} + \left(-\frac{1}{13}\right)\frac{22}{14} + \left(-\frac{5}{13}\right)\frac{43}{17}$$

para obtener $x_2^{(1)}$, hacemos

$$x_2^{(1)} = 2.0122 = \frac{22}{14} + 1.2141\frac{3}{14} + 0\frac{22}{14} + \frac{1}{14}\frac{43}{17}$$

y para obtener $x_3^{(1)}$, hacemos

$$x_3^{(1)} = 3.0517 = \frac{43}{17} + 1.2141\frac{4}{17} + 2.0122\frac{2}{17} + 0\frac{43}{17}$$

y así siguiendo, obtenemos las demás aproximaciones.

En σ ponemos la suma de los elementos de cada renglón. Por ejemplo, para la iteración 1, obtenemos

$$6.2780 = 1.2141 + 2.0122 + 3.0517.$$

Para calcular los valores que van en la columna $\sum$ hacemos, por ejemplo,

$$6.2440 = 1.2141\frac{1391}{3094} + 2.0122\frac{126}{3094} + \frac{43}{17}\left(-\frac{969}{3094}\right) + \frac{19828}{3094}$$

$$5.9744 = 0.9793\frac{1391}{3094} + 1.9992\frac{126}{3094} + 3.0517\left(-\frac{969}{3094}\right) + \frac{19828}{3094}$$

y así siguiendo.

Se observa que estos valores difieren de los correspondientes en la columna σ, hecho que surge debido a los errores de redondeo y no a errores de cálculo.

El proceso iterativo se detiene cuando para $\varepsilon > 0$ dado suficientemente pequeño se verifica que

$$\left| x^{(k+1)} - x^{(k)} \right| \leq \varepsilon .$$

En este ejemplo vemos que con sólo tres iteraciones podemos decir que

$$x_1 \to 1, \quad x_2 \to 2, \quad x_3 \to 3$$

que es la solución del sistema dado. Con el mismo número de iteraciones usando el método de Jacobi no era tan fácil de predecir.

Observaciones

1. En el caso de que la matriz *A* de los coeficientes del sistema *Ax = B* es simétrica y definida positiva, existe además otra importante condición suficiente de convergencia del método de Gauss – Seidel.

"Si la matriz A es simétrica y definida positiva, el método de Gauss-Seidel para el sistema $Ax = B$ converge para un valor inicial $x^{(0)}$ arbitrario".

Este tipo de matrices aparece mucho en la práctica: en problemas de mínimos cuadrados (sistemas normales) y, en general, en determinados problemas de mecánica.

Ejemplo 28. Consideremos, por ejemplo, el siguiente sistema

$$\begin{cases} 8.467x_1 + 5.137x_2 + 3.141x_3 + 2.063x_4 = 29.912 \\ 5.137x_1 + 6.421x_2 + 2.617x_3 + 2.003x_4 = 25.058 \\ 3.141x_1 + 2.617x_2 + 4.128x_3 + 1.628x_4 = 16.557 \\ 2.063x_1 + 2.003x_2 + 1.628x_3 + 3.446x_4 = 12.690 \end{cases}$$

Se puede constatar que no se cumple la condición (b) pues, por ejemplo

$$\text{para } j = 1 \quad \left|\frac{a_{21}}{a_{22}}\right| + \left|\frac{a_{31}}{a_{33}}\right| + \left|\frac{a_{41}}{a_{44}}\right| = \left|\frac{5.137}{6.421}\right| + \left|\frac{3.141}{4.128}\right| + \left|\frac{2.063}{3.446}\right| > 1$$

ni tampoco se cumple la condición (a) pues, por ejemplo

$$\text{para } i = 1 \quad \left|\frac{a_{12}}{a_{11}}\right| + \left|\frac{a_{13}}{a_{11}}\right| + \left|\frac{a_{14}}{a_{11}}\right| = \left|\frac{5.137}{8.467}\right| + \left|\frac{3.141}{8.467}\right| + \left|\frac{2.063}{8.467}\right| > 1$$

No obstante, debido a que la matriz formada por los coeficientes del sistema es simétrica y definida positiva, el procedimiento de Gauss - Seidel es convergente. En efecto, tomemos como aproximación inicial o de orden cero al vector nulo, es decir, $x^{(0)} = (0,0,0,0)^t$.

Usando las fórmulas (43) obtenemos

$x_1^{(1)} = 3.527743;\ x_2^{(1)} = 1.0761789;\ x_3^{(1)} = 0.6405519;\ x_4^{(1)} = 0.6394323$

$x_1^{(2)} = 2.4864240;\ x_2^{(2)} = 1.4527539;\ x_3^{(2)} = 0.9458066;\ x_4^{(2)} = 0.9027476$

$x_1^{(3)} = 2.0805547;\ x_2^{(3)} = 1.5709097;\ x_3^{(3)} = 1.0758803;\ x_4^{(3)} = 1.015579$

$x_1^{(4)} = 1.9331190;\ x_2^{(4)} = 1.6006458;\ x_3^{(4)} = 1.1247069;\ x_4^{(4)} = 1.0635111$

$\vdots$

$x_1^{(20)} = 1.8741337;\ x_2^{(20)} = 1.5969776;\ x_3^{(20)} = 1.1413522;\ x_4^{(20)} = 1.0930918$

que es la solución del sistema dado.

2. Si bien no es simple decir cuando una matriz dada es simétrica y definida positiva, sí resulta fácil transformar el sistema $Ax = B$ para que la condición sea satisfecha.

Es suficiente multiplicar a ambos miembros de $Ax = B$ por A^t por izquierda, obteniendo así el sistema

$$A^t Ax = A^t B$$

y como sabemos del Álgebra Lineal, la matriz $A^t A$ es simétrica y definida positiva.

Esta operación se llama *normalización* y permite aplicar el método de Gauss-Seidel a un sistema cualquiera. (En la observación 4 veremos numéricamente el proceso de normalización).

3. Al igual que en el método de Jacobi, en el método de Gauss - Seidel no es necesario tomar como aproximación de orden cero la columna C de términos constantes. Suele ser común también tomar como aproximación de orden cero el vector nulo, o sea, $x^{(0)} = (0,0,...,0)^t$. En el ejemplo 28 dado en la observación 1 trabajamos con dicha aproximación inicial.

Ejemplo 29. Resolvamos el sistema dado en el ejemplo 26, pero ahora tomando $x^{(0)} = (0, 0, 0)^t$.

	A			B		
	13	1	5	30		
	3	-14	1	-22		
	-4	-2	17	43		
	M			C		
	0	-1/13	-5/13	30/13		
	3/14	0	1/14	22/14		
	4/17	2/17	0	43/17		
S	1391/3094	26/3094	-969/3094	19828/3094	σ	$\sum$
$x^{(0)}$	0	0	0			
$x^{(1)}$	30/13	2.0659	3.3154		7.6890	7.5302
$x^{(2)}$	0.8737	1.9954	2.9696		5.8387	5.8441
$x^{(3)}$	1.0211	2.0023	3.0051		6.0285	6.0190
$x^{(4)}$	0.9980	1.9998	2.9994		5.9972	5.9974
$\vdots$	$\vdots$	$\vdots$	$\vdots$			
$\vec{x}$	1	2	3			

4. Como hemos dicho anteriormente, las condiciones de convergencia del procedimiento de Gauss - Seidel (al igual que en el método de Jacobi) son sólo suficientes, como se muestra en el siguiente ejemplo.

Ejemplo 30. Analicemos el siguiente sistema

$$\begin{cases} 3x_1 + 2x_2 = 7 \\ 4x_1 + x_2 = 6 \end{cases}$$

que sabemos tiene las raíces $x_1 = 1,\ x_2 = 2$.

Es inmediato que no se verifican ni las condiciones (a) y (b) del teorema antes enunciado, ni la matriz formada con los coeficientes del sistema es simétrica y definida positiva.

Permutemos las dos ecuaciones de manera tal que a_{11} sea preponderante

$$\begin{cases} 4x_1 + x_2 = 6 \\ 3x_1 + 2x_2 = 7 \end{cases}$$

Siguen sin verificarse las tres condiciones suficientes de convergencia, pero aplicando el método de Gauss - Seidel se pueden obtener las soluciones:

	A		*B*		
	4	1	6		
	3	2	7		
	M		*C*		
	0	-1/4	3/2		
	-3/2	0	7/2		
S	-3/2	-1/4	10/2	σ	$\sum$
$x^{(0)}$	0	0			
$x^{(1)}$	3/2	5/4		11/4	11/4
$x^{(2)}$	19/16	55/32		93/32	93/32
$x^{(3)}$	137/128	485/256		759/256	759/256
$x^{(4)}$	1051/1024	4015/2048		6117/2048	6117/2048
$\vdots$	$\vdots$	$\vdots$			
$\vec{x}$	1	2			

Luego, el método iterativo de Gauss - Seidel es convergente (aunque no se verificaban ninguna de las tres condiciones de convergencia).

Si resolvemos el sistema tal como está dado al principio, el proceso iterativo de Seidel es divergente:

	A		B		
	3	2	7		
	4	1	6		
	M		C		
	0	-2/3	7/3		
	-4	0	6		
S	-4	-2/3	25/3	σ	$\sum$
$x^{(0)}$	0	0			
$x^{(1)}$	7/3	-10/3		-1	-1
$x^{(2)}$	41/9	-110/9		-69/9	-69/9
$x^{(3)}$	283/27	-970/27		-687/27	-687/27
$\vdots$	$\vdots$	$\vdots$			

A medida que k crece, $x^{(k)}$ también crece indefinidamente sin converger a ningún valor.

Ahora bien, si para este caso que diverge usamos la observación 2, entonces obtendremos una matriz simétrica y definida positiva y, por lo tanto, el método de Gauss - Seidel será convergente (aunque no se verifiquen las condiciones suficientes de convergencia (a) y (b)).

Entonces, dado el sistema original

$$\begin{cases} 3x_1 + 2x_2 = 7 \\ 4x_1 + x_2 = 6 \end{cases}$$

hacemos $A^t Ax = A^t B$, donde

$$A = \begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix}, \qquad A^t = \begin{pmatrix} 3 & 4 \\ 2 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 7 \\ 6 \end{pmatrix}, \qquad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Obtenemos

$$A^t A = \begin{pmatrix} 3 & 4 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix} = \begin{pmatrix} 25 & 10 \\ 10 & 5 \end{pmatrix} \qquad A^t B = \begin{pmatrix} 3 & 4 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} 7 \\ 6 \end{pmatrix} = \begin{pmatrix} 45 \\ 20 \end{pmatrix}$$

de donde, resulta

$$\begin{pmatrix} 25 & 10 \\ 10 & 5 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 45 \\ 20 \end{pmatrix}.$$

Luego, tenemos el sistema

$$\begin{cases} 25x_1 + 10x_2 = 45 \\ 10x_1 + 5x_2 = 20 \end{cases}$$

La matriz formada con los coeficientes de este sistema es simétrica y definida positiva. Resolvemos entonces este sistema por el método de Gauss - Seidel:

	A		*B*		
	25	10	45		
	10	5	20		
	M		*C*		
	0	-2/5	9/5		
	-2	0	4		
S	-2	-2/5	29/5	σ	$\sum$
$x^{(0)}$	0	0			
$x^{(1)}$	9/5	2/5		11/5	11/5
$x^{(2)}$	41/25	18/25		59/25	59/25
$x^{(3)}$	9/25	82/25		91/25	91/25
$x^{(4)}$	61/125	378/125		439/125	439/125
$x^{(5)}$	369/625	1762/625		2131/625	2131/625
$x^{(6)}$	2101/3125	8298/3125		10399/3125	10399/3125
$x^{(7)}$	11529/15625	39442/15625		50971/15625	50971/15625
$x^{(8)}$	61741/78125	189018/78125		250759/78125	250759/78125
$x^{(9)}$	325089/390625	912322/390625		1237411/390625	1237411/390625
$\vdots$	$\vdots$	$\vdots$			
$\vec{x}$	1	2			

5. Dado que el procedimiento de Gauss - Seidel es de iteración, si se produce un error en la determinación de alguna incógnita éste no invalida el cálculo, pues en los pasos subsiguientes se corrigen automáticamente los resultados. Sin embargo, la rapidez de convergencia depende en gran parte de la buena elección de la primera aproximación.

6. Como todo proceso iterativo, en los programas correspondientes a los métodos de Jacobi y Gauss - Seidel se debe ingresar la cantidad máxima de iteraciones que deseamos se repitan dichos procedimientos como así también, la cota de error admisible en las aproximaciones.

3.3. Elementos de juicio.

En el siguiente cuadro se muestra un resumen de los elementos de juicio implicados en la solución de sistemas de ecuaciones lineales.

Comparación de las características de los diferentes métodos para encontrar la solución de sistemas de ecuaciones lineales

Método	N° máximo aproxima-do de ecuacio-nes*	Estabilidad	Precisión	Alcance de las aplicaciones	Esfuerzo de progra-mación	Comentarios
Regla de Cramer	3	---	Se ve afectado por errores de redondeo.	Limitado	---	Requiere de un gran esfuerzo computacional para sistemas de más de 3 ecuaciones.
Eliminación Gaussiana (con pivoteo parcial)	40	---	Se ve afectado por errores de redondeo.	General	Moderado	El algoritmo de solución más básico.
Gauss-Jordan normalizado (con pivoteo parcial)	40	---	Se ve afectado por errores de redondeo.	General	Moderado	Permite calcular la matriz inversa.

Crout (con pivoteo parcial)	40	---	Se ve afectado por errores de redondeo en menor grado que los tres métodos anteriores.	General	Moderado	De uso frecuente en los programas, pues puede economizarse espacio para almacena-miento. Permite calcular la matriz inversa.
Cholesky	40	---	Se ve afectado por errores de redondeo en menor grado que los tres primeros métodos.	Apropiado sólo para matrices simétricas y definidas positivas.	Moderado	De uso frecuente en los programas, pues puede economizarse espacio para almacena-miento.
Jacobi	1000	Puede no converger si no hay dominancia diagonal.	Excelente	Apropiado sólo para sistemas dominantes diagonal-mente	Fácil	---
Gauss-Seidel	1000	Puede no converger si no hay dominancia diagonal o si la matriz formada con los coeficientes del sistema no es simétrica y definida positiva.	Excelente	Apropiado sólo para sistemas dominantes diagonalment e y para matrices simétricas y definidas positivas.	Fácil	Más rápido que el método de Jacobi.

***** El límite superior depende de la computadora y del grado de dispersión de las ecuaciones. No se toma en cuenta el condicionamiento de los sistemas.

3.4. Algoritmo de eliminación Gaussiana con sustitución hacia atrás. Pseudocódigo.

Para resolver el sistema lineal de $n \times n$

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n = a_{i,n+1}, \quad i = 1, 2, \ldots, n:$$

ENTRADA número de incógnitas y de ecuaciones n; matriz aumentada $A = (a_{ij})$ donde $1 \le i \le n$ y $1 \le j \le n + 1$.
SALIDA solución $x_1, x_2, \ldots, x_n$ o mensaje que el sistema no tiene solución única.
Paso 1 Para $i = 1, \ldots, n - 1$ seguir los Pasos 2-4. (Proceso de eliminación).
Paso 2 Sea p el menor entero con $i \le p \le n$ y $a_{pi} \neq 0$.
Si p no puede encontrarse entonces
SALIDA ("No existe solución única");
PARAR.
Paso 3 Si $p \neq i$ entonces realizar
para $k = i, \ldots, n + 1$
tomar $z = a_{pk}$;
$a_{pk} = a_{ik}$;
$a_{ik} = z$.
Paso 4 Para $j = i + 1, \ldots, n$ seguir los Pasos 5 y 6.
Paso 5 Tomar $m_{ji} = a_{ji} / a_{ii}$.
Paso 6 Para $k = i + 1, \ldots, n + 1$
tomar $a_{jk} = a_{jk} - m_{ji}\, a_{ik}$.
Paso 7 Si $a_{nn} = 0$ entonces SALIDA ("No existe solución única");
PARAR.
Paso 8 Tomar $x_n = a_{n,n+1} / a_{nn}$. (Empieza la sustitución hacia atrás).
Paso 9 Para $i = n - 1, \ldots, 1$

$$\text{tomar } x_i = [a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}x_j] / a_{ii}\,.$$

Paso 10 SALIDA ($x_1, \ldots, x_n$); (Procedimiento completado satisfactoria-mente).
PARAR.

3.5. Sistemas lineales sobredeterminados.

Vamos a considerar ahora el problema de resolver un sistema de ecuaciones lineales

$$A\vec{x} = \vec{b} \tag{44}$$

donde A es una matriz de orden m x n ($m > n$), $\vec{b}$ es un m-vector dado y $\vec{x}$ es un n-vector desconocido.

Ejemplo 31. Para ilustrar las ideas usaremos el siguiente sistema:

$$\begin{pmatrix} 2 & 2 \\ 3 & 4.5 \\ 4 & 8 \\ 6 & 18 \\ 9 & 40.5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 11.86 \\ 19.34 \\ 27.84 \\ 47.94 \\ 85.82 \end{pmatrix}. \tag{45}$$

Puesto que hay mas ecuaciones que incógnitas es imposible, en general, hallar un vector $\vec{x}$ que satisfaga exactamente a todas las ecuaciones. Los sistemas de este tipo son conocidos como *sistemas incompatibles* o también son llamados *sistemas sobredeterminados*. Los sistemas sobredeterminados surgen en el trabajo experimental o computacional cada vez que se generan más resultados que los que se requerirían si fuera alcanzable la precisión. En otras palabras, con el objeto de reducir la influencia de los errores que resultan de mediciones uno usualmente toma un número más grande de mediciones que el número de incógnitas.

Nuestro primer objetivo, entonces, será definir qué debemos entender por una "solución" y después de hacer esto examinaremos el problema de computarla.

3.5.1. Solución por mínimos cuadrados.

Primero definimos el *vector residual*

$$\vec{r}_{\vec{x}} = \vec{r}(\vec{x}) = \vec{b} - A\vec{x}. \tag{46}$$

Luego definimos como la solución del sistema sobredeterminado (44), al vector $\vec{x}^*$ que minimiza la norma l_2 ó euclídea del vector residual; esto es

$$\left\| \vec{r}(\vec{x}^*) \right\| = \min_{\vec{x}} \left\| \vec{r}(\vec{x}) \right\| = \min_{\vec{x}} \left[\sum_{i=1}^{m} \left| r_i(\vec{x}) \right|^2 \right]^{\frac{1}{2}} \tag{47}$$

El vector $\vec{x}^*$ se dice que es *la solución por mínimos cuadrados del sistema* (44). El término mínimos cuadrados viene del hecho que la norma de un vector es la raíz cuadrada de la suma de los cuadrados de sus componentes ($\left\| \vec{a} \right\| = \left(\vec{a}^t \vec{a} \right)^{\frac{1}{2}} = \left(\sum_{i=1}^{n} a_i^2 \right)^{\frac{1}{2}}$, siendo $\vec{a} = (a_1, a_2, \cdots, a_n)^t$).

Es decir que la solución por mínimos cuadrados de un sistema de ecuaciones sobredeterminado $A\vec{x} = \vec{b}$ es un vector $\vec{x}^*$ que hace que la raíz

cuadrada de la suma de los cuadrados de las componentes del vector residual sea mínima.

Debemos observar que la definición de $\vec{x}^*$ en la ecuación (47) depende de la norma particular del vector utilizada. Ocurre que nosotros estamos usando la norma euclídea o norma l_2 que produce la solución por mínimos cuadrados, pero hay otras normas que a veces son usadas en algunas aplicaciones. Así, por ejemplo, si la norma subyacente es l_∞ o la *norma uniforme*

$$\|\vec{r}\|_\infty = \max_{1\le i\le m} |r_i|$$

entonces $\vec{x}^*$ se llama *solución minimax*, que satisface la relación

$$\|\vec{r}(\vec{x}^*)\|_\infty = \min_{\vec{x}} \max_{1\le i\le m} |r_i(\vec{x})| \ ;$$

otra posibilidad es la norma l_1

$$\|\vec{r}\|_1 = \sum_{i=1}^{m} |r_i| \ .$$

Cada una de estas definiciones conducen a diferentes métodos para computar a $\vec{x}^*$. Como una regla general, la solución por mínimos cuadrados es uno de los métodos más fáciles (y más baratos) para computar $\vec{x}^*$, razón por lo cual es uno de los más comúnmente usado. Es, por otra parte, el único método que analizaremos aquí. Digamos, no obstante, que existen muy buenas subrutinas ampliamente disponibles para los otros dos casos.

Volviendo a la solución por mínimos cuadrados: la solución $\vec{x}^*$ está caracterizada por el siguiente resultado.

Sea A una matriz real m x n ($m > n$) dada y $\vec{b}$ un m-vector dado. Entonces, si $\vec{x}$ satisface

$$A^t(\vec{b} - A\vec{x}^*) = \vec{0} \tag{48}$$

tenemos, para cualquier vector $\vec{y}$, la siguiente desigualdad

$$\|\vec{b} - A\vec{x}^*\| \le \|\vec{b} - A\vec{y}\|. \tag{49}$$

En efecto, si ponemos $\vec{r}_{\vec{x}^*} = \vec{b} - A\vec{x}^*$ y $\vec{r}_{\vec{y}} = \vec{b} - A\vec{y}$, entonces $\vec{r}_{\vec{y}} = (\vec{b} - A\vec{x}^*) + (A\vec{x}^* - A\vec{y}) = \vec{r}_{\vec{x}^*} + A(\vec{x}^* - \vec{y})$, de donde, $\vec{r}_{\vec{y}}^{\,t} = \vec{r}_{\vec{x}^*}^{\,t} + \left[A\left(\vec{x}^* - \vec{y}\right)\right]^t$.

Usando esto y la ecuación (48) que puede ser escrita como $A^t\vec{r}_{\vec{x}^*} = \vec{0}$, obtenemos

$$\begin{aligned}\vec{r}_{\vec{y}}^{\,t}\vec{r}_{\vec{y}} &= \left[\vec{r}_{\vec{x}^*}^{\,t} + \left[A\left(\vec{x}^* - \vec{y}\right)\right]^t\right]\ \left[\vec{r}_{\vec{x}^*} + A\left(\vec{x}^* - \vec{y}\right)\right] = \\ &= \vec{r}_{\vec{x}^*}^{\,t}\vec{r}_{\vec{x}^*} + \left[A\left(\vec{x}^* - \vec{y}\right)\right]^t A\left(\vec{x}^* - \vec{y}\right) + \vec{r}_{\vec{x}^*}^{\,t} A\left(\vec{x}^* - \vec{y}\right) + \left[A\left(\vec{x}^* - \vec{y}\right)\right]^t \vec{r}_{\vec{x}^*} = \\ &= \vec{r}_{\vec{x}^*}^{\,t}\vec{r}_{\vec{x}^*} + \left[A\left(\vec{x}^* - \vec{y}\right)\right]^t A\left(\vec{x}^* - \vec{y}\right) + 0\end{aligned}$$

y así, $\left\|\vec{r}_{\vec{y}}\right\|^2 = \left\|\vec{r}_{\vec{x}^*}\right\|^2 + \left\|A\left(\vec{x}^* - \vec{y}\right)\right\|^2 \geq \left\|\vec{r}_{\vec{x}^*}\right\|^2$, lo que prueba la desigualdad (49).

De la ecuación (48) se sigue que la solución por mínimos cuadrados $\vec{x}^*$ satisface

$$A^t A\vec{x}^* = A^t\vec{b}. \qquad (50)$$

Ahora A^tA es una matriz simétrica *n* x *n* y la ecuación (50) nos da *n* ecuaciones lineales con *n* incógnitas. Las ecuaciones del sistema (50) se llaman *ecuaciones normales*.

Observemos que A^t no puede ser cancelada de ambos miembros de la ecuación (50), por cuanto esto sería lo mismo que multiplicar por $\left(A^t\right)^{-1}$ que no está definida ya que *A* es rectangular. Las ecuaciones normales (50) se pueden resolver por cualquiera de los métodos analizados para tal fin (siempre que A^tA sea una matriz no singular).

Observemos que si *A* es de rango pleno, esto es, si *r*(*A*) = *n*, entonces $\vec{x}^*$ es única. Esto es evidente de la parte final de la demostración del resultado que caracteriza la solución porque: $A\left(\vec{x}^* - \vec{y}\right) = \vec{0}$ si y sólo si $\vec{x}^* - \vec{y} = \vec{0}$, esto es, si y sólo si $\vec{x}^* = \vec{y}$.

Además, si *r*(*A*) = *n*, entonces A^tA es no singular y las ecuaciones normales tienen solución única.

Por otra parte, si *r*(*A*) < *n*, entonces las ecuaciones normales tendrán un número infinito de soluciones, cada una de las cuales minimiza el residual.

El algoritmo para computar la solución por mínimos cuadrados es el siguiente (implementación directa de (50)):

i) Formamos A^tA y $A^t\vec{b}$.

ii) Resolvemos el sistema $A^tA\vec{x}^* = A^t\vec{b}$.

Para el sistema (45) dado en el ejemplo 31, las ecuaciones normales son

$$\begin{pmatrix} 146 & 522 \\ 522 & 2052.5 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1253.12 \\ 4672.1 \end{pmatrix}$$

cuya solución es

$$\vec{x}^* = \begin{pmatrix} 4.9 \\ 1.03 \end{pmatrix}$$

con los residuales

$$\vec{r}_{\vec{x}^*} = \begin{pmatrix} 0 \\ 5\text{x}10^{-3} \\ 0 \\ 0 \\ 5\text{x}10^{-3} \end{pmatrix}$$

Como una medida de cuán bien $\vec{x}^*$ satisface al sistema usaremos *el error de la raíz cuadrática media* que se define

$$ERMS = \left[\frac{1}{m}\left(\vec{r}^{\,t}_{\vec{x}^*}\ \vec{r}_{\vec{x}^*}\right)\right]^{\frac{1}{2}} = 3.1622776\text{x}10^{-3}$$

que puede ser interpretado como el error "promedio" de cada ecuación por mínimos cuadrados.

Si ocurriera que el valor de *ERMS* no fuera particularmente pequeño, de acuerdo a la teoría debemos interpretar que cualquier otro $\vec{x}$ tendrá un valor mayor para *ERMS*.

3.6. Sistemas de ecuaciones no lineales.

El método de Newton desarrollado en el Capítulo anterior se puede extender para obtener la solución de las ecuaciones no lineales simultáneas, y es quizás el método más conocido para tal fin.

Consideremos de una forma general un sistema no lineal de ecuaciones

$$\begin{aligned} f_1(x_1, x_2, ..., x_n) &= 0 \\ f_2(x_1, x_2, ..., x_n) &= 0 \\ &\vdots \\ f_n(x_1, x_2, ..., x_n) &= 0 \end{aligned} \tag{51}$$

con los primeros miembros reales.

Escribimos el sistema (51) en forma mas reducida considerando el conjunto de argumentos $x_1, x_2, \cdots, x_n$ como un vector *n*-dimensional

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Anágolamente, el conjunto de funciones $f_1, f_2, \cdots, f_n$ es también un vector *n*-dimensional (función vector)

$$\vec{f} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix}.$$

El sistema (51) puede, por consiguiente, escribirse abreviadamente

$$\vec{f}(\vec{x}) = \vec{0}. \tag{51'}$$

Ejemplo 32. Los siguientes son sistemas de ecuaciones no lineales:

$$\vec{f}(\vec{x}) = \begin{pmatrix} f_1(\vec{x}) \\ f_2(\vec{x}) \end{pmatrix} = \begin{pmatrix} 4x_1^2 + 9x_2^2 - 16x_1 - 54x_2 + 61 \\ x_1 x_2 - 2x_1 - 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \vec{0}$$

$$\vec{f}(\vec{x}) = \begin{pmatrix} f_1(\vec{x}) \\ f_2(\vec{x}) \\ f_3(\vec{x}) \end{pmatrix} = \begin{pmatrix} x_2 x_3^2 + x_3 e^{x_2} sen(x_1 x_3) \\ x_1 x_3^2 - e^{x_2} \cos(x_1 x_3) \\ 2x_1 x_2 x_3 + x_1 e^{x_2} sen(x_1 x_3) \end{pmatrix} = \vec{0}.$$

Consideremos el caso particular de dos ecuaciones con dos incógnitas. Sean x_n, y_n las raíces aproximadas del sistema

$$\begin{aligned} f(x,y) &= 0 \\ g(x,y) &= 0 \end{aligned} \tag{52}$$

donde f y g son funciones continuamente diferenciables. Haciendo

$$\begin{aligned} x &= x_n + h_n \\ y &= y_n + k_n \end{aligned} \tag{53}$$

(h_n y k_n son las correcciones, es decir el error de la raíz) tendremos

$$f(x_n + h_n, y_n + k_n) = 0$$

$$g(x_n + h_n, y_n + k_n) = 0$$

de donde, utilizando la fórmula de Taylor y limitándonos a los términos lineales en h_n y k_n, tendremos

$$0 = f(x_n + h_n, y_n + k_n) \approx f(x_n, y_n) + \frac{\partial f}{\partial x}(x_n, y_n)h_n + \frac{\partial f}{\partial y}(x_n, y_n)k_n$$

$$0 = g(x_n + h_n, y_n + k_n) \approx g(x_n, y_n) + \frac{\partial g}{\partial x}(x_n, y_n)h_n + \frac{\partial g}{\partial y}(x_n, y_n)k_n$$

o bien

$$\begin{aligned} \frac{\partial f}{\partial x}(x_n, y_n)h_n + \frac{\partial f}{\partial y}(x_n, y_n)k_n &= -f(x_n, y_n) \\ \frac{\partial g}{\partial x}(x_n, y_n)h_n + \frac{\partial g}{\partial y}(x_n, y_n)k_n &= -g(x_n, y_n) \end{aligned} \tag{54}$$

El sistema obtenido en la ecuación (54) es un sistema lineal de dos ecuaciones con dos incógnitas (h_n y k_n). Si el Jacobiano

$$J(x_n, y_n) = \begin{vmatrix} \dfrac{\partial f}{\partial x}(x_n, y_n) & \dfrac{\partial f}{\partial y}(x_n, y_n) \\ \dfrac{\partial g}{\partial x}(x_n, y_n) & \dfrac{\partial g}{\partial y}(x_n, y_n) \end{vmatrix} \neq 0$$

entonces tendremos para el sistema (54) (por la Regla de Cramer)

$$h_n = \frac{\begin{vmatrix} -f(x_n, y_n) & \dfrac{\partial f}{\partial y}(x_n, y_n) \\ -g(x_n, y_n) & \dfrac{\partial g}{\partial y}(x_n, y_n) \end{vmatrix}}{J(x_n, y_n)} = -\frac{1}{J(x_n, y_n)} \begin{vmatrix} f(x_n, y_n) & \dfrac{\partial f}{\partial y}(x_n, y_n) \\ g(x_n, y_n) & \dfrac{\partial g}{\partial y}(x_n, y_n) \end{vmatrix}$$

$$k_n = \frac{\begin{vmatrix} \dfrac{\partial f}{\partial x}(x_n, y_n) & -f(x_n, y_n) \\ \dfrac{\partial g}{\partial x}(x_n, y_n) & -g(x_n, y_n) \end{vmatrix}}{J(x_n, y_n)} = -\frac{1}{J(x_n, y_n)} \begin{vmatrix} \dfrac{\partial f}{\partial x}(x_n, y_n) & f(x_n, y_n) \\ \dfrac{\partial g}{\partial x}(x_n, y_n) & g(x_n, y_n) \end{vmatrix}$$

Insertando estas correcciones en la ecuación (53), obtendremos, para $n = 0, 1, 2, \ldots$

$$\begin{aligned} x_{n+1} &= x_n - \frac{1}{J(x_n, y_n)} \begin{vmatrix} f(x_n, y_n) & \dfrac{\partial f}{\partial y}(x_n, y_n) \\ g(x_n, y_n) & \dfrac{\partial g}{\partial y}(x_n, y_n) \end{vmatrix} \\ y_{n+1} &= y_n - \frac{1}{J(x_n, y_n)} \begin{vmatrix} \dfrac{\partial f}{\partial x}(x_n, y_n) & f(x_n, y_n) \\ \dfrac{\partial g}{\partial x}(x_n, y_n) & g(x_n, y_n) \end{vmatrix} \end{aligned} \tag{55}$$

Estas fórmulas se conocen con el nombre de *método de Newton para resolver dos ecuaciones no lineales simultáneas*.

Observación. Como en el caso unidimensional, el orden de convergencia es dos. Sin embargo, aquí debemos observar que los

problemas en el uso del método de Newton con sistemas de ecuaciones son completamente diferentes de aquellos para ecuaciones simples. Para ecuaciones simples hemos observado que frecuentemente es útil una buena información a priori sobre la localización de la raíz; cuando esto no es posible, podemos usar un método siempre convergente para obtener una buena aproximación de la raíz. Por lo tanto, en este caso estamos principalmente interesados en la eficiencia de los métodos y comparativamente poco interesados en que el método sea o no convergente. Pero la convergencia en sistemas de ecuaciones es un problema de tal importancia que usualmente quedaremos satisfechos con cualquier orden de convergencia, únicamente si el método es convergente. Frecuentemente, si la aproximación inicial no está completamente cercana a la solución, la iteración no será convergente. Las extensiones de muchos métodos iterativos a sistemas de ecuaciones no lineales están sujetas al mismo tipo de limitaciones sobre la convergencia que para el método de Newton. Al investigar sobre un método para sistemas de ecuaciones no lineales cuya convergencia esté asegurada o cuyas propiedades de convergencia sean al menos razonablemente buenas, una idea obvia es tratar de generalizar el método de la regula falsi a sistemas; pero como no podemos decir nada acerca del signo de $\vec{f}(\vec{x})$, no es posible la generalización siempre convergente. Sin embargo, se puede generalizar el método de la secante a sistemas, pero también está sujeto a limitaciones sobre la convergencia similares a las anteriores (y, por lo tanto, no desarrollaremos su algoritmo).

Nuestras conclusiones son en el sentido de que la solución de ecuaciones no lineales simultáneas es, generalmente, un problema muy difícil y que no se conocen métodos iterativos muy satisfactorios para este problema.

Ejemplo 33. Apliquemos el método de Newton a la resolución de 2 ecuaciones no lineales simultáneas siguiente

$$\begin{cases} f(x,y) = x^2 + y - 11 = 0 \\ g(x,y) = y^2 + x - 7 = 0 \end{cases}$$

dada $(x_0, y_0) = (3.8, -1.8)$ como primera aproximación y siendo la tolerancia de $\varepsilon = 5\text{x}10^{-5}$.

Usaremos el siguiente esquema: (notación: $\frac{\partial f}{\partial x} = f_x, \frac{\partial f}{\partial y} = f_y$, y así siguiendo)

f	f_x	f_y
g	g_x	g_y
D	h_n	k_n
Sol.	x_{n+1}	y_{n+1}

siendo $D = f_x g_y - g_x f_y$, $h_n = \dfrac{f g_y - g f_y}{D}$, $k_n = \dfrac{f_x g - g_x f}{D}$, $x_{n+1} = x_n - h_n$, $y_{n+1} = y_n - k_n$.

Tenemos $f_x = 2x,\ f_y = 1,\ g_x = 1,\ g_y = 2y.$

1.64 0.04	7.6 1	1 -3.6	
-28.36	0.209591 3.590409	0.047109 -1.847109	Iteración 1
0.043928 0.002221	7.180818 1	1 -3.694218	
-27.527507	5.976(-3) 3.584433	1.016(-3) -1.848125	Iteración 2
3.5(-5) -9.85(-7)	7.168866 1	1 -3.69625	
-27.497921	5(-6) 3.584428	1.53(-6) -1.848127	Iteración 3

Como $|x_3 - x_2| = 5\,\text{x}10^{-6} < \varepsilon$ y $|y_3 - y_2| \approx 2\,\text{x}10^{-6} < \varepsilon$ detenemos el proceso iterativo, siendo la solución aproximada: $x = 3.58443$, $y = -1.84813$.

La idea anterior del método de Newton puede extenderse fácilmente a los sistemas no lineales de *n* ecuaciones con *n* incógnitas como el dado en la ecuación (51'): $\vec{f}(\vec{x}) = \vec{0}$.

Supongamos que hemos hallado la aproximación *p*-ésima $\vec{x}^{(p)} = (x_1^{(p)}, x_2^{(p)}, \cdots, x_n^{(p)})$ de una de las raíces separadas $\vec{x} = (x_1, x_2, \cdots, x_n)$ de la ecuación vectorial (51'). La raíz exacta de la ecuación (51') puede representarse entonces como

$$\vec{x} = \vec{x}^{(p)} + \vec{\varepsilon}^{(p)} \qquad (56)$$

donde $\vec{\varepsilon}^{(p)} = (\varepsilon_1^{(p)}, \varepsilon_2^{(p)}, \cdots, \varepsilon_n^{(p)})$ es la corrección (error de la raíz).

Poniendo la ecuación (56) en la ecuación (51'), tenemos

$$\vec{f}(\vec{x}^{(p)} + \vec{\varepsilon}^{(p)}) = \vec{0}\,. \tag{57}$$

Suponiendo que la función $\vec{f}(\vec{x})$ es continuamente diferenciable en un cierto dominio convexo que contiene a $\vec{x}$ y $\vec{x}^{(p)}$, desarrollemos el primer miembro de la ecuación (57) según la fórmula de Taylor para *n* variables limitándonos a los términos lineales

$$\vec{0} = \vec{f}(\vec{x}^{(p)} + \vec{\varepsilon}^{(p)}) \approx \vec{f}(\vec{x}^{(p)}) + \vec{f}'(\vec{x}^{(p)})\vec{\varepsilon}^{(p)} \tag{58}$$

es decir,

$$\vec{f}(\vec{x}^{(p)}) + \vec{f}'(\vec{x}^{(p)})\vec{\varepsilon}^{(p)} = \vec{0} \tag{59}$$

o en forma desarrollada para cada función f_i (i = 1, 2, ..., n)

$$f_i\left(x_1^{(p)}, x_2^{(p)}, \cdots, x_n^{(p)}\right) + \sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j}\left(x_1^{(p)}, x_2^{(p)}, \cdots, x_n^{(p)}\right)\varepsilon_j^{(p)} = 0. \tag{60}$$

De las fórmulas (59) y (60) se deduce que la derivada $\vec{f}'(\vec{x})$ a de ser considerada como la matriz Jacobiana del conjunto de funciones $f_1, f_2, \cdots, f_n$ con respecto a las variables $x_1, x_2, \cdots, x_n$; esto es

$$\vec{f}'(\vec{x}) = W(\vec{x}) = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \cdots & \dfrac{\partial f_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{pmatrix}$$

o abreviadamente

$$\vec{f}'(\vec{x}) = W(\vec{x}) = \left[\frac{\partial f_i}{\partial x_j}\right] \qquad (i, j = 1, 2, ..., n).$$

El sistema (60) es un sistema lineal en las correcciones $\varepsilon_j^{(p)}$ (j = 1, 2, ..., n) con la matriz $W(\vec{x})$, y de aquí que la fórmula (59) puede escribirse como sigue

$$\vec{f}(\vec{x}^{(p)}) + W(\vec{x}^{(p)})\vec{\varepsilon}^{(p)} = \vec{0} \tag{61}$$

de donde, dando por supuesto que la matriz $W(\vec{x}^{(p)})$ es no singular, tenemos

$$\vec{\varepsilon}^{(p)} = -W^{-1}(\vec{x}^{(p)})\vec{f}(\vec{x}^{(p)})$$

en consecuencia,

$$\vec{x}^{(p+1)} = \vec{x}^{(p)} - W^{-1}(\vec{x}^{(p)})\vec{f}(\vec{x}^{(p)}) \quad (p = 0, 1, ...) \tag{62}$$

es decir que obtenemos *el método de Newton para resolver n ecuaciones no lineales simultáneas*.

Resumiendo, digamos que para propósitos computacionales una iteración de este método consta de los siguientes pasos:

1. Calcular $\vec{f}(\vec{x}^{(p)})$ y $W(\vec{x}^{(p)})$.
2. Resolver el sistema lineal $W(\vec{x}^{(p)})\vec{\varepsilon}^{(p)} = -\vec{f}(\vec{x}^{(p)})$ para $\vec{\varepsilon}^{(p)}$ (ver ecuación (61)).
3. Calcular $\vec{x}^{(p+1)} = \vec{x}^{(p)} + \vec{\varepsilon}^{(p)}$.

Observación. Cuando el número de ecuaciones no lineales aumenta se hace más pesado resolver el sistema de ecuaciones lineales resultante para la corrección de las variables (paso 2 anterior). Algún tiempo puede ahorrarse si se observa que el valor de las derivadas no cambia significativamente, de modo que se puede dejar invariable el valor de la matriz $W(\vec{x}^{(0)})$ y resolver cambiando simplemente los términos independientes. Esto no sería otra cosa más que el método de Newton modificado o método de von Mises extendido a sistemas de ecuaciones no lineales. Si no sucede esto (lo que ocurre frecuentemente), seguimos con el problema de resolver el sistema de ecuaciones lineales resultante. Es bien sabido que la Regla de Cramer (que aplicamos para resolver el sistema de dos ecuaciones lineales) nunca debe usarse para la solución de sistemas lineales de orden un poco grande. La solución de tales sistemas se encuentra en forma mucho mas conveniente por medio del método de Crout. Eventualmente podríamos usar cualquiera de los métodos que hemos analizado para tal fin, pero el método de Crout es el más conveniente debido a que el conjunto de ecuaciones debe resolverse en forma repetida, lo que es fácilmente efectuado por dicho método.

Veamos con un ejemplo cómo se utilizan los métodos de Newton y Crout combinados para resolver un sistema de ecuaciones no lineales simultáneas.

Ejemplo 34. Apliquemos el método de Newton y el método de Crout combinados para la resolución de ecuaciones no lineales siguiente

$$\begin{cases} f(x,y,z) = x^2 + y^2 + z^2 - 1 = 0 \\ g(x,y,z) = 2x^2 + y^2 - 4z = 0 \\ h(x,y,z) = 3x^2 - 4y + z^2 = 0 \end{cases}$$

comenzando con la aproximación inicial $x_0 = y_0 = z_0 = 0.5$ y una tolerancia de $\varepsilon = 5\text{x}10^{-4}$.

Tenemos

$$\begin{array}{lll} f_x = 2x & f_y = 2y & f_z = 2z \\ g_x = 4x & g_y = 2y & g_z = -4 \\ h_x = 6x & h_y = -4 & h_z = 2z \end{array}$$

Disponemos los cálculos arrojados por este algoritmo en el siguiente esquema práctico

	1= $f_x(\vec{x}^{(0)})$ 2= $g_x(\vec{x}^{(0)})$ 3= $h_x(\vec{x}^{(0)})$	1= $f_y(\vec{x}^{(0)})$ 1= $g_y(\vec{x}^{(0)})$ -4= $h_y(\vec{x}^{(0)})$	1= $f_z(\vec{x}^{(0)})$ -4= $g_z(\vec{x}^{(0)})$ 1= $h_z(\vec{x}^{(0)})$	0.25= - $f(\vec{x}^{(0)})$ 1.25= - $g(\vec{x}^{(0)})$ 1= - $h(\vec{x}^{(0)})$
Matriz transformada por el método de Crout	1 2 3	1 -1 -7	1 6 40	0.25 -0.75 -0.125
	0.375	0	-0.125	$\vec{\varepsilon}^{(0)}$
	0.875	0.5	0.375	$\vec{x}^{(1)} = \vec{x}^{(0)} + \vec{\varepsilon}^{(0)}$
	1.75 3.5 5.25	1 1 -4	0.75 -4 0.75	-0.15625 -0.28125 -0.4375
Matriz transformada por el método de Crout	1.75 3.5 5.25	0.571429 -1.000002 -7.000002	0.428571 5.499988 36.999929	-0.089286 -0.031251 -0.005068
	-0.085184	-3.377(-3)	-0.005068	$\vec{\varepsilon}^{(1)}$
	0.789816	0.496623	0.369932	$\vec{x}^{(2)}$

	1.579632 3.159264 4.738896	0.993246 0.993246 -4	0.739864 -4 0.739864	-0.007293 -0.014525 -0.021786
Matriz transformada por el método de Crout	1.579632 3.159264 4.738896	0.628783 -0.993245 -6.979737	0.468377 5.516994 37.027441	-4.617 (-3) -6.2 (-5) -9 (-6)
	-4.605 (-3)	-1.2 (-5)	-9 (-6)	$\vec{\varepsilon}^{(2)}$
	0.785211	0.496611	0.369923	$\vec{x}^{(3)}$
	1.570422 3.140844 4.711266	0.993222 0.993222 -4	0.739846 -4 0.739846	-2.2 (-5) -4.3 (-5) -6.8 (-5)
Matriz transformada por el método de Crout	1.570422 3.140844 4.711266	0.632455 -0.993220 -6.979664	0.471113 5.517098 37.027798	-1.4 (-5) -1 (-6) -2 (-7)
	-1.4 (-5)	1 (-7)	-2 (-7)	$\vec{\varepsilon}^{(3)}$
	0.785197	0.496611	0.369923	$\vec{x}^{(4)}$

Como $\left|\varepsilon_x^{(3)}\right| = 1.4(-5) < \varepsilon$, $\left|\varepsilon_y^{(3)}\right| = 1(-7) < \varepsilon$, $\left|\varepsilon_z^{(3)}\right| = 2(-7) < \varepsilon$, entonces podemos tomar $x = 0.7852$, $y = 0.4966$, $z = 0.3699$.

EJERCICIOS PROPUESTOS .....

1. a) Resolver el siguiente conjunto de sistemas lineales $Ax = B$ utilizando el método de eliminación de Gauss (sin pivoteo y con columna de control) y calcular el determinante de la matriz A:

i)

$$A = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

ii)

$$A = \begin{pmatrix} 3 & -0.1 & -0.2 \\ 0.1 & 7 & -0.3 \\ 0.3 & -0.2 & 10 \end{pmatrix}, \quad B = \begin{pmatrix} 7.85 \\ -19.3 \\ 71.4 \end{pmatrix}$$

iii)

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \\ 1 & 16 & 81 & 256 \end{pmatrix}, \quad B = \begin{pmatrix} 2 \\ 10 \\ 44 \\ 190 \end{pmatrix}$$

Efectuar los cálculos con 6 cifras significativas.

b) Repetir el problema dado en el apartado a) utilizando el método de Gauss-Jordan normalizado, y encontrar la inversa de la matriz A en los apartados ii) y iii).
c) Hacer lo mismo que en el apartado b) utilizando el método de Crout.

2. Resolver el siguiente sistema de ecuaciones lineales utilizando el método de Crout (con columna de control) y calcular el determinante de la matriz asociada:

$$\begin{aligned} 10x_2 + x_3 &= 2 \\ x_1 + 3x_2 \quad x_3 &= 6 \\ 2x_1 + 4x_2 + x_3 &= 5 \end{aligned}$$

3. Resolver los siguientes sistemas de ecuaciones lineales simétricos y definidos positivos utilizando el método de Cholesky (con columna de control) y calcular el determinante de la matriz asociada:

a)

$$\begin{aligned} x_1 + 2x_2 - x_3 &= 2 \\ 2x_1 + 5x_2 + x_3 + x_4 &= 6 \\ -x_1 + x_2 + 14x_3 + 5x_4 &= 6 \\ x_2 + 5x_3 + 3x_4 &= 2 \end{aligned}$$

b)

$$\begin{aligned} 8x + 7y + 83z + 235w &= 161.1 \\ 7x + 83y + 235z + 1907w &= 989.1 \\ 83x + 235y + 1907z + 7987w &= 4713.3 \\ 235x + 1907y + 7987z + 55643w &= 32622.9 \end{aligned}$$

Retener en todos los cálculos 4 decimales.

4. Hacer los programas correspondientes a los métodos de Gauss - Jordan y Crout y resolver los sistemas de ecuaciones lineales siguientes:
a) Todos los dados en los ejercicios anteriores.

b)

$$0.3x_1 - 0.5\,x_2 + 0.7\,x_3 + 0.4\,x_4 - 0.8\,x_5 = 1.3$$
$$1.5\,x_1 + 2.4\,x_2 - 3.1\,x_3 + 2.5\,x_4 + 1.6\,x_5 = 2.4$$
$$2.3\,x_1 - 1.8\,x_2 + 2.2\,x_3 - 0.7\,x_4 - 2.5\,x_5 = 1.3$$
$$1.1x_1 + 1.3\,x_2 - 0.5\,x_3 + 1.4\,x_4 + 2.7\,x_5 = -1.2$$
$$2.6\,x_1 - 3.4\,x_2 + 1.4\,x_3 - 2.3\,x_4 - 5.1\,x_5 = 2.3$$

5. Resolver los siguientes sistemas de ecuaciones lineales utilizando los métodos iterativos de Jacobi y Gauss - Seidel (con columna de control):

a)

$$10\,x_1 + x_2 + x_3 = 46.31$$
$$2x_1 + 12x_2 + x_3 = 65.20$$
$$x_1 + 2\,x_2 + 15x_3 = 81.32$$

Retener en todos los cálculos 4 decimales. Especificar el número de iteraciones necesarias para obtener la solución con una tolerancia $\varepsilon = 1 \times 10^{-3}$.

b)

$$80\,x_1 + 0.1x_2 + x_3 = 48$$
$$0.4x_1 + 20x_2 - 0.8x_3 = -4$$
$$0.1x_1 - 0.6x_2 + 10x_3 = 3$$

Retener en todos los cálculos 4 dígitos significativos. Especificar el número de iteraciones necesarias para obtener la solución con una tolerancia $\varepsilon = 1 \times 10^{-3}$.

c)

$$2\,x_1 + 12\,x_2 + x_3 - 4\,x_4 = 13$$
$$2\,x_1 + x_2 - 3\,x_3 + 9\,x_4 = 31$$
$$13\,x_1 + 5\,x_2 - 3\,x_3 + x_4 = 18$$
$$3\,x_1 - 4\,x_2 + 10\,x_3 + x_4 = 29$$

Retener en todos los cálculos 4 decimales. Especificar el número de iteraciones necesarias para obtener la solución con una tolerancia $\varepsilon = 1 \times 10^{-2}$.

6. Hacer el programa correspondiente al método de Jacobi y resolver, siempre que sea posible, los sistemas dados en los ejercicios 1 a) y 5.

7. Encontrar la solución por mínimos cuadrados y el error de la raíz cuadrática media de los siguientes sistemas de ecuaciones lineales:

a)

$$\begin{aligned} x + y &= -2.5 \\ 3x + 4y &= 12 \\ 2x - 3y &= -6 \\ 3x + 2y &= 6 \\ 4x + y &= 4 \end{aligned}$$

b)

$$\begin{aligned} x_1 - x_2 &= 2 \\ x_1 + x_2 &= 4 \\ 2x_1 + x_2 &= 8 \\ x_1 + 2x_2 &= 4 \\ 2x_1 - x_2 &= 5 \\ x_1 - 2x_2 &= 2 \end{aligned}$$

c)

$$\begin{pmatrix} 1.0 & 0 & 0 \\ 1.0 & 0.6 & 0.565 \\ 1.0 & 1.2 & 0.932 \\ 1.0 & 1.8 & -0.751 \\ 1.0 & 2.4 & 0.675 \\ 1.0 & 3.0 & 0.141 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 5.00 \\ 6.66 \\ 7.53 \\ 7.10 \\ 5.30 \\ 2.36 \end{pmatrix}$$

8. Resolver los siguientes sistemas de ecuaciones no lineales utilizando los métodos de Newton y Crout combinados:

a)

$$f(x,y) = y - x^2 + y^2 = 0$$
$$g(x,y) = x - x^2 - y^2 = 0$$

tomando como aproximaciones iniciales $x_0 = y_0 = 0.5$ y siendo la tolerancia $\varepsilon = 1 \times 10^{-5}$.

b)

$$f(x,y) = 4x^3 - 27xy^2 + 25 = 0$$
$$g(x,y) = 4x^2 - 3xy^3 - 1 = 0$$

tomando como aproximaciones iniciales $x_0 = 1.1$, $y_0 = 1.0$ y siendo la tolerancia $\varepsilon = 5 \times 10^{-4}$.

c)

$$f(x,y,z) = x^2 + y^2 + z^2 - 1 = 0$$
$$g(x,y,z) = 2x^2 + y^2 - 4z - 0.5 = 0$$
$$h(x,y,z) = 3x^2 - 4y + z^2 - 0.1 = 0$$

tomando como aproximaciones iniciales $x_0 = y_0 = z_0 = 0.5$ y siendo la tolerancia $\varepsilon = 5 \times 10^{-4}$.

9. Considerar el sistema que se muestra en la Figura 1. Hay tres bloques atados por una cuerda de peso despreciable apoyados sobre una superficie lisa inclinada 45º respecto a la horizontal. El coeficiente de fricción entre el plano y la masa de 100 kg es de 0.25 y entre las masas de 50 y 20 kg es de 0.375.

Expresar las ecuaciones del movimiento en forma matricial (sumando las fuerzas en dirección paralela al plano y usando la segunda ley de Newton) y resolver el sistema que resulta, utilizando:

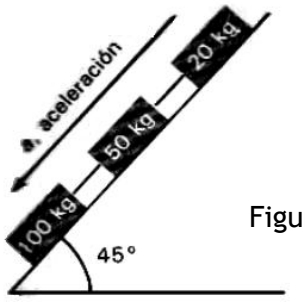


Figura 1

a) El método de Gauss-Jordan.
b) El método de Crout.
c) El método de Jacobi (cota de error admisible $\varepsilon = 5\text{x}10^{-4}$).
d) El método de Gauss-Seidel (cota de error admisible $\varepsilon = 5\text{x}10^{-4}$).
e) Según los resultados obtenidos, expresar sus conclusiones.

10. La mayor parte de los diferentes campos de la ingeniería manejan distribuciones de temperaturas en materiales sólidos. La distribución de temperatura en estado estacionario bidimensional se define por la ecuación de Laplace:

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0 \qquad (1)$$

en donde T es la temperatura y x e y son las coordenadas. Las derivadas de la ecuación (1) se aproximan usando diferencias divididas finitas. La Figura 2 muestra una malla bidimensional, esquema útil en las aproximaciones desarrolladas para la ecuación (1).

Las aproximaciones por diferencias divididas finitas de las derivadas son:

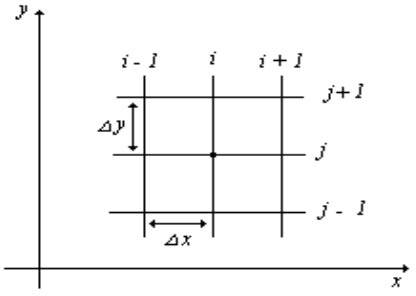


Figura 2

$$\frac{\partial T}{\partial x} \approx \frac{\Delta T}{\Delta x} = \frac{T_{i+1,j} - T_{i,j}}{\Delta x}$$

de donde,

$$\frac{\partial^2 T}{\partial x^2} \approx \frac{T_{i+1,j} - 2T_{i,j} + T_{i-1,j}}{\Delta x^2}$$

y de manera similar

$$\frac{\partial^2 T}{\partial y^2} \approx \frac{T_{i,j+1} - 2T_{i,j} + T_{i,j-1}}{\Delta y^2}.$$

Suponiendo que $\Delta x = \Delta y$, la ecuación de Laplace se puede aproximar como

$$T_{i+1,j} + T_{i-1,j} + T_{i,j+1} + T_{i,j-1} - 4T_{i,j} = 0 \qquad (2)$$

la cual es aplicable a cada nodo i, j de la Figura 2. Al aplicar la ecuación (2) a cada nodo resulta un sistema de ecuaciones acopladas, ya que la

temperatura en varias posiciones aparece en más de una ecuación. Esto produce un sistema de ecuaciones lineales.

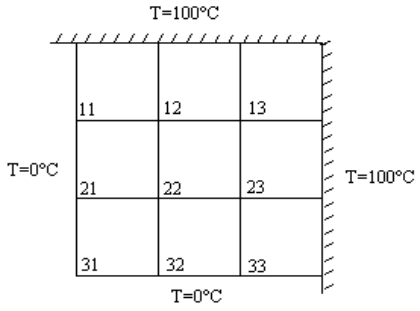


Figura 3

Considérese la placa plana de la Figura 3. Los lados de la placa se mantienen a temperaturas constantes de 0º C y 100º C, como se muestra en dicha figura. La distribución de la temperatura dentro de la placa se puede aproximar en nueve puntos internos aplicando la ecuación de Laplace en cada punto.

Plantear el sistema de ecuaciones lineales generado y resolverlo utilizando:

a) El método de Gauss-Jordan.
b) El método de Cholesky.
c) El método de Jacobi ($\varepsilon = 5x10^{-4}$).
d) El método de Gauss-Seidel ($\varepsilon = 5x10^{-4}$).
e) Según los resultados obtenidos, expresar sus conclusiones.

11. Los ingenieros electricistas a menudo deben encontrar las corrientes que circulan y los voltajes que existen en una red compleja de resistores. A continuación se muestra un problema típico.

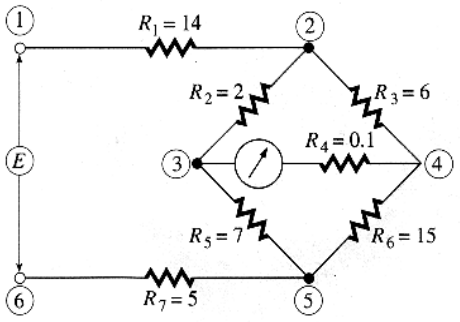


Figura 4

Siete resistores están conec-tados y al circuito se aplica un voltaje en los puntos 1 y 6 (véase la Figura 4). Aunque se tiene especial interés en encontrar la corriente que circula a través del amperímetro, el método numérico utilizado puede proporcionar los voltajes en cada punto numerado (estos puntos se denominan nodos) y la corriente que pasa por cada una de las ramas del circuito. Dos leyes están implicadas:

Ley de Kirchhoff: *La suma de todas las corrientes que fluyen hacia un nodo es cero*.

Ley de Ohm: *La corriente que pasa por un resistor es igual al voltaje que pasa por ese resistor dividido entre su resistencia*.

Es posible plantear once ecuaciones usando estas leyes y a partir de ellas despejar once cantidades desconocidas (los cuatro voltajes y las siete corrientes). Para $v_1 = 5$ voltios y $v_6 = 0$ voltios:

a) Escribir las once ecuaciones.
b) Resolver el sistema de ecuaciones lineales resultante para encontrar el voltaje en cada nodo y las corrientes que circulan en cada rama del circuito, utilizando los distintos métodos numéricos estudiados y los correspondientes programas. Luego,
   i) Escribir la solución obtenida, según el método numérico empleado.
   ii) Comparar, en general y para esta situación en particular, los distintos métodos numéricos utilizados según sus ventajas y desventajas.
c) Formular y fundamentar una conclusión según los resultados obtenidos.

12. La presión que se requiere para sumir un objeto grande y pesado en un suelo blando y homogéneo situado arriba de un terreno de base dura, puede predecirse mediante la presión que se requiere para introducir objetos más pequeños en el mismo terreno. Específicamente, la presión p para sumir una placa circular de radio r a una distancia d en un terreno blando, donde el terreno de base sólida se halla a una distancia $D > d$ debajo de la superficie, puede aproximarse con una ecuación de la forma

$$p = k_1 e^{k_2 r} + k_3 r$$

donde k_1, k_2 y k_3 son constantes, con $k_2 > 0$, que dependen de d y de la consistencia del suelo, pero no del radio de la placa.

Si queremos determinar el tamaño mínimo de la placa necesario para sostener una gran carga, metemos tres placas pequeñas con radios distintos a la misma distancia y se miden los pesos requeridos para hacerlo; en la Figura 5 se muestran las cargas requeridas para esta maniobra.

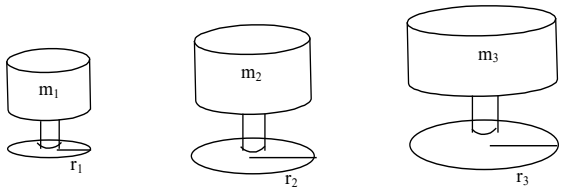


Figura 5

Esto genera tres ecuaciones no lineales del tipo a la anterior, en las tres incógnitas k_1, k_2 y k_3.

a) Calcular los valores de k_1, k_2 y k_3 si se supone que una placa cuyo radio es de 1 plg requiere una presión de 10 lb/plg^2 para enterrarse a 1 pie en un campo fangoso, una placa cuyo radio es de 2 plg requiere una presión de 12 lb/plg^2 para enterrarse a 1 pie y una placa de 3 plg de radio requiere una presión de 15 lb/plg^2 para enterrarse esta distancia (suponiendo que el lodo tiene una profundidad de más de 1 pie).
b) Usar los cálculos de la parte a) para predecir el tamaño mínimo de la placa circular que se necesitará para sostener una carga de 500 lb en este campo, con un hundimiento menor a 1 pie.
c) Formular y fundamentar una conclusión según los resultados obtenidos.

……

Capítulo 4

Interpolación y aproximación polinomial. Aproximación por mínimos cuadrados.

4

4.1. Introducción.

La interpolación, que es el cálculo de valores para una función tabulada en puntos que no aparecen en la tabla, es históricamente una tarea fundamental. Los nombres de muchos matemáticos famosos están asociados con procedimientos de interpolación: Newton, Gauss, Bessel, Stirling. La necesidad de interpolar empezó con los primeros estudios de astronomía cuando el movimiento de cuerpos celestes debía determinarse a partir de observaciones periódicas. Hoy en día, los estudiantes rara vez tienen que interpolar para valores de senos, logaritmos y demás funciones no algebraicas a partir de tablas. Sus calculadoras y computadoras calculan los valores usando técnicas que describiremos en este Capítulo. ¿Por qué entonces se dedica tanto tiempo y esfuerzo a un tema que parece casi obsoleto? Hay cuatro razones. Primero, los métodos de interpolación constituyen la base de otros procedimientos: derivación e integración numéricas y métodos de solución de ecuaciones diferenciales ordinarias y ecuaciones diferenciales parciales. Segundo, estos métodos demuestran algunos resultados teóricos importantes sobre los polinomios y la exactitud de los métodos numéricos. Tercero, interpolar con polinomios sirve como una excelente introducción a ciertas técnicas para trazar curvas suaves. Y, por último, la historia en sí es fascinante.

En este Capítulo se comparan varias formas de realizar interpolación y se contrastan estos procedimientos con otras formas para ajustar datos imprecisos y trazar curvas suaves. Se consideran los siguientes procedimientos:

- **Diferencias divididas, diferencias finitas y diferencias centrales.**

Se describen estos eficientes métodos para construir los polinomios de interpolación y obtener valores interpolados. En este caso, los datos están relativamente libres de error. La interpolación polinomial está diseñada para ajustar un polinomio único de n-ésimo orden que pase exactamente por los n+1 puntos exactos. Este polinomio se presenta en diferentes formatos según las abscisas sean o no equidistantes. Los polinomios que utilizan diferencias se adaptan idealmente a aquellos casos en que el orden propio del polinomio se desconoce. Estos métodos se programan fácilmente en un formato que

compara los resultados con órdenes diferentes. Además, se puede incorporar con facilidad una aproximación del error del método. De esta forma, se puede comparar y escoger a partir de los resultados usando varios polinomios de órdenes diferentes.

- **Polinomio de Lagrange.**

Se presenta el polinomio de interpolación de Lagrange: una forma directa, pero tediosa desde la perspectiva del cálculo, para construir un polinomio de interpolación. Este polinomio es apropiado cuando el orden se conoce a priori. Es más simple de programar que los anteriores y no requiere de los cálculos y almacenamiento de diferencias.

- **Aproximación por mínimos cuadrados.**

Se considera el ajuste de polinomios y otras funciones a datos inexactos. Este problema clásico debe enfrentarse, por ejemplo, cuando se interpretan datos experimentales. Las técnicas de aproximación por mínimos cuadrados (también llamadas regresión con mínimos cuadrados) se usan para desarrollar la "mejor curva" que ajuste todas las tendencias de los datos sin pasar necesariamente a través de algún punto. Se analizan distintas técnicas que modelan diversas situaciones, según la relación existente entre las variables del problema a resolver.

Los fundamentos matemáticos para la interpolación se encuentran en las expansiones de la serie de Taylor y las diferencias divididas y finitas. En la regresión con mínimos cuadrados se requieren conocimientos de estadística.

4.2. Interpolación y aproximación polinomial.

4.2.1. Introducción.

Hasta el momento hemos estado ocupados, fundamentalmente, con el problema de la aproximación a números, tales como las raíces de ecuaciones no lineales o la solución de sistemas de ecuaciones. Volvemos ahora nuestra atención al problema de la aproximación a funciones y, con mayor generalidad, a números tales como las derivadas y las integrales que dependen de una infinidad de valores de una función.

El estudio de la teoría de aproximación involucra dos tipos generales de problemas. Un problema surge cuando se tiene explícitamente una función pero se desea encontrar un tipo de función "más simple", como un polinomio, para determinar los valores aproximados a la función dada. El otro problema en la teoría de aproximación es el que consiste en ajustar

funciones a datos dados y encontrar la "mejor" función dentro de cierta clase que pueda usarse para representar a los datos.

Ejemplo 1. Consideremos el problema de computar valores de la función *sen x.*

Excepto para valores muy especiales de *x*, estos valores no pueden ser calculados exactamente; en efecto, se demuestra en teoría de números que el número *sen x* es trascendental, $\forall\ x \neq 0$ racional. Calcular *sen x* desde su serie de Maclaurin

$$sen\, x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots$$

requerirá infinitos términos de la serie, lo cual no puede hacerse. Si se desean valores de *sen x*, esta función debe ser aproximada (con un error que sea lo más pequeño posible) por una función que pueda ser realmente computada.

Ejemplo 2. Cada 10 años se toma un censo de la población de los EEUU. A continuación se muestra una tabla con los datos en miles de personas de la población de 1940 hasta 1990:

Año	1940	1950	1960	1970	1980	1990
Población (en miles)	132165	151326	179323	203302	226542	249633

Al revisar estos datos podríamos preguntarnos si podrían usarse para estimar razonablemente la población, digamos, en 1965 o incluso en el año 2010. Algunas predicciones de este tipo pueden obtenerse usando una función que ajuste los datos dados.

Pero, ¿cuáles funciones pueden ser realmente calculadas? Toda computadora puede realizar las cuatro operaciones aritméticas básicas más el análisis de alguna condición en términos de desigualdades. Cualquier función que pueda ser calculada por medio de un número finito de adiciones, sustracciones, multiplicaciones y divisiones es racional. Así, es claro que las únicas funciones que pueden ser computadas son las funciones racionales o funciones racionales segmentarias (significa que es posible particionar el intervalo de definición en subintervalos, tal que en cada subintervalo la función es representada por una función racional. Estas funciones racionales pueden ser diferentes en distintos subintervalos (ver Figura 1)).

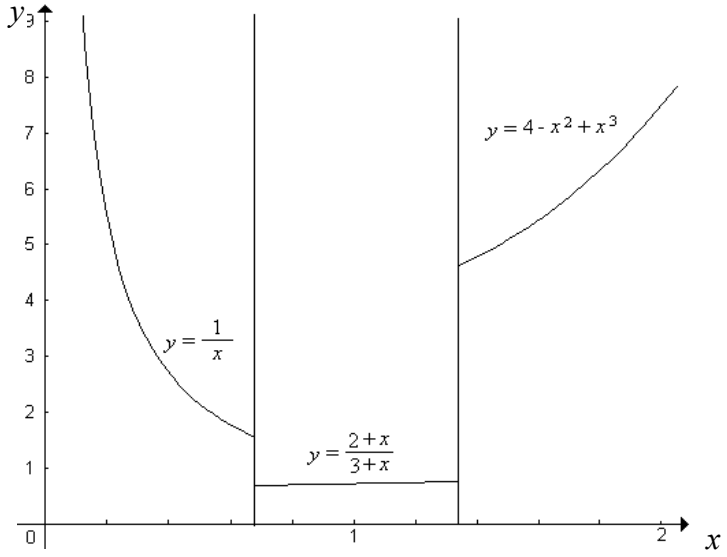


Figura 1

Una funcional racional o racional segmentaria es, entonces, la clase más general de funciones que puede ser evaluada directamente en una computadora digital. (Esto constituye una ligera exageración. Algunas funciones mas generales pueden ser evaluadas usando operaciones lógicas u operaciones sobre la magnitud de un número que es útil en muchas computadoras. Pero para propósitos prácticos lo anterior es cierto). En nuestro estudio nos restringiremos a la *aproximación por polinomios*.

La aproximación por polinomios es una de las ideas más antiguas en el Análisis Numérico y sigue siendo una de las más utilizadas.

La clase de los polinomios algebraicos, esto es, el conjunto de funciones de la forma

$$P(x) = a_0 + a_1 x + \ldots + a_n x^n$$

donde n es un entero no negativo y a_0, a_1, …, a_n son constantes reales, es una de las más útiles y bien conocidas clases de funciones reales de variable real. Un polinomio $P(x)$ se usa como sustituto para la función $f(x)$ por un sin número de razones. Quizá la más importante de todas sea que los polinomios son fáciles de calcular al incluir solamente potencias simples de enteros. Sus derivadas e integrales indefinidas se encuentran también sin mucho esfuerzo y son también polinomios. Las raíces de ecuaciones polinomiales se descubren con menores dificultades que en el caso de otras funciones. Otro aspecto importante para considerar la clase de polinomios en la

aproximación de funciones es que aproximan de manera uniforme a las funciones continuas, esto es, dada una función definida y continua en un intervalo cerrado, existe un polinomio que está tan "cerca" de la función dada como se desee (Figura 2). Este resultado se expresa más precisamente en el teorema siguiente.

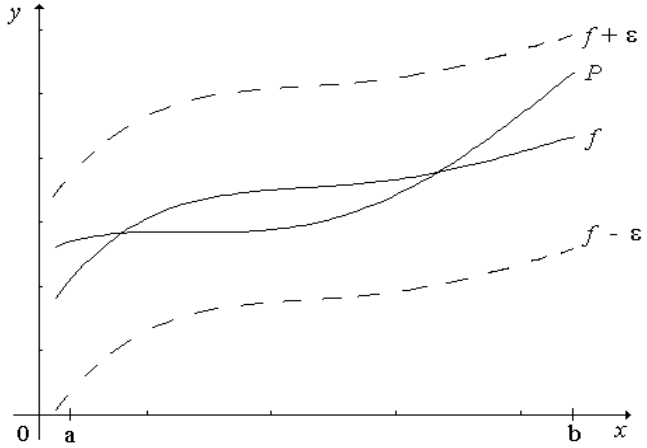


Figura 2

Teorema de aproximación de Weierstrass. Si f está definida y es continua en un intervalo finito y cerrado [a, b], para cada $\varepsilon > 0$ existe un polinomio $P(x)$ definido en [a, b] con la propiedad de que $|f(x) - P(x)| < \varepsilon$ para todo $x \in$ [a, b].

La demostración de este teorema puede encontrarse en cualquier texto elemental de Análisis Real. Se puede probar de muchas formas diferentes. Algunas demostraciones contienen un algoritmo para la construcción del polinomio de aproximación. Dichos polinomios construidos de esta forma no son generalmente adecuados para propósitos de una computación numérica, debido a que su grado es mucho más alto que el necesario y resulta impráctico usarlo. Además, si bien es confortable conocer que algunos polinomios aproximarían a $f(x)$ con una exactitud específica en todo el intervalo [a, b], esto no garantiza que un tal polinomio será encontrado por un algoritmo práctico. Tampoco nos dice nada acerca de la existencia de un polinomio de aproximación satisfactorio para un conjunto de datos dados $\{(x_k, f(x_k)): k = 0, 1, \ldots, n\}$.

En resumen, las razones antes expuestas hace que la clase de polinomios se use con frecuencia para aproximar otras funciones que se conoce o se supone son continuas. Dichas razones nos permiten entender el por qué de la popularidad de los polinomios como sustitutos.

Un ejemplo familiar es la aproximación de $f(x)$ por los N primeros términos de su desarrollo de Taylor. Como los polinomios de Taylor tienen la propiedad de que toda la información usada está concentrada en un punto, digamos x_0, esto generalmente limita el uso de la aproximación con polinomios de Taylor a situaciones donde las aproximaciones se necesitan en puntos muy cercanos a x_0. Para propósitos de cómputos ordinarios es más eficiente utilizar métodos que incluyan información de varios puntos. La construcción de este tipo de polinomios es lo que se considera en este Capítulo. Los polinomios de Taylor son útiles sobre intervalos pequeños para funciones cuyas derivadas existen y son fáciles de evaluar, pero obviamente este no es siempre el caso. El uso primordial de los polinomios de Taylor en el Análisis Numérico no es con propósitos de aproximación, sino para la derivación de técnicas numéricas. En consecuencia, el polinomio de Taylor es frecuentemente de poca utilidad y se deben buscar métodos alternativos de aproximación. El material que veremos se concentra en encontrar polinomios aproximantes que puedan determinarse simplemente especificando algunos puntos en el plano por los cuales deben pasar.

4.2.2. Criterios de aproximación.

La diferencia $f(x) - P(x)$ es *el error en la aproximación* y la idea central es, desde luego, mantener este error razonablemente pequeño. La simplicidad de los polinomios permite que esta meta se alcance de diversas maneras, de las cuales consideraremos *la colocación*.

El polinomio de colocación es aquel que coincide (se coloca) con $f(x)$ en ciertos puntos específicos; sin embargo, no garantiza nada con referencia a la primera derivada o a derivadas de órdenes superiores. Generalmente se utilizan cuando sólo nos interesan ciertos puntos de la función original y ésta no es sencilla de evaluar; también cuando no conocemos la función original y sólo contamos con una muestra de los puntos reales.

Teorema. Si $x_0, x_1, \ldots, x_n$ son $(n + 1)$ números reales distintos y f es una función cuyos valores están dados en esos números (o argumentos), entonces existe un único polinomio de colocación P de grado a lo más n para los argumentos $x_0, x_1, \ldots, x_n$; es decir, tal que $f(x) = P(x)$ para estos argumentos.

Demostración. *Existencia*. Se probará más adelante exhibiendo realmente un polinomio de tales características.

Unicidad. La probaremos suponiendo que el polinomio de colocación no es único. Supongamos entonces que existe otro polinomio de grado a lo más n, $Q(x)$, que verifica que $f(x) = Q(x)$ para los argumentos $x_0, x_1, \ldots, x_n$. Ya que el grado de $D(x) = P(x) - Q(x)$ puede ser a lo más n, este polinomio

puede tener a lo más n ceros en tanto que no sea el polinomio cero. En vista de que los x_i son distintos, $D(x)$ tiene $n + 1$ ceros; pero entonces $D(x)$ debe valer cero. Por consiguiente, $P \equiv Q$.

4.3. El polinomio de interpolación. Planteamiento del problema.

El significado de la palabra "colocación" es similar al de "aproximación" y al de "interpolación" tratándose de este tema, ya que garantiza que el polinomio resultante tocará a la función original $f(x)$ en los puntos dados como datos, también llamados *puntos muestreados*. Se dice que la "interpolación" es el arte de leer entre las líneas de una tabla. *La interpolación polinomial* (ajustar un polinomio a los puntos dados) es uno de los temas más importantes en los métodos numéricos, ya que la mayoría de los demás modelos numéricos se basan en la interpolación polinomial. Por ejemplo, los modelos de integración numérica y de diferenciación numérica se obtienen, respectivamente, integrando y derivando fórmulas de interpolación polinomial. La interpolación polinomial se puede expresar en varias formas alternativas que pueden transformarse entre sí. Entre éstas se encuentran *las series de potencias*, *la interpolación de Newton*, *la interpolación de Gauss*, *la interpolación de Lagrange*, entre otras.

Supongamos que tenemos $(n + 1)$ pares de datos $(x_0, f(x_0))$, $(x_1, f(x_1))$, …, $(x_n, f(x_n))$ que representan $(n + 1)$ puntos de la gráfica de una función $y = f(x)$. Las abscisas x_k, $k = 0, 1, \ldots, n$, se suponen distintas y los espacios entre ellas son arbitrarios. Para abreviar escribimos $f(x_k) = f_k$, para $k = 0, 1, \ldots, n$.

Deseamos aproximar $f(x)$ mediante una función $P(x)$ que sea fácilmente manipulable matemáticamente y que pueda evaluarse en cualquier x dentro de un intervalo I que contiene a los x_k, para $k = 0, 1, \ldots, n$. Posteriormente, el valor de $P(x)$ se utiliza para aproximar a $f(x)$ cuando no se conoce explícitamente. Dado que tenemos $(n + 1)$ valores de la función f_k, para $k = 0, 1, \ldots, n$, podemos imponer $(n+1)$ condiciones para determinar los coeficientes en la aproximación polinomial.

Esto significa que podemos determinar un polinomio $P(x)$ de grado máximo n con los coeficientes determinados por las $(n + 1)$ condiciones

$$P(x_k) = f_k, \qquad k = 0, 1, \ldots, n.$$

El polinomio de interpolación es aquel polinomio $P(x)$ de grado máximo n que se aproxima a $f(x)$ sobre un intervalo que contiene a $x_0, x_1, \ldots, x_n$ y que satisface: $P(x_k) = f_k$, para $k = 0, 1, \ldots, n$. Las abscisas x_k en este contexto son llamadas *puntos de interpolación* (o *puntos de coincidencia de*

la interpolación o *nodos*) y las ordenadas f_k son *los valores interpolados* (Figura 3).

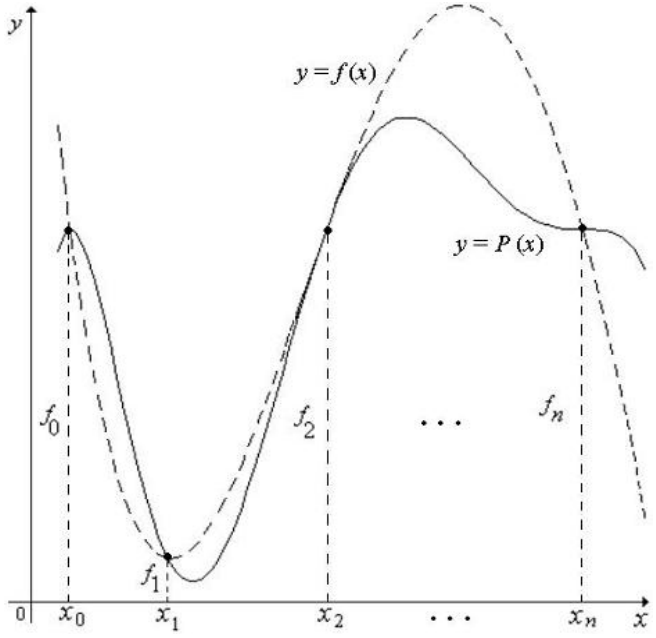


Figura 3

Supongamos que existe un polinomio $P(x)$ de la forma $P(x) = a_0 + a_1x + \ldots + a_nx^n$ que satisface las restricciones impuestas $P(x_k) = f_k$, para $k = 0, 1, \ldots, n$ (es decir que suponemos que el polinomio de grado a lo más n que pasa a través de los $(n + 1)$ pares de datos se puede expresar mediante potencias de x, donde $a_0, a_1, \ldots, a_n$ son coeficientes). El ajuste de la serie de potencias a los $(n + 1)$ pares de datos dados da un sistema de ecuaciones lineales

$$a_0 + a_1x_0 + a_2x_0^2 + \ldots + a_nx_0^n = f_0$$
$$a_0 + a_1x_1 + a_2x_1^2 + \ldots + a_nx_1^n = f_1$$
$$\vdots$$
$$a_0 + a_1x_n + a_2x_n^2 + \ldots + a_nx_n^n = f_n$$

o escrito matricialmente

$$\begin{pmatrix} 1 & x_0 & x_0^{\,2} & \cdots & x_0^{\,n} \\ 1 & x_1 & x_1^{\,2} & \cdots & x_1^{\,n} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^{\,2} & \cdots & x_n^{\,n} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \end{pmatrix}$$

que tiene una solución única para cualquier elección de f_0, f_1, ... , f_n (Teorema). (La matriz de los coeficientes se conoce como *matriz de Vandermonde*). Por consiguiente, el determinante de la matriz de Vandermonde es distinto de cero para los distintos nodos x_0, x_1, …, x_n. Esta solución única nos dará valores para los coeficientes a_0, a_1, ... , a_n con los que construiremos un polinomio de la forma

$$P(x) = a_0 + a_1x + ... + a_nx^n$$

Aunque los coeficientes pueden determinarse resolviendo las ecuaciones simultáneas por medio de un programa computacional, dicho intento no es deseable por dos razones:

1. Se necesita un programa que resuelva un conjunto de ecuaciones lineales.
2. La solución de la computadora quizá no sea precisa (realmente las potencias de x_k pueden ser números muy grandes, y si es así, el efecto de los errores por redondeo será importante).

Por fortuna, existen mejores métodos para determinar una interpolación polinomial sin resolver las ecuaciones lineales, como *la fórmula de interpolación de Newton*, *la fórmula de interpolación de Lagrange*, entre otras.

Es importante que tengamos presente que existe uno y sólo un polinomio de grado máximo n asociado con los datos (suponiendo, claro está, que las $(n + 1)$ abscisas son distintas). Sin embargo, existe ciertamente la posibilidad de expresar este polinomio de maneras distintas y de llegar a él a través de distintos algoritmos, con sus ventajas y sus desventajas. Dichos algoritmos generan el mismo polinomio en formas diferentes. Es decir que P como función o sea como conjunto de pares ordenados $(x, P(x))$ es único. Lo que hay son varias formas de representar a P por una fórmula explícita. Cada fórmula sugiere cierto algoritmo para calcular P e independientemente de la fórmula de interpolación, todas las interpolaciones polinomiales que se ajustan a los mismos datos son matemáticamente idénticas.

Observemos que la interpolación polinomial no proporciona siempre una solución satisfactoria con la exactitud prefijada al problema de aproximación propuesto, ya que la coincidencia de los valores de la función

$f(x)$ con los valores del polinomio $P(x)$ aun en el caso de que los puntos de interpolación o nodos x_k y x_{k+1} estén próximos, no garantiza una desviación $f(x) - P(x)$ pequeña en el intervalo $[x_k, x_{k+1}]$ (ver Figura 4).

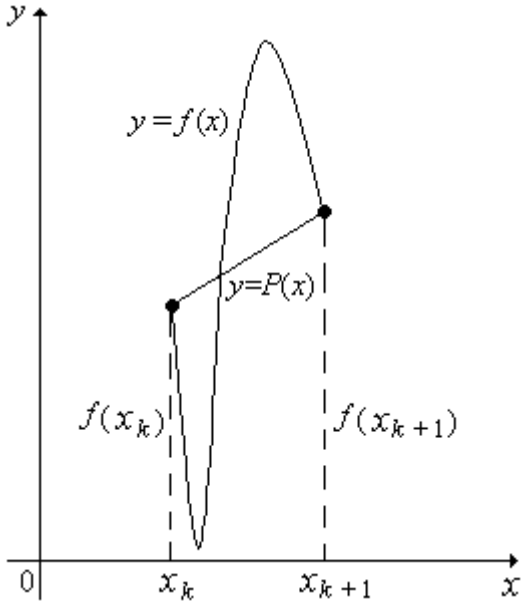


Figura 4

En resumen, la interpolación polinomial consiste en determinar el único polinomio de grado a lo más n que se ajusta a los $n + 1$ pares de datos dados. Este polinomio proporciona una fórmula para estimar valores intermedios entre valores conocidos.

Se discutirán varios algoritmos para la construcción del polinomio de interpolación, los cuales se engloban básicamente en dos clases:

1. *f* entra a través de su valor en un punto y por diferencias.
2. *f* entra a través de sus valores (u ordenadas) en todos los puntos de interpolación.

Para deducir los algoritmos que se incluyen dentro de la clase 1 necesitamos definir previamente los conceptos de *diferencias divididas*, *diferencias finitas* y *diferencias centrales*.

4.3.1. Diferencias divididas.

Supongamos dados los puntos $(x_i, f(x_i))$ para $i = 0, 1, ..., n$. La primera diferencia dividida de la función f se define como sigue

$$f(x_0x_1)=\frac{f(x_1)-f(x_0)}{x_1-x_0}$$

$$f(x_1x_2)=\frac{f(x_2)-f(x_1)}{x_2-x_1}$$

$$\vdots$$

$$f(x_{n-1}x_n)=\frac{f(x_n)-f(x_{n-1})}{x_n-x_{n-1}}$$

Análogamente, la segunda diferencia dividida se define

$$f(x_0x_1x_2)=\frac{f(x_1x_2)-f(x_0x_1)}{x_2-x_0}$$

$$f(x_1x_2x_3)=\frac{f(x_2x_3)-f(x_1x_2)}{x_3-x_1}$$

$$\vdots$$

$$f(x_{n-2}x_{n-1}x_n)=\frac{f(x_{n-1}x_n)-f(x_{n-2}x_{n-1})}{x_n-x_{n-2}}$$

En general, la *n*-ésima diferencia dividida, por definición, es

$$f(x_0x_1x_2\ldots x_n)=\frac{f(x_1x_2\ldots x_n)-f(x_0x_1\ldots x_{n-1})}{x_n-x_0}$$

En esta definición se observa que los (n - 1) argumentos del primer término del segundo miembro coinciden con los (n - 1) últimos argumentos del segundo término, y el denominador es la diferencia entre aquellos valores que no son comunes en los dos términos.

Notación. $f(x_0\, x_1 \ldots x_n) = f[x_0\, x_1 \ldots x_n] = f[x_0, x_1, \ldots, x_n]$

Dada una tabla con los valores $(x_i, f(x_i))$, podemos construir con ella un cuadro sinóptico o tabla de las diferencias divididas, como se muestra a continuación

x	$f(x)$	Primeras diferencias divididas	Segundas diferencias divididas	Terceras diferencias divididas	...
x_0	f_0				
		$f(x_0\, x_1)$			
x_1	f_1		$f(x_0\, x_1\, x_2)$		
		$f(x_1\, x_2)$		$f(x_0\, x_1\, x_2\, x_3)$	...
x_2	f_2		$f(x_1\, x_2\, x_3)$		
		$f(x_2\, x_3)$			
x_3	f_3			$\vdots$	
			$\vdots$		
$\vdots$	$\vdots$	$\vdots$			
				$f(x_{n-3}\, x_{n-2}\, x_{n-1}\, x_n)$	...
x_{n-1}	f_{n-1}		$f(x_{n-2}\, x_{n-1}\, x_n)$		
		$f(x_{n-1}\, x_n)$			
x_n	f_n				

La información a la izquierda de la línea vertical ya se conoce, mientras que las cantidades a la derecha se deben calcular (según las expresiones dadas antes).

Ejemplo 3. Dados los valores $(x_i,\ f(x_i))$ que se muestran a continuación, construyamos el cuadro de las diferencias divididas correspondiente.

x_i	$f(x_i)$	$f(x_i\, x_{i+1})$	$f(x_i\, x_{i+1}\, x_{i+2})$	$f(x_i\, x_{i+1}\, x_{i+2}\, x_{i+3})$	$f(x_i\, x_{i+1}\, x_{i+2}\, x_{i+3}\, x_{i+4})$
2	31				
		15			
0	1		4		
		3		3	
-1	-2		13		0
		68		3	
5	406		25		0
		193		3	
4	213		46		
		239			
6	691				

Se observa que los valores x_i han sido dados en un orden cualquiera. Si damos los valores en forma creciente, el cuadro de las diferencias es el siguiente

x_i	f_i	Primera	Segunda	Tercera	Cuarta
-1	-2				
		3			
0	1		4		
		15		3	
2	31		19		0
		91		3	
4	213		34		0
		193		3	
5	406		46		
		285			
6	691				

Se observa, además, que no es necesario que los datos estén igualmente espaciados.

4.3.1.1. Propiedades de las diferencias divididas.

La diferencia dividida es independiente del orden en que se tomen los valores de las abscisas, es decir, la diferencia dividida es una función simétrica de sus argumentos.

En efecto, para probar la simetría de la primera diferencia dividida debemos probar que $f(x_{n-1}\ x_n) = f(x_n\ x_{n-1})$, lo cual resulta evidente de la definición, pues

$$f(x_{n-1}x_n) = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} = \frac{f(x_{n-1}) - f(x_n)}{x_{n-1} - x_n} = f(x_n x_{n-1})$$

pero también puede observarse partiendo del hecho que

$$f(x_n x_{n-1}) = \frac{f(x_{n-1})}{x_{n-1} - x_n} + \frac{f(x_n)}{x_n - x_{n-1}}$$

y

$$f(x_{n-1}x_n) = \frac{f(x_n)}{x_n - x_{n-1}} + \frac{f(x_{n-1})}{x_{n-1} - x_n}$$

o sea que al intercambiar x_n con x_{n-1} y $f(x_n)$ con $f(x_{n-1})$ simplemente se invierte el orden de los dos términos a la derecha. En particular, $f(x_0\ x_1) = f(x_1\ x_0)$.

En el ejemplo anterior: $f(2\ \ 0) = f(0\ \ 2) = 15$, $f(0\ \ -1) = f(-1\ \ 0) = 3$.

Este procedimiento puede aplicarse a diferencias de orden más alto. Tomemos la segunda diferencia dividida $f(x_0\ x_1\ x_2)$:

$$f(x_0x_1x_2) = \frac{f(x_1x_2) - f(x_0x_1)}{x_2 - x_0} = \frac{1}{x_2 - x_0}\left[\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}\right] =$$
$$= \frac{1}{x_2 - x_0}\left[\frac{x_1f(x_2) - x_1f(x_1) - x_0f(x_2) + x_0f(x_1) - x_2f(x_1) + x_2f(x_0) + x_1f(x_1) - x_1f(x_0)}{(x_2 - x_1)(x_1 - x_0)}\right] =$$
$$= \frac{1}{x_2 - x_0}\left[\frac{f(x_0)(x_2 - x_1) + f(x_1)(x_0 - x_2) + f(x_2)(x_1 - x_0)}{(x_2 - x_1)(x_1 - x_0)}\right] =$$
$$= \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)}.$$

Intercambiando cualesquiera dos argumentos x_j y x_k y los valores f_j y f_k correspondientes, sólo se intercambian en estas condiciones los términos que contienen a f_j y f_k a la derecha dejando que el resultado no sufra cambio. Puesto que cualquier permutación de los argumentos x_k puede ser afectada por intercambios sucesivos de pares, la diferencia dividida es invariante bajo las permutaciones (tanto de los números x_k como de los f_k).

Así, $f(x_0\ x_1\ x_2) = f(x_2\ x_1\ x_0) = f(x_0\ x_2\ x_1) = f(x_2\ x_0\ x_1) = ...$, en otros términos, la segunda diferencia dividida es una función simétrica de sus argumentos.

Los resultados anteriores sugieren que para cualquier entero positivo n se puede escribir

$$f(x_0x_1...x_n) = \frac{f_0}{(x_0 - x_1)(x_0 - x_2)...(x_0 - x_n)} + \frac{f_1}{(x_1 - x_0)(x_1 - x_2)...(x_1 - x_n)} + ... +$$
$$+ \frac{f_n}{(x_n - x_0)(x_n - x_1)...(x_n - x_{n-1})} \qquad (1)$$

donde el coeficiente de f_i es

$$C_i^{(n)} = \frac{1}{(x_i - x_0)(x_i - x_1)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_n)}, \quad i = 0, 1, ..., n \quad (2)$$

expresiones estas que generalizan los resultados previos.

La prueba es por inducción. Ya tenemos probado este resultado para n =1 y 2. Supongámoslo cierto para $n = k$. Entonces, por definición

$$f(x_0 x_1 ... x_{k+1}) = \frac{1}{x_{k+1} - x_0}\left[f(x_1 x_2 ... x_{k+1}) - f(x_0 x_1 ... x_k)\right]. \qquad (3)$$

Puesto que hemos supuesto verdadero nuestro resultado para diferencias de orden k, el coeficiente de f_i a la derecha para $i = 1, 2, ..., k$, será

$$\frac{1}{(x_{k+1} - x_0)}\left[\frac{1}{(x_i - x_1)(x_i - x_2)...(x_i - x_{k+1})} - \frac{1}{(x_i - x_0)(x_i - x_1)...(x_i - x_k)}\right]$$

donde se entiende que el factor $(x_i - x_i)$ no se incluye en los productos del denominador. Pero este coeficiente se reduce a

$$\frac{1}{(x_{k+1} - x_0)}\left[\frac{x_i - x_0 - x_i + x_{k+1}}{(x_i - x_0)(x_i - x_1)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_k)(x_i - x_{k+1})}\right] = \qquad (4)$$

$$= \frac{1}{(x_i - x_0)(x_i - x_1)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_k)(x_i - x_{k+1})} = C_i^{(k+1)}$$

para $i = 1, 2, ..., k$, como se quería.

Para $i = 0$ ó $i = k + 1$ el coeficiente de f_i queda de una pieza en vez de dos, pero en ambos casos se observa fácilmente que será el que exige la ecuación (2) con $n = k + 1$; esto es

$$\frac{1}{(x_0 - x_1)(x_0 - x_2)...(x_0 - x_{k+1})} = C_0^{(k+1)}, \qquad \frac{1}{(x_{k+1} - x_0)(x_{k+1} - x_1)...(x_{k+1} - x_k)} = C_{k+1}^{(k+1)}$$

pues de la ecuación (3), para $i = 0$ el coeficiente de f_i es

$$\frac{1}{(x_{k+1} - x_0)}\left[\frac{-1}{(x_0 - x_1)(x_0 - x_2)...(x_0 - x_k)}\right] = \frac{1}{(x_0 - x_1)(x_0 - x_2)...(x_0 - x_k)(x_0 - x_{k+1})}$$

y para $i = k + 1$ el coeficiente de f_i es

$$\frac{1}{(x_{k+1} - x_0)}\left[\frac{1}{(x_{k+1} - x_1)(x_{k+1} - x_2)...(x_{k+1} - x_k)}\right] = \frac{1}{(x_{k+1} - x_0)(x_{k+1} - x_1)...(x_{k+1} - x_k)}.$$

Esto completa la inducción. Así, las ecuaciones (1) y (2) son ciertas para cualquier entero positivo n.

Esto prueba que la n-ésima diferencia dividida es simétrica, pues del resultado anterior se desprende que si cualquier par de argumentos se intercambia, digamos x_j y x_k, los términos que incluyen a f_j y a f_k a la derecha se intercambian y no hay ningún otro cambio.

Por lo tanto, $f(x_0\, x_1 \ldots x_n)$ puede expresarse como diferencia entre dos diferencias divididas de orden $(n - 1)$ teniendo cada una de ellas, de las n abscisas, $(n - 1)$ comunes y dividiendo esta diferencia por la resta de las abscisas que no son comunes.

Así, por ejemplo,

$$f(x_0x_1x_2x_3x_4) = \frac{f(x_1x_2x_3x_4) - f(x_0x_1x_2x_3)}{x_4 - x_0} = \frac{f(x_0x_2x_3x_4) - f(x_0x_1x_3x_4)}{x_2 - x_1} = \\ = \frac{f(x_1x_2x_3x_4) - f(x_0x_2x_3x_4)}{x_1 - x_0}.$$

La ecuación (1) puede escribirse en forma sintética del siguiente modo

$$f(x_0x_1 \ldots x_n) = \sum_{i=0}^{n} \frac{f(x_i)}{\prod_{\substack{j=0 \\ j \neq i}}^{n} (x_i - x_j)}$$

donde $\prod_{\substack{j=0 \\ j \neq i}}^{n} (x_i - x_j)$ es la productoria o producto de los factores

$$(x_i - x_0)\,(x_i - x_1) \ldots (x_i - x_{i-1})\,(x_i - x_{i+1}) \ldots (x_i - x_n)$$

Si nuestra función $f(x)$ es, en particular, un polinomio de grado n, siendo por definición $f(x_0x) = \frac{f(x) - f(x_0)}{x - x_0}$, tenemos que en tal caso la primera diferencia dividida es un polinomio de grado n - 1.

El mismo argumento puede repetirse para la segunda diferencia dividida. En efecto, $f(x_0x_1x) = \frac{f(x_1x) - f(x_0x_1)}{x - x_0}$, y como el numerador es un polinomio de grado n - 1, entonces la segunda diferencia dividida es un polinomio de grado n -2.

Continuando en esta forma, la n-ésima diferencia dividida es un polinomio de grado 0 (o sea, una constante) y las diferencias divididas de orden mayores que n son todas nulas.

En el ejemplo 3, la tercera diferencia dividida es constante y, por lo tanto, $f(x)$ es un polinomio de grado 3.

Ejemplo 4. Sea $f(x) = x^n$, $n \in N$. Entonces, de la definición de primera diferencia dividida de la función f, se tiene

$$f(x_0 x) = \frac{x^n - x_0^n}{x - x_0} = x^{n-1} + x^{n-2}x_0 + x^{n-3}x_0^2 + \ldots + x^2 x_0^{n-3} + x x_0^{n-2} + x_0^{n-1}$$

de donde, la primera diferencia dividida es una función homogénea en x_0 x de grado n - 1. Análogamente,

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

$$= \frac{1}{x - x_0}\left[\left(x^{n-1} + x^{n-2}x_1 + \ldots + x x_1^{n-2} + x_1^{n-1}\right) - \left(x_1^{n-1} + x_1^{n-2}x_0 + \ldots + x_1 x_0^{n-2} + x_0^{n-1}\right)\right] =$$

$$= \frac{x^{n-1} - x_0^{n-1}}{x - x_0} + x_1 \frac{x^{n-2} - x_0^{n-2}}{x - x_0} + x_1^2 \frac{x^{n-3} - x_0^{n-3}}{x - x_0} + \ldots + x_1^{n-2} \frac{x - x_0}{x - x_0} =$$

$$= x^{n-2} + x^{n-3}x_0 + x^{n-4}x_0^2 + \ldots + x^2 x_0^{n-4} + x x_0^{n-3} + x_0^{n-2} + x_1\left(x^{n-3} + x^{n-4}x_0 + \ldots + x_0^{n-3}\right) +$$
$$+ x_1^2\left(x^{n-4} + x^{n-5}x_0 + \ldots + x_0^{n-4}\right) + \ldots + x_1^{n-2}$$

de modo que la segunda diferencia dividida $f(x_0\ x_1\ x)$ es una función homogénea de grado n - 2 en $x_0\ x_1\ x$. En general, $f(x_0\ x_1\ \ldots\ x_{k-1}\ x)$ será una función homogénea de grado n - k en $x_0\ x_1\ \ldots\ x_{k-1}\ x$ y haciendo $n = k$, $f(x_0\ x_1\ \ldots\ x_{n-1}\ x)$, es decir, la n-ésima diferencia dividida será de grado 0, esto es, una constante.

Como corolario podemos afirmar que las diferencias divididas de orden $(n+1)$ de x^n son ceros.

Observación. A menudo, las diferencias divididas de orden superior tienden a cero pero no llegan a valer exactamente cero y esto se debe, por lo general, a errores de redondeo en los datos.

4.3.1.2 Fórmula fundamental de interpolación de Newton con diferencias divididas.

Sea f definida sobre algún intervalo [a, b]. Dados $(n + 1)$ puntos distintos de [a, b], x_0, x_1, ..., x_n y los correspondientes valores $f(x_0)$, $f(x_1)$, ..., $f(x_n)$ de la función $f(x)$ que suponemos con derivada continua hasta el orden $(n + 1)$ en [a, b], nos proponemos determinar el polinomio de interpolación

(colocación) de Newton con diferencias divididas que nos permitirá calcular valores de la función $f(x)$ en puntos x distintos de $x_0, x_1, ..., x_n$. (Sabemos que hay un único polinomio P de grado máximo n que interpola a f en los $(n + 1)$ puntos o nodos: $P(x_k) = f_k$, $0 \leq k \leq n$).

De la definición de diferencia dividida

$$f(xx_0) = \frac{f(x_0) - f(x)}{x_0 - x}$$

De aquí, deducimos que

$$f(x) = f(x_0) + (x - x_0) f(x\, x_0) \tag{5}$$

Siendo además por definición

$$f(xx_0x_1) = \frac{f(x_0x_1) - f(xx_0)}{x_1 - x}$$

surge que

$$f(x\, x_0) = f(x_0\, x_1) + (x - x_1) f(x\, x_0\, x_1) \tag{6}$$

Reemplazando $f(x\, x_0)$ en la ecuación (5) nos da

$$f(x) = f(x_0) + (x - x_0) f(x_0\, x_1) + (x - x_0)\,(x - x_1) f(x\, x_0\, x_1) \tag{7}$$

La interpretación de la ecuación (7) nos dice que la función $f(x)$ es igual a un polinomio que pasa por los puntos $(x_0, f(x_0))$, $(x_1, f(x_1))$ (tomando $f(x_0x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$) más un error que está dado por la expresión

$$E_1(x) = (x - x_0)\,(x - x_1) f(x\, x_0\, x_1)$$

Se observa que $P_{01}(x) = f(x_0) + (x - x_0) f(x_0\, x_1)$ es una recta que pasa por los puntos señalados, pues

$$P_{01}(x) = f(x_0) + (x - x_0)\frac{f(x_1) - f(x_0)}{x_1 - x_0} \quad \text{y} \quad P_{01}(x_0) = f(x_0),\ P_{01}(x_1) = f(x_1)$$

Cuando reemplazamos el valor de la función por el valor del polinomio en el intervalo (x_0, x_1) hacemos interpolación lineal. La aproximación $f(x) \approx P_{01}(x)$ es exacta para todos los valores de x si y sólo si

$f(x)$ es una función lineal, o sea si $f(x) = mx + n$. Para cualquier otra función $f(x)$ la aproximación es exacta en $x = x_0$ y en $x = x_1$, puesto que $E_1(x_0) = 0$ y $E_1(x_1) = 0$.

Introduciendo la abscisa x_2, podemos escribir por definición

$$f(xx_0x_1x_2) = \frac{f(x_0x_1x_2) - f(xx_0x_1)}{x_2 - x}$$

surge de aquí que

$$f(x\,x_0\,x_1) = f(x_0\,x_1\,x_2) + (x - x_2)\,f(x\,x_0\,x_1\,x_2) \tag{8}$$

Reemplazando la ecuación (8) en la (7) nos da

$$f(x) = f(x_0) + (x\text{-}x_0)f(x_0\,x_1) + (x\text{-}x_0)(x\text{-}x_1)f(x_0\,x_1\,x_2) + (x\text{-}x_0)(x\text{-}x_1)(x\text{-}x_2)f(x\,x_0\,x_1\,x_2) \tag{9}$$

La ecuación (9) nos dice que la función $f(x)$ es igual a un polinomio

$$P_{012}(x) = f(x_0) + (x\text{-}x_0)\,f(x_0\,x_1) + (x\text{-}x_0)\,(x\text{-}x_1)\,f(x_0\,x_1\,x_2)$$

más un error

$$E_2(x) = (x\text{-}x_0)\,(x\text{-}x_1)(x\text{-}x_2)\,f(x\,x_0\,x_1\,x_2)$$

Se observa además que $P_{012}(x)$ es un polinomio de orden 2 que pasa por los puntos $(x_0, f(x_0))$, $(x_1, f(x_1))$, $(x_2, f(x_2))$, pues

$$P_{012}(x) = f(x_0) + (x - x_0)\frac{f(x_1) - f(x_0)}{x_1 - x_0} + (x - x_0)(x - x_1)\left[\frac{f(x_2) - f(x_0)}{(x_2 - x_1)(x_2 - x_0)} + \frac{f(x_0) - f(x_1)}{(x_1 - x_2)(x_0 - x_1)}\right]$$

(en el último término de esta expresión usamos que

$$f(x_0x_1x_2) = f(x_1x_0x_2) = \frac{f(x_0x_2) - f(x_1x_0)}{x_2 - x_1} = \frac{f(x_2) - f(x_0)}{(x_2 - x_1)(x_2 - x_0)} + \frac{f(x_0) - f(x_1)}{(x_1 - x_2)(x_0 - x_1)})$$

Luego, $P_{012}(x_0) = f(x_0)$, $P_{012}(x_1) = f(x_1)$, $P_{012}(x_2) = f(x_2)$.

La aproximación $f(x) \approx P_{012}(x)$ será exacta para todo valor de x si y sólo si $f(x)$ es un polinomio de segundo grado, es decir si $f(x) = mx^2 + nx + p$, y para cualquier función $f(x)$ la aproximación es exacta en x_0, x_1, x_2, puesto que el error $E_2(x)$ se anula en x_0, x_1, x_2.

Introduciendo las sucesivas abscisas y haciendo las sustituciones correspondientes, podemos generalizar la ecuación (9) y escribir

$$f(x) = f(x_0) + (x\text{-}x_0) f(x_0\, x_1) + (x\text{-}x_0)\,(x\text{-}x_1) f(x_0\, x_1\, x_2) + (x\text{-}x_0)\,(x\text{-}x_1)(x\text{-}x_2) f(x_0\, x_1\, x_2\, x_3)$$
$$+ \ldots + (x\text{-}x_0)(x\text{-}x_1) \ldots (x\text{-}x_{n\text{-}1}) f(x_0\, x_1 \ldots x_n) + (x\text{-}x_0)(x\text{-}x_1)\ldots(x\text{-}x_n) f(x\, x_0\, x_1 \ldots x_n) \qquad (10)$$

La ecuación (10) se conoce con el nombre de *fórmula de interpolación de Newton con diferencias divididas*. Se puede aplicar a puntos que tengan igual o distinta separación y pueden estar dados en cualquier orden. En esta fórmula la función $f(x)$ se reemplaza por un polinomio de grado máximo n que pasa por los puntos $(x_i, f(x_i))$ para $i = 0, 1,..., n$, más un término error. Se puede expresar la ecuación (10) escribiendo sintéticamente

$$f(x) = \sum_{i=0}^{n} \left[\prod_{s=0}^{i-1} (x - x_s) f(x_0 x_1 \ldots x_i) \right] + \prod_{s=0}^{n} (x - x_s) f(x x_0 x_1 \ldots x_n) \qquad (11)$$

Para que la ecuación (11) sea válida debemos convenir que

$$\prod_{s=0}^{-1} (x - x_s) = 1.$$

La ecuación (10) puede también obtenerse fácilmente sin más que observar que por definición valen las igualdades

$$f(x) = f(x_0) + (x\text{-}x_0) f(x\, x_0)$$
$$f(x\, x_0) = f(x_0\, x_1) + (x\text{-}x_1) f(x\, x_0\, x_1)$$
$$f(x\, x_0\, x_1) = f(x_0\, x_1\, x_2) + (x\text{-}x_2) f(x\, x_0\, x_1\, x_2)$$
$$f(x\, x_0\, x_1\, x_2) = f(x_0\, x_1\, x_2\, x_3) + (x\text{-}x_3) f(x\, x_0\, x_1\, x_2\, x_3)$$
$$\cdot$$
$$\cdot$$
$$\cdot$$
$$f(x\, x_0\, x_1\, x_2 \ldots x_{n\text{-}1}) = f(x_0\, x_1 \ldots x_n) + (x\text{-}x_n) f(x\, x_0\, x_1 \ldots x_n)$$

Multiplicando la segunda igualdad por $(x\text{-}x_0)$, la tercera por $(x\text{-}x_0)\,(x\text{-}x_1)$, la cuarta por $(x\text{-}x_0)(x\text{-}x_1)(x\text{-}x_2)$, y así siguiendo hasta llegar a la última igualdad que la multiplicamos por $(x\text{-}x_0)(x\text{-}x_1) \ldots (x\text{-}x_{n\text{-}1})$ y sumando, obtenemos

$$f(x) = f(x_0) + (x\text{-}x_0) f(x_0\, x_1) + (x\text{-}x_0)\,(x\text{-}x_1) f(x_0\, x_1\, x_2) + (x\text{-}x_0)\,(x\text{-}x_1)(x\text{-}x_2) f(x_0\, x_1\, x_2\, x_3)$$
$$+ \ldots + (x\text{-}x_0)(x\text{-}x_1) \ldots (x\text{-}x_{n\text{-}1}) f(x_0\, x_1 \ldots x_n) + (x\text{-}x_0)(x\text{-}x_1)\ldots(x\text{-}x_n) f(x\, x_0\, x_1 \ldots x_n)$$

Ejemplo 5. Ilustraremos el uso de la ecuación (10).

x_i	f_i	Primera	Segunda	Tercera	Cuarta
-2	64				
		-21			
1	1		23		
		94		16	
3	189		119		3
		451		40	
4	640		319		3
		1408		58	
6	3456		551		
		3061			
7	6517				

Calcularemos el valor de la función en $x = 2$. Elegimos las abscisas del siguiente modo

$$x_0 = -2,\ x_1 = 1,\ x_2 = 3,\ x_3 = 4,\ x_4 = 6$$

y reemplazando en la ecuación (10), obtenemos

$$P_4(2) = 64 + (2+2)(-21) + (2+2)(2-1)23 + (2+2)(2-1)(2-3)16 + (2+2)(2-1)(2-3)(2-4)3 = 32$$

Análogamente, para $x = -1.5$ y $x = 5$ con la misma elección de abscisas anterior, obtenemos

$$P_4(-1.5) = 64+0.5(-21)+0.5(-2.5)23+0.5(-2.5)(-4.5)16+0.5(-2.5)(-4.5)(-5.5)3 = 21.9375$$

$$P_4(5) = 64+(5+2)(-21)+(5+2)(5-1)23+(5+2)(5-1)(5-3)16+(5+2)(5-1)(5-3)(5-4)3 = 1625$$

Si para $x = 2$ consideramos las abscisas en el siguiente modo

$$x_0 = 1,\ x_1 = 3,\ x_2 = 4,\ x_3 = 6,\ x_4 = 7$$

obtenemos

$$P_4(2) = 1 + (2-1)94 + (2-1)(2-3)119 + (2-1)(2-3)(2-4)40 + (2-1)(2-3)(2-4)(2-6)3 = 32$$

Si para $x = 5$ consideramos las abscisas en el siguiente modo

$$x_0 = 4,\ x_1 = 6,\ x_2 = 3,\ x_3 = 7,\ x_4 = 1$$

obtenemos

$$P_4(5) = 640+(5-4)1408+(5-4)(5-6)319+(5-4)(5-6)(5-3)58+(5-4)(5-6)(5-3)(5-7)3 = 1625$$

Si para $x = 6.5$ la elección de las abscisas es

$$x_0 = 7,\ x_1 = 6,\ x_2 = 4,\ x_3 = 3,\ x_4 = 1$$

obtenemos

$$P_4(6.5) = 6517 + (6.5-7)\ 3061 + (6.5-7)\ (6.5-6)\ 551 + (6.5-7)\ (6.5-6)\ (6.5-4)\ 5\ 8 +$$

$$+ (6.5-7)\ (6.5-6)\ (6.5-4)(6.5-3)\ 3 = 4805.9375$$

Si para $x = 3.7$ la elección de las abscisas es

$$x_0 = 3,\ x_1 = 4,\ x_2 = 1,\ x_3 = 6,\ x_4 = -2$$

obtenemos

$$P_4(3.7) = 189 + (3.7-3)\ 451 + (3.7-3)\ (3.7-4)\ 119 + (3.7-3)\ (3.7-4)\ (3.7-1)\ 40 +$$

$$+ (3.7-3)\ (3.7-4)\ (3.7-1)(3.7-6)\ 3 = 460.9423$$

El ejemplo muestra que cuando se interpola en el comienzo de la tabla el recorrido de la misma se hace en diagonal descendente, cuando se interpola al final el recorrido se hace en diagonal ascendente, y cuando se interpola en el medio el recorrido se hace en zig-zag.

En general, conviene elegir como origen el punto tabulado más próximo al valor a interpolar. Este ejemplo muestra que la cuarta diferencia dividida es constante e igual a 3, por consiguiente, la función tabulada es un polinomio de orden o grado 4. En este caso particular, no tiene importancia la elección del origen ya que no se comete ningún error al interpolar. En este ejemplo vimos que obteníamos los mismos resultados interpolando en $x = 2$ y $x = 5$ eligiendo distintos orígenes.

Si se desea conocer la expresión analítica del polinomio tabulado en este ejemplo, podemos escribir

$$f(x) = 64 + (x+2)(-21) + (x+2)(x-1)23 + (x+2)(x-1)(x-3)16 + (x+2)(x-1)(x-3)(x-4)3 =$$

$$= (((3(x-4) + 16)(x-3) + 23)(x-1) -21)(x+2) + 64 = 3x^4 - 2x^3$$

Observación. Si se tiene un polinomio cualquiera, por ejemplo

$$a_1x^6 + a_2x^5 + a_3x^4 + a_4x^3 + a_5x^2 + a_6x + a_7$$

y se desea calcular dicho polinomio para distintos valores de x, podemos escribir el polinomio en la forma siguiente

$$(((((a_1x + a_2)x + a_3)x + a_4)x + a_5)x + a_6)x + a_7$$

Se observa que los paréntesis están balanceados; existen, en este caso particular, cinco paréntesis de apertura o izquierdo y cinco paréntesis de cierre o derechos. Si el cálculo para los diferentes valores de x lo hacemos con el polinomio escrito en la forma original, realizamos 26 multiplicaciones más 6 sumas. Escribiéndolo en la forma indicada sólo realizamos 6 multiplicaciones y 6 sumas. Este recurso es muy usado en computación cuando se trata de tabular series de potencias o polinomios de grado elevado, ya que además del ahorro considerable de operaciones evita las condiciones de underflow y overflow.

4.3.1.3. El error de la fórmula de interpolación.

La ecuación (10) nos dice que la función $f(x)$ es aproximada por un polinomio $P_n(x)$ de grado n como máximo. El valor del polinomio $P_n(x)$ coincide con el valor de la función en las $n + 1$ abscisas $x_0, x_1, ..., x_n$. El error de una interpolación polinomial lo definimos como $E(x) = f(x) - P_n(x)$, y según la ecuación (11) está dado por

$$E(x) = \prod_{s=0}^{n} (x - x_s) f(x x_0 x_1 ... x_n) \qquad (12)$$

Hemos supuesto que $f(x)$ tiene derivadas continuas en el intervalo considerado hasta el orden $(n + 1)$, es decir, existe $f^{(n+1)}(x)$. Mostraremos ahora que para un valor x comprendido entre x_0 y x_n

$$f(x x_0 x_1 ... x_n) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \quad (\xi \text{ dependiendo de } x, x_0, x_1, x_2, ..., x_n)$$

Observando la ecuación (12) se comprueba que $E(x)$ se anula en los $(n+1)$ puntos de interpolación $x_0, x_1, ..., x_n$, de donde, por el teorema de Rolle la derivada primera $E^{(1)}(x)$ tiene entonces garantizados n ceros distintos en el intervalo comprendido entre x_0 y x_n; la derivada segunda $E^{(2)}(x)$ tiene entonces garantizados $(n - 1)$ ceros distintos en el intervalo comprendido entre x_0 y x_n. Al continuar aplicando el teorema de Rolle en esta forma, se llegará al final que la derivada n-ésima $E^{(n)}(x)$ en el intervalo comprendido entre x_0 y x_n, tiene al menos un cero, digamos $x = \xi$; esto es, $E^{(n)}(\xi) = 0$ $(\xi \in \text{Int}(x_0, x_1, ..., x_n))$.

Derivando n veces la ecuación (10) y en dicha derivada n-ésima hacemos $x = \xi$, se obtiene

$$f^{(n)}(\xi) = n!\, f(x_0\, x_1 \ldots x_n) \tag{13}$$

donde $\xi \in \text{Int}\,(x_0, x_1, \ldots, x_n)$. Se observa que la derivada n-ésima de todos los términos del segundo miembro de la ecuación (10) se anulan salvo el $(n+1)$-ésimo que es un polinomio de grado n multiplicado por la diferencia dividida de orden n. Se sabe además que la derivada n-ésima de un polinomio de grado n es $n!$.

De la ecuación (13) se deduce

$$f\left(x_0 x_1 \ldots x_n\right) = \frac{f^{(n)}\left(\xi\right)}{n!}, \quad \xi \in \text{Int}\,(x_0, x_1, \ldots, x_n) \tag{14}$$

Puesto que la ecuación (14) es válida cualquiera sea n, podemos escribir

$$f\left(x x_0 x_1 \ldots x_n\right) = \frac{f^{(n+1)}\left(\xi\right)}{\left(n+1\right)!},\ \xi \text{ dependiendo de } x, x_0, x_1, \ldots, x_n \tag{15}$$

Reemplazando este valor en la ecuación (11), nos da

$$f\left(x\right) = \sum_{i=0}^{n}\left[\prod_{s=0}^{i-1}\left(x - x_s\right) f\left(x_0\, x_1 \ldots x_i\right)\right] + \prod_{s=0}^{n}\left(x - x_s\right)\frac{f^{(n+1)}\left(\xi\right)}{\left(n+1\right)!} \tag{16}$$

Para poder calcular el error cometido en una interpolación se requiere conocer la expresión analítica de $f(x)$. El punto ξ es desconocido, pero podemos hacer una estimación o acotación del mismo.

Ejemplo 6. Calculemos $ln(1.6)$ a partir de la siguiente tabla de diferencias divididas

x	$ln(1+x)$	Primera	Segunda	Tercera	Cuarta
0.4	0.3364722				
		0.689929			
0.5	0.4054651		-0.21371		
		0.625816		0.0823418	
0.7	0. 5306283		-0.1807733		-0.033617
		0.571584		0.0655333	
0.8	0.5877867		-0.15456		
		0.540672			
0.9	0.6418539				

Eligiendo como origen $x_0 = 0.7$, $x_1 = 0.8$, $x_2 = 0.5$, $x_3 = 0.9$, $x_4 = 0.4$, obtenemos

$$P_4(0.6) = 0.5306283 + (-0.1)\ 0.571584 + (-0.1)(-0.2)\ (-0.1807733) +$$

$$+ (-0.1)(-0.2)(0.1)\ 0.0655333 + (-0.1)(-0.2)(0.1)(-0.3)(-0.033617) = 0.4700057$$

Como en la tabla hemos interpolado haciendo uso hasta la diferencia dividida de cuarto orden, el error cometido está dado por

$$E(0.6) = \prod_{s=0}^{4}(x - x_s)\frac{f^{(5)}(\xi)}{5!} = (-0.1)(-0.2)(0.1)(-0.3)(0.2)\frac{f^{(5)}(\xi)}{5!}$$

siendo $f(x) = ln(1 + x)$ y

$$f^{(1)}(x) = \frac{1}{1+x}; f^{(2)}(x) = \frac{-1}{(1+x)^2}; f^{(3)}(x) = \frac{2}{(1+x)^3}; f^{(4)}(x) = \frac{-6}{(1+x)^4}; f^{(5)}(x) = \frac{24}{(1+x)^5}.$$

Esta derivada quinta debe calcularse en un punto ξ entre 0.4 y 0.9; como dicho punto es desconocido, la calculamos en los dos extremos del intervalo, es decir, en $x = 0.4$ y $x = 0.9$:

$$f^{(5)}(0.4) = \frac{24}{5.37824} = 4.462426369, \qquad f^{(5)}(0.9) = \frac{24}{24.76099} = 0.969266576$$

de modo que el error quedará acotado por

$$-0.0000045 \le E(0.6) \le -0.0000009$$

El error verdadero es $E(0.6) = f(0.6) - P_4(0.6) = 0.4700036 - 0.4700057 = -2.1 \times 10^{-6}$

El valor exacto de $f(0.6)$ se da sólo hasta la séptima decimal por lo que está sujeto a errores de redondeo de a lo más $\pm 5 \times 10^{-8}$. Por lo tanto, el error verdadero no tiene significado excepto para ilustrar el efecto del error de redondeo de una resta.

Así, concluimos que los resultados de la interpolación son exactos dentro de los errores de redondeo.

En este caso conocemos la expresión analítica de $f(x)$, pero si la desconociéramos o bien si sólo se sabe que los valores de la función responden a datos experimentales o también si las derivadas son difíciles de

calcular, entonces la ecuación (12) nos puede ser de utilidad para hacer una estimación del error. La ecuación (12) nos dice que

$$E(x) = \prod_{s=0}^{n} (x - x_s) f(xx_0x_1 \ldots x_n)$$

y según la ecuación (15)

$$f(xx_0x_1 \ldots x_n) = \frac{f^{(n+1)}(\xi)}{(n+1)!}, \qquad \xi \in \text{Int}\,(x_0, x_1, \ldots, x_n)$$

Así, si las derivadas son difíciles de calcular o bien desconocidas, entonces pueden ser estimadas por las diferencias de orden alto de acuerdo a la ecuación (15), donde uno pone $x = x_{n+1}$, un punto donde $f(x_{n+1})$ es conocida pero no es usada en la interpolación; o sea que la ecuación (15) sería puesta

$$f(x_0x_1 \ldots x_nx_{n+1}) \approx \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

y entonces, la ecuación (12) sería puesta

$$E(x) \approx \prod_{s=0}^{n} (x - x_s) f(xx_0x_1 \ldots x_{n+1})$$

Así, si la fórmula de interpolación es considerada como los $n + 1$ términos de una gran expansión (en donde el número de puntos de interpolación varía), entonces el término residuo (resto) es estimado por el primer término despreciado.

En resumen, como

$$P_n(x) = f(x_0) + (x - x_0) f(x_0\, x_1) + \ldots + (x - x_0)(x - x_1) \ldots (x - x_{n-1}) f(x_0\, x_1 \ldots x_n)$$

el error de esta ecuación es aproximadamente igual al término que se añadiría a la misma si la interpolación se extendiera para ajustarse a otro punto más, x_{n+1}; es decir, el error es

$$E(x) \approx (x-x_0)(x-x_1) \ldots (x-x_n) f(x_0\, x_1 \ldots x_{n+1})$$

En el ejemplo 6, usemos hasta la diferencia de orden 3 para obtener el polinomio $P_3(x)$ de grado 3 y la cuarta diferencia dividida usémosla para obtener una estimación del error cometido.

Así,

$$P_3(0.6) = 0.4699855$$

y

$$E(0.6) = \prod_{s=0}^{3}(x - x_s) f(x_0 x_1 x_2 x_3 x_4) = (-0.1)(-0.2)(0.1)(-0.3)(-0.033617) = 0.0000202$$

Para $x = 0.6$ el error verdadero es

$$E(0.6) = f(0.6) - P_3(0.6) = 0.4700036 - 0.4699855 = 0.0000181.$$

4.3.1.4. Diferencias divididas con abscisas repetidas.

En la definición de las diferencias divididas hemos supuesto que todas las abscisas son diferentes. Sin embargo, si la expresión $f(x_0\ x_0+\varepsilon)$ tiende a un límite cuando $\varepsilon \to 0$, a este límite lo designamos $f(x_0\ x_0)$. Esto es,

$$f(x_0 x_0) = \lim_{\varepsilon \to 0} f(x_0\ x_0 + \varepsilon)$$

Llamando $x = x_0 + \varepsilon$, entonces $x - x_0 = \varepsilon$. Si $\varepsilon \to 0$, se tiene que $x - x_0 \to 0$ o bien $x \to x_0$. Luego,

$$f(x_0 x_0) = \lim_{\varepsilon \to 0} f(x_0\ x_0 + \varepsilon) = \lim_{x \to x_0} f(x_0 x) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$$

En forma análoga podemos generalizar para las diferencias divididas de mayor orden, si observamos la ecuación (14) que volvemos a escribir

$$f(x_0 x_1 \ldots x_n) = \frac{f^{(n)}(\xi)}{n!}. \qquad (17)$$

Si en esta expresión suponemos que las abscisas x_0, x_1, ..., x_n son todas coincidentes con x_0, entonces se transforma (en el límite) en

$$f(\underbrace{x_0 x_0 \ldots x_0}_{n+1 \text{ abscisas}}) = \frac{f^{(n)}(x_0)}{n!}$$

de modo que

$$f(x_0 x_0 x_0) = \frac{f^{(2)}(x_0)}{2!}, \quad f(x_0 x_0 x_0 x_0) = \frac{f^{(3)}(x_0)}{3!}, \ldots$$

Teniendo presente la ecuación (17) y si todas las abscisas coinciden con x_0, la ecuación (10) se transforma en

$$f(x)=f(x_0)+(x-x_0)f^{(1)}(x_0)+(x-x_0)^2\frac{f^{(2)}(x_0)}{2!}+\ldots+(x-x_0)^n\frac{f^{(n)}(x_0)}{n!}+(x-x_0)^{n+1}\frac{f^{(n+1)}(\xi)}{(n+1)!},$$
$$x_0<\xi<x \qquad (18)$$

y esta es la fórmula de Taylor que nos permite calcular el valor de la función en un entorno del punto x_0 cuando se conoce el valor de la función en x_0 y el de las derivadas sucesivas en x_0.

Volvemos a considerar el ejemplo 5. Siendo constante la diferencia dividida cuarta y agregando los argumentos 7, 7, 7 podemos completar la tabla y, por lo tanto, calcular $\frac{f^{(3)}(7)}{3!}$, $\frac{f^{(2)}(7)}{2!}$ y $\frac{f^{(1)}(7)}{1!}$.

x	$f(x)$	Primera	Segunda	Tercera	Cuarta
-2	64				
		-21			
1	1		23		
		94		16	
3	189		119		3
		454		40	
4	640		319		3
		1408		58	
6	3456		551		3
		3061		70	
7	6517		761		3
		3822		79	
7			840		3
				82	
7					
7					

Si se agregan los argumentos 7, 7, 7, al ser constante la diferencia dividida de orden 4 completamos esa columna agregando 3, 3, 3.

El primer valor a calcular en la columna de la tercera diferencia dividida se obtiene fácilmente de

$$3 = \frac{x-58}{7-3}, \text{ de donde, } x = 70,$$

el segundo sale de

$$3 = \frac{x-70}{7-4}, \text{ de donde, } x = 79,$$

y finalmente, el tercero de

$$3 = \frac{x-79}{7-6}, \text{ de donde, } x = 82.$$

Este valor 82 es el valor de $f(7777)$ o lo que es lo mismo decir que

$$f(7\,7\,7\,7) = \frac{f^{(3)}(7)}{3!} = 82.$$

Análogamente, el primer valor a calcular en la columna de la segunda diferencia dividida se obtiene considerando que

$$70 = \frac{x-551}{7-4}, \text{ de donde, } x = 761.$$

El segundo valor se obtiene de

$$79 = \frac{x-761}{7-6}, \text{ de donde, } x = 840.$$

Este valor 840 es el valor de $f(777)$, o sea, $f(7\,7\,7) = \frac{f^{(2)}(7)}{2!} = 840.$

Para la columna de la primera diferencia dividida se hace

$$761 = \frac{x-3061}{7-6}, \text{ de donde, } x = 3822$$

y, por lo tanto,

$$f(7\,7) = \frac{f^{(1)}(7)}{1!} = 3822.$$

4.3.2. Interpolación con diferencia finita.

Hasta el momento hemos supuesto que las abscisas x_0, x_1, …, x_n están dadas en un orden cualquiera y, en general, no son equidistantes.

Suponemos ahora que los valores de la función están tabulados para abscisas que se acomodan consecutivamente e igualmente espaciadas, con paso h, de modo que si x_0 es un valor dado de x y h una longitud fija o paso constante, los restantes puntos de interpolación se designan por $x_r = x_0 + rh$, para $r = 1, 2, ..., n$. Los valores correspondientes de la función se designan por $f(x_r) = f_r$, $r = 0, 1, \ldots, n$. De acuerdo a esto, se tiene

$$x_0 \quad , \quad x_1 = x_0 + h \quad , \quad x_2 = x_0 + 2h \quad , \quad \ldots , \quad x_n = x_0 + nh$$

$$f(x_0) = f_0 \quad , \quad f(x_1) = f_1 \quad , \quad f(x_2) = f_2 \quad , \quad \ldots , \quad f(x_n) = f_n$$

4.3.2.1. Tabla de diferencias hacia adelante.

La diferencia hacia adelante de orden 1 de $f(x)$ en $x = x_r$ se indica con Δf_r y se define mediante

$$\Delta f_r = f_{r+1} - f_r = f(x_r + h) - f(x_r)$$

(en la última igualdad usamos: $x_{r+1} = x_0 + (r+1)h = x_0 + rh + h = x_r + h$).

Análogamente se define la diferencia hacia adelante de orden 2

$$\Delta^2 f_r = \Delta(\Delta f_r) = \Delta f_{r+1} - \Delta f_r = f_{r+2} - f_{r+1} - f_{r+1} + f_r = f_{r+2} - 2f_{r+1} + f_r = f(x_r + 2h) - 2f(x_r + h) + f(x_r)$$

Del mismo modo, la diferencia hacia adelante de orden 3 es

$$\Delta^3 f_r = \Delta(\Delta^2 f_r) = \Delta^2 f_{r+1} - \Delta^2 f_r = f_{r+3} - 2f_{r+2} + f_{r+1} - f_{r+2} + 2f_{r+1} - f_r = f_{r+3} - 3f_{r+2} + 3f_{r+1} - f_r =$$

$$= f(x_r + 3h) - 3f(x_r + 2h) + 3f(x_r + h) - f(x_r)$$

En general, la diferencia hacia adelante de orden k se define como

$$\Delta^k f_r = \Delta^{k-1} f_{r+1} - \Delta^{k-1} f_r \, , \quad \forall\, k \in N.$$

La diferencia hacia adelante de orden cero es, por definición,

$$\Delta^0 f_r = f_r$$

La tabla de diferencias hacia adelante que se muestra a continuación es un medio conveniente para evaluar tales diferencias, para un conjunto dado de datos:

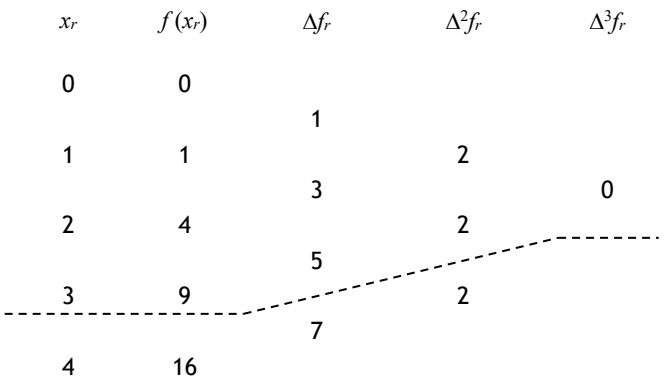
x_r	f_r	Δf_r	$\Delta^2 f_r$	$\Delta^3 f_r$	$\Delta^4 f_r$	$\Delta^5 f_r \ldots$
x_0	f_0					
		Δf_0				
x_1	f_1		$\Delta^2 f_0$			
		Δf_1		$\Delta^3 f_0$		
x_2	f_2		$\Delta^2 f_1$		$\Delta^4 f_0$	
		Δf_2		$\Delta^3 f_1$		$\Delta^5 f_0 \ldots$
x_3	f_3		$\Delta^2 f_2$		$\Delta^4 f_1$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_{n-1}	f_{n-1}		$\Delta^2 f_{n-2}$			
		Δf_{n-1}				
x_n	f_n					

Uno debe saber lo siguiente acerca de la tabla anterior de diferencias: si f_r se toma como $f_r = f(x_r)$, donde $f(x)$ es un polinomio de orden, digamos, n y los x_r tienen igual separación, entonces la columna para la diferencia de orden n se convierte en una constante y la siguiente columna, $(n + 1)$-ésima diferencia, se anula. Si esto ocurre, sabemos que los datos pertenecen a un polinomio de orden n. Sin embargo, si una columna de diferencias tiene uno o más valores anormalmente grandes es probable que existan algunos errores humanos en el proceso de desarrollo de la tabla o en el conjunto de datos.

Puede ocurrir también que las diferencias de orden superior tienden a anularse, pero quizá no lleguen a valer exactamente cero. A menudo, la causa se debe a errores de redondeo en los datos. Esto puede ocurrir aunque los datos pertenezcan a un polinomio de orden bajo.

Ejemplo 7. Formemos la tabla de diferencias dado el siguiente cuadro de valores:

$x = 0, \quad 1, \quad 2, \quad 3, \quad 4$
$f(x) = 0, \quad 1, \quad 4, \quad 9, \quad 16$



La diferencia de orden 3 es nula, entonces se sabe que la diferencia de orden 2 (en este ejemplo es igual a 2) es constante. Por consiguiente, una vez conocido el valor de la diferencia segunda se puede agregar siempre un valor más al último calculado en la columna de las diferencias primeras por simple adición (en el ejemplo, 7 = 5 + 2). Conocida una nueva diferencia primera ($\Delta = 7$), se puede obtener un nuevo valor en la columna de $y = f(x)$ ($7 + 9 = 16$ que corresponde a $x = 4$). En este ejemplo vemos que los valores de $f(x)$ del cuadro corresponden a la función $y = f(x) = x^2$, así que con esta metodología podemos construir rápidamente una tabla de cuadrados sin más que simples adiciones.

Cualquier polinomio de grado n puede ser tabulado en esta forma, una vez que a partir de $(n + 1)$ valores se ha logrado calcular el valor constante de su diferencia de orden n, Δ^n.

En cambio, las funciones trascendentes como e^x, $ln\ x$, $sen\ x$, $cos\ x$, $tg\ x$, $Sh\ x$, $Ch\ x$, $Th\ x$, ... tienen infinitas diferencias sucesivas. Pero considerando sólo un cierto número de cifras decimales y un intervalo h suficientemente pequeño, las diferencias sucesivas van disminuyendo y pueden considerarse prácticamente nulas de una en adelante. Así, por ejemplo, una tabla de logaritmos de 5 decimales (como la de Hoüel), da la diferencia primera prácticamente constante y la diferencia segunda prácticamente nula.

Ejemplo 8. Sean dados los valores de x_r y $f(x_r)$ y calculemos hasta la diferencia de orden 5.

x_r	f_r	Δf_r	$\Delta^2 f_r$	$\Delta^3 f_r$	$\Delta^4 f_r$	$\Delta^5 f_r$
0.55	1.8181818					
		-0.1515151				
0.60	1.6666667		0.0233099			
		-0.1282052		-0.0049948		
0.65	1.5384615		0.0183151		0.0013317	
		-0.1098901		-0.0036631		-0.0004158
0.70	1.4285714		0.0146520		0.0009159	
		-0.0952381		-0.0027472		-0.0002696
0.75	1.3333333		0.0119048		0.0006463	
		-0.0833333		-0.0021009		-0.0001794
0.80	1.25		0.0098039		0.0004669	
		-0.0735294		-0.0016340		
0.85	1.1764706		0.0081699			
		-0.0653595				
0.90	1.1111111					

4.3.2.2. Fórmula de interpolación de Newton hacia adelante en puntos igualmente espaciados.

La ecuación (10) de interpolación de Newton con diferencias divididas nos servirá para deducir todas las fórmulas de interpolación con diferencias finitas.

Siendo $x_r = x_0 + rh$, para $r = 1, 2, ..., n$, entonces

$\Delta f(x_0) = f_1 - f_0 = (x_1 - x_0) f(x_0 x_1) = h f(x_0 x_1)$, de donde, $f(x_0 x_1) = \dfrac{\Delta f(x_0)}{h}$

Del mismo modo,

$$\Delta^2 f(x_0) = \Delta f_1 - \Delta f_0 = h f(x_1 x_2) - h f(x_0 x_1) = h\,[(x_2 - x_0) f(x_0 x_1 x_2)] = h.2h f(x_0 x_1 x_2) = 2h^2 f(x_0 x_1 x_2)$$

de donde,

$$f(x_0 x_1 x_2) = \frac{\Delta^2 f(x_0)}{2h^2}$$

Análogamente,

$$\Delta^3 f_0 = \Delta^2 f_1 - \Delta^2 f_0 = 2h^2\,[f(x_1 x_2 x_3) - f(x_0 x_1 x_2)] = 2h^2\,[(x_3 - x_0) f(x_0 x_1 x_2 x_3)] =$$

$$= 2h^2\, 3h f(x_0 x_1 x_2 x_3) = 3!\, h^3 f(x_0 x_1 x_2 x_3)$$

de donde,

$$f(x_0 x_1 x_2 x_3) = \frac{\Delta^3 f(x_0)}{3! h^3}$$

En general, podemos escribir

$$\Delta^n f_0 = \Delta^{n-1} f_1 - \Delta^{n-1} f_0 = (n-1)!\, h^{n-1} [f(x_1 x_2 \ldots x_n) - f(x_0 x_1 \ldots x_{n-1})] =$$

$$= (n-1)!\, h^{n-1} [(x_n - x_0) f(x_0 x_1 \ldots x_n)] = n!\, h^n f(x_0 x_1 \ldots x_n)$$

de donde,

$$f(x_0 x_1 \ldots x_n) = \frac{\Delta^n f_0}{n! h^n}$$

Reemplazando en la ecuación (10) por estos valores, se obtiene

$$f(x) = f(x_0) + (x - x_0)\frac{\Delta f_0}{h} + (x - x_0)(x - x_1)\frac{\Delta^2 f_0}{2! h^2} + (x - x_0)(x - x_1)(x - x_2)\frac{\Delta^3 f_0}{3! h^3} + \quad (19)$$
$$+ \ldots + (x - x_0)(x - x_1)\ldots(x - x_{n-1})\frac{\Delta^n f_0}{n! h^n} + E(x)$$

donde,

$$E(x) = (x - x_0)(x - x_1)\ldots(x - x_n)\frac{f^{(n+1)}(\xi)}{(n+1)!}$$

con ξ un punto cualquiera dentro del intervalo que contiene a las incógnitas y los datos.

Esta ecuación se conoce como *fórmula de interpolación de Newton hacia adelante en puntos equidistantes.* Esta se puede simplificar más aun definiendo una nueva cantidad *t*: $t = (x - x_0)/h$ (distancia desde *x* a x_0 en unidades de *h*). Usando esta definición, se tiene

$$x - x_0 = th$$
$$x - x_1 = x - x_0 - h = th - h = h\,(t - 1)$$
$$x - x_2 = x - x_0 - 2h = th - 2h = h\,(t - 2)$$
$$\vdots$$
$$x - x_n = x - x_0 - nh = th - nh = h\,(t - n)$$

valores que pueden sustituirse en la ecuación (19) para dar

$$f(x)=f(x_0)+t\,\Delta f_0+t(t-1)\frac{\Delta^2 f_0}{2!}+t(t-1)(t-2)\frac{\Delta^3 f_0}{3!}+\ldots+t(t-1)(t-2)\ldots(t-n+1)\frac{\Delta^n f_0}{n!}+$$
$$+t(t-1)(t-2)\ldots(t-n)h^{n+1}\frac{f^{(n+1)}(\xi)}{(n+1)!}. \qquad (20)$$

Para el cálculo del error en la expresión anterior se requiere el conocimiento de la función y también saber que es diferenciable. Usualmente, éste no es el caso. El error de la interpolación se evalúa entonces de la misma forma que en la interpolación de Newton con diferencias divididas, donde consideramos que

$$f(x_0x_1\ldots x_{n+1})\approx\frac{f^{(n+1)}(\xi)}{(n+1)!}$$

(en donde uno pone $x = x_{n+1}$, un punto donde $f(x_{n+1})$ es conocida pero no es usada en la interpolación) y como

$$f(x_0x_1\ldots x_{n+1})=\frac{\Delta^{n+1}f_0}{(n+1)!h^{n+1}}$$

entonces

$$\frac{\Delta^{n+1}f_0}{(n+1)!h^{n+1}}\approx\frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Luego, el último término de la ecuación (20) que nos da el error de la interpolación de Newton hacia adelante es aproximadamente igual a

$$E(x)\approx\frac{t\,(t\text{-}1)\,(t\text{-}2)\ldots(t\text{-}n)}{(n+1)!}\Delta^{n+1}f_0$$

(es decir que es aproximadamente igual al término que se añadiría a la ecuación $P_n(x) = f(x_0) + t\,\Delta f_0 + t\,(t-1)\,\frac{\Delta^2 f_0}{2!} + \ldots + t\,(t-1)\ldots(t-n+1)\,\frac{\Delta^n f_0}{n!}$ si la interpolación se extendiera para ajustarse a otro punto más, x_{n+1}).

Ejemplo 9. Consideremos el ejemplo 7.

De la ecuación (20) y como $t=(x-x_0)/h=(x-0)/1=x$, podemos escribir

$$P_2(x) = 0 + t\,1 + \frac{t(t-1)}{2!}2 = t + t\,(t\text{ - }1) = x + x\,(x\text{ - }1) = x^2 = f(x)$$

y $E(x) = 0$.

Ejemplo 10. Consideremos el ejemplo 8. Queremos evaluar el polinomio de interpolación en $x = 0.57$ y obtener una estimación del error cometido.

Para ello escribimos, según la ecuación (20)

$$P_5(x)=1.8181818+\frac{t}{1!}\,(-0.1515151)+\frac{t(t-1)}{2!}\,0.0233099+\frac{t(t-1)(t-2)}{3!}\,(-0.0049948)+$$

$$+\frac{t(t-1)(t-2)(t-3)}{4!}\,0.0013317+\frac{t(t-1)(t-2)(t-3)(t-4)}{5!}\,(-0.0004158)$$

y

$$E(x) = \frac{0.05^6}{6!}\;t(t\text{-}1)(t\text{-}2)(t\text{-}3)(t\text{-}4)(t\text{-}5)\;\frac{720}{\xi^7}$$

siendo ξ un cierto valor intermedio entre 0.55 y 0.80. (En la última igualdad usamos: $f(x)=\frac{1}{x}$, $f^{(1)}(x)=-\frac{1}{x^2}$, $f^{(2)}(x)=\frac{2}{x^3}$, $f^{(3)}(x)=-\frac{6}{x^4}$, $f^{(4)}(x)=\frac{24}{x^5}$, $f^{(5)}(x)=-\frac{120}{x^6}$, $f^{(6)}(x)=\frac{720}{x^7}$). (Los restantes cálculos quedan a cargo del lector).

Ejemplo 11. Estimemos el error resultante de la aproximación de la función $f(x) = sen\,x$ por el polinomio de interpolación de quinto grado $P_5(x)$ que coincide con la función dada para los valores $x = 0º, 5º, 10º, 15º, 20º, 25º$.

Como $f^{(6)}(x) = -sen\,x$, entonces $\left|f^{(6)}(x)\right| \le 1$. Además, $h = 5º = \pi/36$, de donde, $t = \frac{x}{\pi/36} = \frac{36}{\pi}x$. Luego,

$$|E(x)| = |sen\,x - P_5(x)| \le \frac{(\pi/36)^6}{6!}\,|\,t(t\text{ - }1)(t\text{ - }2)(t\text{ - }3)(t\text{ - }4)(t\text{ - }5)|\;1$$

$$\le \frac{(\pi/36)^6}{6!}(36/\pi)^6\,|\,x(x-\frac{\pi}{36})(x-\frac{\pi}{18})(x-\frac{\pi}{12})(x-\frac{\pi}{9})(x-\frac{5\pi}{36})\,|$$

$$\leq \frac{1}{6!} | x(x-\frac{\pi}{36})(x-\frac{\pi}{18})(x-\frac{\pi}{12})(x-\frac{\pi}{9})(x-\frac{5\pi}{36}) |$$

Por ejemplo, para x = 12º30' = 0.218166 radianes, se obtiene

$$|E(x)| = |sen\, x - P_5(x)| \leq \frac{1}{6!} \left| -1.552697 \text{ x } 10^{-6} \right| = 2.156523 \text{ x } 10^{-9} \approx 2.2 \text{ x } 10^{-9}$$

esto es,

$$|E(x)| \leq 2.2 \text{ x } 10^{-9}.$$

La ecuación (20) es apropiada para interpolar al comienzo de la tabla, pues así $t = (x - x_0) / h$ es chico. Para interpolar al final de la tabla necesitamos introducir la noción de diferencia finita hacia atrás.

4.3.2.3.Diferencias hacia atrás.

La ecuación (20) resuelve el problema de interpolar en una tabla de intervalo constante h un valor correspondiente al valor x de la variable comprendida entre x_0 y x_n, pero como las diferencias Δf_0, $\Delta^2 f_0$, ... que aparecen en la fórmula se refieren al primer valor f_0, ésta se emplea generalmente cuando x es próxima a x_0. Cuando la interpolación debe efectuarse para un valor de x próximo a x_n interesa conocer otras fórmulas en las que se utilicen las diferencias sucesivas relativas al último valor f_n.

Se define, entonces, la diferencia hacia atrás de orden uno como

$$\nabla f_k = f_k - f_{k-1} = f(x_k) - f(x_k - h)$$

(en la última igualdad usamos: $x_{k-1} = x_0 + (k-1)h = x_0 + kh - h = x_k - h$).

En forma análoga, la diferencia hacia atrás de orden 2 se define como

$$\nabla^2 f_k = \nabla f_k - \nabla f_{k-1} = f_k - f_{k-1} - f_{k-1} + f_{k-2} = f_k - 2 f_{k-1} + f_{k-2}$$

De la misma forma, la diferencia hacia atrás de orden 3 se define como

$$\nabla^3 f_k = \nabla^2 f_k - \nabla^2 f_{k-1} = f_k - 2 f_{k-1} + f_{k-2} - f_{k-1} + 2 f_{k-2} - f_{k-3} = f_k - 3 f_{k-1} + 3 f_{k-2} - f_{k-3}$$

En general, la diferencia hacia atrás de orden p se define como

$$\nabla^p f_k = \nabla^{p-1} f_k - \nabla^{p-1} f_{k-1}, \quad \forall\, p \in N.$$

La diferencia hacia atrás de orden cero es, por definición, $\nabla^0 f_k = f_k$

Se observa de acuerdo con las definiciones anteriores que

$$\nabla^0 f_k = \Delta^0 f_k,\ \nabla f_k = \Delta f_{k-1},\ \nabla^2 f_k = \Delta^2 f_{k-2},\ \nabla^3 f_k = \Delta^3 f_{k-3} \text{ y, en general, } \nabla^r f_k = \Delta^r f_{k-r}$$

(r entero no negativo y k entero tal que $k \geq r$).

En una tabla de valores, se tiene

x_k	$f(x_k)$	∇f_k	$\nabla^2 f_k$	$\nabla^3 f_k$	$\nabla^4 f_k$	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_{n-4}	f_{n-4}					
		∇f_{n-3}				
x_{n-3}	f_{n-3}		$\nabla^2 f_{n-2}$			
		∇f_{n-2}		$\nabla^3 f_{n-1}$		
x_{n-2}	f_{n-2}		$\nabla^2 f_{n-1}$		$\nabla^4 f_n$	...
		∇f_{n-1}		$\nabla^3 f_n$		
x_{n-1}	f_{n-1}		$\nabla^2 f_n$			
		∇f_n				
x_n	f_n					

4.3.2.4 Fórmula de interpolación de Newton hacia atrás en puntos igualmente espaciados

Para deducir la fórmula de interpolación correspondiente a partir de la fórmula con diferencias divididas (10) introducimos las abscisas $x_0, x_1, ..., x_n$ en orden inverso; es decir, se reordenan las abscisas como $x_n, x_{n-1}, ..., x_1, x_0$. El polinomio de interpolación es siempre el mismo: por $(n + 1)$ puntos distintos pasa un polinomio y sólo uno de grado máximo n. De este modo, la ecuación (10) da

$$f(x) = f(x_n) + (x-x_n) f(x_n x_{n-1}) + (x-x_n)(x-x_{n-1}) f(x_n x_{n-1} x_{n-2}) + ... + (x-x_n)(x-x_{n-1})...(x-x_1) f(x_n x_{n-1} ... x_1 x_0) + \\ + (x-x_n)(x-x_{n-1}) \ldots (x-x_1)(x-x_0) \frac{f^{(n+1)}(\xi)}{(n+1)!} \qquad (21)$$

Por definición, se tiene

$$f(x_n\, x_{n-1}) = \frac{f\,(x_{n-1}) - f(x_n)}{x_{n-1} - x_n} = \frac{f\,(x_n) - f\,(x_{n-1})}{x_n - x_{n-1}} = \frac{\nabla f(x_n)}{h}$$

$$f(x_n\, x_{n-1}\, x_{n-2}) = \frac{f\,(x_{n-1}x_{n-2}) - f(x_n x_{n-1})}{x_{n-2} - x_n} = \frac{f\,(x_n x_{n-1}) - f\,(x_{n-1}x_{n-2})}{x_n - x_{n-2}} =$$

$$= \frac{1}{2h}\left[\frac{\nabla f(x_n)}{h} - \frac{\nabla f(x_{n-1})}{h}\right] = \frac{\nabla^2 f(x_n)}{2!h^2}$$

$$f(x_n\, x_{n-1}\, x_{n-2}\, x_{n-3}) = \frac{f\,(x_{n-1}x_{n-2}x_{n-3}) - f(x_n x_{n-1}x_{n-2})}{x_{n-3} - x_n} = \frac{f(x_n x_{n-1}x_{n-2}) - f\,(x_{n-1}x_{n-2}x_{n-3})}{x_n - x_{n-3}} =$$

$$= \frac{1}{3h}\left[\frac{\nabla^2 f_n}{2h^2} - \frac{\nabla^2 f_{n-1}}{2h^2}\right] = \frac{\nabla^3 f_n}{3!h^3}$$

De modo que, en general,

$$f(x_n\, x_{n-1}\, x_{n-2}\, \ldots\, x_{n-r}) = \frac{\nabla^r f_n}{r!h^r}$$

y

$$f(x_n\, x_{n-1}\, x_{n-2}\, \ldots\, x_{n-(r+1)}) = \frac{f(x_{n-1}x_{n-2}\ldots x_{n-r-1}) - f(x_n x_{n-1}\ldots x_{n-r})}{x_{n-r-1} - x_n} =$$

$$= \frac{f(x_n x_{n-1}\ldots x_{n-r}) - f\,(x_{n-1}x_{n-2}\ldots x_{n-r-1})}{x_n - x_{n-r-1}} = \frac{1}{(r+1)h}\left[\frac{\nabla^r f_n}{r!h^r} - \frac{\nabla^r f_{n-1}}{r!h^r}\right] = \frac{1}{(r+1)!}\,\frac{\nabla^{r+1} f_n}{h^{r+1}}.$$

(En la expresión anterior hemos usado: $x_n = x_0 + nh$, $x_{n-r-1} = x_0 + (n\text{-}r\text{-}1)\,h$, de donde, $x_n - x_{n-r-1} = (r+1)\,h$).

Reemplazando estas fórmulas en la ecuación (21), obtenemos

$$f(x) = f(x_n) + (x\text{-}x_n)\frac{\nabla f_n}{h} + (x\text{-}x_n)\,(x\text{-}x_{n-1})\frac{\nabla^2 f_n}{2!h^2} + (x\text{-}x_n)\,(x\text{-}x_{n-1})\,(x\text{-}x_{n-2})\frac{\nabla^3 f_n}{3!h^3} + \ldots +$$

$$+ (x\text{-}x_n)\,(x\text{-}x_{n-1})\ldots(x\text{-}x_1)\frac{\nabla^n f_n}{n!h^n} + E\,(x) \qquad (22)$$

donde,

$$E(x) = (x\text{-}x_n)(x\text{-}x_{n\text{-}1})\ldots(x\text{-}x_1)(x\text{-}x_0)\ \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

con ξ un valor comprendido entre las abscisas x, x_0, x_1, ..., x_n.

La ecuación (22) puede simplificarse si se introduce una nueva variable: $t = (x - x_n)/h$. Esta definición se puede usar para desarrollar la siguiente expresión simplificada de los términos en la ecuación (22)

$$\begin{aligned}
x - x_n &= th \\
x - x_{n\text{-}1} &= x - (x_n - h) = x - x_n + h = th + h = h(t+1) \\
x - x_{n\text{-}2} &= x - (x_n - 2h) = x - x_n + 2h = th + 2h = h(t+2) \\
&\vdots \\
x - x_0 &= x - (x_n - nh) = x - x_n + nh = th + nh = h(t+n)
\end{aligned}$$

los cuales pueden sustituirse en la ecuación (22) para dar

$$f(x) = f(x_n) + t\nabla f_n + t(t+1)\frac{\nabla^2 f_n}{2!} + t(t+1)(t+2)\frac{\nabla^3 f_n}{3!} + \ldots + t(t+1)\ldots(t+n\text{-}1)\frac{\nabla^n f_n}{n!} +$$
$$+ t(t+1)(t+2)\ldots(t+n\text{-}1)(t+n)\,h^{n+1}\,\frac{f^{(n+1)}(\xi)}{(n+1)!} \qquad (23)$$

ecuación conocida como *fórmula de interpolación de Newton hacia atrás en puntos equidistantes.* Se utiliza para interpolar en la parte final de una tabla.

Si la función es desconocida o bien sus derivadas son difíciles de calcular, para obtener una estimación del error se procede de manera análoga al caso de interpolación de Newton hacia adelante, en donde ahora

$$\frac{\nabla^{n+1} f_n}{(n+1)!h^{n+1}} \approx \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

y así, el error de la interpolación de Newton hacia atrás es aproximadamente igual a

$$E(x) \approx \frac{t(t+1)(t+2)\ldots(t+n)}{(n+1)!}\nabla^{n+1} f_n$$

4.3.3. Comentarios adicionales.

Antes de proseguir con otro tema deberíamos preguntarnos por qué se aborda el caso de los datos igualmente espaciados si los métodos de Newton con diferencias divididas y Lagrange (Párrafo 4.3.5) son compatibles con los

datos espaciados arbitrariamente. Antes del advenimiento de las computadoras digitales, estos métodos tuvieron gran utilidad en la interpolación de tablas con datos igualmente espaciados. De hecho, se desarrolló un esquema conocido como *tabla de diferencias* para facilitar la implementación de estas técnicas (las ya vistas son un ejemplo de estas tablas).

Sin embargo, y debido a que las fórmulas son un subconjunto de los esquemas de Newton y de Lagrange compatibles con la computadora, la necesidad de puntos equiespaciados se fue perdiendo, aunque su importancia en temas posteriores nos muestra la razón de por qué deben ser incluidas en nuestro estudio. En particular, se pueden emplear en la derivación de fórmulas de integración numérica que emplean comúnmente datos equiespaciados.

Vimos que la ecuación que nos da el error en los polinomios de interpolación de Newton es

$$E(x) = (x - x_0)(x - x_1) \dots (x - x_n) f(x_0 x_1 \dots x_n x)$$

Suponiendo que la diferencia dividida no varía demasiado a lo largo del rango de datos, entonces el error es proporcional al producto: $(x - x_0)(x - x_1)\dots (x - x_n)$. Obviamente, mientras más cercanos estén los puntos base a las x menor será la magnitud de este producto. Además, como ya hemos visto, en el caso en que la función sea desconocida o bien cuando es conocida pero sus derivadas son difíciles de obtener, entonces el error en las fórmulas de interpolación de Newton se representa mediante un término adicional que resulta de un punto extra de los datos.

La *extrapolación* es el proceso de calcular un valor de $f(x)$ que cae fuera del rango de los puntos base conocidos x_0, x_1, ..., x_n. Como ya se dijo, la interpolación más exacta usualmente se obtiene cuando las incógnitas caen cerca de los puntos base. Obviamente, esto no sucede cuando las incógnitas caen fuera del rango y, por lo tanto, el error en la extrapolación puede ser muy grande.

Como se muestra en la Figura 5, la naturaleza abierta en los extremos de la extrapolación, que se basa en el ajuste de una parábola a través de los primeros tres puntos, representa un paso en la incógnita porque el proceso extiende la curva más allá de la región conocida.

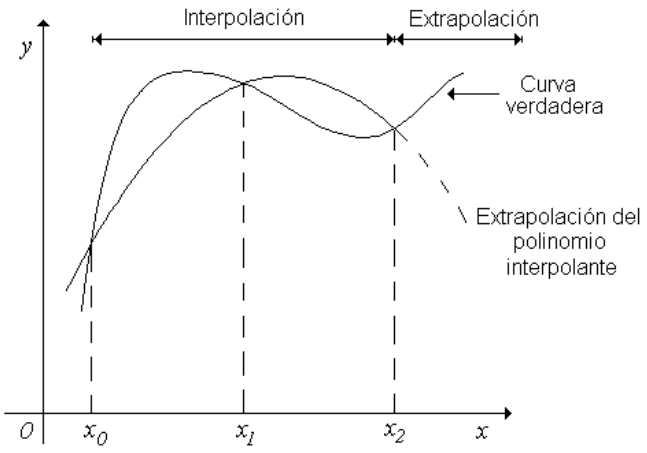


Figura 5

Como tal, la curva verdadera diverge fácilmente de la predicción. Por lo tanto, se debe tener cuidado extremo en casos donde se deba extrapolar.

Es así entonces que la ecuación (20) se usa también para calcular valores de $f(x)$ donde el argumento x está fuera del margen de la tabla pero siempre que $x < x_0$, y la ecuación (23) se utiliza también con el mismo fin pero siempre que $x > x_n$.

Ejemplo 12. Con los datos del ejemplo 8 nos proponemos evaluar el polinomio de interpolación en $x = 0.93$ y estimar el error para este valor.

$$\text{Siendo } x = 0.93 = x_n + h\,t = 0.90 + 0.05\,t, \text{ entonces } t = \frac{0.93 - 0.90}{0.05} = 0.6.$$

Luego,

$$P_5(0.93) = 1.1111111 + 0.6(-0.0653595) + 0.6(1.6)\frac{0.0081699}{2} +$$
$$+ 0.6(1.6)(2.6)\left(\frac{-0.0016340}{3!}\right) + 0.6(1.6)(2.6)(3.6)\frac{0.0004669}{4!} +$$
$$+ 0.6(1.6)(2.6)(3.6)(4.6)\left(\frac{-0.0001794}{5!}\right) = 1.0752502.$$

La estimación del error está dada por

$$E(0.93) = \frac{0.05^6\ (0.6)(1.6)(2.6)(3.6)(4.6)(5.6)}{6!}\ \frac{720}{\xi^7}$$

donde ξ es un punto comprendido entre 0.65 y 0.90.

Luego, efectuando operaciones (con $\xi \geq 0.65$ y $\xi \leq 0.90$), resulta

$$0.0000076 \leq E(0.93) \leq 0.0000738.$$

Dado que en este caso es $f(x) = \frac{1}{x}$, entonces

$$f(0.93) = \frac{1}{0.93} = 1.0752688.$$

Por lo tanto, el error exacto es

$$E(0.93) = 1.0752688 - 1.0752502 = 0.0000186.$$

Ejemplo 13. Lo mismo que en el ejemplo 12 pero para $x = 0.54$.

Como $x = x_0 + ht$, entonces $t = \frac{0.54 - 0.55}{0.05} = -0.20$.

$$P_5(0.54) = 1.8181818 + (-0.2)(-0.1515151) + (-0.2)(-1.2)\frac{0.0233099}{2} +$$
$$+ (-0.2)(-1.2)(-2.2)\left(\frac{-0.0049948}{3!}\right) + (-0.2)(-1.2)(-2.2)(-3.2)\frac{0.0013317}{4!} +$$
$$+ (-0.2)(-1.2)(-2.2)(-3.2)(-4.2)\left(\frac{-0.0004158}{5!}\right) = 1.8518399.$$

La estimación del error está dada por

$$E(0.54) = \frac{0.05^6\ (-0.2)(-1.2)(-2.2)(-3.2)(-4.2)(-5.2)}{6!}\ \frac{720}{\xi^7}$$

con ξ comprendido entre 0.55 y 0.80. Haciendo los cálculos, resulta

$$0.0000027 \leq E(0.54) \leq 0.0000397.$$

Como $f(0.54) = \frac{1}{0.54} = 1.8518519$, entonces el error exacto es

$$E(0.54) = 1.2 \text{ x } 10^{-5} = 0.000012.$$

Ejemplo 14. Lo mismo que en el ejemplo 12 pero para $x = 0.79$.

Usaremos la fórmula de Newton hacia adelante tomando $x_0 = 0.65$, x_1=0.70, $x_2 = 0.75$, $x_3 = 0.80$, $x_4 = 0.85$, $x_5 = 0.90$. Aquí, $t = \dfrac{0.79 - 0.65}{0.05} =$ 2.8. Entonces,

$$P_5(0.79)=1.5384615 + 2.8(-0.1098901) + 2.8(1.8)\frac{0.0146520}{2} + 2.8(1.8)(0.8)\left(\frac{-0.0027472}{3!}\right) +$$

$$+ 2.8(1.8)(0.8)(-0.2)\frac{0.0006463}{4!} + 2.8(1.8)(0.8)(-0.2)(-1.2)\left(\frac{-0.0001794}{5!}\right) = 1.2658230.$$

La estimación de error nos da

$$E(0.79) = \frac{0.05^6\ (2.8)(1.8)(0.8)(-0.2)(-1.2)(-2.2)}{6!}\ \frac{720}{\xi^7}$$

con ξ comprendido entre 0.65 y 0.90. Luego, $-6.79 \text{ x } 10^{-7} \le E(0.79) \le -6.96 \text{ x } 10^{-8}$.

El error exacto es $E(0.79) = -2.16 \text{ x } 10^{-7}$.

4.3.4. Diferencias centrales.

En la construcción de las fórmulas de interpolación de Newton con diferencias finitas sólo se utilizaron aquellos valores de la función que caían en un lado del valor inicial elegido; estas fórmulas son, por lo tanto, de naturaleza lateral (de un solo lado).

Cuando la interpolación debe efectuarse próxima a la mitad o en el centro de una tabla, es conveniente seleccionar una abscisa como origen x_0 y designar las abscisas hacia adelante como x_1, x_2, x_3, x_4, … y las abscisas hacia atrás como x_{-1}, x_{-2}, x_{-3}, x_{-4}, …, de modo que se tiene

x	$f(x)$	$\delta f(x)$	$\delta^2 f(x)$	$\delta^3 f(x)$	$\delta^4 f(x)$	$\delta^5 f(x)$	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_{-4}	f_{-4}						
		$\delta f_{-7/2}$					
x_{-3}	f_{-3}		$\delta^2 f_{-3}$				
		$\delta f_{-5/2}$		$\delta^3 f_{-5/2}$			
x_{-2}	f_{-2}		$\delta^2 f_{-2}$		$\delta^4 f_{-2}$		
		$\delta f_{-3/2}$		$\delta^3 f_{-3/2}$		$\delta^5 f_{-3/2}$	
x_{-1}	f_{-1}		$\delta^2 f_{-1}$		$\delta^4 f_{-1}$		
		$\delta f_{-1/2}$		$\delta^3 f_{-1/2}$		$\delta^5 f_{-1/2}$	
x_0	f_0		$\delta^2 f_0$		$\delta^4 f_0$	...	
		$\delta f_{1/2}$		$\delta^3 f_{1/2}$		$\delta^5 f_{1/2}$	
x_1	f_1		$\delta^2 f_1$		$\delta^4 f_1$		
		$\delta f_{3/2}$		$\delta^3 f_{3/2}$		$\delta^5 f_{3/2}$	
x_2	f_2		$\delta^2 f_2$		$\delta^4 f_2$		
		$\delta f_{5/2}$		$\delta^3 f_{5/2}$			
x_3	f_3		$\delta^2 f_3$				
		$\delta f_{7/2}$					
x_4	f_4						
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Las diferencias centrales se definen de modo análogo a las diferencias hacia adelante y hacia atrás, pero ahora se usa información tanto hacia adelante como hacia atrás del punto en cuestión. Sin embargo, aunque las diferencias centrales de orden par pueden asociarse a una elección particular de x, las diferencias de orden impar deben ahora asociarse con los valores adyacentes de x. La primera diferencia central se define como

$$\delta f_{r+1/2} = f_{r+1} - f_r$$

El subíndice $r + 1/2$ es una notación convencional y no representa ninguna cantidad en $x = x_r + h/2$. Se opta por esta notación por el hecho que la diferencia primera se coloca en el medio de los argumentos x_r y x_{r+1}. Se observa además que δf_r carece de significado cuando r es entero. Sin embargo, $\delta^2 f_r$ sí tiene sentido libre de ambigüedad, denotando

$$\delta^2 f_r = \delta f_{r+1/2} - \delta f_{r-1/2} = f_{r+1} - f_r - f_r + f_{r-1} = \Delta^2 f_{r-1} = \nabla^2 f_{r+1}$$

Las diferencias centrales de orden mayor se definen en forma semejante con r entero

$$\delta^{2n} f_r = \delta^{2n-1} f_{r+1/2} - \delta^{2n-1} f_{r-1/2}$$

$$\delta^{2n+1} f_{r+1/2} = \delta^{2n} f_{r+1} - \delta^{2n} f_r, \quad n \in N$$

Como hemos señalado, vale que

$$\delta^2 f_r = \Delta^2 f_{r-1} = \nabla^2 f_{r+1}$$

y, en general, puede probarse que

$$\delta^r f_k = \Delta^r f_{k-r/2} = \nabla^r f_{k+r/2}$$

con $r \in N$, r par.

Puede verificarse fácilmente que

$$\Delta \nabla f_k = \nabla \Delta f_k$$

En efecto,

$$\Delta \nabla f_k = \Delta (f_k - f_{k-1}) = f_{k+1} - f_k - (f_k - f_{k-1}) = f_{k+1} - 2 f_k + f_{k-1} = \delta^2 f_k$$

$$\nabla \Delta f_k = \nabla (f_{k+1} - f_k) = f_{k+1} - f_k - (f_k - f_{k-1}) = f_{k+1} - 2 f_k + f_{k-1} = \delta^2 f_k$$

Del mismo modo

$$\delta^3 f_{3/2} = \Delta^3 f_0 = \nabla^3 f_3$$

pues, por definición

$$\delta^3 f_{3/2} = \delta^2 f_2 - \delta^2 f_1 = \delta f_{5/2} - \delta f_{3/2} - \delta f_{3/2} + \delta f_{1/2} = f_3 - f_2 - f_2 + f_1 - f_2 + f_1 + f_1 - f_0 =$$

$$= f_3 - 3 f_2 + 3 f_1 - f_0 = \Delta^3 f_0 = \nabla^3 f_3$$

4.3.4.1. Fórmula de interpolación de Gauss hacia adelante con diferencias centrales.

No es muy usada en Análisis Numérico, pero permite deducir fácilmente las importantes fórmulas de Stirling y de Bessel.

En la fórmula de interpolación con diferencias divididas (10) introducimos los argumentos en el orden siguiente: x_0, x_1, x_{-1}, x_2, x_{-2}, x_3, x_{-3}, …, de modo que dicha fórmula se escribe

$$f(x) = f(x_0) + (x-x_0) f(x_0 x_1) + (x-x_0)(x-x_1) f(x_0 x_1 x_{-1}) + (x-x_0)(x-x_1)(x-x_{-1}) f(x_0 x_1 x_{-1} x_2) +$$

$$+ (x-x_0)(x-x_1)(x-x_{-1})(x-x_2) f(x_0 x_1 x_{-1} x_2 x_{-2}) + (x-x_0)(x-x_1)(x-x_{-1})(x-x_2)(x-x_{-2}) f(x_0 x_1 x_{-1} x_2 x_{-2} x_3) + \ldots$$

De acuerdo a la definición de diferencias centrales, se tiene

$$f(x_0\, x_1) = \frac{f(x_1) - f(x_0)}{h} = \frac{\delta\, f_{1/2}}{h}$$

$$f(x_0\, x_1\, x_{-1}) = f(x_{-1}\, x_0\, x_1) = \frac{f(x_0 x_1) - f(x_{-1} x_0)}{x_1 - x_{-1}} = \left[\frac{\delta\, f_{1/2}}{h} - \frac{\delta\, f_{-1/2}}{h}\right]\frac{1}{2h} = \frac{\delta^{\,2} f_0}{2!\,h^2}$$

$$f(x_0\, x_1\, x_{-1}\, x_2) = f(x_{-1}\, x_0\, x_1\, x_2) = \frac{f(x_0 x_1 x_2) - f(x_{-1} x_0 x_1)}{x_2 - x_{-1}} = \left[\frac{\delta^{\,2} f_1}{2h^2} - \frac{\delta^{\,2} f_0}{2h^2}\right]\frac{1}{3h} = \frac{1}{3!\,h^3}\delta^3 f_{1/2}$$

$$f(x_0 x_1 x_{-1} x_2\, x_{-2}) = f(x_{-2} x_{-1} x_0\, x_1\, x_2) =$$

$$= \frac{f(x_{-1} x_0 x_1 x_2) - f(x_{-2} x_{-1} x_0 x_1)}{x_2 - x_{-2}} = \left[\frac{\delta^{\,3} f_{1/2}}{3!\,h^3} - \frac{\delta^3 f_{-1/2}}{3!\,h^3}\right]\frac{1}{4h} = \frac{1}{4!\,h^4}\delta^4 f_0$$

y así sucesivamente

$$f(x_{-n}\, x_{-n+1} \ldots x_{-1}\, x_0\, x_1 \ldots x_n\, x_{n+1}) = \frac{\delta^{\,2n+1} f_{1/2}}{(2n+1)!\,h^{2n+1}}$$

$$f(x_{-n}\, x_{-n+1}\, x_{-n+2} \ldots x_{-1}\, x_0\, x_1 \ldots x_n) = \frac{\delta^{\,2n} f_0}{(2n)!\,h^{2n}}$$

Reemplazando en la fórmula las diferencias divididas por las diferencias centrales, se tiene

$$f(x) = f(x_0) + \frac{(x - x_0)}{h}\delta\, f_{1/2} + \frac{(x - x_0)(x - x_1)}{2!\,h^2}\delta^2 f_0 + \frac{(x - x_0)(x - x_1)(x - x_{-1})}{3!\,h^3}\delta^3 f_{1/2} +$$
$$+ \frac{(x - x_0)(x - x_1)(x - x_{-1})(x - x_2)}{4!\,h^4}\delta^4 f_0 + \frac{(x - x_0)(x - x_1)(x - x_{-1})(x - x_2)(x - x_{-2})}{5!\,h^5}\delta^5 f_{1/2} + \ldots \quad (24)$$

La ecuación (24) puede simplificarse si se introduce la variable t definida como la distancia desde x a x_0 en unidades de h, o sea, $t = (x - x_0)/h$, o bien, $x = x_0 + ht$; siendo además $x_k = x_0 + kh$ y $x_{-k} = x_0 - kh$, surge que, $x - x_k = h(t - k)$ y $x - x_{-k} = h(t + k)$.

Reemplazando en la ecuación (24) resulta

$$f(x) = f(x_0) + t\delta\, f_{1/2} + \frac{t(t-1)}{2!}\delta^2\, f_0 + \frac{t(t^2-1^2)}{3!}\delta^3\, f_{1/2} + \frac{t(t^2-1^2)(t-2)}{4!}\delta^4\, f_0 +$$

$$+\frac{t(t^2-1^2)(t^2-2^2)}{5!}\delta^5 f_{1/2}+\ldots+\frac{t(t^2-1^2)(t^2-2^2)\ldots(t^2-(m-1)^2)(t-m)}{(2m)!}\delta^{2m} f_0$$

ó

$$+\frac{t(t^2-1^2)(t^2-2^2)\ldots(t^2-m^2)}{(2m+1)!}\delta^{2m+1} f_{1/2}+E(x). \qquad (25)$$

Si en la ecuación (25) se conserva hasta la *n*-ésima diferencia, donde $n = 2m$, o sea hasta una diferencia par, el error está dado por

$$E(x)=\frac{h^{(2m+1)}t(t^2-1^2)(t^2-2^2)\ldots(t^2-m^2)}{(2m+1)!}f^{(2m+1)}(\xi) \qquad (26)$$

Si en la ecuación (25) se conserva hasta la diferencia *n*-ésima, donde $n = 2m + 1$, o sea hasta una diferencia impar, el error está dado por

$$E(x)=\frac{h^{(2m+2)}t(t^2-1^2)(t^2-2^2)\ldots(t^2-m^2)(t-(m+1))}{(2m+2)!}f^{(2m+2)}(\xi) \qquad (27)$$

Si no se conoce la expresión analítica de la función $f(x)$ (considerando para un h suficientemente pequeño que $f^{(2m+1)}(\xi)\approx\frac{\delta^{2m+1}f_{1/2}}{h^{2m+1}}$ y que $f^{(2m+2)}(\xi)\approx\frac{\delta^{2m+2}f_0}{h^{2m+2}}$), la ecuación (26)

puede escribirse

$$E(x)\approx\frac{t(t^2-1^2)(t^2-2^2)\ldots(t^2-m^2)}{(2m+1)!}\delta^{2m+1}f_{1/2}$$

y análogamente (para un h pequeño), la ecuación (27) se transforma en

$$E(x)\approx\frac{t(t^2-1^2)(t^2-2^2)\ldots(t^2-m^2)(t-(m+1))}{(2m+2)!}\delta^{2m+2}f_0$$

La ecuación (25) se conoce con el nombre de *fórmula de interpolación de Gauss hacia adelante con diferencias centrales.* En el cuadro de las diferencias el recorrido que esta fórmula hace es en zig-zag, como se ilustra gráficamente, conservando siempre los argumentos 0 y 1/2:

x	$f(x)$	δf	$\delta^2 f$	$\delta^3 f$	$\delta^4 f$	$\delta^5 f$	...	
⋮	⋮	⋮	⋮	⋮	⋮	⋮		
x_{-3}	f_{-3}							
		$\delta f_{-5/2}$						
x_{-2}	f_{-2}		$\delta^2 f_{-2}$					
		$\delta f_{-3/2}$		$\delta^3 f_{-3/2}$				
x_{-1}	f_{-1}		$\delta^2 f_{-1}$		$\delta^4 f_{-1}$			
		$\delta f_{-1/2}$		$\delta^3 f_{-1/2}$		$\delta^5 f_{-1/2}$		Gauss hacia adelante
x_0	f_0		$\delta^2 f_0$		$\delta^4 f_0$	...		
		$\delta f_{1/2}$		$\delta^3 f_{1/2}$		$\delta^5 f_{1/2}$		
x_1	f_1		$\delta^2 f_1$		$\delta^4 f_1$			
		$\delta f_{3/2}$		$\delta^3 f_{3/2}$				
x_2	f_2		$\delta^2 f_2$					
		$\delta f_{5/2}$						
x_3	f_3							
⋮	⋮	⋮	⋮	⋮	⋮	⋮		

Ejemplo 15. Con los datos del ejemplo 8, interpolemos para $x = 0.79$.

Consideramos $x_0 = 0.75$, $x_1 = 0.80$, $x_{-1} = 0.70$, $x_2 = 0.85$, $x_{-2} = 0.65$, $x_3=0.90$. Entonces, como $x = 0.79 = 0.75 + 0.05t$, resulta $t = 0.8$. Reemplazando en la ecuación (25), se tiene

$$P_5(0.79) = 1.3333333+0.8(-0.0833333)+\frac{0.8(-0.2)}{2}0.0119048+\frac{0.8(-0.36)}{3!}(-0.0021009)+$$

$$+\frac{0.8(-0.36)(-1.2)}{4!}0.0006463+\frac{0.8(-0.36)(-3.36)}{5!}(-0.0001794) = 1.2658230.$$

La estimación del error está dada por

$$E(0.79) = 0.05^6\ \frac{(0.8)(-0.36)(-3.36)(-2.2)}{6!}\ \frac{720}{\xi^7}$$

con ξ comprendido entre 0.65 *y* 0.90, esto es,

$$-0.000000679 \le E(0.79) \le -0.0000000694.$$

4.3.4.2. Fórmula de interpolación de Gauss hacia atrás con diferencias centrales.

Introduciendo las abscisas en el orden siguiente con diferencias centrales: x_0, x_{-1}, x_1, x_{-2}, x_2, x_{-3}, x_3, ... en la fórmula de diferencias divididas (10), se obtiene

$$f(x) = f(x_0) + (x-x_0) f(x_0 x_{-1}) + (x-x_0)(x-x_{-1}) f(x_0 x_{-1} x_1) + (x-x_0)(x-x_{-1})(x-x_1) f(x_0 x_{-1} x_1 x_{-2}) +$$

$$+ (x-x_0)(x-x_{-1})(x-x_1)(x-x_{-2}) f(x_0 x_{-1} x_1 x_{-2} x_2) + \ldots \qquad (28)$$

Reemplazando las diferencias divididas en función de las diferencias centrales, resulta

$$f(x_0 x_{-1}) = f(x_{-1} x_0) = \frac{f(x_0) - f(x_{-1})}{x_0 - x_{-1}} = \frac{\delta f_{-1/2}}{h}$$

$$f(x_0 x_{-1} x_1) = f(x_{-1} x_0 x_1) = \frac{f(x_0 x_1) - f(x_{-1} x_0)}{x_1 - x_{-1}} = \frac{\delta f_{1/2} - \delta f_{-1/2}}{2h^2} = \frac{1}{2! h^2} \delta^2 f_0$$

$$f(x_0 x_{-1} x_1 x_{-2}) = f(x_{-2} x_{-1} x_0 x_1) = \frac{f(x_{-1} x_0 x_1) - f(x_{-2} x_{-1} x_0)}{x_1 - x_{-2}} = \frac{1}{3h}\left[\frac{\delta^2 f_0}{2h^2} - \frac{\delta^2 f_{-1}}{2h^2}\right] = \frac{1}{3! h^3} \delta^3 f_{-1/2}$$

$$f(x_0 x_{-1} x_1 x_{-2} x_2) = f(x_{-2} x_{-1} x_0 x_1 x_2) =$$

$$= \frac{f(x_{-1} x_0 x_1 x_2) - f(x_{-2} x_{-1} x_0 x_1)}{x_2 - x_{-2}} = \frac{1}{4h}\left[\frac{\delta^3 f_{1/2}}{3! h^3} - \frac{\delta^3 f_{-1/2}}{3! h^3}\right] = \frac{1}{4! h^4} \delta^4 f_0$$

Luego,

$$f(x) = f(x_0) + \frac{(x-x_0)}{h} \delta f_{-1/2} + \frac{(x-x_0)(x-x_{-1})}{2! h^2} \delta^2 f_0 + \frac{(x-x_0)(x-x_{-1})(x-x_1)}{3! h^3} \delta^3 f_{-1/2} +$$

$$+ \frac{(x-x_0)(x-x_{-1})(x-x_1)(x-x_{-2})}{4! h^4} \delta^4 f_0 + \frac{(x-x_0)(x-x_{-1})(x-x_1)(x-x_{-2})(x-x_2)}{5! h^5} \delta^5 f_{-1/2} + \ldots \qquad (29)$$

Introduciendo la variable $t = (x - x_0)/h$, o sea, $x = x_0 + ht$, resulta

$$f(x) = f(x_0) + t\delta_{-1/2} + \frac{t(t+1)}{2!}\delta^2 f_0 + \frac{t(t^2-1^2)}{3!}\delta^3 f_{-1/2} + \frac{t(t^2-1^2)(t+2)}{4!}\delta^4 f_0 + \tag{30}$$

$$+ \frac{t(t^2-1^2)(t^2-2^2)}{5!}\delta^5 f_{-1/2} + \ldots + \frac{t(t^2-1^2)(t^2-2^2)\ldots(t^2-(m-1)^2)(t+m)}{(2m)!}\delta^{2m} f_0$$

ó

$$+ \frac{t(t^2-1^2)(t^2-2^2)\ldots(t^2-m^2)}{(2m+1)!}\delta^{2m+1} f_{-1/2} + E(x)$$

Si en la ecuación (30) se retiene la diferencia *n*-ésima, donde $n = 2m$, o sea hasta una diferencia de orden par, el error está dado por

$$E(x) = \frac{h^{(2m+1)} t(t^2-1^2)(t^2-2^2)\ldots(t^2-m^2)}{(2m+1)!} f^{(2m+1)}(\xi) \tag{31}$$

Si en la ecuación (30) se retiene hasta la diferencia *n*-ésima, donde $n = 2m+1$, o sea hasta una diferencia de orden impar, el error está dado por

$$E(x) = \frac{h^{(2m+2)} t(t^2-1^2)(t^2-2^2)\ldots(t^2-m^2)(t+m+1)}{(2m+2)!} f^{(2m+2)}(\xi) \tag{32}$$

Análogamente como se hizo para el error en la fórmula de interpolación de Gauss hacia adelante, si no se conoce la expresión analítica de $f(x)$ la ecuación (31) se transforma en

$$E(x) \approx \frac{t(t^2-1^2)(t^2-2^2)\ldots(t^2-m^2)}{(2m+1)!}\delta^{2m+1} f_{-1/2}$$

y la ecuación (32) puede escribirse

$$E(x) \approx \frac{t(t^2-1^2)(t^2-2^2)\ldots(t^2-m^2)(t+m+1)}{(2m+2)!}\delta^{2m+2} f_0$$

La ecuación (30) se conoce con el nombre de *fórmula de interpolación de Gauss hacia atrás con diferencias centrales.*

El recorrido que hace esta fórmula también es en zig-zag conservando siempre los argumentos 0 y -1/2, y está representado en la tabla siguiente

x	$f(x)$	δf	$\delta^2 f$	$\delta^3 f$	$\delta^4 f$	$\delta^5 f$ ...	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_{-3}	f_{-3}						
		$\delta f_{-5/2}$					
x_{-2}	f_{-2}		$\delta^2 f_{-2}$				
		$\delta f_{-3/2}$		$\delta^3 f_{-3/2}$			
x_{-1}	f_{-1}		$\delta^2 f_{-1}$		$\delta^4 f_{-1}$		
		$\boldsymbol{\delta f_{-1/2}}$		$\boldsymbol{\delta^3 f_{-1/2}}$		$\boldsymbol{\delta^5 f_{-1/2}}$	Gauss hacia atrás
x_0	$\boldsymbol{f_0}$		$\boldsymbol{\delta^2 f_0}$		$\boldsymbol{\delta^4 f_0}$	...	
		$\delta f_{1/2}$		$\delta^3 f_{1/2}$		$\delta^5 f_{1/2}$	
x_1	f_1		$\delta^2 f_1$		$\delta^4 f_1$		
		$\delta f_{3/2}$		$\delta^3 f_{3/2}$			
x_2	f_2		$\delta^2 f_2$				
		$\delta f_{5/2}$					
x_3	f_3						
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$ $\vdots$	

Ejemplo 16. Considerando el ejemplo 8, interpolemos para $x = 0.79$.

Consideramos $x_0 = 0.80$, $x_{-1} = 0.75$, $x_1 = 0.85$, $x_{-2} = 0.70$, $x_2 = 0.90$, $x_{-3} = 0.65$. Entonces, $x = 0.79 = 0.80 + 0.05t$, de donde, $t = -0.20$. Según la ecuación (30), se tiene

$$P_5(0.79) = 1.25 + (-0.2)(-0.0833333) + \frac{(-0.2)(0.8)}{2} 0.0098039 + \frac{(-0.2)(-0.96)}{3!} (-0.0021009) +$$

$$+ \frac{(-0.2)(-0.96)(1.8)}{4!} 0.0004669 + \frac{(-0.2)(-0.96)(-3.96)}{5!} (-0.0001794) = 1.2658230.$$

Obtenemos el mismo valor usando la fórmula de Gauss hacia adelante. Esto se entiende perfectamente, dado que en ambos casos usamos los mismos puntos para determinar el polinomio de interpolación. La estimación del error está dada por

$$E(0.79) = 0.05^6 \frac{(-0.2)(-0.96)(-3.96)(2.98)}{6!} \frac{720}{\xi^7}$$

con ξ comprendido entre 0.65 y 0.90.

Luego,

$$-0.00000073 \le E(0.79) \le -0.000000074.$$

4.3.4.3. Fórmula de interpolación de Stirling.

La fórmula de Stirling se obtiene promediando las dos fórmulas de Gauss y se usa cuando se interpola en el centro para valores de argumentos tales que $x_0 - \frac{h}{2} < x < x_0 + \frac{h}{2}$.

Sumando las ecuaciones (25) y (30) y dividiendo por 2, se obtiene

$$f(x_0 + th) = f(x_0) + \frac{t}{2}(\delta f_{1/2} + \delta f_{-1/2}) + \frac{t}{2!}\frac{1}{2}\left[(t-1)+(t+1)\right]\delta^2 f_0 +$$
$$+ \frac{t(t^2-1^2)}{3!}\frac{1}{2}(\delta^3 f_{1/2} - \delta^3 f_{-1/2}) + \frac{t(t^2-1^2)}{4!}\frac{\left[(t-2)+(t+2)\right]}{2}\delta^4 f_0 +$$
$$+ \frac{t(t^2-1^2)(t^2-2^2)}{5!}\frac{1}{2}(\delta^5 f_{1/2} + \delta^5 f_{-1/2}) + \ldots +$$

$$+ \frac{t(t^2-1^2)(t^2-2^2)\ldots(t^2-(m-1)^2)}{(2m)!}\frac{\left[(t-m)+(t+m)\right]}{2}\delta^{2m} f_0 \qquad (33)$$

ó

$$+ \frac{t(t^2-1^2)(t^2-2^2)\ldots(t^2-m^2)}{(2m+1)!}\frac{(\delta^{2m+1} f_{1/2} + \delta^{2m+1} f_{-1/2})}{2} + E(x)$$

Es usual encontrar en los textos la letra μ para indicar el promedio. Así,

$$\mu\delta f_0 = \frac{1}{2}(\delta f_{1/2} + \delta f_{-1/2}), \qquad \mu\delta^3 f_0 = \frac{1}{2}(\delta^3 f_{1/2} + \delta^3 f_{-1/2}).$$

Con esta notación, la fórmula de Stirling se escribe

$$f(x_0+th) = f(x_0) + t\,\mu\delta f_0 + \frac{t^2}{2}\delta^2 f_0 + \frac{t(t^2-1^2)}{3!}\mu\delta^3 f_0 + \frac{t^2(t^2-1^2)}{4!}\delta^4 f_0 +$$

$$+ \frac{t(t^2-1^2)(t^2-2^2)}{5!}\mu\delta^5 f_0 + \ldots + \frac{t(t^2-1^2)(t^2-2^2)\ldots(t^2-(m-1)^2)}{(2m)!}\delta^{2m} f_0 \qquad (34)$$

ó

$$+\frac{t(t^2-1^2)(t^2-2^2)...(t^2-m^2)}{(2m+1)!}\mu\delta^{2m+1}f_0 + E(x)$$

Si la ecuación (34) termina con una diferencia $n = 2m$ de orden par, el error está dado por

$$E(x) = h^{(2m+1)}\frac{t(t^2-1^2)(t^2-2^2)...(t^2-m^2)}{(2m+1)!}f^{(2m+1)}(\xi) \qquad (35)$$

Si la ecuación (34) termina con una diferencia $n = 2m + 1$ de orden impar, el error está dado por

$$E(x) = h^{(2m+2)}\frac{t(t^2-1^2)(t^2-2^2)...(t^2-m^2)}{(2m+2)!}\left[\frac{(t-m-1)f^{(2m+2)}(\xi_1)+(t+m+1)f^{(2m+2)}(\xi_2)}{2}\right] (36)$$

En la ecuación (35) ξ es un punto comprendido entre x, x_0 y $x_{\pm1}, x_{\pm2},\ldots,x_{\pm m}$.

Análogamente, en la ecuación (36) ξ_1 y ξ_2 son puntos interiores del intervalo $x_0, x_{\pm 1}, x_{\pm 2}, \ldots, x_{\pm m}, x_{\pm (m+1)}$ y x.

La ecuación (34) es conocida como *la fórmula de interpolación de Stirling*. El recorrido que hace dicha fórmula en la tabla de las diferencias es horizontal con argumento 0:

$$x_0 \quad f_0 \quad \frac{\delta f_{-1/2}}{\delta f_{1/2}} \quad P \quad \delta^2 f_0 \quad \frac{\delta^3 f_{-1/2}}{\delta^3 f_{1/2}} \quad P \quad \delta^4 f_0 \quad \frac{\delta^5 f_{-1/2}}{\delta^5 f_{1/2}} \quad P$$

La letra *P* indica que debe hacerse el promedio entre el valor superior e inferior, en este caso el promedio de las diferencias impares.

Análogamente como hicimos en Gauss, si no se conoce la expresión analítica de $f(x)$, entonces la ecuación (35) puede escribirse

$$E(x) \approx \frac{t(t^2-1^2)(t^2-2^2)...(t^2-m^2)}{2(2m+1)!}(\delta^{2m+1}f_{1/2} + \delta^{2m+1}f_{-1/2})$$

y la ecuación (36)

$$E(x) \approx \frac{t(t^2-1^2)(t^2-2^2)...(t^2-m^2)}{2(2m+2)!}\left[(t-m-1)+(t+m+1)\right]\delta^{2m+2} f_0$$

Ejemplo 17. Considerando el ejemplo 8, nos proponemos calcular el valor de la función en 0.77 y 0.72.

Para $x = 0.77$, consideramos $x_0 = 0.75$, $x_1 = 0.80$, $x_{-1} = 0.70$, $x_2 = 0.85$, $x_{-2} = 0.65$, $x_3 = 0.90$, $x_{-3} = 0.60$. Entonces, $x = 0.77 + 0.05t$, de donde, $t = 0.4$.

Luego, observando el cuadro y haciendo el promedio de las diferencias impares, se tiene

$$P_5(0.77) = 1.3333333 + 0.4(-0.0892857) + \frac{0.16}{2}\,0.0119048 + \frac{0.4(-0.84)}{3!}\,(-0.0024241) +$$

$$+\frac{0.16(-0.84)}{4!}\,0.0006463 + \frac{0.4(-0.84)(-3.84)}{5!}\,(-0.0002245) = 1.2987011.$$

Para hacer una estimación del error suponemos $\xi_1 = \xi_2$ en la ecuación (36) de modo que

$$E(x) \approx 0.05^6\,\frac{t(t^2-1^2)(t^2-2^2)}{6!}\frac{720}{\xi^7}$$

con ξ comprendido entre 0.60 y 0.90.

Luego,

$$0.0000000 \le E(0.77) \le 0.0000003.$$

Análogamente, para $x = 0.72 = x_0 + ht = 0.70+0.05t$ resulta $t = 0.4$. Observando la tabla de las diferencias y haciendo el promedio de las diferencias impares en la horizontalidad de $x = 0.70$, se tiene

$$P_5(0.72) = 1.4285714 + 0.4(-0.1025641) + \frac{0.16}{2}\,0.0146520 + \frac{0.4(-0.84)}{3!}\,(-0.0032052) +$$

$$+\frac{0.16(-0.84)}{4!}\,0.0009159 + \frac{0.4(-0.84)(-3.84)}{5!}\,(-0.0003427) = 1.3888886.$$

4.3.4.4. Fórmula de interpolación de Bessel.

Cuando se trata de interpolar para valores de x tales que $x_0 + \frac{h}{2} < x < x_0 + h$, se suele también usar la fórmula de Bessel. Esta fórmula es de uso frecuente en Astronomía. Puede obtenerse promediando la suma de las fórmulas de Gauss hacia adelante y de Gauss hacia atrás, pero escribiendo esta última tomando como origen x_1.

La fórmula de Gauss hacia atrás (30) es

$$f(x) = f(x_0) + t\delta_{-1/2} + \frac{t(t+1)}{2!}\delta^2 f_0 + \frac{t(t+1)(t-1)}{3!}\delta^3 f_{-1/2} + \frac{t(t+1)(t-1)(t+2)}{4!}\delta^4 f_0 +$$

$$+ \frac{t(t+1)(t-1)(t+2)(t-2)}{5!}\delta^5 f_{-1/2} + \ldots$$

Si en lugar de elegir como origen x_0 elegimos x_1, todos los subíndices deben correrse una unidad y t reemplazado por (t -1), pues se tiene que $x = x_0 + ht = = x_1 - h + ht = x_1 + h(t - 1)$, de modo que la ecuación (30) se transforma en

$$f(x) = f(x_1) + (t-1)\delta f_{1/2} + \frac{t(t-1)}{2!}\delta^2 f_1 + \frac{t(t-1)(t-2)}{3!}\delta^3 f_{1/2} + \frac{(t-1)t(t-2)(t+1)}{4!}\delta^4 f_1 +$$

$$+ \frac{(t-1)t(t-2)(t+1)(t-3)}{5!}\delta^5 f_{1/2} + \ldots$$

Sumando a ésta la ecuación (25)

$$f(x) = f_0 + t\delta f_{1/2} + \frac{t(t-1)}{2!}\delta^2 f_0 + \frac{t(t-1)(t+1)}{3!}\delta^3 f_{1/2} + \frac{t(t-1)(t+1)(t-2)}{4!}\delta^4 f_0 +$$

$$+ \frac{t(t-1)(t+1)(t-2)(t+2)}{5!}\delta^5 f_{1/2} + \ldots$$

y luego promediando, se obtiene

$$f(x) = \frac{1}{2}(f_0 + f_1) + (t-1/2)\delta f_{1/2} + \frac{t(t-1)}{2!}\left(\frac{\delta^2 f_1 + \delta^2 f_0}{2}\right) + \frac{t(t-1)(t-1/2)}{3!}\delta^3 f_{1/2} +$$

$$+\frac{(t-1)t(t-2)(t+1)}{4!}\left(\frac{\delta^4 f_1+\delta^4 f_0}{2}\right)+\frac{t(t-1)(t+1)(t-2)(t-1/2)}{5!}\delta^5 f_{1/2}+\ldots$$

Indicando con

$$\mu f_{1/2}=\frac{1}{2}(f_0+f_1),\qquad \mu\delta^2 f_{1/2}=\frac{(\delta^2 f_0+\delta^2 f_1)}{2},\qquad \mu\delta^4 f_{1/2}=\frac{(\delta^4 f_0+\delta^4 f_1)}{2}$$

se obtiene

$$f(x)=\mu f_{1/2}+(t-1/2)\delta f_{1/2}+\frac{t(t-1)}{2}\mu\delta^2 f_{1/2}+\frac{t(t-1/2)(t-1)}{3!}\delta^3 f_{1/2}+\frac{t^2(t^2-1^2)(t-2)}{4!}\mu\delta^4 f_{1/2}+ \tag{37}$$

$$+\frac{t(t^2-1^2)(t-2)(t-1/2)}{5!}\delta^5 f_{1/2}+\ldots+\frac{t(t^2-1^2)(t^2-2^2)\ldots(t^2-(m-1)^2)(t-m)}{(2m)!}\mu\delta^{2m} f_{1/2}$$

ó

$$+\frac{t(t^2-1^2)(t^2-2^2)\ldots(t^2-(m-1)^2)(t-m)(t-1/2)}{(2m+1)!}\delta^{2m+1} f_{1/2}+E(x)$$

Si la ecuación (37) termina con una diferencia $n = 2m$ de orden par, el error está dado por

$$E(x)=h^{(2m+1)}\frac{t(t^2-1^2)(t^2-2^2)\ldots(t^2-(m-1)^2)(t-m)(t-1/2)}{(2m+1)!}f^{(2m+1)}(\xi) \tag{38}$$

Si la ecuación (37) termina con una diferencia $n = 2m + 1$ de orden impar, el error está dado por

$$E(x)=h^{(2m+2)}\frac{t(t^2-1^2)(t^2-2^2)\ldots(t^2-m^2)(t-(m+1))}{(2m+2)!}f^{(2m+2)}(\xi) \tag{39}$$

Análogamente como hemos hecho antes, si no se conoce la expresión analítica de $f(x)$, entonces la ecuación (38) puede escribirse

$$E(x)\approx\frac{t(t^2-1^2)(t^2-2^2)\ldots(t^2-(m-1)^2)(t-m)(t-1/2)}{(2m+1)!}\delta^{2m+1} f_{1/2}$$

y la ecuación (39)

$$E(x) \approx \frac{t(t^2-1^2)(t^2-2^2)...(t^2-m^2)(t-(m+1))}{(2m+2)!}\left(\frac{\delta^{2m+2}f_0+\delta^{2m+2}f_1}{2}\right)$$

La ecuación (37) se conoce como *la fórmula de interpolación de Bessel*. Se observa en la tabla de las diferencias que esta fórmula recorre la horizontal con argumento 1/2:

$$\begin{array}{lllllll} x_0 & f_0 & & \delta^2 f_0 & & \delta^4 f_0 & \\ & P & \delta f_{1/2} & P & \delta^3 f_{1/2} & P & \delta^5 f_{1/2} \\ x_1 & f_1 & & \delta^2 f_1 & & \delta^4 f_1 & \end{array}$$

La letra P indica que debe hacerse el promedio entre el valor superior e inferior que aparece en la tabla de diferencias.

Ejemplo 18. Consideramos el ejemplo 8 para interpolar en 0.73 y 0.83.

Entonces, para $x = 0.73 = 0.70 + 0.05t$, se tiene, $t = 0.6$. De la tabla de diferencias surge

$$P_5(0.73) = 1.3809524 + 0.1(-0.0952381) + \frac{0.6(-0.4)}{2}0.0132784 + \frac{0.6(-0.4)(0.1)}{3!}(-0.0027472) +$$

$$+\frac{0.6(-0.64)(-1.4)}{4!}0.0007811 + \frac{0.6(-0.64)(-1.4)(0.1)}{5!}(-0.0002696) = 1.3698635.$$

En este caso ($x = 0.73$), hemos considerado $x_0 = 0.70$, $x_1 = 0.75$, $x_{-1} = 0.65$, $x_2 = 0.80$, $x_{-2} = 0.60$, $x_3 = 0.85$.

Si queremos acotar el error, teniendo presente que hemos realizado el cálculo tomando hasta la diferencia quinta se tiene

$$E(x) = h^6 \frac{t(t^2-1^2)(t^2-2^2)(t-3)}{6!}f^{(6)}(\xi) = 0.05^6 \frac{0.6(-0.64)(-3.64)(-2.4)}{6!}\frac{720}{\xi^7}$$

con ξ comprendido entre 0.60 y 0.85.

Luego,

$$-0.0000019 \leq E(0.73) \leq -0.0000001.$$

En forma análoga, para $x = 0.83$ consideramos $x_0 = 0.80$, $x_1 = 0.85$, $x_{-1} = 0.75$, $x_2 = 0.90$, $x_{-2} = 0.70$, $x_3 = 0.95$, así que completando la tabla del ejemplo 8 se tiene

⋮						-0.0001228
					0.0003441	
				-0.0012899		
0.90	1.1111111		0.0068800			
		-0.0584795				
0.95	1.0526316					

y aplicando la fórmula de Bessel, de la tabla de diferencias, surge

$$P_5(0.83)=1.2132353 + 0.1(-0.0735294)+\frac{0.6(-0.4)}{2}0.0089869+\frac{0.6(-0.4)(0.1)}{3!}(-0.0016340)+$$

$$+\frac{0.6(-0.64)(-1.4)}{4!}0.0004055+\frac{0.6(-0.64)(-1.4)(0.1)}{5!}(-0.0001228)=1.2048195.$$

La estimación del error está dada por

$$E(0.83) = 0.05^6\ \frac{0.6(-0.64)(-3.64)(-2.4)}{6!}\frac{720}{\xi^7}$$

con ξ comprendido entre 0.70 y 0.95.

Luego,

$$-0.00000064 \le E(0.83) \le -0.000000075.$$

Observación. Las tablas de diferencias están compiladas ordinariamente para una exactitud de un cierto valor decimal. Si la función $f(x)$ tiene derivadas continuas hasta el orden m, entonces dado un intervalo suficientemente pequeño, $h = \Delta x$, sus diferencias hasta el m-ésimo orden inclusive varían suavemente y la m-ésima diferencia es prácticamente constante dentro de los límites de los decimales dados. Cualquier violación de esta condición en una sección de una tabla, generalmente, indica un error de cálculo (si la función no tiene singularidades). Cuando el esquema de las diferencias muestra regularidad y no existen fluctuaciones en el cálculo de las

diferencias sucesivas, esto pone en evidencia que los datos son correctos y que el polinomio interpolante dará una buena aproximación en el intervalo considerado.

Ejemplo 19. Sea $f(x) = cos\ (x)$ y consideremos la siguiente tabla de diferencias:

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$
0º	1				
		-0.0003427			
1º5'	0.9996573		-0.0006851		
		-0.0010278		0.0000007	
3º	0.9986295		-0.0006844		0.0000005
		-0.0017122		0.0000012	
4º5'	0.9969173		-0.0006832		0.0000004
		-0.0023954		0.0000016	
6º	0.9945219		-0.0006816		0.0000004
		-0.0030770		0.0000020	
7º5'	0.9914449		-0.0006796		
		-0.0037566			
9º	0.9876883				

En este ejemplo se observa que las diferencias cuartas son prácticamente constantes, y esto indica que un polinomio de cuarto grado debe proporcionar un buen ajuste para los propósitos de interpolación en el intervalo $x_0 < x < x_n$.

Puede ser de interés mostrar algunos ejemplos donde el polinomio de interpolación no da una buena aproximación de la función, lo cual se pone de manifiesto al formar el cuadro de diferencias.

Ejemplo 20. Sea $f(x) = cos\ (\pi x)$ y consideremos la siguiente tabla de diferencias:

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$	$\Delta^5 f$
0	1					
		-2				
1	-1		4			
		2		-8		
2	1		-4		16	
		-2		8		-32
3	-1		4		-16	
		2		-8		32
4	1		-4		16	
		-2		8		
5	-1		4			
		2				
6	1					

El cuadro muestra que las diferencias sucesivas van creciendo y los signos se van alternando. El polinomio interpolante no da una buena aproximación de la función.

La misma función fue tabulada en un intervalo más reducido y el polinomio interpolante da, en este caso, una buena representación de la función.

Este ejemplo muestra que para una función periódica el polinomio interpolante no da un buen ajuste para la función, salvo que se trate de un intervalo reducido como ya se indicó.

Ejemplo 21. Sea $f(x) = 1/(3x + 1)$ y consideremos la siguiente tabla de diferencias:

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$	$\Delta^5 f$
-3	-0.125					
		-0.075				
-2	-0.2		-0.225			
		-0.3		2.025		
-1	-0.5		1.8		-6.075	
		1.5		-4.05		13.01786
0	1		-2.25		6.94286	
		-0.75		2.89286		-10.4143
1	0.25		0.64286		-3.47144	
		-0.10714		-0.57858		
2	0.14286		0.06428			
		-0.04286				
3	0.1					

En esta función las diferencias finitas se van incrementando y presentan fluctuaciones de signos. Acá tampoco el polinomio interpolante da una buena aproximación de la función.

Esta función no es continua en el intervalo (-3, 3); la misma es discontinua en $x = -1/3$.

Ejemplo 22. Sea $f(x) = x + 1/x$ y consideremos la siguiente tabla de diferencias:

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$	$\Delta^5 f$
-5	5.2					
		1.8666667				
-3	-3.3333333		-0.5333334			
		1.3333333		3.2000001		
-1	-2		2.6666667		-8.5333335	
		4		5.3333334		17.0666667
1	2		-2.6666667		8.5333335	
		1.3333333		-3.2000001		
3	3.3333333		0.5333334			
		1.8666667				
5	5.2					

La función en este ejemplo no está definida en $x = 0$.

Ejemplo 23. Sea $f(x) = 2 + x^{2/3}$ y consideremos la siguiente tabla de diferencias:

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$	$\Delta^5 f$
-3	4.0800838					
		-0.4926827				
-2	3.5874011		-0.0947184			
		-0.5874011		-0.3178805		
-1	3		-0.4125989		2.7304794	
		-1		2.4125989		-7.5556772
0	2		2		-4.8251978	
		1		-2.4125989		7.5556772
1	3		-0.4125989		-2.7304794	
		0.5874011		0.3178805		
2	3.5874011		-0.0947184			
		0.4926827				
3	4.0800838					

En este ejemplo la función está definida en el intervalo (-3, 3), pero la derivada de la función en $x = 0$ se hace infinita y en este caso el polinomio interpolante no da un buen ajuste de la función.

4.3.5. Fórmula de interpolación de Lagrange.

Para muchos propósitos es conveniente disponer de una fórmula de interpolación que esté expresada explícitamente en función de sus ordenadas en lugar de sus diferencias; es decir, de una fórmula que no requiera cálculos previos en la tabla de diferencias.

Sea entonces f una función definida sobre algún intervalo [*a*, *b*] y sean $x_0, x_1, \cdots, x_n$ $(n+1)$ puntos distintos del intervalo [*a*, *b*], los cuales no se suponen que sean equidistantes y no es necesario que estén numerados es su orden natural. Suponemos *f* continuamente diferenciable $(n+1)$ veces en el intervalo [*a*, *b*]. Sean además conocidos los valores correspondientes de la función $y = f(x)$ para $x_0, x_1, \cdots, x_n$, esto es, conocemos $f(x_0)$, $f(x_1)$, ..., $f(x_n)$.

Se desea construir un polinomio $P_n(x)$ de grado no superior a *n* y que tenga para los puntos especificados $x_0, x_1, \cdots, x_n$ los mismos valores que la función $f(x)$; esto es, tal que

$$P_n(x_k) = f(x_k) = f_k, \qquad k = 0, 1, \ldots, n$$

Debe quedar claro que lo que presentaremos a continuación es una forma alternativa del polinomio de interpolación asociado con una tabla de datos (x_k, f_k) con $0 \le k \le n$, puesto que como sabemos, existe uno y sólo un polinomio de interpolación de grado máximo *n* asociado con los datos, suponiendo que las $(n+1)$ abscisas x_k son distintas. En resumen, veremos un método alternativo para expresar el polinomio de interpolación.

Resolvamos primero un problema particular: construyamos un polinomio $L_k(x)$, tal que

$$L_k(x_j) = \begin{cases} 0, & si \;\; j \neq k \\ 1, & si \;\; j = k \end{cases} \tag{40}$$

Como el polinomio deseado se hace nulo para *n* puntos, $x_0, x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_n$, entonces tiene la forma

$$L_k(x) = C_k (x - x_0)(x - x_1) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n). \tag{41}$$

(descomposición factorial de un polinomio), donde C_k es un coeficiente constante. Estableciendo, $x = x_k$ en la ecuación (41) y teniendo en cuenta que $L_k(x_k) = 1$, tenemos

$$C_k (x_k - x_0)(x_k - x_1) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n) = 1$$

Por consiguiente,

$$C_k = \frac{1}{(x_k - x_0)(x_k - x_1) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n)}$$

Poniendo este valor en la ecuación (41), tenemos

$$L_k(x) = \frac{(x-x_0)(x-x_1)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)} = \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{(x-x_j)}{(x_k-x_j)}. \quad (42)$$

El polinomio así construido satisface las condiciones requeridas: si $x = x_k$, entonces todos los factores tienen el valor 1 y, por lo tanto, el producto también. Por otra parte, si $x = x_j$ donde $j \neq k$, entonces el factor que contiene a $x - x_j$ es cero y el producto se anula.

Además, otro resultado que obtenemos es que como es un producto de $n + 1 - 1 = n$ factores lineales, entonces $L_k(x)$ representa un polinomio de grado n.

Ataquemos ahora la solución del problema general: hallar un polinomio $P_n(x)$ de grado que no exceda a n y que satisfaga que $P_n(x_k) = f_k$ para $k = 0, 1, ..., n$.

Este polinomio es de la forma

$$P_n(x) = \sum_{k=0}^{n} L_k(x) f_k \quad (43)$$

(y este es el polinomio conocido con el nombre de *polinomio de interpolación de Lagrange*).

En efecto, en primer lugar es evidente que el grado del polinomio $P_n(x)$ construido de este modo no es superior a n (por ser suma de polinomios de grado n con factores constantes). En segundo lugar, y por la condición (40), tenemos

$$P_n(x_j) = \sum_{k=0}^{n} L_k(x_j) f_k = L_j(x_j) f_j = f_j , \quad j = 0, 1, \cdots, n.$$

Sustituyendo ahora el valor $L_k(x)$ de la ecuación (42) en la (43), obtenemos

$$P_n(x) = \sum_{k=0}^{n} \frac{(x-x_0)(x-x_1)...(x-x_{k-1})(x-x_{k+1})...(x-x_n)}{(x_k-x_0)(x_k-x_1)...(x_k-x_{k-1})(x_k-x_{k+1})...(x_k-x_n)} f_k =$$

$$= \frac{(x-x_1)(x-x_2)...(x-x_n)}{(x_0-x_1)(x_0-x_2)...(x_0-x_n)} f_0 + \frac{(x-x_0)(x-x_2)...(x-x_n)}{(x_1-x_0)(x_1-x_2)...(x_1-x_n)} f_1 + ... +$$

$$+\frac{(x-x_0)(x-x_1)...(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)...(x_n-x_{n-1})}f_n \qquad (44)$$

La ecuación (44) se llama *fórmula de interpolación de Lagrange.*

Los polinomios $L_k(x)$ dados por la ecuación (42) se llaman *coeficientes de interpolación de Lagrange*. De la ecuación (42) podemos escribir la (44) sintéticamente en la forma

$$P_n(x)=\sum_{k=0}^{n}\prod_{\substack{j=0\\ j\neq k}}^{n}\frac{(x-x_j)}{(x_k-x_j)}f_k$$

Esta ecuación es particularmente larga si el orden n es grande. Sin embargo, su escritura en un programa de computación necesita únicamente un número pequeño de líneas (observando la ecuación (44) se ve que el primer término es f_0 veces el producto $\frac{(x-x_i)}{(x_0-x_i)}$ para toda i, excepto para i = 0; el segundo término es f_1 veces el producto $\frac{(x-x_i)}{(x_1-x_i)}$ para toda i, excepto para i = 1; los otros términos siguen el mismo patrón).

Consideremos dos casos especiales del polinomio de interpolación de Lagrange. Para n = 1, tenemos dos puntos y la fórmula de Lagrange es, entonces, la ecuación de la recta $y = P_1(x)$ que pasa por dichos puntos

$$y=P_1(x)=\frac{x-b}{a-b}f(a)+\frac{x-a}{b-a}f(b)$$

donde a y b son las abscisas de esos puntos. (*Interpolación lineal*).

Para n = 2, tenemos la ecuación de la parábola $y = P_2(x)$ que pasa por estos puntos

$$y=P_2(x)=\frac{(x-b)(x-c)}{(a-b)(a-c)}f(a)+\frac{(x-a)(x-c)}{(b-a)(b-c)}f(b)+\frac{(x-a)(x-b)}{(c-a)(c-b)}f(c)$$

donde a, b y c son las abscisas de los puntos dados. (*Interpolación cuadrática*).

4.3.5.1. El error de la fórmula de interpolación de Lagrange.

Como deseamos usar el polinomio de interpolación de Lagrange para aproximar a la función f en puntos que no pertenecen al conjunto de puntos de interpolación x_k, k = 0, 1, ..., n, estamos interesados en estimar la

diferencia $f(x) - P_n(x)$ para $x \in [a, b]$, intervalo éste que contiene a los puntos de interpolación. Es claro que sin hipótesis posteriores nada en absoluto se puede decir acerca de esa cantidad. Podemos, en efecto, cambiar la función *f* a voluntad en puntos que no sean los puntos de interpolación sin cambiar el polinomio P_n en absoluto (ver Figura 6).

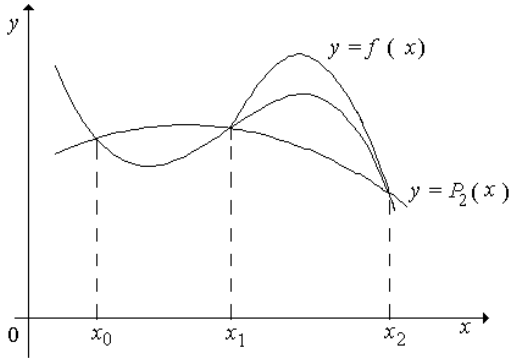


Figura 6

Podemos, sin embargo, hacer una afirmación definida si suponemos un conocimiento cualitativo de las derivadas de la función *f*. Usaremos entonces las hipótesis dadas al comienzo, a saber: sea f una función definida sobre un intervalo [*a*, *b*] y sean $x_0, x_1, \cdots, x_n$ $(n + 1)$ puntos distintos de [*a*, *b*]. No se supone que estos puntos sean equidistantes ni incluso que se enuncien en su orden natural. Suponemos además que f es continuamente diferenciable $(n + 1)$ veces sobre el intervalo [*a*, *b*]. Entonces, para cada $x \in [a, b]$ existe un punto ξ localizado dentro del intervalo [*a*, *b*] que contiene a los puntos $x, x_0, x_1, \cdots, x_n$ tal que

$$E(x) = f(x) - P_n(x) = \frac{1}{(n+1)!} F(x) f^{(n+1)}(\xi), \text{ donde, } F(x) = \prod_{k=0}^{n} (x - x_k) \quad (45)$$

y P_n es el polinomio interpolante dado por la ecuación (44).

Probemos la ecuación (45). Si *x* es uno de los puntos x_k no hay nada que probar ya que ambos miembros de la ecuación (45) se anulan para cualquier ξ. Si *x* es un valor fijo diferente de cualquiera de los puntos x_k, $k = 0, 1, ..., n$, consideremos la función auxiliar $G = G(t)$ definida para *t* en [*a*, *b*] por

$$G(t) = f(t) - P_n(t) - C\,F(t), \text{ donde, } C = \frac{f(x) - P_n(x)}{F(x)}. \quad (46)$$

Para $t = x_k$

$$G(x_k) = f(x_k) - P_n(x_k) - CF(x_k) = f_k - f_k - C\,0 = 0, \quad k = 0, 1, ..., n.$$

Además, para $t = x$

$$G(x) = f(x) - P_n(x) - CF(x) = 0, \text{ por la definición de } C.$$

Por lo tanto, la función G se anula en los $(n + 2)$ números distintos $x, x_0, x_1, \cdots, x_n$. Por el teorema de Rolle, la derivada G' debe tener por lo menos $(n + 1)$ ceros dentro del intervalo $[a, b]$. La derivada G'' debe tener no menos de n ceros y, finalmente, continuando con este razonamiento, la derivada $G^{(n+1)}$ debe tener por lo menos un cero dentro del intervalo $[a, b]$ en consideración. Sea ξ tal cero. Diferenciemos la ecuación (46) $(n + 1)$ veces y hagamos $t = \xi$. Como P_n es un polinomio de grado a lo más n, la derivada $(n + 1)$ de P_n debe ser idénticamente cero. Además, la derivada $(n + 1)$ de CF es $(n + 1)!$, pues F es polinomio con término principal igual a t^{n+1}. Tenemos así

$$0 = G^{(n+1)}(\xi) = f^{(n+1)}(\xi) - C\,(n + 1)!$$

de donde,

$$C = \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

Recordando la definición de C y usando esta última igualdad, resulta

$$f(x) - P_n(x) = \frac{1}{(n+1)!} F(x) f^{(n+1)}(\xi)$$

donde ξ depende de x y cae dentro del intervalo $[a, b]$, resultando así lo que queríamos demostrar.

La fórmula del error dada por la ecuación (45) es un resultado teórico importante ya que es usada para obtener el error de los métodos numéricos de derivación e integración. A pesar de ello, su uso práctico está restringido a funciones cuyas derivadas tengan cotas conocidas (por ejemplo, las funciones trigonométricas o logarítmicas).

Observemos que debido a la equivalencia entre las distintas fórmulas de interpolación, la expresión que nos da el error del polinomio de interpolación es la misma para todas las fórmulas.

Ejemplo 24. Dados los valores

x	0	3	-1	1	5
$f(x)$	0	189	5	1	1625

escribiremos la fórmula de interpolación de Lagrange que se ajusta a los cinco datos.

Ya que el número de datos es cinco, el orden de la fórmula de Lagrange es $n = 4$. La interpolación de Lagrange queda

$$P_4(x)=\frac{(x-3)(x+1)(x-1)(x-5)}{(0-3)(0+1)(0-1)(0-5)}0+\frac{(x-0)(x+1)(x-1)(x-5)}{(3-0)(3+1)(3-1)(3-5)}189+\frac{(x-0)(x-3)(x-1)(x-5)}{(-1-0)(-1-3)(-1-1)(-1-5)}5+$$

$$+\frac{(x-0)(x-3)(x+1)(x-5)}{(1-0)(1-3)(1+1)(1-5)}1+\frac{(x-0)(x-3)(x-1)(x+1)}{(5-0)(5-3)(5-1)(5+1)}1625.$$

Efectuando las operaciones y simplificando, obtenemos

$$P_4(x) = 3x^4 - 2x^3.$$

Ejemplo 25. ¿Con qué grado de exactitud podemos calcular $\sqrt{115}$ mediante la fórmula de interpolación de Lagrange para la función $y = \sqrt{x}$, si elegimos los puntos de interpolación $x_0 = 100$, $x_1 = 121$, $x_2 = 144$?

Tenemos $y'=\frac{1}{2}x^{-\frac{1}{2}}$, $y''=-\frac{1}{4}x^{-\frac{3}{2}}$, $y'''=\frac{3}{8}x^{-\frac{5}{2}}$. Luego,

$$E(115) = (115 - 100)(115 - 121)(115 - 144)\frac{3}{8}\frac{\xi^{-5/2}}{3!}$$

donde ξ es un punto comprendido entre 100 y 144. De aquí,

$$\frac{1}{16}15\text{x}6\text{x}29\frac{1}{\sqrt{144^5}} \le E(115) \le \frac{1}{16}15\text{x}6\text{x}29\frac{1}{\sqrt{100^5}}$$

es decir,

$$6.5556279 \text{ x } 10^{-4} \le E(115) \le 1.63125 \text{ x } 10^{-3}.$$

4.3.5.2. Derivación de la fórmula de Lagrange usando diferencias divididas.

La fórmula de interpolación de Lagrange puede obtenerse también muy fácilmente a partir de la propiedad de simetría de las diferencias divididas y teniendo presente que la diferencia dividida de orden $(n + 1)$ de un polinomio P de grado n es cero. Por lo tanto, si tenemos $(n + 2)$ abscisas $x, x_0, x_1, \cdots, x_n$, entonces obtenemos

$$P(x\, x_0\, x_1 \ldots\, x_n) = 0 = \frac{P(x)}{(x-x_0)(x-x_1)\ldots(x-x_n)} + \frac{P(x_0)}{(x_0-x)(x_0-x_1)\ldots(x_0-x_n)} +$$

$$+ \frac{P(x_1)}{(x_1-x)(x_1-x_0)(x_1-x_2)\ldots(x_1-x_n)} + \ldots + \frac{P(x_n)}{(x_n-x)(x_n-x_0)\ldots(x_n-x_{n-1})}.$$

Despejando $P(x)$ resulta

$$P(x) = \frac{(x-x_1)(x-x_2)\ldots(x-x_n)}{(x_0-x_1)(x_0-x_2)\ldots(x_0-x_n)}P(x_0) + \frac{(x-x_0)(x-x_2)\ldots(x-x_n)}{(x_1-x_0)(x_1-x_2)\ldots(x_1-x_n)}P(x_1) + \ldots +$$

$$+ \frac{(x-x_0)(x-x_1)\ldots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\ldots(x_n-x_{n-1})}P(x_n)$$

y esta última expresión es la fórmula de interpolación de Lagrange ($P(x_k) = f_k$, para $k = 0, 1, \ldots, n$).

4.3.6. Comentarios adicionales.

La representación de los polinomios de interpolación que hemos visto no se da en la forma acostumbrada (series de potencias), $P(x) = a_0x^n + a_1x^{n-1} + \ldots + a_n$. Desde luego que las ecuaciones que dan los polinomios de interpolación pueden ponerse en tal forma (ya que son totalmente equivalentes), pero usualmente no hay ninguna razón para hacerlo.

Para los casos en donde el orden del polinomio se desconozca, el método de Newton tiene ventajas debido a que profundiza en el comportamiento de las diferentes fórmulas de orden superior. Además, la aproximación del error dada por la ecuación

$$E(x) \approx \prod_{k=0}^{n}(x-x_k)f(x_0x_1\ldots x_{n+1})$$

puede integrarse fácilmente en los cálculos de Newton, ya que la aproximación usa una diferencia dividida. De esta forma, desde este punto de vista de cálculo, a menudo se prefiere el método de Newton. Cuando se va a llevar a cabo sólo una interpolación, ambos métodos, el de Newton y el de Lagrange, requieren de un esfuerzo de cálculo similar. Sin embargo, la versión de Lagrange es un poco más fácil de programar. También existen casos en donde la forma de Newton es más susceptible a los errores de redondeo. Debido a esto y a que no requiere calcular y almacenar diferencias, la forma de Lagrange se usa a menudo cuando el orden del polinomio se conoce a priori. Debemos también tener presente que la forma de Lagrange no lleva ninguna información ni implícita ni explícita respecto del error en el polinomio de interpolación. En este sentido, los métodos que usan diferencias se desempeñan mucho mejor. Si las diferencias máximas son prácticamente constantes, entonces el resultado de la interpolación en un sentido estricto es ordinariamente exacto para los lugares decimales dados con los datos tabulados y, por lo tanto, no es necesario estimar el error. Cuando se usa la fórmula de Lagrange no hay posibilidad de seguir el curso de las diferencias; si es posible, debe estimarse el error. Si la función $f(x)$ está tabulada y no se conoce su expresión analítica, estrictamente hablando, es imposible estimar el error en el polinomio de interpolación. Totalmente cierto, ya que para un polinomio dado es teóricamente posible construir una infinidad de funciones distintas que coinciden con el polinomio en el conjunto de puntos dados; para puntos intermedios la desviación del polinomio de interpolación de la función puede ser arbitrariamente grande. No obstante, si la naturaleza de la función es tal que su gráfica es una curva suave, resulta posible entonces determinar aproximadamente el error en el polinomio de interpolación, para un alto grado de viabilidad, en base a los valores de las diferencias de orden superior obtenidos con las fórmulas antes indicadas.

Hemos visto que $E(x) = \prod_{k=0}^{n}(x - x_k)\frac{f^{(n+1)}(\xi)}{(n+1)!}$, es decir que $E(x)$ es el producto de dos factores: uno de ellos depende de las propiedades de la función y no está sujeto a regulación y el otro está determinado exclusivamente por la elección de los puntos de interpolación. Para una elección cuidadosa de los argumentos de interpolación y funciones razonables de manera que la derivada de orden $(n + 1)$ de la función esté acotada, entonces el error de la aproximación tiende a anularse a medida que n crece, es decir, cuando aumentamos el número de datos. Pero si la función es tal que su derivada de orden $(n + 1)$ no está acotada, entonces con el sólo hecho de incrementar n no vamos a obtener mejores resultados. No hay garantía de que el polinomio de interpolación converja a la función exacta al aumentar el número de datos. Así, la interpolación mediante un polinomio

de orden grande debe evitarse o utilizarse con precauciones extremas. Aunque no existe un criterio para determinar el orden óptimo del polinomio de interpolación, generalmente se recomienda utilizar uno con orden relativamente bajo en un pequeño rango de x.

4.4. Aproximación por mínimos cuadrados.

4.4.1. Introducción.

En este Capítulo se exploraron técnicas para interpolar, especialmente para aquellas situaciones en que los datos son conocidos en forma precisa y están lejos de ser lineales. El principio que se utilizó es ajustar una curva polinomial a los puntos. El problema es interesante para muchas aplicaciones. Buena parte del desarrollo proviene del trabajo de Newton y Kepler cuando analizaban los datos sobre las posiciones de estrellas y planetas.

Nuestro objetivo al determinar el comportamiento de una función según se evidencia por la muestra de pares de datos $(x, f(x))$ es múltiple: aproximar otros valores de la función en valores no tabulados de x (interpolación y extrapolación) y estimar la integral de $f(x)$ y de su derivada. Los últimos objetivos conducen hacia formas de resolución de ecuaciones diferenciales ordinarias y ecuaciones diferenciales parciales.

La estrategia que se usa para aproximar valores desconocidos de la función es directa. Se encuentra un polinomio que ajusta un conjunto de puntos selectos $(x_i, f(x_i))$ y se supone que el polinomio y la función se comportan casi igual sobre el intervalo en cuestión. Luego, los valores del polinomio deben ser estimaciones razonables de los valores de la función desconocida.

No obstante, cuando los datos no son "suaves" hay problemas con los polinomios de interpolación, lo que significa que hay irregularidades locales. En tales casos, para seguir las irregularidades sería necesario un polinomio de grado superior, pero se encontrará que tales polinomios aunque se ajustan a la irregularidad se desvían bastante en otras regiones donde la función es suave.

Por ello es que no siempre se desea encontrar un polinomio que se ajuste exactamente a los datos; es decir, hay ciertas situaciones en donde no sería razonable pedir que el polinomio aproximante coincida exactamente con los datos dados. A menudo, los valores que desean ajustarse no son exactos o pueden provenir de un conjunto de mediciones experimentales sujetas a error. Ajustar el polinomio de interpolación es ajustarse a los errores de los datos de entrada y no deseamos esto.

En tales casos se aplica una técnica de aproximación denominada *mínimos cuadrados*. Con base en la teoría estadística, con este método se encuentra un polinomio (o algún otro tipo de función de aproximación) que

con mayor probabilidad se aproxima a los valores verdaderos. Es decir que se obtiene una curva que minimiza la diferencia entre los datos y la curva.

4.4.2. Regresión lineal.

Supóngase que se desea ajustar una curva a un conjunto de datos aproximados como los obtenidos por estudiantes en el laboratorio de física para determinar los efectos de la temperatura sobre una resistencia. Los estudiantes registraron mediciones de temperatura y resistencia como se muestra en la Figura 7, donde la gráfica sugiere una relación lineal.

Quieren determinarse las constantes a y b en la ecuación que relaciona la resistencia R con la temperatura T

$$R = aT + b \tag{47}$$

de modo que en un empleo ulterior sea posible predecir la resistencia a cualquier temperatura. La recta trazada a simple vista representa bastante bien a los datos, pero si éstos vuelven a trazarse y se pide algo más para dibujar una recta, rara vez se obtiene exactamente la misma recta. Uno de los requisitos para ajustar una curva a los datos es que el proceso no sea ambiguo. También sería deseable, en cierto sentido, minimizar las desviaciones de los puntos con respecto a la recta. Las desviaciones se miden por las distancias de los puntos a la recta, y cómo se miden estas distancias depende del hecho que ambas variables estén sujetas a error. Se supondrá que el error al leer las temperaturas en la Figura 7 es despreciable de modo que todos los errores estén en las mediciones de la resistencia y se usarán distancias verticales. (Si ambas estuviesen sujetas a error, sería posible usar distancias perpendiculares. El problema se complica considerablemente más. Sólo se abordará el caso más simple).

T, C°	R, ohms
20.5	765
32.7	826
51.0	873
73.2	942
95.7	1032

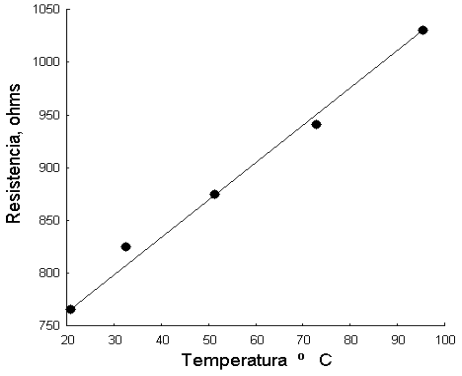


Figura 7

Para minimizar las desviaciones, el criterio acostumbrado es minimizar la suma de los cuadrados de los errores. Este es *el principio de mínimos cuadrados*.

Además de proporcionar un resultado único para un conjunto de datos, el método de mínimos cuadrados también coincide con *el principio de máxima probabilidad* de estadística. Si los errores de medición poseen una distribución denominada normal y si la desviación estándar es constante para todos los datos, entonces puede demostrarse que la recta determinada al minimizar la suma de los cuadrados tiene valores de pendiente y ordenada al origen con máxima probabilidad de ocurrencia.

Sea Y_i un valor experimental y sea y_i un valor de la ecuación

$$y_i = ax_i + b$$

donde x_i es un valor particular de la variable que se supone libre de error. Quieren determinarse los mejores valores de a y b de modo que las y predigan los valores de la función que corresponden a los valores x. Sea $e_i = Y_i - y_i$. El criterio de mínimos cuadrados requiere que

$$S = e_1^2 + e_2^2 + \ldots + e_N^2 = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - ax_i - b)^2$$

sea mínima. N es el número de pares (x, Y). Se alcanza el mínimo al elegir adecuadamente los parámetros a y b, de modo que sean las "variables" del problema. En el mínimo para S, ambas derivadas parciales $\partial S/\partial a$ y $\partial S/\partial b$ son cero. Por lo tanto, recordando que las x_i y las Y_i son puntos de datos no afectados por la elección de los valores de a y b, se tiene

$$\frac{\partial S}{\partial a} = 0 = \sum_{i=1}^{N} 2(Y_i - ax_i - b)(-x_i)$$

$$\frac{\partial S}{\partial b} = 0 = \sum_{i=1}^{N} 2(Y_i - ax_i - b)(-1).$$

Al dividir por -2 cada una de estas ecuaciones y desarrollar la sumatoria, se obtienen las denominadas *ecuaciones normales*

$$\begin{aligned} a\sum x_i^{\,2} + b\sum x_i &= \sum x_i Y_i \\ a\sum x_i + bN &= \sum Y_i. \end{aligned} \tag{48}$$

Todas las sumatorias de la ecuación (48) van desde $i = 1$ hasta $i = N$. Al resolver simultáneamente estas ecuaciones se obtienen los valores para la pendiente y la ordenada al origen a y b, respectivamente.

Para los datos de la Figura 7 se encuentra que

$$N = 5 \quad \sum T_i = 273.1 \quad \sum T_i^{\,2} = 18607.27 \quad \sum R_i = 4438 \quad \sum T_i R_i = 254932.5.$$

Luego, las ecuaciones normales son

$$18607.27a + 273.1b = 254932.5$$

$$273.1a + 5b = 4438$$

A partir de lo anterior se encuentra $a = 3.395$, $b = 702.2$ y así, la ecuación (47) se escribe como

$$R = 3.39T + 702.2$$

La función lineal determinada de esta manera se llama *recta de regresión*.

4.4.3. El ajuste potencial $y = b\,x^{m}$.

Algunas situaciones se modelan mediante una función del tipo $f(x) = bx^m$, donde m es una constante conocida. En estos casos sólo hay que determinar un parámetro.

Supongamos que tenemos N puntos $\{(x_i, Y_i)\}_{i=1}^{N}$ cuyas abscisas son distintas. Entonces, el coeficiente b de la curva potencial óptima en mínimos cuadrados $y = bx^m$ viene dado por

$$\mathrm{b} = \left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i}^{\mathrm{m}} \mathrm{Y_i} \right) / \left(\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i}^{2\mathrm{m}} \right).$$

En efecto, usando la técnica de los mínimos cuadrados lo que hacemos es buscar el mínimo de la función

$$\mathrm{S(b)} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \left(\mathrm{bx_i}^{\mathrm{m}} - \mathrm{Y_i} \right)^2$$

para lo que, en este caso, bastará con resolver $S'(b) = 0$. La derivada es

$$S'(b) = 2\sum_{i=1}^{N}\left(bx_i^{\,m} - Y_i\right)\left(x_i^{\,m}\right) = 2\sum_{i=1}^{N}\left(bx_i^{\,2m} - x_i^{\,m}Y_i\right)$$

luego, el coeficiente b es la solución de la ecuación

$$b = \left(\sum_{i=1}^{N} x_i^{\,m}Y_i\right) \Big/ \left(\sum_{i=1}^{N} x_i^{\,2m}\right).$$

esto es,

$$0 = b\sum_{i=1}^{N} x_i^{\,2m} - \sum_{i=1}^{N} x_i^{\,m}Y_i$$

Ejemplo 26. Con el objetivo de medir la aceleración de la gravedad, se han recogido en la siguiente tabla unos datos experimentales sobre el tiempo que tarda en llegar al suelo un cuerpo, según la altura desde la que se deja caer. La relación funcional es $d = \frac{1}{2}gt^2$, donde d es la distancia de caída medida en metros y t es el tiempo medido en segundos. Vamos a aproximar con estos datos el valor de la aceleración de la gravedad g.

Tiempo, t_i	0.200	0.400	0.600	0.800	1.000
Distancia, d_i	0.1960	0.7850	1.7665	3.1405	4.9075

En la siguiente tabla calculamos los valores de los coeficientes para un ajuste potencial, siendo $m = 2$ el exponente que hemos tomado.

Tiempo, t_i	Distancia, d_i	$d_i t_i^2$	t_i^4
0.200	0.1960	0.00784	0.0016
0.400	0.7850	0.12560	0.0256
0.600	1.7665	0.63594	0.1296
0.800	3.1405	2.00992	0.4096
1.000	4.9075	4.90750	1.0000
		7.68680	1.5664

El coeficiente es $b = 7.68680/1.5664 = 4.9073$ y obtenemos la curva de ajuste $d = 4.9073\ t^2$ con lo cual $g \approx 2b = 9.8146$ m/s^2.

4.4.4. Aplicaciones de la regresión lineal: linealización de relaciones no lineales.

La regresión lineal proporciona una técnica muy poderosa para ajustar datos a una "mejor" línea. Sin embargo, se ha dicho que la relación entre las variables dependiente e independiente es lineal. Este no es siempre el caso y el primer paso en cualquier análisis de regresión es el de trazar y visualizar los datos para decidir si es correcto o aceptable el aplicar un modelo lineal. Por ejemplo, en la Figura 8 se muestran algunos datos que, obviamente, son curvilíneos: en a) se muestran datos mal condicionados en la regresión lineal con mínimos cuadrados, mientras que en b) se observa que una parábola es preferible.

En algunos casos, técnicas como la regresión polinomial (descrita más adelante) serán apropiadas. En otros, se pueden hacer transformaciones que expresen los datos de manera que sean compatibles con la regresión lineal.

Formas conocidas que se intentan son la ecuación elevada a una potencia y la ecuación exponencial (a y b constantes)

$$y = b\,x^a$$

o bien

$$y = b\,e^{ax}$$

El primer modelo tiene gran aplicación en casos de la ingeniería. Como se muestra en la Figura 9 a), la ecuación de potencia (para $a \neq 0$ o 1) es no lineal.

El segundo modelo se usa en muchos ejemplos de la ingeniería caracterizando cantidades que crecen (a positiva) o que decrecen (a negativa) en un promedio proporcional a su magnitud. Un ejemplo es el crecimiento poblacional y la disminución radiactiva. Como se muestra en la Figura 9 b), la ecuación exponencial representa una relación lineal (para $a = 0$) entre y y x.

Un tercer ejemplo de un modelo no lineal es la ecuación de promedio de saturación

$$y = b\frac{x}{a + x}$$

Este modelo, que es particularmente útil en la caracterización de crecimientos poblacionales bajo condiciones limitantes, también representa una relación no lineal entre y y x (Figura 9 c)) que nivela o "satura" conforme x crece.

Las partes d), e) y f) son versiones linealizadas de aquellas, las cuales son transformaciones simples.

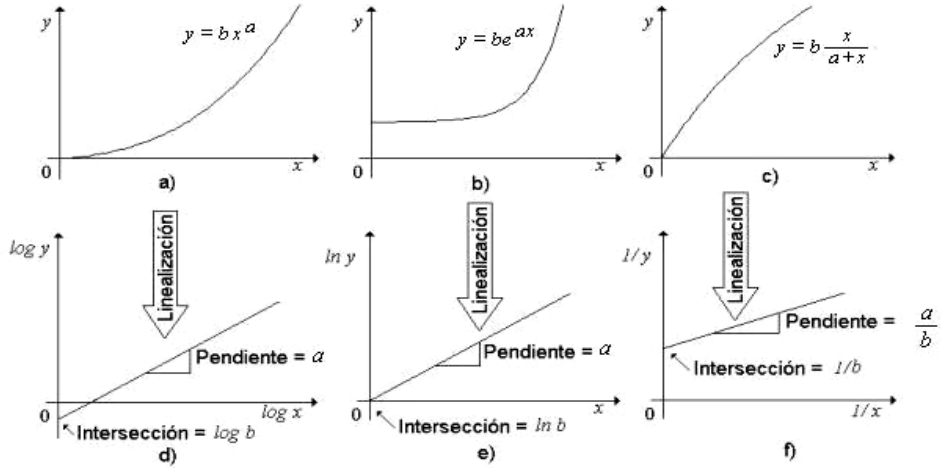


Figura 9

Es posible desarrollar ecuaciones normales para estos tres modelos de manera análoga al desarrollo precedente para una recta por mínimos cuadrados, si las derivadas parciales se igualan a cero. Tales ecuaciones simultáneas no lineales son mucho más difíciles de resolver (se trataron en el Capítulo 3) que las ecuaciones lineales. Por lo tanto, las formas de ecuaciones de potencias y de exponenciales suelen hacerse lineales tomando logaritmos antes de determinar los parámetros.

Por ejemplo, la ecuación $y = b\ x^a$ se puede linealizar tomando logaritmos de base 10 y obtener

$$log\ y = log\ b + a\ log\ x$$

De esta forma, una gráfica logarítmica de *log y* contra *log x* genera una línea recta con una pendiente de *a* y una intersección de *log b* (Figura 9 d)).

La ecuación $y = b\ e^{ax}$ se puede linealizar por medio de logaritmos naturales y obtener

$$ln\ y = ln\ b + ax$$

Por lo tanto, una gráfica semilogarítmica de *ln y* contra *x* genera una línea recta con una pendiente de *a* y una intersección de *ln b* (Figura 9 e)).

Ahora se ajusta la nueva variable $z = log\ y$ o $z = ln\ y$ como una función lineal de *log x* o de *x*, como ya se describió. En este caso no se minimiza la suma de los cuadrados de las desviaciones de *Y* con respecto a la curva, sino más bien las desviaciones de *log Y* o *ln Y*. Después se pueden transformar a su estado original y usarse para propósitos predictivos.

La ecuación $y = b\dfrac{x}{a+x}$ se linealiza invirtiéndola y se obtiene

$$\frac{1}{y} = \frac{a}{b}\frac{1}{x} + \frac{1}{b}$$

Por lo tanto, una gráfica de $1/y$ contra $1/x$ será lineal, con pendiente a/b y una intersección de $1/b$ (Figura 9 f)).

Este modelo en su estado transformado se ajusta usando regresión lineal para evaluar los coeficientes constantes. Luego se puede llevar a su estado original a efectos de usarse para propósitos predictivos.

Encontremos las ecuaciones normales para el caso

$$ln\ y = ln\ b + ax$$

es decir,

$$z = ax + B \qquad \text{(I)}$$

Los datos originales $(x_i,\ Y_i)$ se han transformado con el cambio de variables en $(x_i,\ ln(Y_i))$; a este proceso lo llamamos *método de linealización de los datos*. El problema ahora es calcular la recta de regresión (I) para los nuevos puntos, para lo que planteamos las correspondientes ecuaciones normales

$$\sum x_i^2 a + \sum x_i B = \sum x_i z_i$$

$$\sum x_i\, a + NB = \sum z_i$$

Una vez calculados a y B hallamos el parámetro b de la relación $B = ln\ b$: $b = e^B$.

Mediante un procedimiento similar se pueden encontrar las ecuaciones normales para los otros dos casos.

Ejemplo 27. Consideremos la colección de datos contenidos en las primeras tres columnas de la Tabla 1.

Si x_i se grafica con $ln\ Y_i$, los datos parecen tener una relación lineal, así que es razonable suponer una aproximación de la forma

$$y = b\ e^{ax} \qquad \text{ó} \qquad ln\ y = ln\ b + ax$$

Extendiendo la Tabla 1 y sumando las columnas se obtienen los valores restantes de la Tabla 1.

Tabla 1

i	x_i	Y_i	$ln\ Y_i$	x_i^2	$x_i\ ln\ Y_i$
1	1.00	5.10	1.629	1.0000	1.629
2	1.25	5.79	1.756	1.5625	2.195
3	1.50	6.53	1.876	2.2500	2.814
4	1.75	7.45	2.008	3.0625	3.514
5	2.00	8.46	2.135	4.0000	4.270
	7.50		9.404	11.875	14.422

Las ecuaciones normales son

$$\begin{aligned} 11.875a + 7.50\ ln\ b &= 14.422 \\ 7.50a\ +\ 5\ ln\ b\ \ &=\ \ 9.404 \end{aligned}$$

y resolviendo este sistema, resulta

$$a = 0.5056, \qquad ln\ b = 1.122$$

Como $b = e^{1.122} = 3.071$, la aproximación toma la forma

$$y = 3.071\ e^{0.5056x}$$

la cual, en los puntos correspondientes a los datos, da los valores de la Tabla 2.

Tabla 2

i	x_i	Y_i	$3.071\ e^{0.5056x_i}$
1	1.00	5.10	5.09
2	1.25	5.79	5.78
3	1.50	6.53	6.56
4	1.75	7.45	7.44
5	2.00	8.46	8.44

4.4.5. Polinomios por mínimos cuadrados.

Debido a que los polinomios pueden manipularse fácilmente es común ajustar tales funciones a datos cuya gráfica no es lineal. Para esta situación las ecuaciones normales son lineales, lo cual es una ventaja adicional. En el desarrollo, n se usa como el grado del polinomio y N como el número de pares de datos. Resulta evidente que si $N = n + 1$, entonces el polinomio pasa exactamente por cada punto y son válidos los métodos analizados

previamente en este Capítulo, de modo que en lo sucesivo siempre se tendrá $N > n + 1$.

Se supone la relación funcional

$$y = a_0 + a_1 x + \ldots + a_n x^n \tag{49}$$

con errores definidos como

$$e_i = Y_i - y_i = Y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_n x_i^n$$

Nuevamente se usa Y_i para representar el valor observado o experimental correspondiente a x_i, con x_i libre de error. Se minimiza la suma de cuadrados

$$S = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_n x_i^n)^2.$$

En el mínimo, todas las derivadas parciales $\partial S / \partial a_0$, $\partial S / \partial a_1$, ..., $\partial S / \partial a_n$ desaparecen. Al escribir las ecuaciones para lo anterior se obtienen $n + 1$ ecuaciones

$$\begin{aligned}
a_0 N &+ a_1 \sum x_i + a_2 \sum x_i^2 + \ldots + a_n \sum x_i^n &&= \sum Y_i \\
a_0 \sum x_i &+ a_1 \sum x_i^2 + a_2 \sum x_i^3 + \ldots + a_n \sum x_i^{n+1} &&= \sum x_i Y_i \\
a_0 \sum x_i^2 &+ a_1 \sum x_i^3 + a_2 \sum x_i^4 + \ldots + a_n \sum x_i^{n+2} &&= \sum x_i^2 Y_i \\
&\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \\
a_0 \sum x_i^n &+ a_1 \sum x_i^{n+1} + a_2 \sum x_i^{n+2} + \ldots + a_n \sum x_i^{2n} &&= \sum x_i^n Y_i .
\end{aligned} \tag{50}$$

Al dividir por -2 y reordenando se obtienen $n + 1$ ecuaciones normales a resolver simultáneamente

$$\frac{\partial S}{\partial a_0} = 0 = \sum_{i=1}^{N} 2(Y_i - a_0 - a_1 x_i - \ldots - a_n x_i^n)(-1)$$

$$\frac{\partial S}{\partial a_1} = 0 = \sum_{i=1}^{N} 2(Y_i - a_0 - a_1 x_i - \ldots - a_n x_i^n)(-x_i)$$

$$\vdots$$

$$\frac{\partial S}{\partial a_n} = 0 = \sum_{i=1}^{N} 2(Y_i - a_0 - a_1 x_i - \ldots - a_n x_i^n)(-x_i^n).$$

Se puede demostrar que las ecuaciones normales tienen una solución única siempre que las x_i, para $i = 1, \ldots, N$, sean distintas.

Al expresar tales ecuaciones en forma matricial se observa un patrón interesante en la matriz de coeficientes

$$\begin{pmatrix} N & \sum x_i & \sum x_i^2 & \sum x_i^3 & \ldots & \sum x_i^n \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \ldots & \sum x_i^{n+1} \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \sum x_i^5 & \ldots & \sum x_i^{n+2} \\ \vdots & & & & & \\ \sum x_i^n & \sum x_i^{n+1} & \sum x_i^{n+2} & \sum x_i^{n+3} & \ldots & \sum x_i^{2n} \end{pmatrix} a = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \\ \sum x_i^2 Y_i \\ \vdots \\ \sum x_i^n Y_i \end{pmatrix}$$

Todas las sumatorias de las ecuaciones (50) y (51) van desde 1 hasta N.

No es fácil resolver grandes conjuntos de ecuaciones lineales. En el Capítulo 3 se abordaron los métodos para realizar lo anterior. Estas ecuaciones particulares presentan una dificultad adicional en el sentido de que tienen la propiedad indeseable denominada *mal condicionamiento*. Su resultado es que los errores por redondeo al resolverlas provocan errores insólitamente grandes en las soluciones, que por supuesto son los valores deseados de los coeficientes a_i en la ecuación (49). Hasta $n = 4$ o 5, este problema no es tan grande (es decir, en las soluciones por computadora sólo es aconsejable, pero no esencial, contar con aritmética de precisión doble),

pero más allá de este punto se requieren métodos especiales. Tales métodos especiales usan polinomios ortogonales en una forma equivalente de la ecuación (49). Este tema no se estudiará con mayor detalle. Desde el punto de vista del experimentador, rara vez se requieren funciones más complicadas que los polinomios de cuarto grado y cuando son necesarias, a menudo, el problema puede manipularse ajustando una serie de polinomios a subconjuntos de datos.

La matriz de la ecuación (51) se denomina *matriz normal* para el problema de mínimos cuadrados. También hay otra matriz que corresponde a éste, denominada *matriz de diseño*; es de la forma

$$A = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_N \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_N^2 \\ \vdots & & & & \vdots \\ x_1^n & x_2^n & x_3^n & \cdots & x_N^n \end{pmatrix}$$

Es fácil demostrar que AA^t es justamente la matriz de coeficientes de la ecuación (51). Además, AA^t es una matriz simétrica y definida positiva. También es fácil ver que Ay, donde y es el vector columna de los valores Y, proporciona el lado derecho de la ecuación (51). Esto significa que la ecuación (51) puede escribirse en forma matricial como

$$AA^t a = Ba = Ay$$

Para resolver el sistema puede aplicarse eliminación Gaussiana o cualquiera de los métodos vistos en el Capítulo 3 (aunque sólo para polinomios de grado bajo).

Ejemplo 28. Ajustaremos una cuadrática a los datos de la Tabla 3.

Tabla 3

x_i	0.05	0.11	0.15	0.31	0.46	0.52	0.70	0.74	0.82	0.98	1.17
Y_i	0.956	0.890	0.832	0.717	0.571	0.539	0.378	0.370	0.306	0.242	0.104

En la Figura 10 se muestra una gráfica de los datos (en realidad, los datos son una perturbación de la relación $y = 1 - x + 0.2\,x^2$; es interesante observar cuán bien se aproxima esta función).

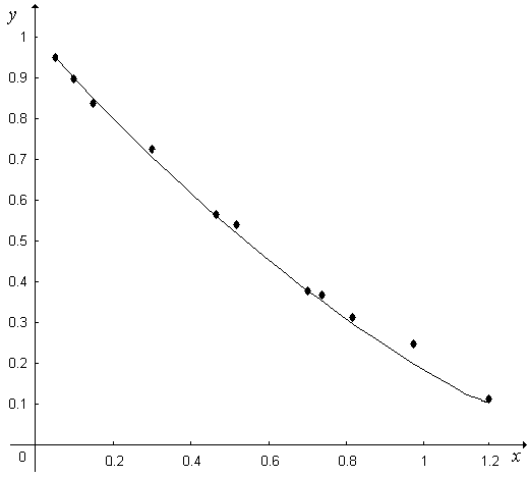


Figura 10

Para establecer las ecuaciones normales se requieren las siguientes sumas

$$\sum x_i = 6.01 \qquad N = 11$$
$$\sum x_i^2 = 4.6545 \qquad \sum Y_i = 5.9050$$
$$\sum x_i^3 = 4.1150 \qquad \sum x_i Y_i = 2.1839$$
$$\sum x_i^4 = 3.9161 \qquad \sum x_i^2 Y_i = 1.3357$$

Es necesario resolver el sistema de ecuaciones

$$11\, a_0 + 6.01\, a_1 + 4.6545\, a_2 = 5.9050$$

$$6.01\, a_0 + 4.6545\, a_1 + 4.1150\, a_2 = 2.1839$$

$$4.6545\, a_0 + 4.1150\, a_1 + 3.9161\, a_2 = 1.3357$$

El resultado es $a_0 = 0.998$, $a_1 = -1.018$, $a_2 = 0.225$ de modo que con el método de mínimos cuadrados se obtiene

$$y = 0.998 - 1.018x + 0.225x^2$$

Compare esto con $y = 1 - x + 0.2x^2$. No es de esperar reproducir los coeficientes exactamente debido a los errores en los datos.

El error

$$\sum_{i=1}^{N}\left|Y_i - a_0 - a_1x_i - a_2x_i^2 - \ldots - a_nx_i^n\right|^2 = \sum_{i=1}^{11}\left|Y_i - 0.998 + 1.018x_i - 0.225x_i^2\right|^2 = 0.0018$$

es el mínimo que puede ser obtenido usando un polinomio cuadrático.

¿Qué grado de polinomio debe usarse?

En el caso general puede preguntarse qué grado de polinomio utilizar. En tanto se usen polinomios de grado superior, por supuesto se reducen las desviaciones de los puntos con respecto a la curva hasta que cuando el grado del polinomio, *n*, es igual a *N* - 1, hay una correspondencia exacta (suponiendo que no hay datos duplicados en el mismo valor *x*) y se tiene un polinomio de interpolación. La respuesta a este problema se encuentra en la estadística. El grado del polinomio de aproximación se incrementa en tanto haya un decremento estadísticamente significativo en la varianza, σ^2, que se calcula como

$$\sigma^2 = \frac{\sum e_i^2}{N - n - 1}. \qquad (52)$$

Para el ejemplo anterior, cuando se hace que el grado del polinomio construido para ajustar los puntos varíe desde 1 hasta 7, se obtienen los resultados que se muestran en la Tabla 4.

Tabla 4

Grado	Ecuación	σ^2 (Ec(52))	$\sum e^2$
1	$y = 0.952 - 0.760x$	0.0010	0.0092
2	$y = 0.998 - 1.018x + 0.225x^2$	0.0002	0.0018
3	$y = 1.004 - 1.079x + 0.351x^2 - 0.069x^3$	0.0003	0.0018
4	$y = 0.998 - 0.838x - 0.522x^2 + 1.040x^3 - 0.454x^4$	0.0003	0.0016
5	$y = 1.031 - 1.704x + 4.278x^2 - 9.477x^3 + 9.394x^4 - 3.290x^5$	0.0001	0.0007
6	$y = 1.038 - 1.910x + 5.952x^2 - 15.078x^3 + 18.277x^4 - 9.835x^5 + 1.836x^6$	0.0002	0.0007
7	$y = 1.032 - 1.742x + 4.694x^2 - 11.898x^3 + 16.645x^4 - 14.346x^5 + 8.141x^6 - 2.293x^7$	0.0002	0.0007

El criterio de la ecuación (52) elige el grado óptimo como igual a 2. Esto no es sorprendente en vista de cómo se obtuvieron los datos. Es importante darse cuenta que el numerador de la ecuación (52), la suma de las desviaciones al cuadrado de los puntos con respecto a la curva, debe disminuir de manera continua a medida que aumenta el grado del polinomio. Lo que hace aumentar a σ^2 es el denominador de la ecuación (52) a medida que se pasa por arriba del grado óptimo. En este ejemplo, tal

comportamiento se observa para $n = 3$. Por arriba de $n = 3$ aparece un segundo efecto. Debido al mal condicionamiento, los coeficientes de los polinomios por mínimos cuadrados se determinan con precisión más deficiente. Esto modifica los incrementos esperados de los valores de σ^2.

A continuación se muestra un algoritmo para obtener un polinomio por mínimos cuadrados:

Dados N pares de datos, (x_i, Y_i), $i = 1, ..., N$, obtener un polinomio por mínimos cuadrados de n-ésimo grado por medio de lo siguiente:

Formar la matriz de coeficientes, M, con $n + 1$ renglones (r) y con $n + 1$ columnas (c) al hacer

$$M_{rc} = \sum_{i=1}^{N} x_i^{r+c-2}.$$

Formar el vector b del lado derecho con $n + 1$ renglones (r) al hacer

$$b_r = \sum_{i=1}^{N} x_i^{r-1} Y_i.$$

Resolver el sistema lineal $Ma = b$ para obtener los coeficientes en

$$y = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n$$

que es el polinomio deseado que ajusta los datos.

4.5. Elementos de juicio.

En el siguiente cuadro se proporciona un resumen de los elementos de juicio relacionados con la interpolación y aproximación polinomial y la aproximación por mínimos cuadrados.

Comparación de las características de los diferentes métodos en la interpolación y aproximación polinomial y en la aproximación por mínimos cuadrados

Método	Error asociado con los datos	Coincidencia con cada uno de los puntos	Número de puntos con los que coincide exactamente	Esfuerzo de programación	Comentarios
Polinomio de interpolación de Newton con diferencias divididas	Pequeño	Exacto	$n + 1$	Fácil	En general, se prefiere para un análisis exploratorio en que el orden propio del polinomio se desconoce. La evaluación de errores es fácil. Se debe preparar una tabla de diferencias divididas.
Polinomio de interpolación de Newton hacia adelante en puntos equidistantes	Pequeño	Exacto	$n + 1$	Fácil	En general, se prefiere para un análisis exploratorio en que el orden propio del polinomio se desconoce. La evaluación de errores es fácil. Se debe preparar una tabla de diferencias finitas. Apropiado para aproximar un valor x que esté cerca del inicio de la tabla.
Polinomio de interpolación de Newton hacia atrás en puntos equidistantes	Pequeño	Exacto	$n + 1$	Fácil	En general, se prefiere para un análisis exploratorio en que el orden propio del polinomio se desconoce. La evaluación de errores es fácil. Se debe preparar una tabla de diferencias finitas. Apropiado para aproximar un valor x que esté cerca del final de la tabla.

Polinomio de interpolación de Gauss hacia adelante con diferencias centrales	Pequeño	Exacto	$n + 1$	Fácil	En general, se prefiere para un análisis exploratorio en que el orden propio del polinomio se desconoce. La evaluación de errores es fácil. Se debe preparar una tabla de diferencias centrales. Apropiado para aproximar un valor x que se encuentre próximo a la mitad de una tabla ($x_0 < x < x_0 + h/2$).
Polinomio de interpolación de Gauss hacia atrás con diferencias centrales	Pequeño	Exacto	$n + 1$	Fácil	En general, se prefiere para un análisis exploratorio en que el orden propio del polinomio se desconoce. La evaluación de errores es fácil. Se debe preparar una tabla de diferencias centrales. Apropiado para aproximar un valor x que se encuentre próximo a la mitad de una tabla ($x_0 - h/2 < x < x_0$).
Polinomio de interpolación de Stirling	Pequeño	Exacto	$n + 1$	Fácil	En general, se prefiere para un análisis exploratorio en que el orden propio del polinomio se desconoce. La evaluación de errores es fácil. Se debe preparar una tabla de diferencias centrales. Apropiado para aproximar un valor x que se encuentre en el centro de una tabla ($x_0 - h/2 < x < x_0 + h/2$).

Polinomio de interpolación de Bassel	Pequeño	Exacto	$n + 1$	Fácil	En general, se prefiere para un análisis exploratorio en que el orden propio del polinomio se desconoce. La evaluación de errores es fácil. Se debe preparar una tabla de diferencias centrales. Apropiado para aproximar un valor x que se encuentre próximo a la mitad de una tabla, pero por debajo de ella ($x_0 + h/2 < x < x_0 + h$).
Polinomio de interpolación de Lagrange	Pequeño	Exacto	$n + 1$	Fácil	En general, se prefiere cuando se conoce el orden del polinomio. Difícil de manejar para los cálculos manuales. No requiere de los cálculos de diferencias.
Regresión lineal	Grande	Aproximado	0	Fácil	Apropiado para mediciones imprecisas. Se utiliza en aquellos casos en donde una variable dependiente y otra independiente se relacionan de manera lineal.
Regresión polinomial	Grande	Aproximado	0	Moderado	Es una extensión de la regresión lineal. Los errores de redondeo se hacen notorios para versiones de orden superior. Se recomienda el uso de doble precisión.

4.6. Algoritmo de la fórmula de diferencia dividida del polinomio de interpolación de Newton. Pseudocódigo.

Para obtener los coeficientes de diferencia dividida del polinomio de interpolación P en los $(n + 1)$ puntos distintos $x_0, x_1, ..., x_n$ para la función f:

ENTRADA los números $x_0, x_1, ..., x_n$; los valores $f(x_0), f(x_1), ..., f(x_n)$ como la primera columna $Q_{00}, Q_{10}, ..., Q_{n0}$ de Q.

SALIDA los números $Q_{00}, Q_{11}, ..., Q_{nn}$ donde $P(x)=\sum_{i=0}^{n} Q_{ii} \prod_{j=0}^{i-1} (x - x_j)$.

Paso 1 Para $i = 1, 2, ..., n$

Para $j = 1, 2, ..., i$

tomar $Q_{ij} = \dfrac{Q_{i,j-1} - Q_{i-1,j-1}}{x_i - x_{i-j}}$.

Paso 2 SALIDA ($Q_{00}, Q_{11}, ..., Q_{nn}$); ($Q_{ii}$ es $f(x_0\, x_1 \ldots x_i)$).

PARAR.

Observación. En el paso 1 se calculan todas las diferencias divididas. Por lo tanto, se puede modificar el formato de salida para producir todas las diferencias divididas.

EJERCICIOS PROPUESTOS .....

1. Calcular la tabla de diferencias divididas para los siguientes datos:

a)

x	3.2	2.7	1
$f(x)$	22.0	17.8	14.2

Redondear los cálculos a 3 decimales.

x	0	2.5069	5.0154	7.5270
$f(x)$	0.3989423	0.3988169	0.3984408	0.3978138

b)

Redondear los cálculos a 7 decimales.

2. Calcular las primeras diferencias divididas para $f(x) = x^2$ y $f(x) = x^3$.

3. Usando la fórmula de interpolación de Newton con diferencias divididas obtener:

a) i) El polinomio de interpolación para los datos de la tabla dada en el ejercicio 1 a).
ii) El nuevo polinomio de interpolación si se añade el punto (4.8, 38.3) a los datos de la tabla dada en el ejercicio 1 a).

b) El polinomio de interpolación para los datos de la tabla dada en el ejercicio 1 b) y estimar $f(-0.1)$, $f(0.1)$ y $f(3.7608)$.

4. Calcular una tabla de diferencias finitas para la función $f(x) = 2x^3 - 2x^2 + 3x - 1$, tomando como valor inicial $x_0 = 0$, utilizando como paso constante $h = 1$.

¿Qué puede decirse con respecto a la diferencia de orden 4 y a las de orden superior a 4?

5. Dada la siguiente tabla de valores:

x	-3	-1	1	3	5	7
$f(x)$	14	4	2	8	22	44

se pide:

a) ¿Existe una parábola que verifique todos los datos de la tabla?. En caso afirmativo, encontrarla.

b) Obtener, en cualquier caso, el polinomio de grado menor o igual que cinco que pasa por los seis puntos de la tabla y utilizarlo para estimar $f(-2)$ y $f(0)$.

6. Dados los puntos P_0 (0, 0), P_1 (1, 1), P_2 (2, 8), P_3 (3, 27) y P_4 (4, 64):

a) Hallar mediante la fórmula de Newton hacia adelante el polinomio de interpolación que une los puntos P_0, P_1, P_2, P_3 y P_4.
b) Hallar mediante la fórmula de Newton hacia atrás el polinomio de interpolación que une los puntos P_0, P_1, P_2, P_3 y P_4.

c) Comparar los resultados de a) y de b).

7. a) Utilizando la fórmula de interpolación de Newton hacia adelante y conociendo $f(x) = 1/x$ para x variando desde 0.55 hasta 0.90 con un espaciado de $h = 0.05$, interpolar en $x = 0.57$. Estimar la exactitud con que puede alcanzarse dicho valor. (Calcular hasta la diferencia de orden 5).

b) Utilizando la fórmula de interpolación de Newton hacia atrás con los datos dados en el apartado a), interpolar en $x = 0.87$ y estimar la exactitud con que puede alcanzarse dicho valor.

8. Utilizando la fórmula de interpolación de Newton más conveniente en cada caso, hallar:

a) $log_{10}1044$, conociendo la tabla de logaritmos de 7 cifras decimales para $f(x) = log_{10} x$ siguiente:

x	1000	1010	1020	1030	1040	1050
$f(x)$	3.0000000	3.0043214	3.0086002	3.0128372	3.0170333	3.0211893

b) $\sqrt{1.01}$ y $\sqrt{1.28}$ con $n = 3$, conociendo la tabla donde se dan los valores de $f(x) = \sqrt{x}$ redondeados a 5 decimales:

x	1.00	1.05	1.10	1.15	1.20	1.25	1.30
$f(x)$	1.00000	1.02470	1.04881	1.07238	1.09544	1.11803	1.14017

9. Hacer los programas correspondientes a las fórmulas de interpolación de Newton hacia adelante y hacia atrás. Utilizando estos programas se pide:

a) Resolver el ejercicio 7 (pero considerando a x desde 0.50). Interpolar además en $x = 0.5650,\ 0.6750,\ 0.9300$.

b) Usando la siguiente tabla que contiene los valores de la integral de probabilidad:

$$\Gamma(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-x^2} dx$$

x	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
$\Gamma(x)$	0.8427	0.8802	0.9103	0.9340	0.9523	0.9661	0.9763	0.9838	0.9891	0.9928	0.9953

calcular $\Gamma(1.43)$.

10. Hallar el polinomio de cuarto grado que tome los valores f de las tablas que se dan a continuación, según los métodos:

a) Gauss hacia delante

x	$f(x)$	δf	$\delta^2 f$	$\delta^3 f$	$\delta^4 f$
1	1				
		-2			
2	-1		4		
		2		-8	
3	**1**		**-4**		**16**
		-2		**8**	
4	-1		4		
		2			
5	1				

b) Gauss hacia atrás

x	$f(x)$	δf	$\delta^2 f$	$\delta^3 f$	$\delta^4 f$
1	1				
		-2			
2	-1		4		
		2		**-8**	
3	**1**		**-4**		**16**
		-2		8	
4	-1		4		
		2			
5	1				

c) Stirling

x	$f(x)$	δf	$\delta^2 f$	$\delta^3 f$	$\delta^4 f$
1	1				
		-2			
2	-1		4		
		2		**-8**	
3	**1**		**-4**		**16**
		-2		**8**	
4	-1		4		
		2			
5	1				

¿Qué conclusión puede obtener a partir de los polinomios construidos con las distintas fórmulas de interpolación?

11. Aplicar la fórmula de Bessel para obtener un polinomio de grado 5 que tome los valores *f* de la siguiente tabla:

x	$f(x)$	δf	$\delta^2 f$	$\delta^3 f$	$\delta^4 f$	$\delta^5 f$
1	0					
		-1				
2	-1		10			
		9		108		
3	**8**		**118**		**216**	
		127		**324**		**120**
4	**135**		**442**		**336**	
		569		660		
5	704		1102			
		1671				
6	2375					

12. Usando las fórmulas de Gauss (hacia adelante y hacia atrás), de Bessel y de Stirling, resolver los siguientes apartados:

a) Estimar $f(1.5708)$, conociendo la siguiente tabla:

x	$f(x)$	δf	$\delta^2 f$	$\delta^3 f$	$\delta^4 f$
1.3	3.6693				
		0.3859			
1.4	4.0552		0.0406		
		0.4265		0.0042	
1.5	4.4817		0.0448		0.0004
		0.4713		0.0048	
1.6	4.9530		0.0496		0.0004
		0.5209		0.0052	
1.7	5.4739		0.0548		
		0.5757			
1.8	6.0496				

b) Estimar $\Gamma(0.5437)$, conociendo la siguiente tabla que nos da los valores de la integral de probabilidad

$$\Gamma(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-x^2}dx$$

y las diferencias

x	$\Gamma(x)$	δf	$\delta^2 f$	$\delta^3 f$
0.51	0.5292437			
		0.0086550		
0.52	0.5378987		-0.0000896	
		0.0085654		-0.0000007
0.53	0.5464641		-0.0000903	
		0.0084751		-0.0000007
0.54	0.5549392		-0.0000910	
		0.0083841		-0.0000007
0.55	0.5633233		-0.0000917	
		0.0082924		-0.0000006
0.56	0.5716157		-0.0000923	
		0.0082001		
0.57	0.5798158			

13. Utilizando la fórmula de interpolación de Lagrange se pide :

a) Encontrar el polinomio de interpolación que pase por los puntos $P_0(0, -2)$, $P_1(1, 6)$, $P_2(3, 40)$.

b) Dada la siguiente tabla de valores:

x	0	-1	3	7
$f(x)$	0	2	4	7

encontrar el polinomio de interpolación y estimar $f(4)$.

c) Dada la siguiente información: $f(-2) = f(2) = 0$ y $f(-1) = f(1) = 1$, estimar $f(0)$.

d) Dada la siguiente tabla de valores:

x	0	3	4.5	7.5	10.5
$f(x)$	1	0.9986295	0.9969173	0.9914449	0.9832549

estimar $f(1.5)$ y $f(6)$.

e) Dada la siguiente información:

sen 0.3 = 0.295520
sen 0.4 = 0.389418
sen 0.5 = 0.479426
sen 0.6 = 0.564642

estimar *sen* 0.42.

14. Hacer el programa correspondiente a la fórmula de interpolación de Lagrange. Utilizando este programa se pide:

a) Resolver los apartados c), d) y e) del ejercicio 13.

b) Conociendo la función $f(x) = cos\ x$ y las abscisas 5.0, 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7 interpolar para $x = 5.127, 5.347$.

15. Considerar los datos presentados en la tabla que se da a continuación. Para encontrar la recta de mínimos cuadrados que aproxima a estos datos, extender la tabla y sumar las columnas. Graficar la recta junto con los puntos que representan a los datos.

x_i	Y_i
1	1.3
2	3.5
3	4.2
4	5.0
5	7.0
6	8.8
7	10.1
8	12.5
9	13.0
10	15.6
11	16.1

16. Calcular y graficar la recta de regresión para los datos siguientes:

(-1, 10), (0, 9), (1, 7), (2, 5), (3, 4), (4, 3), (5, 0), (6, -1).

17. En el cuadro siguiente se dan las longitudes de una varilla de determinado material sometida a diferentes temperaturas

T°C	15	25	35	45	55	65	75
L mm	962.4	962.6	962.7	962.9	963.1	963.3	963.5

Se sabe por Física que $L = a + bT$.
Calcular la línea de regresión.

18. Calcular y graficar la línea de regresión para los datos de la siguiente tabla. Evaluar el error (o la desviación) de la línea ajustada.

i	1	2	3	4	5	6
x_i	0.1	0.4	0.5	0.7	0.7	0.9
Y_i	0.61	0.92	0.99	1.52	1.47	2.03

19. Ajustar un polinomio cuadrático a los datos de la tabla del ejercicio anterior. Evaluar el error (o la desviación) del polinomio ajustado.

20. Ajustar los datos que se muestran en la siguiente tabla con el polinomio de mínimos cuadrados discreto de grado dos. Evaluar el error mínimo que se obtiene usando el polinomio cuadrático. Realizar la gráfica correspondiente.

i	1	2	3	4	5
x_i	0	0.25	0.5	0.75	1.00
Y_i	1.0000	1.2840	1.6487	2.1170	2.7183

21. Determinar y graficar la parábola óptima en mínimos cuadrados para los cuatro puntos (-3, 3), (0, 1), (2, 1) y (4, 3).

22. Calcular la velocidad inicial y la aceleración de un movimiento uniformemente acelerado, conocidos los espacios recorridos en los tiempos siguientes:

t	0	2	3	4	6	9
e	0	11.86	19.34	27.84	47.94	85.82

23. Hallar un ajuste exponencial $y = be^{ax}$ a los cinco datos (0, 1.5), (1, 2.5), (2, 3.5), (3, 5.0) y (4, 7.5). Realizar la gráfica del ajuste exponencial.

24. Hacer el programa para obtener un polinomio por mínimos cuadrados. Utilizando este programa resolver los ejercicios 17, 20 y 21.

25. El valor promedio de una corriente eléctrica oscilante durante un período puede ser cero. Por ejemplo, supóngase que la corriente se describe mediante una senoidal simple

$$i(t) = sen\ (2\pi\, t/T)$$

donde T es el período. El valor promedio de esta función se puede determinar mediante la siguiente ecuación

$$i = \frac{\int_0^T \operatorname{sen}\left(\frac{2\pi t}{T}\right) dt}{T - 0} = \frac{-\cos 2\pi + \cos 0}{T} = 0$$

A pesar de que la corriente es igual a cero, esta corriente es capaz de realizar trabajo y generar calor. Por lo tanto, los ingenieros eléctricos, a menudo, caracterizan esta corriente mediante

$$I_{RMS} = \sqrt{\frac{\int_0^T i^2(t) dt}{T}} \qquad (1)$$

en donde I_{RMS} se conoce como corriente RMS (raíz de la corriente media al cuadrado). Para evitar este resultado nulo, la corriente se eleva al cuadrado antes de calcular el promedio.

En este caso, supóngase que la corriente en un circuito es de

$$i(t) = 10\, e^{-t/T}.\, \operatorname{sen}\left(\frac{2\pi t}{T}\right) \quad \text{para } 0 \le t \le T/2 \qquad (2)$$

$$i(t) = 0 \quad \text{para } T/2 \le t \le T$$

a) Determinar la corriente RMS ajustando un polinomio de segundo grado que coincida con $i^2(t)$ exactamente en $t = 0$, $T/4$, $T/2$. Luego, integrar este polinomio analíticamente y calcular la corriente RMS en el intervalo de 0 a T usando la ecuación (1). Supóngase que $T = 1$s.

b) Graficar el polinomio de segundo orden y la corriente verdadera.

c) A partir de los resultados obtenidos, expresar sus conclusiones.

Observación. Este resultado se podrá comparar cuando se vea integración numérica. Se usarán varias técnicas de integración numérica (utilizando el polinomio de interpolación) para llevar a cabo esos mismos cálculos.

26. En 1601 el astrónomo alemán Johannes Kepler formuló su tercera ley del movimiento planetario, $T = Cx^{3/2}$, donde x es la distancia al sol medida en millones de kilómetros, T es el período orbital medido en días y C es una constante. Las parejas de datos (x, T) observados para los primeros

cuatro planetas, Mercurio, Venus, La Tierra y Marte, son (58, 88), (108, 225), (150, 365) y (228, 687).

a) Obtener el coeficiente *C* por el método de mínimos cuadrados y graficar la curva junto con los puntos que representan a los datos.

b) A partir de los resultados obtenidos, expresar sus conclusiones.

27. Los modelos de crecimiento poblacional son importantes en muchos campos de la ingeniería. La suposición de que la tasa de crecimiento de la población (*dp/dt*) es proporcional a la población actual (*p*) en el tiempo (*t*) es de fundamental importancia en muchos de los modelos, en forma de ecuación

$$\frac{dp}{dt} = kp \qquad (1)$$

donde *k* es un factor de proporcionalidad conocido como la tasa de crecimiento específico y tiene unidades de tiempo $^{-1}$. Si *k* es una constante, entonces se puede obtener la solución de la ecuación (1) de la teoría de ecuaciones diferenciales

$$p(t) = p_0 e^{kt} \qquad (2)$$

donde p_0 es la población en el tiempo $t = 0$. Se observa que $p(t)$ en la ecuación (2) tiende a infinito a medida que *t* crece. Este comportamiento es claramente imposible en los sistemas reales. Por lo tanto, se debe modificar el modelo y hacerlo más realista.

Primero, se debe reconocer que la tasa de crecimiento específico *k* no puede ser constante a medida que la población crece. Esto es porque, cuando *p* tiende a infinito, el organismo que se modela se ve limitado por factores tales como el almacenamiento de comida y producción de desperdicios tóxicos. Una manera de expresar esto matemáticamente es la de usar el modelo de tasa de crecimiento y saturación tal como

$$k = k_{máx} \frac{f}{K + f} \qquad (3)$$

donde $k_{máx}$ es la máxima tasa de crecimiento posible para valores de comida (*f*) abundante y *K* es la constante de semi-saturación. Vemos que cuando $K = f$, $k = k_{máx}$ /2. Por lo tanto, *K* es la cantidad de comida disponible que sostiene una tasa de crecimiento poblacional igual a la mitad

de la tasa máxima. Las constantes K y $k_{máx}$ son valores empíricos basados en medidas experimentales de k para varios valores de f. Como ejemplo, supóngase que la población p representa una levadura empleada en la producción comercial de cerveza y f es la concentración de la fuente de carbono a fermentarse. Las medidas de k contra f de la levadura se muestran en el siguiente cuadro. Estos son los datos usados en la evaluación de las constantes en un modelo de promedio de crecimiento de saturación que caracteriza a la cinética microbial. Se necesita calcular $k_{máx}$ y K de estos datos empíricos.

f, mg/l	k, días^{-1}
7	0.29
9	0.37
15	0.48
25	0.65
40	0.80
75	0.97
100	0.99
150	1.07

a) Usar el procedimiento de mínimos cuadrados lineal para determinar $k_{máx}$ y K, y realizar el gráfico correspondiente.

b) A partir de los resultados obtenidos, expresar sus conclusiones.

Observación. La ecuación diferencial resultante se puede resolver usando otros métodos numéricos cuando se conoce $f(t)$.

28. Se usa la función $f(x) = 1.44 / x^2 + 0.24\ x$ para generar seis parejas de datos (0.25, 23.1), (1.0, 1.68), (1.5, 1.0), (2.0, 0.84), (2.4, 0.826) y (5.0, 1.2576).

Obtener los ajustes mediante polinomios óptimos en mínimos cuadrados para 2, 3, 4 y 5 grados.

a) Graficar, para cada caso, el polinomio óptimo y la función $f(x)$.

b) A partir de los resultados obtenidos, expresar sus conclusiones.

Observación. No deja de ser tentadora la posibilidad de utilizar un polinomio óptimo en el sentido de los mínimos cuadrados para ajustar datos que no son lineales. Pero si los datos no muestran una naturaleza polinomial, puede ocurrir que la curva resultante presente oscilaciones grandes. Este fenómeno llamado *oscilación polinomial*, se hace más pronunciado conforme aumenta el grado del polinomio y por esta razón, no se suelen usar

polinomios de grado 6 o mayor, a no ser que se sepa que la función de la cual provienen los datos es un polinomio.

29. Supóngase que un ingeniero trabaja para una compañía que fabrica computadoras de cierto tipo. Las consideraciones sobre planificación y localización de recursos (horas - hombre, metales, plásticos y componentes electrónicos) requieren que este ingeniero sea capaz de predecir hasta cuándo permanecerán en el mercado las computadoras en función del tiempo. En este caso de estudio, se proporcionan datos que describen el número de computadoras de la compañía que se encuentran en el mercado en diferentes tiempos hasta 60 días. Al ingeniero se le pide que examine estos datos, y usando métodos de extrapolación, calcule cuántas computadoras se tendrán disponibles a los 90 días. Los datos se muestran en el siguiente cuadro:

Tiempo (días)	Número de computadoras en el mercado
0	50000
10	35000
20	31000
30	20000
40	19000
50	12000
60	11000

a) Usar polinomios de interpolación de Newton del primero hasta el sexto grado. Utilizar las curvas resultantes para predicciones en los días 55, 65 y 90.
Graficar el polinomio de sexto orden.

b) Usar polinomios de regresión del primero hasta el sexto grado.
Graficar las curvas de regresión de segundo y tercer orden.

c) A partir de los resultados obtenidos en los apartados a) y b), expresar sus conclusiones.

Observación. Analizando los datos dados en el cuadro anterior, se observa que los mismos no son uniformes. Aunque el número de computadoras decrezca en el tiempo, la taza de decrecimiento varía de intervalo a intervalo comportándose aleatoriamente. Ya que la tendencia sugerida por los datos no es uniforme, los polinomios de orden superior oscilarán para intersecar cada punto. Estas oscilaciones llevarán a interpolaciones y extrapolaciones falsas.

Puesto que la regresión no se restringe para pasar por cada uno de los puntos, algunas veces puede resultar útil para remediar esta situación. Sin embargo, la regresión también llevará a resultados poco consistentes.

La razón principal para que la interpolación y la regresión estén mal condicionadas para este problema, es que ni siquiera se basan en un modelo de la realidad física.

En ambos casos, por ejemplo, no se toma en consideración que más allá de $t = 60$ el número de computadoras debe estar entre 0 y 11000. Para una situación como la analizada aquí se tiene, por lo tanto, que derivar un modelo matemático que tenga una base teórica; esto es, que sea capaz de simular y predecir el número de computadoras en el mercado en función del tiempo. Se puede desarrollar una ecuación diferencial para este propósito, la cual llevará a resultados más satisfactorios.

30. El mástil de un barco tiene un área transversal de 0.876 pulg^2 y se construye de una aleación de aluminio experimental. Se llevan a cabo pruebas para definir la relación entre esfuerzo (fuerza por área) aplicado al material y deformación (deflexión por unidad de longitud). Los resultados de estas pruebas se resumen en el siguiente cuadro:

Número de puntos	Esfuerzo lb/pulg^2	Deformación pies/pie
1	7200	0.0020
2	7500	0.0045
3	8000	0.0060
4	5200	0.0013
5	10000	0.0085
6	1800	0.0005

Es necesario calcular el cambio en la longitud del mástil debido a la deformación causada por la fuerza del viento. La compresión causada por el aire se puede calcular usando la relación:

$$\text{Esfuerzo} = \frac{\text{fuerza en el mástil}}{\text{área de la sección transversal del mástil}}$$

En este caso, se tiene una fuerza del viento de 6440.6 libras y el esfuerzo se calcula mediante:

$$\text{Esfuerzo} = \frac{6440.6 \text{ lb}}{0.876 \text{ pulg}^2} = 7350 \text{ lb/pulg}^2$$

Este esfuerzo puede ser usado para calcular la deformación, la cual a su vez se puede sustituir en la ley de Hooke y calcular el cambio de la longitud del mástil:

$$\Delta l = (\text{deformación})(\text{longitud})$$

en donde la longitud se refiere a la altura del mástil. Por lo tanto, el problema se reduce a la determinación de valores de la deformación de los datos del cuadro dado. Ya que no se dispone de ningún punto para un valor de esfuerzo dado de 7350, el problema necesitará de algún ajuste de curvas.

a) Usar interpolación polinomial de Newton de orden 0 al 5 para calcular la deformación a un esfuerzo de 7350 lb/pulg^2.
Graficar el polinomio de quinto orden que ajusta los datos del cuadro.

b) Usar regresión lineal para ajustar una línea recta a través de los datos. Luego ajustar una línea recta al logaritmo de la deformación contra el logaritmo del esfuerzo (linealización de una ecuación de potencias: $y = bx^a$). Graficar ambas situaciones.

c) A partir de los resultados obtenidos en los apartados a) y b), expresar sus conclusiones.

Observación. Aunque la curva pase muy bien a través de los puntos en la vecindad del esfuerzo de 7350 lb/pulg^2, se observará que ésta oscila en otras partes del rango de datos.

Esto es, debido a que los tres datos se encuentran muy cercanos del valor de 7350 lb/pulg^2 la interpolación no debe variar significativamente en este punto, como es de esperarse. Sin embargo, si se requieren aproximaciones de otras fuerzas las oscilaciones de los polinomios pueden llevar a resultados inexactos. Estos resultados ilustran que la interpolación con polinomios de grado superior está mal condicionada para datos inciertos o con "ruido" del tipo a los de este problema. La regresión proporciona una alternativa que, en general, es más apropiada para estas situaciones. Se observará que la regresión lineal lleva a resultados fuera de la realidad (con deformaciones negativas en un esfuerzo igual a cero). Para evitar este resultado no realista se puede usar la transformación logarítmica. Esta versión mostrará los resultados físicos más realistas (la deformación es cero cuando el esfuerzo es cero) y la curva será un poco más realista (capturará algunas curvaturas sugeridas por los datos).

Se observará que la interpolación polinomial y los dos tipos de regresión, llevan a resultados diferentes de deformación. Debido al realismo físico y al comportamiento más satisfactorio a través del rango completo de los datos, serán más aceptables los resultados proporcionados por el segundo tipo de regresión.

Capítulo 5

5

Derivación e integración numérica.

5

5.1. Derivación numérica.

Hemos utilizado las fórmulas de interpolación para encontrar valores aproximados de la función $f(x)$ en puntos donde ésta no se conoce. Otro uso de las fórmulas de interpolación es aplicarlas a las operaciones fundamentales del análisis: *derivación* e *integración*. Por consiguiente, en lugar de efectuar las operaciones sobre la función, que en general no se conoce salvo en un conjunto discreto de puntos, éstas se realizan sobre el polinomio de interpolación más adecuado a las necesidades del problema.

Antes de dar algunas técnicas conviene aclarar que la operación de derivación es básicamente inestable, debe ser utilizada con mucha precaución y, en general, no esperar mucha precisión ya que tiende a magnificar pequeñas discrepancias o errores entre el polinomio y la función. Todo lo contrario ocurre con la integración. Hablando en términos generales, debe tenerse presente que la derivación aproximada es una operación menos exacta que la interpolación. En efecto, la proximidad de las coordenadas de dos curvas $y = f(x)$, $y = P(x)$ en el intervalo $[a, b]$, no garantiza la proximidad de sus derivadas $f\,'(x)$ y $P\,'(x)$ en dicho intervalo; esto es, no garantiza una pequeña divergencia de las pendientes de las tangentes a las curvas dado el mismo valor de los argumentos (ver Figura 1).

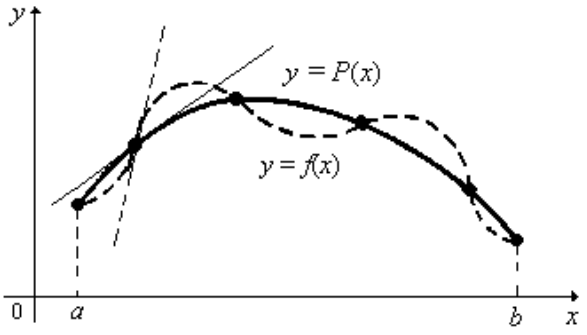


Figura 1

5.1.1. Derivadas de una función dada para valores no equidistantes de la variable.

Si la función está dada para valores no equidistantes de la variable, nos conviene utilizar la fórmula de interpolación de Newton con diferencias divididas o bien la fórmula de interpolación de Lagrange. En *la fórmula de interpolación de Newton con diferencias divididas*

$$f(x) = f(x_0) + (x - x_0) f(x_0 x_1) + (x - x_0)(x - x_1) f(x_0 x_1 x_2) + \\ + (x - x_0)(x - x_1)(x - x_2) f(x_0 x_1 x_2 x_3) + \ldots + (x - x_0)(x - x_1)\ldots(x - x_{n-1}) f(x_0 x_1 \ldots x_n) + \\ + \prod_{i=0}^{n} (x - x_i) f(x\, x_0 x_1 \ldots x_n)$$

si llamamos $\beta_i = (x - x_i) \quad (i = 0,1,\ 2,\ldots,n)$ y $\prod(x) = \prod_{i=0}^{n} (x - x_i)$, se tiene

$$f(x) = f(x_0) + \beta_0 f(x_0 x_1) + \beta_0 \beta_1 f(x_0 x_1 x_2) + \beta_0 \beta_1 \beta_2 f(x_0 x_1 x_2 x_3) + \ldots + \\ + \beta_0 \beta_1 \ldots \beta_{n-1} f(x_0 x_1 \ldots x_n) + \prod(x) f(x\, x_0 x_1 \ldots x_n) \qquad (1)$$

donde,

$$E(x) = \prod(x)\, f(x\, x_0\ x_1\ \ldots\ x_n)$$

Derivando en ambos miembros de (1), obtenemos

$$f'(x) = f(x_0 x_1) + (\beta_0 + \beta_1) f(x_0 x_1 x_2) + (\beta_0 \beta_1 + \beta_0 \beta_2 + \beta_1 \beta_2) f(x_0 x_1 x_2 x_3) + \\ + (\beta_0 \beta_1 \beta_2 + \beta_0 \beta_1 \beta_3 + \beta_0 \beta_2 \beta_3 + \beta_1 \beta_2 \beta_3) f(x_0 x_1 x_2 x_3 x_4) + \ldots + \frac{dE(x)}{dx} \qquad (2)$$

donde,

$$\frac{dE(x)}{dx} = \prod'(x) f(x\, x_0 x_1 x_2 \ldots x_n) + \prod(x) \frac{df(x\, x_0 x_1 \ldots x_n)}{dx}$$

Pero,

$$\frac{df(x\, x_0 x_1 \ldots x_n)}{dx} = f(x\ x\, x_0 x_1 \ldots x_n) = \frac{f^{(n+2)}(\xi_2)}{(n+2)!}$$

de modo que

$$\frac{dE(x)}{dx} = \prod'(x) \frac{f^{(n+1)}(\xi_1)}{(n+1)!} + \prod(x) \frac{f^{(n+2)}(\xi_2)}{(n+2)!} \qquad (3)$$

La ecuación (2) nos permite calcular la derivada de una función para valores no equidistantes de la variable y para una estimación del error cometido usamos la ecuación (3). Si la derivada se calcula en cualquiera de los puntos tabulados, el error queda reducido a

$$\prod'(x)\frac{f^{(n+1)}(\xi_1)}{(n+1)!} \qquad (3')$$

pues en dichos puntos $\prod(x)=0$.

Ejemplo 1. Supongamos dada la siguiente tabla de diferencias

x	$f(x)$	Primera	Segunda	Tercera	Cuarta
0.3	1.1618342				
		0.5956860			
0.4	1.2214028		0.1553133		
		0.6422800		0.0267750	
0.6	1.3498588		0.1660233		0.0034807
		0.6920870		0.0288634	
0.7	1.4190675		0.1804550		0.0037560
		0.7462235		0.0314926	
0.9	1.5683122		0.1962013		0.0041080
		0.8247040		0.0339574	
1.1	1.7332530		0.2131800		
		0.8886580			
1.2	1.8221188				

En este ejemplo, calculamos $f\,'(x)$ en puntos tabulados haciendo uso hasta de la cuarta diferencia dividida y hacemos una estimación del error usando la ecuación (3').

Si $x = 0.4$ y tomamos $x_0 = 0.4, x_1 = 0.6, x_2 = 0.7, x_3 = 0.9, x_4 = 1.1,$ entonces

$$\beta_0 = 0,\ \beta_1 = -0.2,\ \beta_2 = -0.3,\ \beta_3 = -0.5,\ \beta_4 = -0.7$$

Aplicando la ecuación (2), se tiene

$$f'(0.4) \approx 0.6422800 + (-0.2)0.1660233 + (-0.2)(-0.3)0.0288634 + \\ + (-0.2)(-0.3)(-0.5)0.0037560 = 0.6106945.$$

La función tabulada en el ejemplo es $f(x) = e^{\frac{x}{2}}$ de modo que para acotar el error hacemos

$$f^{(5)}(x) = \frac{1}{32} e^{\frac{x}{2}}, \quad \prod'(x) = \beta_1\beta_2\beta_3\beta_4 = (-0.2)(-0.3)(-0.5)(-0.7).$$

Luego,

$$E(\xi) = \frac{1}{5!}\,\frac{1}{32}\,0.021\,e^{\frac{\xi}{2}}, \quad 0.4 \le \xi \le 1.1$$

y efectuando los cálculos

$$0.0000067 \le E \le 0.0000095$$

El valor exacto con 7 decimales de la derivada primera de $f(x) = e^{\frac{x}{2}}$ en $x = 0.4$ es

$$f'(0.4) = \frac{1}{2} e^{0.2} = 0.6107014.$$

Para $x = 0.4$ y tomando ahora $x_0 = 0.3,\ x_1 = 0.4,\ x_2 = 0.6,\ x_3 = 0.7,\ x_4 = 0.9$, obtenemos

$$\beta_0 = 0.1,\ \beta_1 = 0,\ \beta_2 = -0.2,\ \beta_3 = -0.3,\ \beta_4 = -0.5$$

de donde,

$$f'(0.4) \approx 0.5956860 + (0.1)0.1553133 + (0.1)(-0.2)0.0267750 + (0.1)(-0.2)(-0.3)0.0034807 = \\ = 0.6107027$$

y

$$-0.0000012 \le E \le -0.00000091$$

Derivando sucesivamente la ecuación (2), obtenemos

$$f^{(2)}(x) = 2f(x_0\,x_1\,x_2) + 2(\beta_0 + \beta_1 + \beta_2)\,f(x_0\,x_1\,x_2\,x_3) + 2(\beta_0\,\beta_1 + \beta_0\,\beta_2 + \beta_0\,\beta_3 + \beta_1\,\beta_2 + \\ + \beta_1\,\beta_3 + \beta_2\,\beta_3)\,f(x_0\ \ x_1\ \ x_2\ \ x_3\ \ x_4) + \ldots + \frac{d^2E(x)}{dx^2}$$

$$f^{(3)}(x) = 6f(x_0x_1x_2x_3) + 6(\beta_0 + \beta_1 + \beta_2 + \beta_3)f(x_0x_1x_2x_3x_4) + \ldots + \frac{d^3E(x)}{dx^3}$$ y así sucesivamente.

Para encontrar la expresión de la derivada *m*-ésima del error, $\frac{d^m E(x)}{dx^m}$, hacemos uso de la fórmula de Leibnitz de la derivada *m*-ésima de un producto

$$\frac{d^m (UV)}{dx^m} = U\, D^m V + m\ DU\ D^{m-1} V + \frac{m(m-1)}{2} D^2 U\ D^{m-2} V + ... + D^m U\ V$$

y se llega a que

$$\frac{d^m E}{dx^m} = \sum_{k=0}^{m} \frac{m!}{k!(m-k)!(n+k+1)!} f^{(n+k+1)}(\xi_k) \frac{d^{m-k}}{dx^{m-k}} \left[\prod_{i=0}^{n} (x - x_i) \right]$$

suponiendo $m \le n$ y que $f(x)$ posee derivada hasta el orden $(n + m + 1)$.

5.1.2. Derivadas de una función dada para valores equidistantes de la variable

Si la función está dada para valores equidistantes de la variable, derivando *la fórmula de Newton hacia adelante*

$$f(x) = f(x_0) + t\Delta f_0 + \frac{t(t-1)}{2!}\Delta^2 f_0 + ... + \frac{t(t-1)...(t-n+1)}{n!}\Delta^n f_0 + \\ + \frac{h^{n+1} t(t-1)...(t-n)}{(n+1)!} f^{(n+1)}(\xi)$$

y teniendo en cuenta que: $x = x_0 + th$, $f(x) = f(x_0 + th)$, de donde, $\frac{df}{dt} = \frac{df}{dx}\frac{dx}{dt} = f'(x_0 + th)\ h$, se tiene

$$hf'(x_0 + th) = \Delta f_0 + \frac{1}{2}(2t-1)\Delta^2 f_0 + \frac{1}{6}(3t^2 - 6t + 2)\Delta^3 f_0 + \frac{1}{24}(4t^3 - 18t^2 + 22t - 6)\Delta^4 f_0 + \\ + \frac{1}{120}(5t^4 - 40t^3 + 105t^2 - 100t + 24)\Delta^5 f_0 + ... \qquad (4)$$

Si hacemos $t = 0$ y sólo consideramos hasta la diferencia cuarta, se tiene

$$h f'(x_0) = \Delta f_0 - \frac{1}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0 - \frac{1}{4}\Delta^4 f_0 + \frac{h^5}{5} f^{(5)}(\xi)$$

o bien, si se reemplazan las diferencias finitas por los correspondientes valores de la función, se obtiene

$$f'(x_0) = \frac{1}{12h}(-25f_0 + 48f_1 - 36f_2 + 16f_3 - 3f_4) + \frac{h^4}{5}f^{(5)}(\xi). \quad (5)$$

Si hacemos $t = 1$ en la ecuación (4), nos da

$$h\,f'(x_1) = \Delta f_0 + \frac{1}{2}\Delta^2 f_0 - \frac{1}{6}\Delta^3 f_0 + \frac{1}{12}\Delta^4 f_0 - \frac{h^5}{20}f^{(5)}(\xi)$$

o bien, reemplazando las diferencias finitas por los valores de la función que corresponden, se obtiene

$$f'(x_1) = \frac{1}{12h}(-3f_0 - 10f_1 + 18f_2 - 6f_3 + f_4) - \frac{h^4}{20}f^{(5)}(\xi) \quad (6)$$

Si hacemos $t = 2$ en la ecuación (4), nos da

$$h\,f'(x_2) = \Delta f_0 + \frac{3}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0 - \frac{1}{12}\Delta^4 f_0 + \frac{h^5}{30}f^{(5)}(\xi)$$

o bien,

$$f'(x_2) = \frac{1}{12h}(f_0 - 8f_1 + 8f_3 - f_4) + \frac{h^4}{30}f^{(5)}(\xi) \quad (7)$$

Si hacemos $t = 3$ en la ecuación (4), nos da

$$h\,f'(x_3) = \Delta f_0 + \frac{5}{2}\Delta^2 f_0 + \frac{11}{6}\Delta^3 f_0 + \frac{1}{4}\Delta^4 f_0 - \frac{h^5}{20}f^{(5)}(\xi)$$

o bien,

$$f'(x_3) = \frac{1}{12h}(-f_0 + 6f_1 - 18f_2 + 10f_3 + 3f_4) - \frac{h^4}{20}f^{(5)}(\xi) \quad (8)$$

Si hacemos $t = 4$ en la ecuación (4), nos da

$$h\ f'(x_4) = \Delta f_0 + \frac{7}{2}\Delta^2 f_0 + \frac{26}{6}\Delta^3 f_0 + \frac{25}{12}\Delta^4 f_0 + \frac{h^5}{5}f^{(5)}(\xi)$$

o bien,

$$f'(x_4) = \frac{1}{12h}(3f_0 - 16f_1 + 36f_2 - 48f_3 + 25f_4) + \frac{h^4}{5}f^{(5)}(\xi) \qquad (9)$$

Para encontrar las derivadas sucesivas, derivamos la ecuación (4), obteniendo

$$h^2 f''(x_0 + th) = \Delta^2 f_0 + \frac{1}{6}(6t-6)\Delta^3 f_0 + \frac{1}{24}(12t^2 - 36t + 22)\Delta^4 f_0 +$$
$$+ \frac{1}{120}(20t^3 - 120t^2 + 210t - 100)\Delta^5 f_0 + \ldots$$

Si hacemos $t = 0$ y sólo consideramos hasta la diferencia cuarta, se tiene

$$h^2 f''(x_0) = \Delta^2 f_0 - \Delta^3 f_0 + \frac{11}{12}\Delta^4 f_0 - \frac{5}{6}h^5 f^{(5)}(\xi)$$

y así podemos obtener la derivada segunda de la función en los puntos de interpolación, la que se puede expresar en función de las diferencias hacia adelante o con los correspondientes valores de la función (tal como se hizo antes para la derivada primera de f).

Puede obtenerse también

$$h^3 f'''(x_0 + th) = \Delta^3 f_0 + \frac{1}{24}(24t - 36)\Delta^4 f_0 + \frac{1}{120}(60t^2 - 240t + 210)\Delta^5 f_0 + \ldots$$

o tomando $t = 0$

$$h^3 f'''(x_0) = \Delta^3 f_0 - \frac{3}{2}\Delta^4 f_0 + \frac{7}{4}\Delta^5 f_0 + \ldots$$

y así siguiendo.

Ejemplo 2. Supongamos dada la siguiente tabla

x	$e^{\frac{x}{2}}$
0.30	1.1618342
0.45	1.2523227
0.60	1.3498588
0.75	1.4549914
0.90	1.5683122
1.05	1.6904588
1.20	1.8221188

Si aplicamos la ecuación (5)

$$f'(0.30) \approx 0.5809131$$

y la estimación del error es

$$0.0000037 \le E(0.30) \le 0.0000050$$

Aplicando la ecuación (6)

$f'(0.45) \approx 0.6261623$ y $-0.0000012 \le E(0.45) \le -0.00000092.$

Aplicando la ecuación (7)

$f'(0.60) \approx 0.6749287$ y $0.0000006 \le E(0.60) \le 0.0000008.$

Aplicando la ecuación (8)

$f'(0.75) \approx 0.7274967$ y $-0.0000012 \le E(0.75) \le -0.00000092.$

Aplicando la ecuación (9)

$f'(0.90) \approx 0.7841522$ y $0.0000037 \le E(0.90) \le 0.0000050.$

(Hemos usado en los cálculos anteriores que $h = 0.15$ y $0.3 \le \xi \le 0.9$).

Ahora, derivando la fórmula de Newton hacia atrás

$$f(x_n + th) = f(x_n) + t\nabla f_n + \frac{t(t+1)}{2!}\nabla^2 f_n + \frac{t(t+1)(t+2)}{3!}\nabla^3 f_n + \\ + \ldots + \frac{t(t+1)\ldots(t+n-1)}{n!}\nabla^n f_n + h^{n+1}\frac{t(t+1)\ldots(t+n)}{(n+1)!}f^{(n+1)}(\xi)$$

se tiene

$$h\,f'(x_n + th) = \nabla f_n + \frac{1}{2}(2t+1)\nabla^2 f_n + \frac{1}{6}(3t^2+6t+2)\nabla^3 f_n + \frac{1}{24}\left(4t^3+18t^2+22t+6\right)\nabla^4 f_n + \\ + \frac{1}{120}(5t^4+40t^3+105t^2+100t+24)\nabla^5 f_n + \ldots \qquad (10)$$

Si hacemos $t = 0$ y sólo consideramos hasta la diferencia cuarta, se tiene

$$h\,f'(x_n) = \nabla f_n + \frac{1}{2}\nabla^2 f_n + \frac{1}{3}\nabla^3 f_n + \frac{1}{4}\nabla^4 f_n + \frac{h^5}{5}f^{(5)}(\xi)$$

Reemplazando las diferencias por los correspondientes valores de la función, se tiene

$$f'(x_n) = \frac{1}{12h}(25f_n - 48f_{n-1} + 36f_{n-2} - 16f_{n-3} + 3f_{n-4}) + \frac{h^4}{5}f^{(5)}(\xi) \quad (11)$$

Si t = -1, la ecuación (10) nos da

$$h\,f'(x_{n-1}) = \nabla f_n - \frac{1}{2}\nabla^2 f_n - \frac{1}{6}\nabla^3 f_n - \frac{1}{12}\nabla^4 f_n - \frac{h^5}{20}f^{(5)}(\xi)$$

o bien,

$$f'(x_{n-1}) = \frac{1}{12h}(3f_n + 10f_{n-1} - 18f_{n-2} + 6f_{n-3} - f_{n-4}) - \frac{h^4}{20}f^{(5)}(\xi) \quad (12)$$

Análogamente, si $t = -2, -3, -4$ a partir de la ecuación (10) y reemplazando las diferencias por los correspondientes valores de la función, se obtiene, respectivamente

$$f'(x_{n-2}) = \frac{1}{12h}(-f_n + 8f_{n-1} - 8f_{n-3} + f_{n-4}) + \frac{h^4}{30}f^{(5)}(\xi) \quad (13)$$

$$f'(x_{n-3}) = \frac{1}{12h}(f_n - 6f_{n-1} + 18f_{n-2} - 10f_{n-3} - 3f_{n-4}) - \frac{h^4}{20}f^{(5)}(\xi) \quad (14)$$

$$f'(x_{n-4}) = \frac{1}{12h}(-3f_n + 16f_{n-1} - 36f_{n-2} + 48f_{n-3} - 25f_{n-4}) + \frac{h^4}{5}f^{(5)}(\xi) \quad (15)$$

Ejemplo 3. Como aplicación de estas fórmulas, utilicemos los datos de la tabla del ejemplo 2. Obtenemos, entonces

$$f'(1.20) \approx 0.9110543$$
$$f'(1.05) \approx 0.8452307$$
$$f'(0.90) \approx 0.7841553$$
$$f'(0.75) \approx 0.7274969$$
$$f'(0.60) \approx 0.6749246$$

Para encontrar las derivadas sucesivas, derivamos la ecuación (10), obteniendo

$$h^2 f''(x_n + th) = \nabla^2 f_n + \frac{1}{6}(6t+6)\nabla^3 f_n + \frac{1}{24}\left(12t^2 + 36t + 22\right)\ \nabla^4 f_n +$$
$$+ \frac{1}{120}(20t^3 + 120t^2 + 210t + 100)\nabla^5 f_n + \ldots$$

y también

$$h^3 f'''(x_n + th) = \nabla^3 f_n + \frac{1}{24}(24t+36)\nabla^4 f_n + \frac{1}{120}(60t^2 + 240t + 210)\nabla^5 f_n + \ldots$$

y así, por ejemplo, haciendo $t = 0$

$$h^2 f''(x_n) = \nabla^2 f_n + \nabla^3 f_n + \frac{11}{12}\nabla^4 f_n + \frac{5}{6}\nabla^5 f_n + \ldots$$

$$h^3 f'''(x_n) = \nabla^3 f_n + \frac{3}{2}\nabla^4 f_n + \frac{7}{4}\nabla^5 f_n + \ldots$$

5.2. Integración numérica.

Los métodos de integración numérica se pueden utilizar para integrar funciones dadas, ya sea mediante una tabla o en forma analítica. Incluso en el caso en que sea posible la integración analítica, la integración numérica puede ahorrar tiempo y esfuerzo si sólo se desea conocer el valor numérico de la integral.

Analizaremos los métodos numéricos que se utilizan para evaluar integrales de una variable

$$I = \int_a^b f(x)\ dx$$

donde $f(x)$ puede estar dada en forma analítica o mediante una tabla. Cuando la función a integrarse es una función simple y continua tal como un polinomio, una función exponencial o una función trigonométrica, entonces la integral simplemente es una función que se puede evaluar fácilmente usando métodos analíticos aprendidos en Análisis. De acuerdo al teorema fundamental del cálculo integral, la integral

$$I = \int_a^b f(x)\ dx$$

se evalúa como

$$I = \int_a^b f(x)\, dx = F(x)\Big|_a^b$$

donde $F(x)$ es la integral de $f(x)$, esto es, cualquier función tal que $F'(x) = f(x)$.

La nomenclatura del lado derecho queda

$$F(x)\Big|_a^b = F(b) - F(a) \quad \text{(Regla de Barrow).}$$

Esta integración depende del conocimiento de ciertas "reglas" que son meros ejemplos de antidiferenciación; esto es, conocer una función $F(x)$ primitiva de $f(x)$, es decir, tal que $F'(x) = f(x)$. Por consiguiente, la integración analítica depende del conocimiento previo de la respuesta (hay numerosas expresiones del tipo $\int_a^b f(x)\, dx$ que no tienen primitiva). Muchas de estas reglas se resumen en manuales y en tablas de integrales. Sin embargo, muchas funciones de importancia práctica son demasiado complicadas para incluirlas en tales tablas. Una razón por lo que las técnicas que veremos son tan valiosas, es porque proporcionan un medio de evaluar estas integrales sin conocimiento de las reglas.

Es así entonces, que si la función es complicada y continua de tal forma que es difícil o imposible de integrar directamente o bien si la función a integrar está tabulada en donde los valores de x y $f(x)$ se dan en un conjunto de puntos discretos, como en el caso, a menudo, de los datos experimentales, se deben emplear en tales casos métodos aproximados.

Los métodos de integración numérica se obtienen al integrar los polinomios de interpolación. Por consiguiente, las distintas fórmulas de interpolación darán por resultado distintos métodos de integración numérica. Los métodos que se estudian a continuación (a excepción del de Gauss) se refieren a *las fórmulas de Newton - Cotes*, que se basan en las fórmulas de interpolación con puntos de separación uniforme, y se deducen al integrar las fórmulas de interpolación de Newton con diferencias finitas. A su vez, las fórmulas de Newton - Cotes se subdividen en las de *tipo cerrado* y las de *tipo abierto*. Las reglas del trapecio y de Simpson (entre otras) pertenecen al tipo cerrado de las fórmulas de Newton – Cotes y, por ejemplo, la regla del punto medio pertenece al tipo abierto de las fórmulas de Newton - Cotes.

Las fórmulas de Newton - Cotes de tipo cerrado son aquellas en las cuales los puntos finales de los intervalos son abscisas, es decir, los puntos extremos de los intervalos son usados en dichas fórmulas; de otra forma, las fórmulas de Newton - Cotes serán de tipo abierto. Las fórmulas de Newton - Cotes son también conocidas con el nombre de *fórmulas de los n puntos*.

La integración de Gauss se basa en la interpolación polinomial usando las raíces de un polinomio ortogonal, como los polinomios de Legendre. Este método resulta numéricamente estable y se usa con éxito en bastantes laboratorios de cálculo. Se obtiene más precisión que con las fórmulas de Newton - Cotes, aunque presenta la desventaja de que los puntos no están separados uniformemente. Los puntos de interpolación así como ciertos coeficientes llamados *pesos*, son números irregulares que necesitan almacenarse. Esto en la práctica (aunque no en la teoría) limita la aplicabilidad de esta interesante fórmula.

Recordemos que para las funciones que se encuentran sobre el eje x, la integral expresada por la ecuación

$$I = \int_a^b f(x)\, dx$$

representa gráficamente el área bajo la curva $f(x)$ entre $x = a$ y $x = b$ (Figura 2).

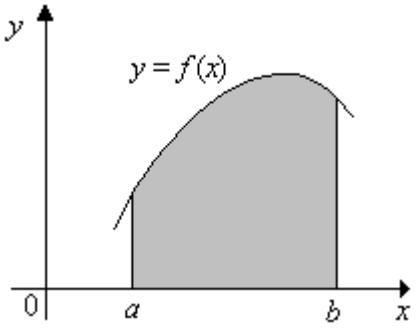


Figura 2

Daremos a continuación distintos métodos que aproximan esta integral.

5.2.1. Regla del trapecio.

La *regla del trapecio* o *regla trapezoidal* es la primera de las fórmulas cerradas de Newton - Cotes. Es un método de integración numérica que se obtiene al integrar la fórmula de interpolación lineal. (Se aproxima la curva $f(x)$ con un polinomio de orden uno en el intervalo $[a, b]$). El resultado de la integración es

$$I = \frac{b-a}{2}\,[f(a) + f(b)] + E \qquad (16)$$

donde el primer término del lado derecho es la regla del trapecio (fórmula de interpolación) y E representa su error.

En efecto, integrando el polinomio de interpolación de Newton hacia adelante de primer orden con término de error, la integral sería

$$I = \int_a^b \left[f(a) + t\Delta f_a + t(t-1)h^2 \frac{f''(\xi)}{2} \right] dx \qquad (17)$$

(ξ es un punto cualquiera dentro del intervalo [a, b]).

Para simplificar el análisis, tomando en consideración que $x = a + th$, entonces $dx = h\,dt$. Debido a que $h = b - a$, los límites de integración a y b corresponden a 0 y 1, respectivamente. Por lo tanto, la ecuación (17) se puede expresar como

$$I = h\int_0^1 \left[f(a) + t\Delta f_a + t(t-1)h^2 \frac{f''(\xi)}{2} \right] dt$$

Se supone que para h pequeña el término $f''(\xi)$ es aproximadamente constante. La ecuación se puede integrar

$$I = h\left[tf(a) + \frac{t^2}{2}\Delta f_a + \left(\frac{t^3}{6} - \frac{t^2}{4} \right) h^2 f''(\xi) \right]_0^1$$

y evaluarse como

$$I = h\left[f(a) + \frac{\Delta f_a}{2} \right] - \frac{1}{12} h^3 f''(\xi)$$

Debido a que

$$\Delta f_a = f(b) - f(a)$$

el resultado se puede escribir como

$$I = h\left[\frac{f(a) + f(b)}{2} \right] - \frac{1}{12} h^3 f''(\xi)$$

o bien

$$I = \underbrace{\frac{b-a}{2}[f(a) + f(b)]}_{\text{Regla del trapecio}} \underbrace{- \frac{1}{12}(b-a)^3 f''(\xi)}_{\text{Error}} \qquad (18)$$

Por lo tanto, el primer término es el de *la regla del trapecio* y el segundo es una *estimación del error*.

Geométricamente, con la regla del trapecio aproximamos el área por debajo de la curva $f(x)$ con el área sombreada por debajo de la recta de interpolación (la cual puede denotarse como $P_1(x)$) y, por lo tanto, el error E es igual al área entre $P_1(x)$ y $f(x)$ (ver Figura 3).

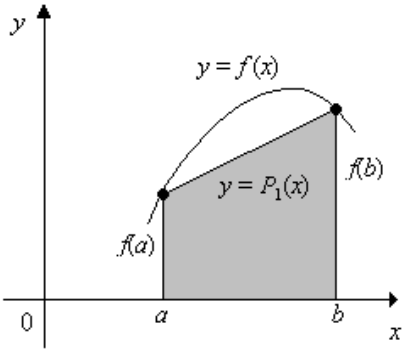


Figura 3

Por lo tanto, la integral se puede representar (recordando fórmulas de la geometría) como

$$I = f(a)(b - a) + (b - a)\frac{[f(b) - f(a)]}{2} + E =$$
$$= (b - a)\frac{[f(b) - f(a)]}{2} + E$$

Cuando se emplea la integral bajo un segmento de línea recta para aproximar la integral bajo la curva, obviamente se incurre en un error que puede ser sustancial. Una estimación del error al aplicar este método está dada (según la ecuación (18)) por

$$E \approx -\frac{1}{12}(b - a)^3 f''(\xi)$$

donde ξ es un punto cualquiera dentro del intervalo $[a, b]$. Esta expresión indica que si la función que se está integrado es lineal, la regla trapezoidal será exacta. De otra manera, ocurrirá un error para funciones con derivadas de segundo y tercer orden (esto es, con curvatura).

5.2.1.1. La regla del trapecio de segmentos múltiples.

Una manera de mejorar la exactitud de la regla trapezoidal es la de dividir el intervalo de integración de a a b en un conjunto de segmentos y

aplicar el método a cada uno de los segmentos. A continuación se suman las áreas de los segmentos individuales y se obtiene la integral sobre el intervalo completo. A la ecuación resultante se la conoce como *regla compuesta del trapecio* o *regla del trapecio de segmentos múltiples*.

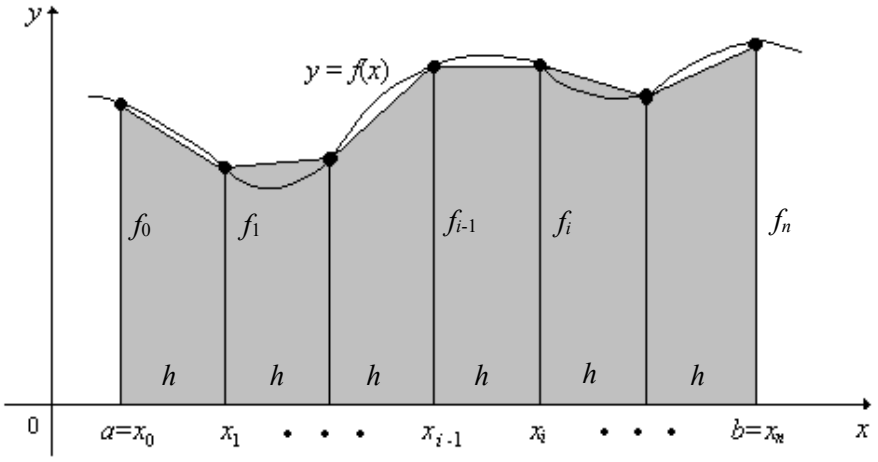


Figura 4

Dividimos el intervalo de integración de a a b en n segmentos con una separación uniforme h; $h = \frac{b-a}{n}$. Tenemos $n + 1$ puntos base igualmente espaciados x_0, x_1, ..., x_n, igualando a y b a x_0 y x_n, respectivamente (ver Figura 4). La integral total se representa como

$$I = \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + ... + \int_{x_{n-1}}^{x_n} f(x)dx$$

Sustituyendo la regla trapezoidal para cada una de las integrales, se obtiene

$$I \approx h\frac{f(x_1)+f(x_0)}{2} + h\frac{f(x_2)+f(x_1)}{2} + ... + h\frac{f(x_n)+f(x_{n-1})}{2} \quad (19)$$

o, agrupando términos

$$I \approx \frac{h}{2}\left[f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n) \right] = h\left[\frac{f(x_0)}{2} + \sum_{i=1}^{n-1} f(x_i) + \frac{f(x_n)}{2} \right] \quad (20)$$

El error en la regla trapezoidal compuesta se obtiene sumando los errores individuales de cada uno de los segmentos, dando

$$E \approx -\frac{1}{12}\frac{(b-a)^3}{n^3}\sum_{i=1}^{n} f''(\xi_i) \qquad (21)$$

donde ξ_i es un punto localizado dentro del segmento *i*.

Este resultado se puede simplificar calculando la media o el valor promedio de la segunda derivada sobre el intervalo completo

$$\frac{\sum_{i=1}^{n} f''(\xi_i)}{n} = \mu$$

Evidentemente, μ cae entre el valor más pequeño *m* y el más grande *M* de la derivada segunda de *f* en el intervalo [*a*, *b*]. Por lo tanto,

$$m \leq \mu \leq M$$

y como *f* '' es continua en [*a*, *b*], toma todos los valores intermedios entre *m* y *M*. Por consiguiente, existe $\xi \in [a, b]$ tal que

$$\mu = f''(\xi)$$

Luego, rescribimos la ecuación (21) como

$$E \approx -\frac{1}{12}\frac{(b-a)^3}{n^3} n f''(\xi)$$

de donde

$$E \approx -\frac{1}{12}\frac{(b-a)^3}{n^2} f''(\xi) \qquad (22)$$

Teniendo en cuenta que $h = \frac{b-a}{n} = \frac{x_n - x_0}{n}$ expresamos la ecuación (22) como

$$E \approx -\frac{h^2}{12}(x_n - x_0) f''(\xi) \qquad (23)$$

La ecuación (23) muestra que el error de la regla del trapecio compuesta es proporcional a h^2 para un intervalo fijo $[a, b]$.

La regla de trapecio compuesta se escribe, a partir de las ecuaciones (20) y (23), como

$$I = h\left[\frac{f(x_0)}{2} + \sum_{i=1}^{n-1} f(x_i) + \frac{f(x_n)}{2}\right] - \frac{h^2}{12}(x_n - x_0) f''(\xi) \qquad (24)$$

donde $\xi \in [a, b] = [x_0, x_n]$.

Ejemplo 4. Utilizando la regla trapezoidal de 4 segmentos, calculemos la integral de $f(x) = e^{x/2}$ desde $a = 0$ hasta $b = 1.00$. Luego, estimaremos el error cometido.

Como $n = 4$, entonces $h = \frac{b-a}{n} = \frac{1}{4} = 0.25$. Así, se tiene la tabla de valores

x	$f(x)$
0	1
0.25	1.1331485
0.50	1.2840254
0.75	1.4549914
1.00	1.6487213

Usando las ecuaciones (20) y (23) obtenemos

$$I = \int_0^1 e^{x/2}\,dx \approx 0.25\,(0.5 + 1.1331485 + 1.2840254 + 1.4549914 + 0.8243606) = 1.2991315$$

y

$$E(\xi) \approx \frac{-(0.25)^2}{12}\frac{e^{\xi/2}}{4}, \quad 0 \le \xi \le 1$$

de donde

$$-0.0021468 \le E \le -0.0013021$$

Esto nos indica que con el paso $h = 0.25$ sólo aseguramos dos decimales.

Hagamos lo mismo que en el caso anterior pero ahora para 8 segmentos. En este caso, $h = 0.125$ y la tabla de valores es

x	$f(x)$
0	1
0.125	1.0644945
0.250	1.1331485
0.375	1.2062302
0.500	1.2840254
0.625	1.3668379
0.750	1.4549914
0.875	1.5488303
1.000	1.6487213

Luego,

$$I = \int_0^1 e^{x/2} dx \approx 1.2978649$$

y

$$E(\xi) \approx \frac{-(0.125)^2}{12} \frac{e^{\xi}}{4}, \quad 0 \le \xi \le 1$$

de donde,

$$-0.0005367 \le E \le -0.0003255$$

lo que nos asegura 3 decimales.

Existen métodos que corrigen errores y mejoran los resultados de la integración numérica en base a la estimación de la misma integral. Conocidos generalmente como *extrapolación de Richardson*, éstos métodos usan dos cálculos de la integral para efectuar un tercer cálculo más exacto.

El cálculo y el error asociado con la regla trapezoidal de segmentos múltiples se representan generalmente como

$$I = I(h) + E(h)$$

donde I es el valor exacto de la integral, $I(h)$ es la aproximación de la integral usando la regla trapezoidal con n segmentos y con tamaño de paso $h = (b - a)/n$ y $E(h)$ es el error (que es proporcional a h^2 para un intervalo fijo $[a, b]$). Si se obtienen dos aproximaciones por separado usando tamaños de paso h_1 y h_2, se tiene que la integral exacta se puede escribir como

$$I = I(h_1) + E(h_1) = I(h_2) + E(h_2) \qquad (25)$$

(Recordemos que el error de la regla trapezoidal de segmentos múltiples se representa por la ecuación

$$E \approx -\frac{b-a}{12} h^2 f''(\xi)\text{, con } n = (b - a)/h,\ \ \xi \in [a, b])$$

Puesto que el error de la regla del trapecio compuesta es proporcional a h^2 (se supone que f'' es una constante que depende del tamaño del paso), los errores con los intervalos h_1 y h_2 se pueden escribir, respectivamente, como

$$E(h_1) \approx C\, h_1^2,\ E(h_2) \approx C\, h_2^2\text{, donde } C \text{ es una constante.} \qquad (26)$$

A partir de la ecuación (25) tenemos que

$$E(h_1) - E(h_2) = I(h_2) - I(h_1) \qquad (27)$$

Al sustituir la ecuación (26) en la ecuación (27), obtenemos

$$Ch_1^2 - Ch_2^2 = I(h_2) - I(h_1)$$

y despejando C

$$C = \frac{I(h_2)-I(h_1)}{h_1^2 - h_2^2}$$

Así, la segunda ecuación (26) da el siguiente valor aproximado de $E(h_2)$

$$E(h_2) \approx h_2^2\ \frac{I(h_2)-I(h_1)}{h_1^2 - h_2^2}$$

o bien

$$E(h_2) \approx \frac{I(h_1)-I(h_2)}{1-(h_1/h_2)^2} \qquad (28)$$

el cual se puede sustituir en la ecuación (25), o sea en $I = I(h_2) + E(h_2)$, dando

$$I \approx I(h_2) + \frac{I(h_1)-I(h_2)}{1-(h_1/h_2)^2} \qquad (29)$$

Este cálculo tiene el importante efecto de quitar el término f'' de los cálculos. En la ecuación (28) se ha desarrollado una expresión que calcula el error en términos del valor de la integral y el tamaño del paso, y en la

ecuación (29) se obtiene una estimación mejorada de la integral. El valor de I no es exacto, puesto que tampoco lo es la ecuación (26), pero se puede demostrar que el error de la ecuación (29) es proporcional a h^4, término que tiene un orden dos veces mayor que al de $I(h)$. Por lo tanto, la ecuación (29) da un resultado más exacto que el de $I(h_1)$ o el de $I(h_2)$. En el caso especial en que el intervalo se divide en dos partes ($h_2 = h_1/2$), la ecuación (29) se transforma en

$$I \approx I(h_2) + \frac{I(h_1) - I(h_2)}{1 - \left(h_1 \Big/ \frac{h_1}{2} \right)^2} = I(h_2) + \frac{I(h_1) - I(h_2)}{(1-4)}$$

o bien

$$I \approx I(h_2) + \frac{1}{3}[I(h_2) - I(h_1)] \qquad (30)$$

Ejemplo 5. Como aplicación de la ecuación (30), usemos las dos estimaciones de la integral obtenidas en el ejemplo 4. Los datos y resultados eran

Segmentos	h	Integral
4	0.25	1.2991315
8	0.125	1.2978649

Utilizando esta información junto con la ecuación (30), calculamos una mejor estimación de la integral, a saber

$$I \approx 1.2978649 + \frac{1}{3}[1.2978649 - 1.2991315] = 1.2978649 - 0.0004222 = 1.2974427$$

valor muy aproximado al exacto que es 1.2974426.

Notemos que si la función es conocida pero calcular sus derivadas es algo complicado, o bien sólo se conocen valores discretos de la función, entonces utilizando la ecuación (28) podemos obtener una estimada del error.

5.2.2. Regla de Simpson.

Además de aplicar la regla trapezoidal con segmentos cada vez más finos, otra manera de obtener una estimación más exacta de una integral es la de usar polinomios de orden superior para conectar los puntos. Por ejemplo, si hay un punto medio extra entre $f(a)$ y $f(b)$, entonces se pueden conectar los tres puntos con una parábola (ver Figura 5). La fórmula resultante de calcular la integral bajo este polinomio se llama *regla de Simpson* y es la segunda fórmula de integración de Newton - Cotes.

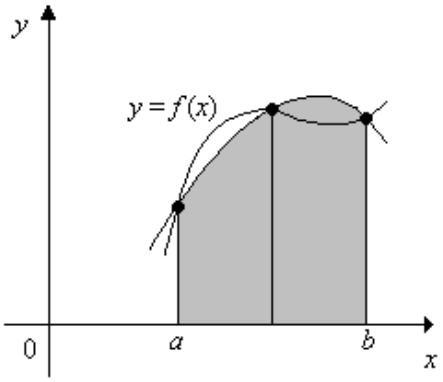


Figura 5

Como se hizo en la regla del trapecio, integraremos el polinomio de Newton hacia adelante de tercer orden con término de error:

$$I = \int_{x_0}^{x_2} \left[f(x_0) + \Delta f(x_0)\, t + \frac{\Delta^2 f(x_0)}{2} t(t-1) + \frac{\Delta^3 f(x_0)}{6} t(t-1)(t-2) + \right.$$
$$\left. + \frac{f^{(4)}(\xi)}{24} t(t-1)(t-2)(t-3) h^4 \right] dx$$

en donde a y b se denominan como x_0 y x_2, respectivamente.

Así, cuando se hacen las simplificaciones y la sustitución, la integral va desde $t = 0$ hasta $t = 2$ ($x = a + th$ y $h = (b - a)/2$)

$$I = h \int_0^2 \left[f(x_0) + \Delta f(x_0)\, t + \frac{\Delta^2 f(x_0)}{2} t(t-1) + \frac{\Delta^3 f(x_0)}{6} t(t-1)(t-2) + \right.$$
$$\left. + \frac{f^{(4)}(\xi)}{24} t(t-1)(t-2)(t-3) h^4 \right] dt = h \left[t f(x_0) + \frac{t^2}{2} \Delta f(x_0) + \right.$$
$$\left. + \left(\frac{t^3}{6} - \frac{t^2}{4} \right) \Delta^2 f(x_0) + \left(\frac{t^4}{24} - \frac{t^3}{6} + \frac{t^2}{6} \right) \Delta^3 f(x_0) + \left(\frac{t^5}{120} - \frac{t^4}{16} + \frac{11 t^3}{72} - \frac{t^2}{8} \right) h^4 f^{(4)}(\xi) \right]_0^2 =$$
$$= h \left[2 f(x_0) + 2 \Delta f(x_0) + \frac{\Delta^2 f(x_0)}{3} + 0\, \Delta^3 f(x_0) - \frac{1}{90} h^4 f^{(4)}(\xi) \right] \qquad (31)$$

Nótese el resultado significativo de que el coeficiente de la tercera diferencia finita es cero (por ello el polinomio se ha escrito hasta el término

de cuarto orden en vez de hasta términos de tercer orden como se esperaría que fuese). Debido a que $\Delta f(x_0) = f(x_1) - f(x_0)$ y $\Delta^2 f(x_0) = f(x_2) - 2f(x_1) + f(x_0)$, la ecuación (31) se puede escribir como

$$I = \underbrace{\frac{h}{3}\left[f(x_0) + 4f(x_1) + f(x_2)\right]}_{\text{Regla de Simpson}} \underbrace{- \frac{1}{90} h^5 f^{(4)}(\xi)}_{\text{Error}} \qquad (32)$$

donde $x_0 = a$, $x_2 = b$ y $h = (b - a)/2$.

Por lo tanto, el primer término es la regla de Simpson y el segundo es una estimación del error. Debido a que la tercera diferencia se anula, se obtiene el resultado significativo de que la fórmula tiene exactitud de tercer orden aunque esté basada únicamente en tres puntos.

5.2.2.1. La regla de Simpson de segmentos múltiples.

Así como en la regla trapezoidal, la regla de Simpson se puede mejorar dividiendo el intervalo de integración en segmentos de igual anchura $h = \frac{b-a}{n}$, con n par. Observemos que, como se ilustra en la Figura 6, se debe usar un número par de segmentos para implementar el método ($n = 2m$).

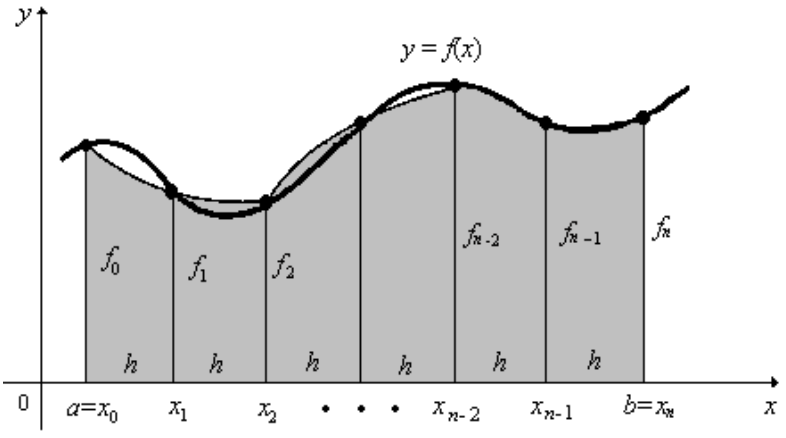


Figura 6

La integral total se representa como

$$I = \int_{x_0}^{x_2} f(x)dx + \int_{x_2}^{x_4} f(x)dx + \cdots + \int_{x_{n-2}}^{x_n} f(x)dx$$

Sustituyendo la regla de Simpson en cada una de las integrales individuales, se obtiene

$$I \approx \frac{h}{3}\left[f(x_0)+4f(x_1)+f(x_2)\right]+\frac{h}{3}\left[f(x_2)+4f(x_3)+f(x_4)\right]+\cdots+$$
$$+\frac{h}{3}\left[f(x_{n-2})+4f(x_{n-1})+f(x_n)\right]=\frac{h}{3}\left[f(x_0)+4\sum_{\substack{i=1\\ i\text{ impar}}}^{n-1} f(x_i)+2\sum_{\substack{i=2\\ i\text{ par}}}^{n-2} f(x_i)+f(x_n)\right] \quad (33)$$

La ecuación (33) se llama *regla de Simpson de segmentos múltiples* o *regla compuesta de Simpson.*

Una estimación del error en la regla de Simpson de segmentos múltiples se obtiene de la misma manera que como se hizo en la regla compuesta del trapecio; es decir, sumando los errores individuales de cada uno de los segmentos y promediando la derivada:

$E \approx -\frac{1}{90}h^5\sum_{i=1}^{m} f^{(4)}(\xi_i)$, siendo $m=\frac{n}{2}$ el número de arcos de parábolas.

Multiplicando y dividiendo por *n* la expresión anterior, resulta

$$E \approx -\frac{1}{90}\frac{h^5 n}{2m}\sum_{i=1}^{m} f^{(4)}(\xi_i)=\frac{-(b-a)h^4}{180}\frac{1}{m}\sum_{i=1}^{m} f^{(4)}(\xi_i)$$

y por lo tanto (tomando el promedio de la derivada cuarta en el intervalo completo)

$$E \approx \frac{-(b-a)h^4}{180} f^{(4)}(\xi), \quad \xi \in [a,\, b]=[x_0, x_n] \quad (34)$$

expresión que nos dice que para un dominio fijo el error es proporcional a h^4.

5.2.2.2. Obtención de la regla de Simpson en forma geométrica.

Hemos obtenido que (según la ecuación (32))

$$I=\int_a^b f(x)dx=\frac{h}{3}\left[f(x_0)+4f(x_1)+f(x_2)\right]+E, \text{ donde, } h=\frac{b-a}{2}=\frac{x_2-x_0}{2}$$

Geométricamente, esta fórmula se obtiene sustituyendo la curva $y = f(x)$ por la parábola $y = P_2(x) = mx^2 + nx + c$ que pasa por tres puntos dados $A_0(x_0, y_0)$, $A_1(x_1, y_1)$, $A_2(x_2, y_2)$ (ver Figura 7). Se supone que los puntos A_0, A_1, A_2 tienen abscisas equidistantes:

$$x_1 - x_0 = x_2 - x_1 = h$$

De aquí, $x_2 - x_0 = 2h$, o sea que $\dfrac{x_2 - x_0}{2} = h$ y x_1 es el punto medio entre x_0 y x_2, esto es $x_1 = (x_0 + x_2)/2$.

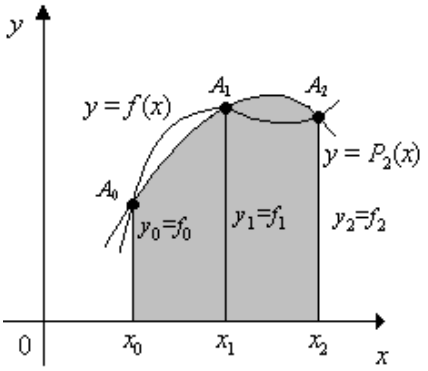


Figura 7

Las tres condiciones que deben satisfacer los coeficientes de la parábola $y = P_2(x) = mx^2 + nx + c$ que pasa por A_0, A_1, A_2 son

$$y_0 = P_2(x_0) = mx_0^2 + nx_0 + c$$

$$y_1 = P_2(x_1) = m\left(\frac{x_0 + x_2}{2}\right)^2 + n\left(\frac{x_0 + x_2}{2}\right) + c$$

$$y_2 = P_2(x_2) = mx_2^2 + nx_2 + c$$

El área comprendida entre el eje x y la parábola $P_2(x)$ limitada por las rectas $x = x_0$, $x = x_2$ es

$$A=\int_{x_0}^{x_2}(mx^2+nx+c)dx=\left[m\frac{x^3}{3}+n\frac{x^2}{2}+cx\right]_{x_0}^{x_2}=\frac{m}{3}(x_2^3-x_0^3)+\frac{n}{2}(x_2^2-x_0^2)+c(x_2-x_0)=$$

$$=\frac{m}{3}\left[(x_2-x_0)(x_2^2+x_2x_0+x_0^2)\right]+\frac{n}{2}(x_2-x_0)(x_2+x_0)+c(x_2-x_0)=$$

$$=\frac{(x_2-x_0)}{6}\left[2mx_2^2+2mx_2x_0+2mx_0^2+3nx_2+3nx_0+6c\right]=$$

$$=\frac{(x_2-x_0)}{6}\left[mx_2^2+nx_2+c+mx_0^2+nx_0+c+mx_2^2+2mx_2x_0+mx_0^2+2nx_2+2nx_0+4c\right]=$$

$$=\frac{(x_2-x_0)}{6}\left[y_2+y_0+m(x_2+x_0)^2+2n(x_2+x_0)+4c\right]=$$

$$=\frac{(x_2-x_0)}{6}\left[y_2+y_0+4y_1\right]=\frac{h}{3}\left[f_0+4f_1+f_2\right].$$

De aquí,

$$I=\underbrace{\frac{h}{3}\left[f_0+4f_1+f_2\right]}_{\text{Regla de Simpson}}+\underbrace{E}_{\text{Error}}$$

donde E es igual al área entre $P_2(x)$ y $f(x)$.

Es así entonces que con la regla de Simpson aproximamos el área por debajo de la curva $f(x)$ con el área sombreada por debajo de la parábola de interpolación $P_2(x)$.

Ejemplo 6. Utilicemos la ecuación (33) con n = 4 para calcular la integral de $f(x)=\dfrac{ln(1+x)}{1+x^2}$ desde a = 0 hasta b = 1.

Como $n=4$ y $h=\dfrac{b-a}{n}$, entonces h = 0.25. Luego, la tabla de valores es

x	$f(x)$
0	0
0.25	0.2100175
0.50	0.3243721
0.75	0.3581541
1.00	0.3465736

Usando la ecuación (33), obtenemos

$$I = \int_0^1 \frac{ln(1+x)}{1+x^2}\,dx \approx \frac{0.25}{3}\left[0.3465736 + 4(0.2100175 + 0.3581541) + 2(0.3243721)\right] = 0.2723337$$

Si para la misma función consideráramos $n = 8$, entonces $h = 0.125$ y la tabla de valores es

x	$f(x)$
0	0
0.125	0.1159710
0.25	0.2100175
0.375	0.2791923
0.5	0.3243721
0.625	0.3491292
0.75	0.3581541
0.875	0.3560261
1.000	0.3465736

Utilizando la ecuación (33), obtenemos

$$I = \int_0^1 \frac{ln(1+x)}{1+x^2}\,dx \approx \frac{0.125}{3}\left[0.3465736 + 4(0.1159710 + 0.2791923 + 0.3491292 + 0.3560261) + \right.$$
$$\left. + 2(0.2100175 + 0.3243721 + 0.3581541\ )\ \right] = 0.2722056.$$

Tal como en el caso de la regla compuesta del trapecio, podemos utilizar la extrapolación de Richardson para corregir errores y mejorar los resultados de la integración numérica, sólo que aquí debemos tener presente que como según la ecuación (34)

$$E \approx -\frac{(b-a)}{180} h^4 f^{(4)}(\xi), \quad \xi \in [a, b]$$

entonces es proporcional a h^4 y, por lo tanto, las ecuaciones correspondientes a (29) y (30) para la regla compuesta de Simpson son, respectivamente

$$I \approx I(h_2) + \frac{I(h_1) - I(h_2)}{1 - (h_1/h_2)^4} \tag{35}$$

o bien, si $h_2 = h_1/2$

$$I \approx I(h_2) + \frac{1}{15}\left[I(h_2) - I(h_1)\right] \tag{36}$$

y se puede demostrar que el error de la ecuación (36) es proporcional a h^6.

Ejemplo 7. Utilicemos la ecuación (36) para obtener una mejor aproximación de la integral calculada en el ejemplo 6:

Segmentos	h	Integral
4	0.25	0.2723337
8	0.125	0.2722056

Luego,
$I \approx 0.2722056 + \dfrac{(0.2722056 - 0.273337)}{15} = 0.2721971$, valor muy próximo al valor exacto que es

$$I = \int_0^1 \frac{ln(1+x)}{1+x^2}\,dx = \frac{\pi}{8} ln\,2 = 0.2721983$$

5.2.3. Comentarios adicionales.

1. El error de la regla compuesta de Simpson es proporcional a h^4 por lo que es dos órdenes más grande que el de la regla compuesta del trapecio. Debido al alto orden del error, la regla compuesta de Simpson tiende a la solución exacta en forma más rápida que la regla compuesta del trapecio cuando h se reduce.

2. **Fórmulas de Newton – Cotes.** Los métodos de integración numérica que se obtienen al integrar las fórmulas de interpolación de Newton reciben el nombre de fórmulas de integración de Newton - Cotes. Como se dijo previamente, la regla del trapecio y la regla de Simpson son miembros de esta familia, las cuales se dividen en fórmulas cerradas y abiertas.

Escribimos las fórmulas cerradas de Newton - Cotes en la forma

$$\int_a^b f(x)dx = \alpha\, h\left[w_0 f_0 + w_1 f_1 + w_2 f_2 + \cdots + w_n f_n\right] + E \qquad (37)$$

donde α y w_i (i = 0, 1, ..., n) son las constantes que aparecen en la Tabla 1 y $f_k = f_k(x)$, $x_k = a + k\,h$, $h = \dfrac{b-a}{n}$.

La ecuación (37) recibe el nombre de fórmula cerrada debido a que el dominio de integración está cerrado por el primer y último datos, esto es, los límites de integración *a* y *b* son puntos de interpolación.

Tabla 1

Segmentos (*n*)	Puntos	α	w_i , $i = 0, 1, \ldots, n$	*E*
1	2	$\frac{1}{2}$	1 1	$-\frac{1}{12}h^3 f^{(2)}$
2	3	$\frac{1}{3}$	1 4 1	$-\frac{1}{90}h^5 f^{(4)}$
3	4	$\frac{3}{8}$	1 3 3 1	$-\frac{3}{80}h^5 f^{(4)}$
4	5	$\frac{2}{45}$	7 32 12 32 7	$-\frac{8}{945}h^7 f^{(6)}$
5	6	$\frac{5}{288}$	19 75 50 50 75 19	$-\frac{275}{12096}h^7 f^{(6)}$
6	7	$\frac{1}{140}$	41 216 27 272 27 216 41	$-\frac{9}{1400}h^9 f^{(8)}$
7	8	$\frac{7}{17280}$	751 3577 1323 2989 2989 1323 3577 751	$-\frac{8183}{518400}h^9 f^{(8)}$
8	9	$\frac{14}{14175}$	989 5888 -928 10946 -4540 10946 -928 5888 989	$-\frac{2368}{467775}h^{11} f^{(10)}$
9	10	$\frac{9}{89600}$	2857 15741 1080 19344 5788 5788 19344 1080 15741 2857	$-\frac{173}{14620}h^{11} f^{(10)}$
10	11	$\frac{5}{299376}$	16067 106300 -48525 272400 -260550 427368 -260550 272400 -48525 106300 16067	$-\frac{1346350}{326918592}h^{13} f^{(12)}$

Se observa que los valores de *w* para *n* grande, también son grandes y cambian de signo. La resta de números grandes puede provocar errores de redondeo. Por esta razón, no son recomendables las fórmulas de Newton - Cotes de orden superior. Además, las fórmulas de orden superior (esto es, mayores de cuatro puntos) rara vez se usan en la práctica. Se puede mejorar la exactitud usando una versión de segmentos múltiples, en vez de optar por las fórmulas de más puntos.

Además, cuando la función se conoce y se requiere de exactitud muy alta, los métodos de integración de Romberg o de Gauss, analizados más adelante, ofrecen alternativas viables y atractivas.

Al observar los valores de *E* se constata que las fórmulas que relacionan un número par de segmentos son exactas si *f*(*x*) es un polinomio

de grado $n+1$ o menor, mientras que las fórmulas que relacionan un número impar de segmentos son exactas si $f(x)$ es un polinomio de grado n o menor. En particular, el error para el caso de tres puntos (Simpson) es $-\frac{1}{90}h^5 f^{(4)}(\xi)$, mientras que para cuatro puntos (llamada regla de los 3/8) es $-\frac{3}{80}h^5 f^{(4)}(\xi)$. Comparando estos errores cuando ambas fórmulas las aplicamos a la evaluación de la misma integral, se observa que en el primer caso $h=\frac{b-a}{2}$ y en el segundo $h=\frac{b-a}{3}$; por lo tanto, el coeficiente del error en la fórmula de Simpson ($E=-\left(\frac{b-a}{2}\right)^5 \frac{1}{90}f''(\xi)$) es $\frac{1}{2880}$ y en la regla de los 3/8 ($E=-\left(\frac{b-a}{3}\right)^5 \frac{3}{80}f^{(4)}(\xi)$) es $\frac{1}{6480}$. Esto indica que la fórmula que tiene como error $-\frac{3}{80}h^5 f^{(4)}(\xi)$ es ligeramente superior a la de Simpson, pero implica considerar una ordenada más.

3. **Integración usando intervalos desiguales.** Las fórmulas vistas hasta el momento se han basado en puntos igualmente espaciados. En la práctica existen muchos casos en donde esta suposición no se cumple y se debe tratar con diferentes tamaños de segmentos. Por ejemplo, los datos derivados experimentalmente, a menudo, son de este tipo. En estos casos, un método es aplicar la regla trapezoidal a cada uno de los segmentos y sumar los resultados:

$$I \approx h_1 \frac{f(x_1)+f(x_0)}{2} + h_2 \frac{f(x_2)+f(x_1)}{2} + \cdots + h_n \frac{f(x_n)+f(x_{n-1})}{2} \quad (38)$$

donde h_i es el ancho del segmento i.

Notemos que este fue el mismo planteamiento usado en la regla trapezoidal de segmentos múltiples. La única diferencia entre las ecuaciones (19) y (38) es que las h de la primera son constantes.

Ejemplo 8. La información del siguiente recuadro se generó usando el polinomio $f(x)=0.2+25x-200x^2+675x^3-900x^4$ con valores de x desigualmente espaciados. Utilicemos la ecuación (38) para determinar la integral de estos datos.

x	$f(x)$
0.00	0.20000000
0.12	1.30972928
0.22	1.30524128
0.32	1.74339328
0.36	2.07490304
0.40	2.45600000
0.44	2.84289496
0.54	3.50729696
0.64	3.18192896
0.70	2.36300000
0.80	0.23200000

La respuesta correcta es 1.64053334.

Aplicando la ecuación (38) a estos datos, se obtiene

$$I \approx 0.12\frac{1.30972928+0.2}{2}+0.10\frac{1.30524128+1.30972928}{2}+\cdots+0.1\frac{0.232+2.363}{2}=$$

$$=0.09058376+0.13074853+\cdots+0.12975=1.56480098$$

Notemos que algunos segmentos adyacentes son de igual ancho (por ejemplo, desde x = 0.12 a 0.32, entre otros) y, por lo tanto, se pueden evaluar usando la regla de Simpson. Esto, en general, lleva a resultados más exactos.

5.2.4. Fórmulas de integración abierta.

Dentro de la clase de fórmulas de integración abierta de Newton – Cotes, veremos una muy sencilla de deducir que es *la regla del punto medio* y su extensión a segmentos múltiples. Las fórmulas abiertas rara vez se usan en la integración. Sin embargo, tienen aplicación directa con los métodos de paso múltiple en la solución de ecuaciones diferenciales ordinarias.

5.2.4.1. Regla del punto medio.

Este método está basado en una interpolación constante o de grado cero y por ello es de esperar que la regla del trapecio sea más exacta que la del punto medio. Sin embargo, se probará que el error cometido en esta última es menor que el error cometido usando la primera. Otra ventaja de la

regla del punto medio con respecto a la del trapecio es que se debe calcular el valor de la función en un solo punto, pero si tenemos dada la función *f* por tabla y sólo conocemos el valor de la función en los extremos del intervalo [*a*, *b*], entonces no podremos usar la regla del punto medio, ya que en esta se requiere el valor de la función en el punto medio del intervalo.

Geométricamente, aproximamos el área por debajo de la curva $f(x)$ por el área del rectángulo sombreado (ver Figura 8).

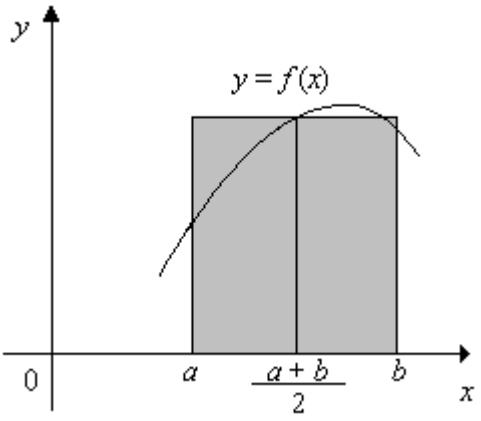


Figura 8

Así, la integral $I = \int_a^b f(x)dx$ puede expresarse

$$I = (b - a) f\left(\frac{a+b}{2}\right) + E$$

Obtengamos esta expresión (y una estimación de *E*) en forma analítica. Llamemos $a = x_0$, $b = x_1$, $y_0 = \frac{x_0 + x_1}{2}$ y consideremos el desarrollo de la serie de Taylor de $f(x)$ alrededor de y_0 de orden 2

$$f(x) = f(y_0) + (x - y_0) f'(y_0) + \frac{(x - y_0)^2}{2} f''(\xi_0) \qquad (39)$$

(supuesto que f tenga derivadas continuas) con ξ_0 un punto comprendido entre y_0 y x.

Integremos la ecuación (39) sobre $[x_0, x_1]$

$$I=\int_{x_0}^{x_1} f(x)dx=\int_{x_0}^{x_1}\left[f(y_0)+(x-y_0)f'(y_0)+\frac{(x-y_0)^2}{2}f''(\xi_0)\right]dx=$$

$$=f(y_0)(x_1-x_0)+\frac{f'(y_0)}{2}\left[(x-y_0)^2\right]_{x_0}^{x_1}+\frac{f''(\xi_0)}{6}\left[(x-y_0)^3\right]_{x_0}^{x_1}=$$

$$=f(y_0)(x_1-x_0)+\frac{f'(y_0)}{2}\left[(x_1-y_0)^2-(x_0-y_0)^2\right]+\frac{f''(\xi_0)}{6}\left[(x_1-y_0)^3-(x_0-y_0)^3\right]=$$

(llamamos h al tamaño del paso de x_0 a x_1)

$$=f(y_0)h+\frac{f'(y_0)}{2}\left[\left(x_1-\frac{x_0+x_1}{2}\right)^2-\left(x_0-\frac{x_0+x_1}{2}\right)^2\right]+\frac{f''(\xi_0)}{6}\left[\left(x_1-\frac{x_0+x_1}{2}\right)^3-\left(x_0-\frac{x_0+x_1}{2}\right)^3\right]=$$

$$=f(y_0)h+\frac{f'(y_0)}{2}\left[\left(\frac{x_1-x_0}{2}\right)^2-\left(\frac{x_0-x_1}{2}\right)^2\right]+\frac{f''(\xi_0)}{6}\left[\left(\frac{x_1-x_0}{2}\right)^3-\left(\frac{x_0-x_1}{2}\right)^3\right]=$$

$$=f(y_0)h+\frac{f''(\xi_0)}{48}\left[(x_1-x_0)^3+(x_1-x_0)^3\right]=f(y_0)h+\frac{f''(\xi_0)}{24}(x_1-x_0)^3=f(y_0)h+\frac{h^3}{24}f''(\xi_0)$$

Luego,

$$I=\underbrace{hf\left(\frac{x_0+x_1}{2}\right)}_{\text{Regla del punto medio}}+\underbrace{\frac{h^3}{24}f''(\xi_0)}_{\text{Error}} \qquad (40)$$

Similarmente como se hizo con los métodos anteriores, la regla del punto medio se puede mejorar dividiendo el intervalo de integración de igual anchura $h=\dfrac{b-a}{n}=\dfrac{x_n-x_0}{n}$ (Figura 9).

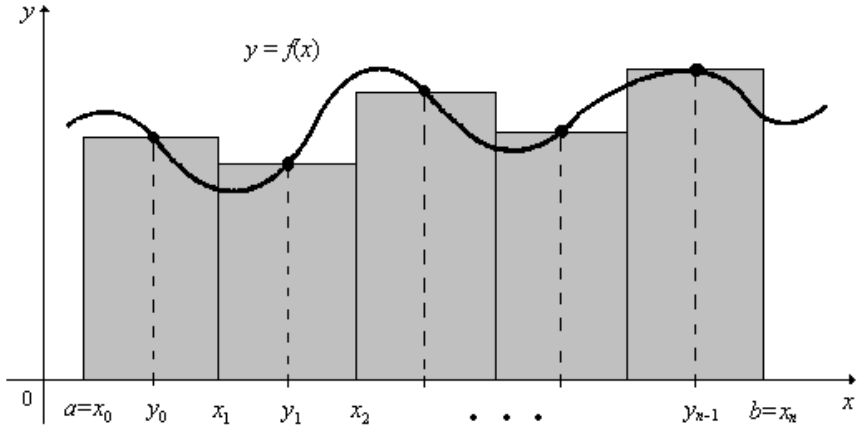


Figura 9

La integral total se expresa como

$$I = \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \cdots + \int_{x_{n-1}}^{x_n} f(x)dx$$

Sustituyendo la regla del punto medio en cada una de las integrales individuales, se obtiene

$$I \approx hf\left(\frac{x_0 + x_1}{2}\right) + hf\left(\frac{x_1 + x_2}{2}\right) + \ldots + hf\left(\frac{x_{n-1} + x_n}{2}\right) = \tag{41}$$

$$= h\left[f(y_0) + f(y_1) + f(y_2) + \cdots + f(y_{n-1})\right], \text{donde}, y_i = \frac{x_i + x_{i+1}}{2} \ (i = 0, 1, ..., n\text{-}1)$$

La ecuación (41) se llama *regla del punto medio de segmentos múltiples* o *regla compuesta del punto medio.*

Una estimación del error en la regla compuesta del punto medio se obtiene sumando los errores individuales de cada uno de los segmentos y promediando la derivada

$$E \approx \frac{h^3}{24}\sum_{i=0}^{n-1} f''(\xi_i) = \frac{nh^3}{24}\frac{1}{n}\sum_{i=0}^{n-1} f''(\xi_i) = \frac{(b-a)}{24}h^2\frac{1}{n}\sum_{i=0}^{n-1} f''(\xi_i)$$

y, por lo tanto, tomando el promedio de la derivada segunda en el intervalo completo

$$E \approx \frac{(b-a)}{24}h^2 f^{(2)}(\xi), \qquad \xi \in [a, b] \tag{42}$$

expresión que nos dice que para un dominio fijo el error es proporcional a h^2.

De las ecuaciones (42) y (23) (ecuación esta última que nos da la estimación del error en la regla compuesta del trapecio) podemos decir que la regla del punto medio (y la compuesta) es más exacta que la regla del trapecio (y la compuesta) en la mayoría de los casos, aunque ésta última prevalece sobre la otra si se emplea la extrapolación de Richardson. La regla del punto medio es cerca de dos veces más exacta que la del trapecio, lo que sorprende a mucha gente. La diferencia entre los valores obtenidos por ambas reglas puede ser usada para estimar el error en cada una de ellas. Sin embargo, la estimación no es infalible; es posible para ambas reglas dar los mismos valores, pero que los mismos sean incorrectos.

Si tenemos dada $f(x)$ por tabla usaremos la regla del trapecio y aunque conozcamos la expresión analítica de $f(x)$ resulta más económico, en cuanto al número de operaciones a realizar, usar la regla del trapecio y no la regla del punto medio. Por ello que en la práctica es común usar la regla del trapecio en la mayoría de los casos.

Ejemplo 9. Utilicemos la regla del punto medio de 4 segmentos para calcular la integral de $f(x) = e^{x/2}$, desde $a = 0$ hasta $b = 1$ y estimar el error. (Observemos que es lo mismo que se calcula en el ejemplo 4 pero allí con la regla del trapecio).

Como $n = 4$, entonces $h = (b - a)/n = 1/4 = 0.25$.

La tabla de valores es

x	0	0.25	0.50	0.75	1
$f(x)$	1	1.1331485	1.2840254	1.4549914	1.6487213

A partir de las ecuaciones (41) y (42), obtenemos

$$I = \int_0^1 e^{x/2}\,dx \approx 0.25\,[1.0644945 + 1.2062302 + 1.3668379 + 1.5488303] = 1.2965982$$

y

$$E(\xi) = \frac{1}{4} e^{\xi/2}\,\frac{(0.25)^2}{24},\ \xi \in [0, 1]$$

de donde,

$$6.5 \times 10^{-4} \le E \le 1.1 \times 10^{-3}$$

Logramos un mejor resultado que el obtenido en el ejemplo 4 para este mismo paso h, aunque aquí también sólo aseguramos dos decimales. Por supuesto que usando la ecuación (30), o sea extrapolando, obtuvimos una muy buena aproximación al valor de la integral I (ejemplo 5).

5.2.5. Fórmula de Euler Maclaurin.

Esta fórmula (al igual que las anteriores) nos permite calcular en forma aproximada la integral definida $I = \int_{x_0}^{x_n} f(x)dx$ y como veremos, da un

conjunto de correcciones a la fórmula compuesta del trapecio que dependen de los valores de las derivadas impares en los extremos del intervalo.

Sea $F(x)$ una función primitiva de $f(x)$, es decir, $F^{(1)}(x) = f(x)$, $F^{(2)}(x) = f^{(1)}(x)$, $F^{(3)}(x) = f^{(2)}(x)$, ...

El desarrollo en serie de Taylor de $F(x)$ será

$$F(x + h) - F(x) = h f(x) + \frac{h^2}{2!} f^{(1)}(x) + \frac{h^3}{3!} f^{(2)}(x) + \frac{h^4}{4!} f^{(3)}(x) + \frac{h^5}{5!} f^{(4)}(x) + \ldots \quad (43)$$

Haciendo sucesivamente $x = x_0, x_1, x_2, \ldots, x_{n-1}$ en la ecuación (43), se obtiene

$$F(x_0 + h) - F(x_0) = h f(x_0) + \frac{h^2}{2!} f^{(1)}(x_0) + \frac{h^3}{3!} f^{(2)}(x_0) + \frac{h^4}{4!} f^{(3)}(x_0) + \frac{h^5}{5!} f^{(4)}(x_0) + \ldots$$

$$F(x_1 + h) - F(x_1) = h f(x_1) + \frac{h^2}{2!} f^{(1)}(x_1) + \frac{h^3}{3!} f^{(2)}(x_1) + \frac{h^4}{4!} f^{(3)}(x_1) + \frac{h^5}{5!} f^{(4)}(x_1) + \ldots$$

$$F(x_2 + h) - F(x_2) = h f(x_2) + \frac{h^2}{2!} f^{(1)}(x_2) + \frac{h^3}{3!} f^{(2)}(x_2) + \frac{h^4}{4!} f^{(3)}(x_2) + \frac{h^5}{5!} f^{(4)}(x_2) + \ldots$$

$$F(x_3 + h) - F(x_3) = h f(x_3) + \frac{h^2}{2!} f^{(1)}(x_3) + \frac{h^3}{3!} f^{(2)}(x_3) + \frac{h^4}{4!} f^{(3)}(x_3) + \frac{h^5}{5!} f^{(4)}(x_3) + \ldots$$

$$\vdots$$

$$F(x_{n-1} + h) - F(x_{n-1}) = h f(x_{n-1}) + \frac{h^2}{2!} f^{(1)}(x_{n-1}) + \frac{h^3}{3!} f^{(2)}(x_{n-1}) + \frac{h^4}{4!} f^{(3)}(x_{n-1}) + \frac{h^5}{5!} f^{(4)}(x_{n-1}) + \ldots$$

Sumando miembro a miembro estas igualdades, se obtiene

$$F(x_n) - F(x_0) =$$

$$= h \sum_{i=0}^{n-1} f(x_i) + \frac{h^2}{2!} \sum_{i=0}^{n-1} f^{(1)}(x_i) + \frac{h^3}{3!} \sum_{i=0}^{n-1} f^{(2)}(x_i) + \frac{h^4}{4!} \sum_{i=0}^{n-1} f^{(3)}(x_i) + \frac{h^5}{5!} \sum_{i=0}^{n-1} f^{(4)}(x_i) + \ldots \quad (44)$$

(Recordemos que las abscisas x_i son equidistantes y además que $x_{i+1} = x_i + h$, para $i = 0, 1, \ldots, n - 1$).

En forma análoga podemos hacer el desarrollo en serie de Taylor de $f(x)$ y se obtiene

$$f(x+h) - f(x) = h f^{(1)}(x) + \frac{h^2}{2!} f^{(2)}(x) + \frac{h^3}{3!} f^{(3)}(x) + \frac{h^4}{4!} f^{(4)}(x) + \frac{h^5}{5!} f^{(5)}(x) + \ldots \quad (45)$$

Haciendo sucesivamente $x = x_0, x_1, x_2, \ldots, x_{n-1}$ en la ecuación (45) y sumando obtenemos

$f(x_n) - f(x_0) =$

$$= h\sum_{i=0}^{n-1} f^{(1)}(x_i) + \frac{h^2}{2!}\sum_{i=0}^{n-1} f^{(2)}(x_i) + \frac{h^3}{3!}\sum_{i=0}^{n-1} f^{(3)}(x_i) + \frac{h^4}{4!}\sum_{i=0}^{n-1} f^{(4)}(x_i) + \frac{h^5}{5!}\sum_{i=0}^{n-1} f^{(5)}(x_i) + \ldots \quad (46)$$

Haciendo el desarrollo en serie de Taylor de $f^{(1)}(x)$, se tiene

$$f^{(1)}(x+h) - f^{(1)}(x) = hf^{(2)}(x) + \frac{h^2}{2!}f^{(3)}(x) + \frac{h^3}{3!}f^{(4)}(x) + \frac{h^4}{4!}f^{(5)}(x) + \frac{h^5}{5!}f^{(6)}(x) + \ldots \quad (47)$$

Dando sucesivamente a x los valores x_0, x_1, x_2, ..., x_{n-1} en la ecuación (47) y sumando se obtiene

$f^{(1)}(x_n) - f^{(1)}(x_0) =$

$$= h\sum_{i=0}^{n-1} f^{(2)}(x_i) + \frac{h^2}{2!}\sum_{i=0}^{n-1} f^{(3)}(x_i) + \frac{h^3}{3!}\sum_{i=0}^{n-1} f^{(4)}(x_i) + \frac{h^4}{4!}\sum_{i=0}^{n-1} f^{(5)}(x_i) + \frac{h^5}{5!}\sum_{i=0}^{n-1} f^{(6)}(x_i) + \ldots \quad (48)$$

Fórmulas semejantes se obtienen considerando $f^{(2)}(x), f^{3)}(x)$, ...

$f^{(2)}(x_n) - f^{(2)}(x_0) =$

$$= h\sum_{i=0}^{n-1} f^{(3)}(x_i) + \frac{h^2}{2!}\sum_{i=0}^{n-1} f^{(4)}(x_i) + \frac{h^3}{3!}\sum_{i=0}^{n-1} f^{(5)}(x_i) + \frac{h^4}{4!}\sum_{i=0}^{n-1} f^{(6)}(x_i) + \frac{h^5}{5!}\sum_{i=0}^{n-1} f^{(7)}(x_i) + \ldots \quad (49)$$

$f^{(3)}(x_n) - f^{(3)}(x_0) =$

$$= h\sum_{i=0}^{n-1} f^{(4)}(x_i) + \frac{h^2}{2!}\sum_{i=0}^{n-1} f^{(5)}(x_i) + \frac{h^3}{3!}\sum_{i=0}^{n-1} f^{(6)}(x_i) + \frac{h^4}{4!}\sum_{i=0}^{n-1} f^{(7)}(x_i) + \frac{h^5}{5!}\sum_{i=0}^{n-1} f^{(8)}(x_i) + \ldots \quad (50)$$

Sumando a la ecuación (44) la ecuación (46) multiplicada por $c_1\ h$, la ecuación (48) por $c_2\ h^2$, la ecuación (49) por $c_3\ h^3$, la ecuación (50) por $c_4\ h^4$, ... se obtiene

$$F(x_n) - F(x_0) + c_1 h\left[f(x_n) - f(x_0)\right] + c_2 h^2\left[f^{(1)}(x_n) - f^{(1)}(x_0)\right] + c_3 h^3\left[f^{(2)}(x_n) - f^{(2)}(x_0)\right] +$$

$$+ c_4 h^4\left[f^{(3)}(x_n) - f^{(3)}(x_0)\right] + \ldots =$$

$$= h\sum_{i=0}^{n-1} f(x_i) + h^2\sum_{i=0}^{n-1} f^{(1)}(x_i)\left[\frac{1}{2!}+c_1\right] + h^3\sum_{i=0}^{n-1} f^{(2)}(x_i)\left[\frac{1}{3!}+\frac{c_1}{2!}+c_2\right]+$$

$$+h^4\sum_{i=0}^{n-1} f^{(3)}(x_i)\left[\frac{1}{4!}+\frac{c_1}{3!}+\frac{c_2}{2!}+c_3\right]+h^5\sum_{i=0}^{n-1} f^{(4)}(x_i)\left[\frac{1}{5!}+\frac{c_1}{4!}+\frac{c_2}{3!}+\frac{c_3}{2!}+c_4\right]+\ldots \quad (51)$$

Las constantes c_1, c_2, c_3, ... las determinamos de modo que anulen los corchetes del segundo miembro de la ecuación (51), obteniéndose

$$c_1=-\frac{1}{2},\ c_2=\frac{1}{12},\ c_3=0,\ c_4=-\frac{1}{720},\ c_5=0,\ c_6=\frac{1}{30240},\ c_7=0,\ c_8=-\frac{1}{1209600},\ldots$$

Reemplazando estas constantes en la ecuación (51), resulta

$$F(x_n) - F(x_0) = h\sum_{i=0}^{n-1} f(x_i) + \frac{h}{2}\left[f(x_n)-f(x_0)\right] - \frac{h^2}{12}\left[f^{(1)}(x_n)-f^{(1)}(x_0)\right] + \quad (52)$$
$$+\frac{h^4}{720}\left[f^{(3)}(x_n)-f^{(3)}(x_0)\right] - \frac{h^6}{30240}\left[f^{(5)}(x_n)-f^{(5)}(x_0)\right] + \frac{h^8}{1204600}\left[f^{(7)}(x_n)-f^{(7)}(x_0)\right]+\ldots$$

Dado que $F(x)$ es una primitiva de $f(x)$, por la Regla de Barrow, se tiene

$$\int_{x_0}^{x_n} f(x)dx = F(x_n) - F(x_0)$$

y, por consiguiente, puede escribirse

$$\int_{x_0}^{x_n} f(x)dx = h\left[\frac{f(x_0)}{2} + f(x_1) + f(x_2) + \cdots + f(x_{n-1}) + \frac{f(x_n)}{2}\right] - \frac{h^2}{12}\left[f^{(1)}(x_n)-f^{(1)}(x_0)\right] +$$
$$+\frac{h^4}{720}\left[f^{(3)}(x_n)-f^{(3)}(x_0)\right] - \frac{h^6}{30240}\left[f^{(5)}(x_n)-f^{(5)}(x_0)\right] + \quad (53)$$
$$+\frac{h^8}{1209600}\left[f^{(7)}(x_n)-f^{(7)}(x_0)\right]-\cdots$$

Como se observa, el primer término del segundo miembro de la ecuación (53) es la fórmula compuesta del trapecio a la cual se le aplican correcciones que dependen de las derivadas de orden impar en los extremos del intervalo.

La relación (53) se conoce con el nombre de *fórmula de Euler Maclaurin* o también como *regla compuesta del trapecio con términos de corrección.*

Ejemplo 10. Calculemos con la fórmula de Euler Maclaurin la $\int_0^1 e^{-x^2} dx$ eligiendo $h = 0.25$.

Como $h = 0.25$, se tiene la tabla de valores

x	$f(x)$
0	1
0.25	0.9394131
0.50	0.7788008
0.75	0.5697828
1	0.3678794

Como $f(x) = e^{-x^2}$, entonces

$f^{(1)}(x) = -2x\ e^{-x^2}$, $f^{(1)}(0) = 0$, $f^{(1)}(1) = -0.7357589$

$f^{(2)}(x) = e^{-x^2}(4x^2 - 2)$

$f^{(3)}(x) = e^{-x^2}(12x - 8x^3)$, $f^{(3)}(0) = 0$, $f^{(3)}(1) = 1.4715178$

$\vdots$

Luego, utilizando la ecuación (53), se obtiene

$$\int_0^1 e^{-x^2} dx = 0.25(0.5 + 0.9394131 + 0.7788008 + 0.5697828 + 0.1839397) -$$

$$-\frac{0.0625}{12}(-0.7357589) + \frac{0.0039063}{720}1.4715178 + \ldots \approx$$

$$\approx 0.7429481 + 0.0038321 + 0.000080 = 0.7468242$$

La ecuación (53) puede aplicarse siempre que se conozca la expresión analítica de $f(x)$. En el caso de tenerse sólo valores discretos de la función, las expresiones de las derivadas en la ecuación (53) pueden reemplazarse en función de las diferencias finitas sucesivas obteniéndose la *fórmula de Gregory.*

En efecto, derivando sucesivamente la ecuación (4)

$$hf^{(1)}(x_0+ht) = \Delta f_0 + \frac{1}{2}(2t-1)\Delta^2 f_0 + \frac{1}{6}(3t^2 - 6t + 2)\Delta^3 f_0 + \frac{1}{24}(4t^3 - 18t^2 + 22t - 6)\Delta^4 f_0 + \ldots$$

se tiene

$$h^2 f^{(2)}(x_0 + ht) = \Delta^2 f_0 + \frac{1}{3!}(6t - 6)\Delta^3 f_0 + \frac{1}{4!}(12t^2 - 36t + 22)\Delta^4 f_0 +$$

$$+\frac{1}{5!}(20\,t^3 - 120\,t^2 + 210\,t - 100)\,\Delta^5 f_0 + \ldots$$

$$h^3 f^{(3)}(x_0 + h\,t) = \Delta^3 f_0 + \frac{1}{4!}(24\,t - 36)\,\Delta^4 f_0 + \frac{1}{5!}(60\,t^2 - 240\,t + 210)\,\Delta^5 f_0 + \ldots$$

$$h^4 f^{(4)}(x_0 + h\,t) = \Delta^4 f_0 + \frac{1}{5!}(120\,t - 240)\,\Delta^5 f_0 + \ldots \qquad (54)$$

$$h^5 f^{(5)}(x_0 + h\,t) = \Delta^5 f_0 + \ldots$$

Haciendo $t = 0$ en las derivadas impares, se obtiene

$$h f^{(1)}(x_0) = \Delta f_0 - \frac{1}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0 - \frac{1}{4}\Delta^4 f_0 + \frac{1}{5}\Delta^5 f_0 - \ldots$$

$$h^3 f^{(3)}(x_0) = \Delta^3 f_0 - \frac{3}{2}\Delta^4 f_0 + \frac{7}{4}\Delta^5 f_0 - \ldots$$

$$h^5 f^{(5)}(x_0) = \Delta^5 f_0 - \ldots$$

Análogamente, derivando sucesivamente la ecuación (10)

$$hf^{(1)}(x_n + ht) = \nabla f_n + \frac{1}{2}(2t+1)\nabla^2 f_n + \frac{1}{6}(3t^2+6t+2)\nabla^3 f_n + \frac{1}{24}(4t^3+18t^2+22t+6)\nabla^4 f_n + \ldots$$

se tiene

$$h^2 f^{(2)}(x_n + ht) = \nabla^2 f_n + \frac{1}{3!}(6\,t + 6)\nabla^3 f_n + \frac{1}{4!}(12\,t^2 + 36\,t + 22)\nabla^4 f_n +$$

$$+\frac{1}{5!}(20\,t^3 + 120\,t^2 + 210\,t + 100)\nabla^5 f_n + \ldots$$

$$h^3 f^{(3)}(x_n + h\,t) = \nabla^3 f_n + \frac{1}{4!}(24\,t+36)\nabla^4 f_n + \frac{1}{5!}(60\,t^2 + 240\,t + 210)\nabla^5 f_n + \ldots \qquad (55)$$

$$h^4 f^{(4)}(x_n + h\,t) = \nabla^4 f_n + \frac{1}{5!}(120\,t + 240)\nabla^5 f_n + \ldots$$

$$h^5 f^{(5)}(x_n + h\,t) = \nabla^5 f_n + \ldots$$

Haciendo $t = 0$ en las derivadas impares, se obtiene

$$hf^{(1)}(x_n) = \nabla f_n + \frac{1}{2}\nabla^2 f_n + \frac{1}{3}\nabla^3 f_n + \frac{1}{4}\nabla^4 f_n + \frac{1}{5}\nabla^5 f_n + \ldots$$

$$h^3 f^{(3)}(x_n) = \nabla^3 f_n + \frac{3}{2}\nabla^4 f_n + \frac{7}{4}\nabla^5 f_n + \ldots$$
$$h^5 f^{(5)}(x_n) = \nabla^5 f_n + \ldots$$

Reemplazando en la ecuación (53) los valores de las derivadas impares en x_0 y en x_n por las relaciones anteriores, se obtiene la *fórmula de Gregory*

$$\begin{aligned}\int_{x_0}^{x_n} f(x)dx = h\left[\frac{f_0}{2} + f_1 + \cdots + f_{n-1} + \frac{f_n}{2}\right] - \frac{h^2}{12}\left(\frac{\nabla f_n - \Delta f_0}{h}\right) - \frac{h^2}{12}\left(\frac{\nabla^2 f_n + \Delta^2 f_0}{2h}\right) - \\ - \frac{h^2}{12}\left(\frac{\nabla^3 f_n - \Delta^3 f_0}{3h}\right) - \frac{h^2}{12}\left(\frac{\nabla^4 f_n + \Delta^4 f_0}{4h}\right) - \frac{h^2}{12}\left(\frac{\nabla^5 f_n - \Delta^5 f_0}{5h}\right) + \\ + \frac{h^4}{720}\left(\frac{\nabla^3 f_n - \Delta^3 f_0}{h^3}\right) + \frac{h^4}{720}\frac{3}{2h^3}\left(\nabla^4 f_n + \Delta^4 f_0\right) + \\ + \frac{h^4}{720}\frac{7}{4}\left(\frac{\nabla^5 f_n - \Delta^5 f_0}{h^3}\right) - \frac{h^6}{30240}\left(\frac{\nabla^5 f_n - \Delta^5 f_0}{h^5}\right) = \\ = h\left[\frac{f_0}{2} + f_1 + \cdots + f_{n-1} + \frac{f_n}{2}\right] - \frac{h}{12}\left(\nabla f_n - \Delta f_0\right) - \frac{h}{24}\left(\nabla^2 f_n + \Delta^2 f_0\right) - \\ - \frac{19}{720}h\left(\nabla^3 f_n - \Delta^3 f_0\right) - \frac{3}{160}h\left(\nabla^4 f_n + \Delta^4 f_0\right) - \frac{863}{60480}h\left(\nabla^5 f_n - \Delta^5 f_0\right) - \ldots\end{aligned} \tag{56}$$

Ejemplo 11. Usando la ecuación (56), calculemos la $\int_0^1 \frac{arctg\, x}{1+x}dx$ eligiendo $h = 0.125$.

Dado que en todos los términos del segundo miembro de la ecuación (56) aparece el factor h, al tabular la función la multiplicamos por h. Así, resulta

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$	$\Delta^4 f$	$\Delta^5 f$
0	0					
0.125	0.01381722	0.01381722	-0.00313657			
0.250	0.02449787	0.01068065	-0.002563	0.00057357		
0.375	0.03261552	0.00811765	-0.00209587	0.00046713	-0.00010644	
0.5	0.03863730	0.00602178	-0.0016899	0.00040597	-0.00006116	0.00004528
0.625	0.04296918	0.00433188	-0.0013367	0.0003532	-0.00005277	0.00000839
0.750	0.04596436	0.00299518	-0.00103754	0.00029916	-0.00005404	0.00000127
0.875	0.04792200	0.00195764	-0.00079225	0.00024529	-0.00005387	0.00000017
1.000	0.04908739	0.00116539				

Por lo tanto, utilizando la ecuación (56), obtenemos

$$I \approx 0.27096715 + 0.00105432 + 0.00016370 + 0.00000866 + 0.00000301 + 0.00000064 =$$
$$= 0.27219748$$

5.2.6. Integración de Romberg y cuadratura de Gauss.

Hemos mencionado que las funciones a integrarse numéricamente tienen, en general, dos formas: una tabla de valores o una ecuación.

La forma de los datos tiene una influencia importante en el esquema que se va a usar para evaluar la integral. Para el caso de información tabular se está limitado al número de puntos datos. En contraste, si se dispone de la función analíticamente, entonces se pueden generar tantos valores de $f(x)$ como sean necesarios para alcanzar una exactitud aceptable.

Estudiaremos a continuación dos métodos que están expresamente diseñados para analizar casos en que se conoce la función. Ambos métodos aprovechan la facilidad de generar valores de la función en el desarrollo de esquemas eficientes de la integración numérica. El primero de ellos se basa en la extrapolación de Richardson, método que combina dos aproximaciones de integración numérica en la obtención de un tercer valor que es más exacto. El algoritmo que implementa la extrapolación de Richardson en su forma más eficiente se llama *integración de Romberg*. Este método es recursivo y se usa para generar una aproximación a la integral dentro de una tolerancia de error especificada. El segundo de los métodos es llamado

cuadratura o integración Gaussiana. Las fórmulas de este método emplean valores de x contenidos dentro de a y b de tal forma que resulta una integral mucho más exacta. (Recordemos que los valores de $f(x)$ en las fórmulas de Newton - Cotes cerradas se determinan en valores específicos de x. Por ejemplo, si se usa la regla trapezoidal para determinar una integral se está restringiendo a tomar el promedio pesado de $f(x)$ en los extremos de los intervalos).

5.2.6.1. Integración de Romberg.

Hemos visto las reglas trapezoidal y de Simpson de segmentos múltiples en las que, para una función analítica, las ecuaciones de error indican que aumentando el número n de segmentos se genera una aproximación más exacta a la integral. Sin embargo, notemos que para valores muy grandes de n el error puede empezar a crecer si los errores de redondeo comienzan a dominar.

También observemos que se necesita un número muy grande de segmentos (y, por lo tanto, esfuerzo de cálculo) para alcanzar niveles altos de exactitud. Como una consecuencia de estos inconvenientes, las reglas trapezoidal y de Simpson de segmentos múltiples algunas veces son inadecuadas en problemas donde se necesita gran eficiencia y pocos errores.

La integración de Romberg es un método diseñado para evitar estos inconvenientes. Es muy similar a los métodos analizados anteriormente, en el sentido de que está basado en la aplicación sucesiva de la regla trapezoidal. Sin embargo, mediante manipulaciones matemáticas se obtienen mejores resultados con menos esfuerzo.

Hemos visto que la ecuación

$$I \approx I(h_2) + \left[\frac{1}{(h_1/h_2)^2 - 1}\right]\left[I(h_2) - I(h_1)\right] \qquad (57)$$

proporciona una forma de combinar dos aplicaciones de la regla trapezoidal con un error de orden h^2 y calcula una estimación de orden h^4. Este planteamiento es un subconjunto de un método más general que combina integrales para obtener mejores estimaciones, conocido con el nombre de *método de integración de Romberg*. El procedimiento que se le atribuye a Romberg consiste, básicamente, en utilizar el método de extrapolación de Richardson reiteradamente sobre la regla trapezoidal cuando la función es conocida en puntos equidistantes del intervalo de integración. Para comprender fácilmente el método de Romberg conviene ir desarrollando un ejemplo y recordar la fórmula de Euler Maclaurin (53) que podemos escribir así

$$I = \int_a^b f(x)dx = \text{trapecios} + c_1 h^2 + c_2 h^4 + c_3 h^6 + \cdots = \\ = h\left[\frac{f_0}{2} + f_1 + \cdots + f_{n-1} + \frac{f_n}{2}\right] + \sum_{j=1}^{\infty} c_j h^{2j} \tag{58}$$

con $h = \dfrac{(b-a)}{n}$ y donde las c_j dependen únicamente de a, b y $f(x)$.

Ejemplo 12. Debemos calcular la integral de la función $f(x) = cosh\ x\ e^x$ desde $a = 0$ hasta $b = 1$, esto es, debemos calcular $I = \int_0^1 cosh\ x\ e^x dx$

Comencemos utilizando la regla trapezoidal de un segmento. Los valores de la función son

$f(0) = 1$
$f(1) = 4.1945280$

Luego,

$$I \approx 1\,\frac{1 + 4.1945280}{2} = 2.5972640 = T_1$$

aproximación ésta en el caso de considerar un segmento ($h = b - a = 1$).

Utilicemos ahora la regla trapezoidal de 2 segmentos: $n = 2$ ($h = 0.5$) y los valores de la función son

$f(0) = 1$
$f(0.5) = 1.8591409$
$f(1) = 4.1945280$

Luego,

$$I \approx 0.5\left[\frac{1}{2} + 1.8591409 + \frac{4.1945280}{2}\right] = 2.2282025 = T_2$$

Combinamos ahora T_1 obtenida con paso $h_1 = 1$ y T_2 obtenida con paso $h_2 = \frac{1}{2}$ para obtener una estimación mejorada, que llamaremos S_2, utilizando la ecuación (57) adecuadamente:

$$I \approx 2.2282025 + \left[\frac{1}{2^2 - 1}\right]\left[2.2282025 - 2.5972640\right] = 2.1051820 = S_2$$

Este no es el valor exacto de la integral sino un valor mejorado.

Utilicemos ahora la regla trapezoidal de $2^2 = 4$ segmentos: $n = 4$ ($h = 0.25$) y los valores de la función son

$f(0) = 1$
$f(0.25) = 1.3243606$
$f(0.5) = 1.8591409$
$f(0.75) = 2.7408445$
$f(1) = 4.1945280$

Luego,

$$I \approx 0.25\left[\frac{1}{2} + 1.3243606 + 1.8591409 + 2.7408445 + \frac{4.1945280}{2}\right] = 2.1304025 = T_3$$

Combinamos ahora T_2 obtenida con paso $h_2 = \frac{1}{2}$ y T_3 obtenida con paso $h_3 = \frac{1}{2^2}$ para obtener una estimación mejorada, que llamaremos S_3, utilizando la ecuación (57) adecuadamente:

$$I \approx 2.1304025 + \left[\frac{1}{2^2 - 1}\right]\left[2.1304025 - 2.2282025\right] = 2.0978025 = S_3$$

Este no es el valor exacto de la integral sino un valor mejorado.

Tenemos así dos estimaciones mejoradas, S_2 y S_3, en base a tres estimaciones de reglas trapezoidales, T_1, T_2 y T_3. Estas dos estimaciones S_2 y S_3 pueden a la vez combinarse para obtener todavía una mejor estimación. Los valores S_2 y S_3 coinciden con los valores que se obtienen utilizando la fórmula de Simpson con pasos $h = \frac{1}{2}, \frac{1}{4},$ respectivamente, pues

$$I \approx \frac{0.5}{3}\left[1 + 4(1.8591409) + 4.1945280\right] = 2.105182 = S_2$$

$$I \approx \frac{0.25}{3}\left[1 + 4(1.3243606 + 2.7408445) + 2(1.8591409) + 4.1945280\right] = 2.0978025 = S_3$$

De acuerdo con esto y teniendo presente que el error en la regla de Simpson compuesta es proporcional a h^4, obtenemos a partir de la ecuación (57)

$$I \approx I(h_2) + \left[\frac{1}{(h_1/h_2)^4 - 1}\right]\left[I(h_2) - I(h_1)\right] \qquad (59)$$

(ecuación (35) obtenida anteriormente). En este caso particular, $I(h_1) = S_2$, $I(h_2)=S_3$, $h_1 = \frac{1}{2}$, $h_2 = \frac{1}{4} = \frac{h_1}{2}$, de donde, obtenemos la estimación mejorada

$$I \approx S_3 + \frac{1}{2^4 - 1}\left[S_3 - S_2\right] = 2.0978025 + \frac{1}{15}\left[2.0978025 - 2.1051820\right] = 2.0973105 = R_3$$

Este valor R_3 no es el valor exacto sino un valor mejorado.

Utilizamos ahora la regla trapezoidal de $2^3 = 8$ segmentos: $n = 8$ ($h = 0.125$) y los valores de la función son

$f(0) = 1$
$f(0.125) = 1.1420127$
$f(0.250) = 1.3243606$
$f(0.375) = 1.5585000$
$f(0.5) = 1.8591409$
$f(0.625) = 2.2451715$
$f(0.750) = 2.7408445$
$f(0.875) = 3.3773013$
$f(1) = 4.1945280$

Luego,

$$I \approx 0.125\left[\frac{1}{2} + 1.1420127 + 1.3243606 + 1.5585000 + 1.8591409 + 2.2451715 + \right.$$
$$\left. + 2.7408445 + 3.3773013 + \frac{4.1945280}{2}\right] = 2.1055744 = T_4$$

Combinamos ahora T_4 obtenida con paso $h_4 = \frac{1}{2^3}$ y T_3 obtenida con paso $h_3 = \frac{1}{2^2}$ para obtener una estimación mejorada, que llamaremos S_4, utilizando la ecuación (57) adecuadamente:

$$I \approx T_4 + \frac{1}{2^2 - 1}\left[T_4 - T_3\right] = 2.1055744 + \frac{1}{3}\left[2.1055744 - 2.1304025\right] = 2.0972984 = S_4$$

valor éste que no es el valor exacto de la integral sino un valor mejorado.

Combinamos ahora S_3 y S_4 para obtener una estimación mejorada, que llamaremos R_4, utilizando la ecuación (59) adecuadamente (se puede mostrar, al igual como se hizo con S_2 y S_3, que S_4 coincide con el valor que se obtiene usando la fórmula de Simpson con $h = 0.125$):

$$I \approx 2.0972984 + \frac{1}{15}\left[2.0972984 - 2.0978025\right] = 2.0972648 = R_4$$

valor éste que no es el valor exacto de la integral sino un valor mejorado.

Pero ahora tenemos dos estimaciones mejoradas, R_3 y R_4, que fueron obtenidas en base a tres estimaciones de reglas de Simpson, S_2, S_3 y S_4, por lo tanto, podemos combinar R_3 y R_4 para obtener una estimación mejor, a la que vamos a llamar W_4, y usaremos la ecuación

$$I \approx R_4 + \left[\frac{1}{2^6 - 1}\right]\left[R_4 - R_3\right] = 2.0972648 + \frac{1}{63}\left[2.0972648 - 2.0973105\right] = 2.0972641 = W_4$$

valor éste que no es el valor exacto de la integral sino uno mejorado.

En resumen, hasta el momento, en este caso particular, hemos calculado cuatro veces la fórmula del trapecio utilizando distintos números de segmentos:

T_1 con paso $h_1 = \frac{1}{2^0} = 1$

T_2 con paso $h_2 = \frac{1}{2^1} = 0.5$

T_3 con paso $h_3 = \frac{1}{2^2} = 0.25$

T_4 con paso $h_4 = \frac{1}{2^3} = 0.125$

De aquí, y recordando la regla compuesta del trapecio que se puede escribir como

$$\int_a^b f(x)dx = \text{trapecios} + h^2 C$$

obtenemos

$$I = T_1 + (1)^2 C$$
$$I = T_2 + (0.5)^2 C$$
$$I = T_3 + (0.25)^2 C$$
$$I = T_4 + (0.125)^2 C$$

A partir de estos T_i (i = 1, 2, 3, 4), realizando tres extrapolaciones obtenemos los S_i (i = 2, 3, 4), valores éstos que coinciden con los que se obtienen utilizando la fórmula de Simpson con pasos $\frac{1}{2}, \frac{1}{2^2}, \frac{1}{2^3}$, respectivamente; esto es,

S_2 con paso $h_2 = \frac{1}{2^1} = 0.5$

S_3 con paso $h_3 = \frac{1}{2^2} = 0.25$

S_4 con paso $h_4 = \frac{1}{2^3} = 0.125$

De acuerdo con esto podemos decir ahora que el valor exacto de la integral puede escribirse

$$I = S_2 + (0.5)^2 C_2$$
$$I = S_3 + (0.25)^2 C_2$$
$$I = S_4 + (0.125)^2 C_2$$

A partir de estos S_i (i = 2, 3, 4) obtuvimos los R_i (i = 3, 4) extrapolando dos veces.

Para obtener R_3 tuvimos que calcular T_3 o sea que usamos $h_3 = \frac{1}{2^2} = 0.25$, de modo que podemos escribir, de acuerdo con la ecuación (58)

$$I = R_3 + (0.25)^6 C_3 = R_3 + \left(\frac{1}{2^2}\right)^6 C_3$$

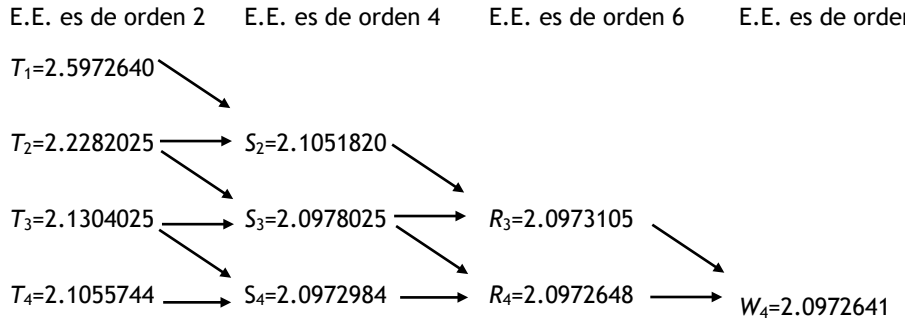
$$I = R_4 + (0.125)^6 C_3 = R_4 + \frac{1}{2^6}\left(\frac{1}{2^2}\right)^6 C_3$$

Multiplicando por 2^6 la segunda ecuación y restando la primera, se obtiene

$$I \approx \frac{2^6 R_4 - R_3}{2^6 - 1} = R_4 + \frac{1}{2^6 - 1}\left[R_4 - R_3\right]$$

ecuación ésta que llamamos W_4. Se puede probar que el error de esta nueva estimación es de orden 8.

La secuencia de estimaciones generadas usando este método en el ejemplo en cuestión puede presentarse en un esquema gráfico como se muestra a continuación (E.E.: error en la estimación)



Se observa que el valor de R_4 y W_4 coinciden hasta la sexta decimal y el proceso puede darse por finalizado.

5.2.6.1.1. Algoritmo de la integración de Romberg.

Este algoritmo está definido por las siguientes relaciones de recurrencia

$$(R)\begin{cases} T_{n,1} = T\left[\dfrac{(b-a)}{2^{n-1}}\right] \quad n = 1,\ 2,\ 3, \cdots \quad (a \text{ y } b \text{ son los límites de integración}) \\ T_{n,k} = T_{n,k-1} + \dfrac{T_{n,k-1} - T_{n-1,k-1}}{2^{2k-2} - 1} \quad k = 2,\ 3,\ \ldots,\ n \end{cases}$$

Estos números se disponen en un arreglo triangular, llamado *esquema de Romberg*, de la siguiente forma

$$\begin{array}{llll} T_{1,1} & & & \\ T_{2,1} & T_{2,2} & & \\ T_{3,1} & T_{3,2} & T_{3,3} & \\ T_{4,1} & T_{4,2} & T_{4,3} & T_{4,4} \end{array}$$

y así sucesivamente; pero nos preguntamos, ¿hasta cuándo?

Se requiere un criterio de terminación o de paro para valorar los resultados obtenidos. Al igual como se hizo en los procesos iterativos, se compara la nueva estimación con la anterior. Cuando el cambio en los valores anterior y actual, en módulo, está bajo un criterio de error preespecificado ε, los cálculos se terminan. Esto es, el valor $T_{n,k}$ es aceptado como una aproximación de I cuando $\left|T_{n,k-1} - T_{n-1,k-1}\right|$ no excede el error prefijado. Así, uno extrapola hasta que dos valores en la misma columna concuerden con la exactitud deseada.

La ventaja del procedimiento de Romberg radica en el hecho de que no hay que tomar ninguna decisión previa respecto al paso óptimo; el número de pasos es variable, depende si alcanzamos la aproximación deseada. (Recordemos que la integración de Romberg está diseñada para casos en que la función por integrar se conoce, ya que el conocimiento de la función permite las evaluaciones necesarias para las implementaciones iniciales de la regla trapezoidal. Los datos en forma tabular rara vez se encuentran en forma necesaria para llevar a cabo evaluaciones sucesivas).

Ejemplo 13. Utilizando el método de integración de Romberg, calculemos la $\int_0^{\pi/2} sen\, x\, dx$ con un error que no exceda a $1x\ 10^{-5}$.

Para lograr la exactitud deseada se ha tenido que utilizar la regla trapezoidal para $n = 1, 2, 4$ y 8 segmentos, siendo entonces

$$h_1 = \frac{\pi}{2}, \quad h_2 = \frac{\pi}{2^2}, \quad h_3 = \frac{\pi}{2^3}, \quad h_4 = \frac{\pi}{2^4}$$

respectivamente.

La tabla siguiente sólo contiene algunos de los valores del integrando para estos h_i ($i = 1, 2, 3, 4$):

x	$sen\ x$
0	0
⋮	⋮
$\frac{\pi}{8}$	0.38268383
⋮	⋮
$\frac{\pi}{4}$	0.70710678
⋮	⋮
$\frac{3\pi}{8}$	0.92387953
⋮	⋮
$\frac{\pi}{2}$	1

Luego, utilizando las relaciones de recurrencia (R) y disponiendo los cálculos según el esquema de Romberg, se tiene

0.785398			
0.948059	1.002279		
0.987116	1.000135	0.999992	
0.996785	1.000008	1.000000	1.000000

Como

$$\left|T_{4,3} - T_{3,3}\right| = \left|1.000000 - 0.999992\right| = 8\text{x}10^{-6} < 1\text{x}10^{-5}$$

entonces, aceptamos a $T_{4,4} = 1.000000$ como una aproximación de

$$I = \int_0^{\pi/2} sen\ x\ dx$$

esto es, $I \approx 1.000000$ y el proceso ha finalizado.

5.2.6.2. Integración o cuadratura Gaussiana.

Hasta antes de integración de Romberg, hemos analizado un conjunto de fórmulas de integración numérica conocidas como las ecuaciones de Newton - Cotes. Una característica de estas fórmulas (excepto el caso

especial visto en el párrafo 5.2.3, punto 3), es que la estimación de la integral se basa en puntos igualmente espaciados. Por consiguiente, la posición de los puntos base usados en estas ecuaciones está predeterminada o fija.

Por ejemplo, como se puede ver en la Figura 10 (a), la base de la regla trapezoidal es tomar el área bajo la línea recta que une los valores de la función evaluada en los extremos del intervalo. La fórmula usada para calcular este área es

$$I \approx (b-a)\frac{f(a)+f(b)}{2}$$

donde a y b son los límites de integración y (b - a) es el ancho del intervalo de integración. Debido a que la regla trapezoidal debe pasar a través de los puntos límites, existen casos como en la Figura 10 (a) en donde la fórmula genera un error muy grande.

Ahora supongamos que la situación de fijar los puntos base se elimina y se va a evaluar libremente el área bajo la línea recta que une dos puntos cualesquiera de la curva. Colocando estos puntos de manera inteligente, se puede definir una línea recta que balancee los errores positivos y negativos. De ahí que, como en la Figura 10 (b), se llegará a un valor más exacto de la integral.

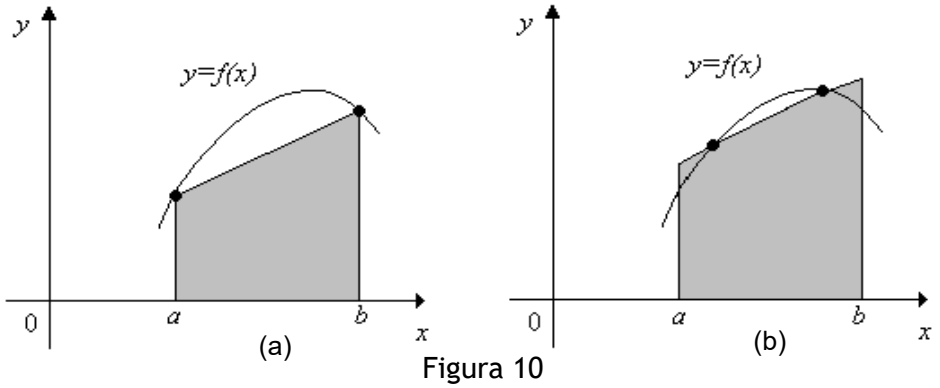


Figura 10

La *cuadratura Gaussiana* es el nombre de uno de estos métodos que implementa esta estrategia. Las fórmulas particulares de cuadratura Gaussiana que describiremos se llaman *fórmulas de Gauss - Legendre*. Estas fórmulas utilizan puntos de Legendre (raíces de polinomios de Legendre) y no se pueden utilizar para integrar una función dada en forma de tabla con intervalos de separación uniforme, debido a que los puntos de Legendre no están separados de esa manera; sin embargo, son más adecuadas para integrar funciones analíticas. La ventaja de estas fórmulas es que su precisión es mayor que la de las fórmulas de Newton - Cotes.

Antes de analizar la expresión general de las fórmulas de Gauss - Legendre (o simplemente Gauss) revisemos los términos del error en las fórmulas de Newton - Cotes. La ecuación que nos da el error de la regla del trapecio nos indica que dicho error es proporcional a f''. Si se usa la regla del trapecio para integrar cada una de las funciones $f = 1, x, x^2, x^3, \ldots$, entonces los resultados serán exactos para $f = 1$ y $f = x$, pero existirán errores para x^2 y las potencias superiores de x. La ecuación que nos da el error de la regla de Simpson nos indica que dicho error es proporcional a f^{iv}, por lo que es exacta si integramos $f = 1, x, x^2$ y x^3. En términos más generales, la fórmula cerrada de Newton - Cotes de orden impar n es exacta si el polinomio tiene orden menor o igual que n, pero cuando n es par la fórmula resulta exacta cuando el integrando es un polinomio de orden menor o igual que $n + 1$.

Sin embargo, el siguiente ejemplo muestra que una integración numérica con dos puntos se puede hacer exacta para el caso de los polinomios de orden tres si se optimizan los valores x de los datos.

Ejemplo 14. La fórmula de integración con dos puntos se puede hacer exacta cuando se integra un polinomio de orden tres. Queremos determinar estos puntos.

Consideremos

$$I = \int_{-1}^{1} f(x)dx \tag{60}$$

y escribamos una fórmula de integración con dos puntos como sigue

$$I = w_1 f(x_1) + w_2 f(x_2) + E \tag{61}$$

donde w_k, $k = 1, 2$ son los coeficientes incógnitas (pesos), x_k, $k = 1, 2$ son puntos indeterminados y E es el término error.

Ya que w_k y x_k son ambos indeterminados, requerimos que $E = 0$ (de modo que I es exacto) para $f(x) = 1, x, x^2, x^3$. Introduciendo cada uno de los $f(x) = 1, x, x^2, x^3$ en la ecuación (61) obtenemos cuatro ecuaciones por resolver:

$$w_1 + w_2 = \int_{-1}^{1} 1\, dx = 2$$

$$w_1 x_1 + w_2 x_2 = \int_{-1}^{1} x\, dx = 0 \tag{62}$$

$$w_1 x_1^2 + w_2 x_2^2 = \int_{-1}^{1} x^2\,dx = \frac{2}{3}$$

$$w_1 x_1^3 + w_2 x_2^3 = \int_{-1}^{1} x^3\,dx = 0$$

donde en el lado derecho aparecen los valores exactos.

Las ecuaciones (62) se resuelven simultáneamente, dando

$$w_1 = w_2 = 1$$

$$x_1 = \frac{-1}{\sqrt{3}} = \text{-0.577350269...}$$

$$x_2 = \frac{1}{\sqrt{3}} = 0.577350269...$$

Con estos pesos y puntos la ecuación (62) es exacta para un polinomio de grado menor o igual a tres, y reemplazándolos en dicha ecuación se obtiene

$$I \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

llamada *fórmula de Gauss - Legendre de dos puntos*.

Nótese que los límites de integración usados en estas ecuaciones van desde –1 a 1. Esto se hizo para simplificar la aritmética y hacer la formulación tan general como sea posible. Un simple cambio de la variable se puede usar para trasladar otros límites de integración en esta forma.

La fórmula de integración que se deriva del ejemplo anterior es el miembro más simple de la cuadratura de Gauss y el método utilizado para obtenerla se denomina *método de los coeficientes indeterminados*. No es sencillo derivar las fórmulas de Gauss - Legendre con más de dos puntos por medio de la extensión del enfoque de este ejemplo. Por lo tanto, en lo que resta de esta exposición se da una fórmula general de Gauss - Legendre. Se puede demostrar que dicha fórmula de orden n es exacta cuando se integra un polinomio de orden menor o igual a $2n$ - 1.

La fórmula de Gauss - Legendre difiere en forma significativa de la de Newton - Cotes, ya que los n puntos se obtienen mediante las raíces del polinomio de Legendre $P_n(x) = 0$, donde $P_n(x)$ es el polinomio de Legendre de orden n, $P_n(x) = \frac{1}{2^n n!}\frac{d^n (x^2-1)^n}{dx^n}$, con $P_0(x) = 1$, fórmula

que se debe a Olindo Rodríguez y que permite generar los polinomios de Legendre. Sin duda alguna, x_1 y x_2 determinados en el ejemplo anterior son raíces de $P_2(x) = 0$.

La *cuadratura de Gauss - Legendre* o *fórmula de Gauss - Legendre* que se extiende sobre el intervalo [-1, 1] está dada por

$$\int_{-1}^{1} f(x)dx \approx \sum_{k=1}^{n} w_k f(x_k) \tag{63}$$

donde n es el número de puntos, w_k son los pesos y x_k son los puntos o abscisas, que son las raíces del polinomio de Legendre $P_n(x)$ dadas en la siguiente tabla

Tabla 2

N	Grado del polinomio para el cual (63) es exacta	Raíces del polinomio de Legendre: x_k	Pesos o coeficientes Gaussianos: w_k
2	3	$\frac{-1}{\sqrt{3}}$ = -0.5773502692 $\frac{1}{\sqrt{3}}$ = 0.5773502692	1. 1.
3	5	$-\sqrt{\frac{3}{5}}$ = -0.7745966642 0 $\sqrt{\frac{3}{5}}$ = 0.7745966642	$\frac{5}{9}$ = 0.5555555556 $\frac{8}{9}$ = 0.8888888889 $\frac{5}{9}$ = 0.5555555556
4	7	-0.8611363116 -0.3399810436 0.3399810436 0.8611363116	0.3478548451 0.6521451549 0.6521451549 0.3478548451
5	9	-0.9061798459 -0.5384693101 0 0.5384693101 0.9061798459	0.2369268851 0.4786286705 0.5688888889 0.4786286705 0.2369268851
6	11	-0.9324695142 -0.6612093865 -0.2386191861 0.2386191861 0.6612093865 0.9324695142	0.1713244924 0.3607615730 0.4679139346 0.4679139346 0.3607615730 0.1713244924

La suma de los coeficientes Gaussianos siempre da dos. Se observa además que para cada k, se tiene $-x_k = x_{n-k+1}$, $w_k = w_{n-k+1}$ y que las abscisas centrales tienen mayor peso. Además, la función no se calcula en los extremos del intervalo. Se habla, por ejemplo, de fórmula de Gauss por 4 puntos, por 5 puntos, y así siguiendo.

La fórmula de Gauss - Legendre se basa, entonces, en la integración de un polinomio ajustado a los puntos dados por las raíces de un polinomio de Legendre. El orden de exactitud de esta fórmula es aproximadamente el doble del de la fórmula cerrada de Newton - Cotes al utilizar el mismo número de puntos. Debido a que los coeficientes o pesos son positivos y no se realizan restas de números grandes como en el caso de la fórmula de Newton - Cotes, no existen errores graves de redondeo, en general. Como la fórmula de Gauss - Legendre requiere de evaluaciones de la función en puntos que no están uniformemente espaciados dentro del intervalo de integración, entonces dicha fórmula no es aplicable a los casos en que la función se desconoce.

El error de la fórmula de Gauss - Legendre se especifica generalmente mediante

$$E = \frac{2^{2n+3}\left[(n+1)!\right]^4}{(2n+3)\left[(2n+2)!\right]^3} f^{(2n+2)}(\xi) \tag{64}$$

donde n es el número de puntos menos uno y $f^{(2n+2)}(\xi)$ es la $(2n+2)$ - ésima derivada de la función después del cambio de variable y ξ se localiza dentro del intervalo de -1 a 1.

La fórmula de integración de Gauss - Legendre puede aplicarse a cualquier intervalo arbitrario $[a, b]$ con la transformación

$$x = \left[z - \frac{(a+b)}{2} \right] \frac{2}{(b-a)} \tag{65}$$

donde z es la coordenada original en $a \leq z \leq b$ y x es la coordenada normalizada en $-1 \leq x \leq 1$. La transformación de x en z es

$$z = \frac{a+b}{2} + \frac{b-a}{2} x \tag{66}$$

Por medio de esta transformación, la integral se puede escribir como

$$\int_a^b f(z)dz = \int_{-1}^{1} f(z)\left(\frac{dz}{dx}\right)dx \approx \frac{b-a}{2}\sum_{k=1}^{n} w_k f(z_k) \qquad (67)$$

donde $\frac{dz}{dx} = \frac{b-a}{2}$. Los valores z_k se obtienen al sustituir x en la ecuación (66) por las abscisas o raíces del polinomio de Legendre; a saber

$$z_k = \frac{a+b}{2} + \frac{b-a}{2}x_k \qquad (68)$$

Ejemplo 15. Supongamos que $n = 2$, $a = 0$ y $b = 2$. Puesto que las abscisas x_k para $n = 2$ en la coordenada normalizada x, $-1 \leq x \leq 1$, son ± 0.57735 (Tabla 2), los puntos correspondientes en z son

$$z_1 = 1 + (-0.57735) = 0.42265$$
$$z_2 = 1 + (0.57735) = 1.57735$$

La derivada es $\frac{dz}{dx} = \frac{b-a}{2} = 1$. Por lo tanto, la fórmula de Gauss - Legendre se escribe como

$$\int_0^2 f(z)dz = \int_{-1}^{1} f(z)\left(\frac{dz}{dx}\right)dx \approx (1)\left[(1)f(0.42265) + (1)f(1.57735)\right]$$

Ejemplo 16. Calculemos con la fórmula de Gauss por 3 puntos la $\int_{-1}^{1} \sqrt{2+x}\,dx$

Usando la fórmula (63) y los datos tabulados para $n = 3$ obtenemos reteniendo 6 decimales (en los datos y en los cálculos)

$$\int_{-1}^{1} \sqrt{2+x}\,dx \approx 0.555556(1.106979) + 0.888889(1.414214) + 0.555556(1.665712) = 2.797464$$

5.2.6.2.1. Comentarios adicionales.

Es frecuente encontrar en los textos como fórmula de integración de Gauss en lugar de la ecuación (63) la ecuación

$$\int_0^1 f(x)dx \approx \sum_{k=1}^{n} p_k f(x_k) \tag{69}$$

en donde ahora las abscisas x_k están dentro del intervalo [0, 1], los p_k son los pesos o coeficientes Gaussianos y esta ecuación es exacta cuando se integra un polinomio de orden menor o igual a $2n - 1$ (n es el número de puntos).

La fórmula (69) puede aplicarse a cualquier intervalo arbitrario [a, b]. En efecto, si x es la coordenada en $0 \le x \le 1$ y $a \le z \le b$, la transformación de x en z es

$$z = a + (b - a)x \tag{70}$$

Por medio de esta transformación, la integral se puede escribir como

$$\int_a^b f(z)dz = \int_0^1 f(z)\left(\frac{dz}{dx}\right)dx \approx (b-a)\sum_{k=1}^{n} p_k f(z_k) \tag{71}$$

donde $\frac{dz}{dx} = b - a$. Los valores z_k se obtienen al sustituir x en la ecuación (70) por las abscisas o raíces del polinomio de Legendre; a saber

$$z_k = a + (b - a)x_k \tag{72}$$

Damos a continuación la tabla de las raíces x_k del polinomio de Legendre $P_n(x)$, pero ahora referidas al intervalo [0, 1], y los pesos correspondientes:

Tabla 3

N	Grado del polinomio para el cual (69) es exacta	Raíces del polinomio de Legendre: x_k	Pesos o coeficientes Gaussianos: p_k
2	3	0.2113248654 0.7886751346	0.5 0.5
3	5	0.1127016654 0.5 0.8872983346	0.2777777778 = 5/18 0.4444444444 = 4/9 0.2777777778 = 5/18
4	7	0.0694318442 0.3300094782 0.6699905218 0.9305681558	0.1739274226 0.3260725774 0.3260725774 0.1739274226
5	9	0.0469100770 0.2307653449 0.5 0.7692346551 0.9530899230	0.1184634425 0.2393143352 0.2844444444 0.2393143352 0.1184634425

6	11	0.0337652429 0.1693953068 0.3806904070 0.6193095930 0.8306046932 0.9662347571	0.0856622462 0.1803807865 0.2339569673 0.2339569673 0.1803807865 0.0856622462
7	13	0.0254460438 0.1292344072 0.2970774243 0.5 0.7029225757 0.8707655928 0.9745539562	0.0647424831 0.1398526957 0.1909150253 0.2089795918 0.1909150253 0.1398526957 0.0647424831

(Observemos que los pesos son la mitad de los dados en la Tabla 2 y las raíces son $x_k = \frac{1}{2}(t_k + 1)$, donde t_k son las raíces de la Tabla 2).

Ejemplo 17. Supongamos que $n = 2$, $a = 0$ y $b = 2$. Puesto que las abscisas para $n = 2$ en la coordenada x, $0 \le x \le 1$, son (según la Tabla 3) 0.21132 y 0.78868, los puntos correspondientes en z son, usando la ecuación (72)

$$z_1 = 0 + 2(0.21132) = 0.42264$$
$$z_2 = 0 + 2(0.78868) = 1.57736$$

La derivada es $\frac{dz}{dx} = b - a = 2$. Por lo tanto, la fórmula de Gauss - Legendre se escribe como

$$\int_0^2 f(z)dz = \int_0^1 f(z)\left(\frac{dz}{dx}\right)dx \approx (2)\left[(0.5)f(0.42264) + (0.5)f(1.57736)\right]$$

El error en la fórmula de Gauss - Legendre se especifica generalmente como

$$E = \frac{\left[(n+1)!\right]^4}{(2n+3)\left[(2n+2)!\right]^3} f^{(2n+2)}(\xi) \qquad (73)$$

donde n es el número de puntos menos uno y $f^{(2n+2)}(\xi)$ es la $(2n+2)$–ésima derivada de la función después del cambio de variable y ξ está dentro del [0, 1].

Ejemplo 18. Calculemos con la fórmula de Gauss por 5 puntos la $\int_0^1 \frac{sen\, x}{x} dx$

Usando la fórmula (69) y los datos de la Tabla 3 para $n = 5$, obtenemos

$$\int_0^1 \frac{sen\,x}{x}dx \approx 0.1184634425\frac{sen(0.0469100770)}{0.0469100770} + 0.23931433525\frac{sen(0.2307653449)}{0.2307653449} +$$

$$+ 0.2844444444\frac{sen(0.5)}{0.5} + 0.23931433525\frac{sen(0.7692346551)}{0.7692346551} +$$

$$+ 0.1184634425\frac{sen(0.9530899230)}{0.9530899230} = 0.946083073$$

Ejemplo 19. Calculemos con la fórmula de Gauss por 3 puntos la $\int_{-1}^{1}\sqrt{2+z}dz$ con los datos de la Tabla 3 reteniendo 6 decimales.

Aquí es $a = -1$, $b = 1$ y el cambio de variable según (70) es

$z = a + (b - a)x$, de donde, $z = -1 + 2x$ y la derivada es $\frac{dz}{dx} = b - a = 2.$

Luego,

$$\int_{-1}^{1}\sqrt{2+z}\,dz = 2\int_0^1\sqrt{1+2x}\,dx \approx 2\sum_{k=1}^{3}p_k\sqrt{1+2x_k} =$$

$$= 2\left[0.277778\sqrt{1+2(0.112702)} + 0.444444\sqrt{1+2(0.5)} + 0.277778\sqrt{1+2(0.887298)}\right] =$$

$$= 2\left[0.307495 + 0.628539 + 0.462698\right] = 2.797464$$

resultado éste que coincide (como era de esperarse) con el que se obtuvo cuando se evaluó la misma integral, pero según los datos de la Tabla 2.

Análogamente se puede calcular la integral dada en el ejemplo 18 realizando la transformación para referir la integral al intervalo [-1, 1] y usar luego los datos de la Tabla 2.

5.2.7. Elementos de juicio.

El siguiente cuadro muestra un resumen de los elementos de juicio relacionados con la integración numérica o cuadratura.

Comparación de los diferentes métodos de integración numérica. Las comparaciones se basan en la experiencia general y no toman en consideración el comportamiento de las funciones especiales

Método	Puntos necesarios para su aplicación	Puntos necesarios para *n* aplicaciones	Error de truncamiento	Aplicación	Trabajo de programación	Conclusiones
Regla del trapecio	2	$n + 1$	$\approx h^3 f^{(2)}(\quad)$	Amplia	Fácil	Cerrado. Sencillo. Necesita un gran número de sub-intervalos para una buena precisión.
Regla de Simpson	3	$2n + 1$	$\approx h^5 f^{(4)}(\quad)$	Amplia	Fácil	Cerrado. Sencillo. Más precisión que la regla del trapecio. Sólo con un número par de intervalos.
Regla del punto medio	2	$n + 1$	$\approx h^3 f^{(2)}(\quad)$	Amplia	Fácil	Abierto. Sencillo. El mismo orden de precisión que la regla del trapecio.
Integración de Romberg	3	Requiere conocer $f(x)$		Requiere conocer $f(x)$	Moderado	Inapropiado para datos tabulados. Más precisión que las fórmulas anteriores.
Cuadratura Gaussiana	≥ 2	Moderado		Requiere conocer $f(x)$	Fácil	Inapropiado para datos tabulados. Buena precisión. No se utilizan los valores de la función en los extremos. Los puntos no están separados uniformemente.

5.2.8. Algoritmo de la regla del trapecio compuesta. Pseudocódigo.

Para aproximar la integral $I = \int_a^b f(x)dx$:

ENTRADA los puntos extremos a, b; entero positivo n; la función $f(x)$ o los valores de las ordenadas $f(a), f(x_1), ..., f(x_{n-1}), f(b)$ (lo que conozca).
SALIDA aproximación XI de I.
Paso 1 Tomar $h = (b - a) / n$.
Paso 2 Tomar $XI0 = (f(a) + f(b)) / 2$;
$XI1 = 0$. (Suma de $f(x_i)$, $i = 1, 2, ..., n - 1$).
Paso 3 Para $i = 1, 2, ..., n - 1$ seguir los Pasos 4 y 5.
Paso 4 Tomar $X_i = a + ih$.
Paso 5 Tomar $XI1 = XI1 + f(X_i)$.
Paso 6 Tomar $XI = h\ (XI0 + XI1)$.
Paso 7 SALIDA (XI);
PARAR.

EJERCICIOS PROPUESTOS .....

1. Calcular la derivada primera de la función $f(x) = e^{x/2}$ en los puntos x = 0.7, 1.2 y estimar el error cometido cuando se conocen los siguientes valores:

x	f(x)	$f(x_i x_{i+1})$	$f(x_i x_{i+1} x_{i+2})$	$f(x_i x_{i+1} x_{i+2} x_{i+3})$	$f(x_i x_{i+1} x_{i+2} x_{i+3} x_{i+4})$
0.3	1.1618342				
		0.5956860			
0.4	1.2214028		0.1553133		
		0.6422800		0.0267750	
0.6	1.3498588		0.1660233		0.0034807
		0.6920870		0.0288634	
0.7	1.4190675		0.1804550		0.0037560
		0.7462235		0.0314926	
0.9	1.5683122		0.1962013		0.0041080
		0.8247040		0.0339574	
1.1	1.7332530		0.2131800		
		0.8886580			
1.2	1.8221188				

2. a) Derivando la fórmula de interpolación de Newton hacia adelante:

i) Calcular hasta la derivada de orden 3 de $f(x)$ en el punto $x = 1$ a partir de la siguiente tabla:

x	$f(x)$	Δf	$\Delta^2 f$	$\Delta^3 f$
1.00	1.00000			
		0.02470		
1.05	1.02470		-0.00059	
		0.02411		0.00005
1.10	1.04881		-0.00054	
		0.02357		0.00004
1.15	1.07238		-0.00050	
		0.02307		0.00002
1.20	1.09544		-0.00048	
		0.02259		0.00003
1.25	1.11803		-0.00045	
		0.02214		
1.30	1.14017			

ii) Calcular la derivada primera de $f(x) = log_{10} x$ en el punto $x = 50$ y estimar el error cometido a partir de la siguiente tabla:

x	f(x)	Δf	Δ^2f	Δ^3f
50	1.6990			
		0.0414		
55	1.7404		-0.0036	
		0.0378		0.0005
60	1.7782		-0.0031	
		0.0347		
65	1.8129			

iii) La siguiente tabla contiene los tiempos t y la velocidad v correspondientes de un móvil. Calcular la aceleración, derivada de la velocidad, en los instantes 0, 60 y 120 segundos:

t	0	60	120	180	240
v	0.0000	0.0824	0.2747	0.6502	1.3851

b) Derivando la fórmula de interpolación de Newton hacia atrás:

i) Hacer lo mismo que en el apartado a) i), pero en el punto $x = 1.30$.

ii) Hacer lo mismo que en el apartado a) ii), pero en el punto $x = 65$.

iii) Hacer lo mismo que en el apartado a) iii), pero en los instantes 120, 180 y 240 segundos.

3. Utilizando la fórmula del trapecio compuesta estimar:

a) $\int_{1.0}^{1.4} f(x)dx$, conociendo la siguiente tabla de valores:

x	1.0	1.1	1.2	1.3	1.4
f(x)	0.010	0.252	0.586	1.024	1.578

b) $\int_{1.0}^{2.0} \sqrt{x}\ dx$, dividiendo el intervalo (1, 2) en 10 partes iguales. Estimar el error cometido. Trabajar con tres cifras decimales redondeadas.

c) $\int_0^1 \frac{dx}{1+x^2}$, eligiendo h = 0.25, 0.125. Realizar la extrapolación correspondiente. Trabajar con siete cifras decimales redondeadas.

4. Hacer el programa correspondiente a la fórmula del trapecio compuesta y resolver:

a) El ejercicio 3 b) y c).

b) (i) $\int_0^5 x\sqrt{25-x^2}\ dx$, $n = 6$

(ii) $\int_0^{1/2} cos^2\ x\ dx$, $n = 6$

(iii) $\int_0^1 (1+x)^x dx$, $n = 40$

(iv) $\int_0^1 cosh(x^2)dx$, $n = 100$

(Con n se indica el número de subintervalos en que se divide el intervalo dado).

5. Utilizando la fórmula de Simpson compuesta estimar:

a) La integral dada en el ejercicio 3 a).

b) $\int_0^1 cos\sqrt{x}\ dx$, dividiendo el intervalo de integración en cuatro subintervalos y utilizando en los cálculos tres cifras decimales redondeadas. Estimar el error cometido.

c) $\int_0^1 e^{-x^2} dx$, eligiendo h = 0.25, 0.125. Realizar la extrapolación correspondiente.

6. Hacer el programa correspondiente a la fórmula de Simpson compuesta y resolver:

a) El ejercicio 5 b) y c).
b) El ejercicio 4 b).

7. Utilizando la fórmula de Euler - Maclaurin estimar:

a) $\int_2^{10} ln\ x\ dx$, $n = 8$

b) $\int_{0.2}^1 (sen\ x - lnx + e^x)\ dx$, $n = 8$

Retener en todos los cálculos a lo sumo 5 cifras decimales redondeadas.

c) El valor de π siendo $\frac{\pi}{4} = \int_0^1 \frac{dx}{1+x^2}$, $h = 0.25$

Retener en todos los cálculos a lo sumo 7 cifras decimales redondeadas.

8. Utilizando la fórmula de Gregory estimar:

a) $\int_0^{0.5} \frac{dx}{\sqrt{1-x^2}}$, $n = 5$, conociendo la siguiente tabla de diferencias:

x	f(x)	Δ f	Δ^2f	Δ^3f	Δ^4f
0	1.0000000				
		0.00503781			
0.1	1.00503781		0.01054510		
		0.001558291		0.00153611	
0.2	1.02062072		0.01208121		0.00152317
		0.002766412		0.00305928	
0.3	1.04828484		0.01514049		0.00260671
		0.004280461		0.00566599	
0.4	1.09108945		0.02080648		
		0.006361109			
0.5	1.15470074				

b) $\int_0^{\pi/2} sen\, x\, dx, \quad n = 6$, conociendo la siguiente tabla de diferencias:

x	sen x	Δ	Δ^2	Δ^3	Δ^4	Δ^5	Δ^6
0	0.00000						
		0.25882					
π/12	0.25882		-0.01764				
		0.24118		-0.01644			
2π/12	0.50000		-0.03407		0.00233		
		0.20711		-0.01411		0.00095	
3π/12	0.70711		-0.04819		0.00328		-0.0002
		0.15892		-0.01083		0.00074	
4π/12	0.86603		-0.05902		0.00402		1
		0.09990		-0.00681			
5π/12	0.96593		-0.06583				
		0.03407					
6π/12	1.00000						

9. Utilizando la regla compuesta del punto medio con $n + 2$ subintervalos, estimar las siguientes integrales:

a) $\int_0^2 \frac{2}{x^2+4}\, dx, \qquad n = 6$

b) $\int_3^5 \frac{1}{\sqrt{x^2-4}}\, dx, \qquad n = 8$

c) $\int_0^2 e^{2x}\, sen\, 3x\, dx, \quad n = 8$

10. Utilizando el método de Romberg estimar:

a) $\int_1^3 \frac{dx}{x}$, $\varepsilon \leq 1\text{x}10^{-3}$. Trabajar con redondeo a 6 cifras decimales.

b) $\int_{1.0}^{1.5} e^{-x^2}\ dx$, $\varepsilon \leq 0.5\text{x}10^{-4}$. Trabajar con redondeo a 5 cifras decimales.

c) $\int_0^{\pi/2} sen\ x\ dx$, $\varepsilon \leq 5\text{x}10^{-5}$. Trabajar con redondeo a 6 cifras decimales.

11. Hacer el programa correspondiente al método de Romberg y resolver:

a) El ejercicio 10.
b) Para $\varepsilon \leq 5\text{x}10^{-6}$

(i) $\int_0^1 (1+x)^x\ dx$ (ii) $\int_0^1 cosh\ x\ e^x\ dx$

12. Utilizando la fórmula de Gauss estimar:

a) $\int_0^1 \sqrt{1+2x}\ dx$, por 3 puntos. Retener en los cálculos 5 cifras decimales redondeadas.

b) $\int_{-1}^1 \sqrt{2+x}\ dx$, por 3 puntos. Retener en los cálculos 6 cifras decimales redondeadas.

c) $\int_0^{\pi/2} sen\ t\ dt$, por 2 y 4 puntos. Retener en los cálculos 5 cifras decimales redondeadas.

d) $\frac{2}{\sqrt{\pi}} \int_0^4 e^{-t^2}\ dt$, por 4 y 6 puntos. Retener en los cálculos 6 decimales redondeados.

13. Hacer el programa correspondiente a la fórmula de Gauss y resolver:

a) El ejercicio 12.

b) Considerando 7 puntos

(i) $\int_0^1 \frac{cosx}{\sqrt{x}}\,dx$ (ii) $\int_0^1 \frac{ln(1+x)}{1+x^2}\,dx$

14. El modelo de Debye para calcular la capacidad calórica de un sólido considera la función

$$\Phi(x) = \int_0^x \frac{t^3}{e^t - 1} dt$$

Utilizar los métodos numéricos de integración (que se puedan aplicar) para calcular valores de $\Phi(x)$ para $x = 1$, 5 y 10.

15. El área de la superficie del sólido de revolución que se obtiene al girar alrededor del eje 0X la región limitada por la curva $y = f(x)$, siendo $a \leq x \leq b$ viene dada por

$$\text{Área} = 2\pi \int_a^b f(x)\sqrt{1 + (f'(x))^2}\,dx$$

Para $f(x) = e^{-x}$, $0 \leq x \leq 1$, calcular el área usando:

a) La regla del punto medio para 2, 4, 6, 8, 20 y 100 segmentos.
b) La regla del trapecio para 2, 4, 6, 8, 20 y 100 segmentos.
c) La regla de Simpson para 2, 4, 6, 8 y 20 segmentos.
d) La regla de Romberg con una cota de error de $5x10^{-7}$.
e) La integración de Gauss-Legendre para 2, 4 y 6 segmentos.
f) A partir de los resultados obtenidos, expresar sus conclusiones.

16. El valor promedio de una corriente eléctrica oscilante durante un período puede ser cero. Por ejemplo, supóngase que la corriente se describe mediante una senoidal simple:

$$i(t) = sen\,(2\pi\, t/T)$$

donde T es el período. El valor promedio de esta función se puede determinar mediante la siguiente ecuación

$$i = \frac{\int_0^T sen\left(\frac{2\pi t}{T}\right)dt}{T-0} = \frac{-\cos 2\pi + \cos 0}{T} = 0$$

A pesar de que la corriente es igual a cero, esta corriente es capaz de realizar trabajo y generar calor. Por lo tanto, los ingenieros eléctricos, a menudo, caracterizan esta corriente mediante

$$I_{RMS} = \sqrt{\frac{\int_0^T i^2(t)dt}{T}}$$

en donde I_{RMS} se conoce como corriente RMS (raíz de la corriente media al cuadrado). Para evitar este resultado nulo, la corriente se eleva al cuadrado antes de calcular el promedio.

En este caso, supóngase que la corriente en un circuito es de

$$i(t) = 10\, e^{-t/T}\, sen\left(\frac{2\pi t}{T}\right) \quad \text{para} \quad 0 \le t \le T/2$$

$$i(t) = 0 \quad \text{para} \quad T/2 \le t \le T$$

Calcular la corriente RMS de la forma de onda mostrada en la Figura 1 usando la regla del trapecio, la regla de Simpson, la integración de Romberg y la cuadratura gaussiana para T = 1s y hasta alcanzar el valor verdadero de 15.4126081. A partir de los resultados obtenidos, expresar sus conclusiones.

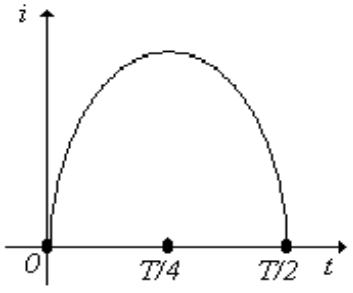


Figura 1

17. Se construye una hoja acanalada para techado, usando una máquina que comprime una hoja plana, de aluminio, y la transforma en una hoja cuya sección transversal tiene la forma de onda de la función seno (ver Figura 2).

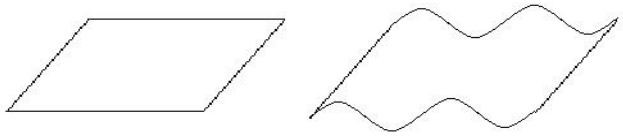


Figura 2

Se necesita una hoja corrugada de 4 pies de largo, cuyas ondas tienen una altura de 1 pulg, desde la línea central, y cada onda tiene aproximadamente una período de 2 π pulg.

El problema de calcular la longitud de la primera hoja plana consiste en determinar la longitud de la onda dada por $f(x) = sen\ x$ de $x = 0$ pulg a $x = 48$ pulg.

Por el cálculo, sabemos que esta longitud es

$$L = \int_0^{48} \sqrt{1 + (f'(x))^2}\, dx$$

Determinar la longitud de la onda usando:

a) La regla del punto medio para 2, 10, 40 y 100 segmentos.
b) La regla del trapecio para 2, 10, 40 y 100 segmentos.
c) La regla de Simpson para 2, 10, 40 y 100 segmentos.
d) La integración de Romberg con una cota de error de $5x10^{-5}$.
e) A partir de los resultados obtenidos, expresar sus conclusiones.

18. Un problema que se encuentra a menudo es determinar la cantidad de calor necesaria para elevar la temperatura de un material. La característica necesaria para realizar este cálculo es la capacidad calorífica *c*. Este parámetro representa la cantidad de calor necesaria para elevar una unidad de masa a una unidad de temperatura. Si *c* es la constante sobre el rango de temperaturas que se van a examinar, el calor necesario ΔH (en calorías) se calcula como

$$\Delta H = mc\ \Delta T$$

donde *c* tiene unidades de calorías por gramo por grado centígrado, *m* es la masa (en gramos) y ΔT es el cambio de temperatura (en grados centígrados). Por ejemplo, la cantidad de calor necesaria para elevar 20 g de agua de 5 a 10° C es igual a

$$\Delta H = (20)1(10\text{-}5) = 100 \text{ cal}$$

donde la capacidad calorífica del agua es aproximadamente 1 cal/g/°C. Tal valor es adecuado cuando ΔT es pequeña. Sin embargo, en rangos mayores de temperatura, la capacidad calorífica no es constante, y de hecho varía en función de la temperatura. Por ejemplo, la capacidad calorífica de un material aumenta con la temperatura de acuerdo a relaciones tales como

$$c(T) = 0.132 + 1.56 \text{ x } 10^{-4}\ T + 2.64 \text{ x } 10^{-7}\ T^2$$

En este caso se pide calcular el calor necesario para elevar 1000 g de este material de –100 a 200° C.

Una manera de calcular el valor promedio de $c(T)$ es mediante la fórmula:

$$\bar{c}(T) = \int_{T_1}^{T_2} \frac{c(T)}{T_2 - T_1} dT$$

En este caso, $c(T)$ es una cuadrática simple y, por lo tanto, ΔH se puede determinar analíticamente. Se debe:

a) Determinar la expresión analítica de ΔH y obtener el valor numérico exacto de ΔH.

b) Comparar los resultados obtenidos en el apartado a) con:
   i) La regla del punto medio para 2, 4, 30, 60, 300 y 6000 segmentos.
   ii) La regla trapezoidal para 2, 4, 30, 60, 300 y 6000 segmentos.
   iii) La regla de Simpson para 2 y 4 segmentos.
   iv) La integración de Romberg con una cota de error de $\varepsilon = 5 \text{ x } 10^{-2}$.
   v) A partir de los resultados obtenidos, expresar sus conclusiones.

……

Apéndice: Octave

A.1. Introducción.

A.1.1 ¿Qué es Octave?

Octave es un programa que ha sido desarrollado dentro del *proyecto GNU*. Es un lenguaje de alto nivel diseñado originalmente para realizar cálculos numéricos en la computadora. Tiene una interfase de línea de comando para resolver problemas lineales y no lineales, y un lenguaje de programación similar a su contraparte comercial MATLAB, con el que es prácticamente compatible.

Este apéndice se refiere a la versión 2.1.50 de Octave, aunque cabe señalar que éstas se renuevan periódicamente por medio de los aportes de la comunidad de usuarios.

A.1.2. Instalación.

Existen varias versiones de Octave, todas disponibles en forma gratuita en Internet. La página principal de Octave es **http://www.octave.org**. La versión original se utiliza en LINUX (un sistema operativo gratuito). A pesar de ello, existen varias versiones para Windows (el sistema operativo propietario de Microsoft). Dos de las páginas desde donde se puede obtener la distribución para Windows son

http://prdownloads.sourceforge.net/octave/octave-2.1.50-inst.exe
http://math.furman.edu/~tlewis/math13/downloads.html

Una vez obtenido el instalador, simplemente se debe ejecutar el mismo y se iniciará el proceso de instalación de Octave.

A.1.3 Notación utilizada

El texto que aparece en *itálica* representa la estructura general de lo que se está describiendo. Cuando aparece algo entre <*corchetes*> significa que debe ser remplazado por un nombre en particular (así <*nombre_de_función*> se reemplazará, por ejemplo, por **seno(x)**). El texto que aparece en **negrita** representa un código de Octave listo para ser ejecutado.

A.1.4. Ejecutando Octave por primera vez.

Para ejecutar el programa, simplemente se debe hacer click en el ícono correspondiente. Cuando el software se inicia, se muestra una ventana donde podemos escribir los comandos que necesitemos. Octave muestra un mensaje inicial y un prompt indicando que está esperando órdenes del usuario.

Tipeando los comandos a utilizar, el software responderá a cada uno de ellos mostrando los resultados a través de la pantalla. Es decir, cada vez que tipeamos algo, el software nos responderá, excepto que finalicemos lo que escribimos con un punto y coma (**;**). Además, Octave puede graficar nuestros datos y resultados en otra ventana.

Para poder obtener información de Octave es necesario conocer el nombre de la orden que se quiere usar. Tipeando **help** y luego presionando **ENTER**, nos mostrará todos los operadores, palabras reservadas, funciones, variables predefinidas (built-in) y ficheros de funciones. Si ya se conoce el nombre del comando, simplemente hay que pasarlo como parámetro. Por ejemplo,

>> help sqrt

De todas formas, el texto completo de ayuda está disponible a partir del momento que se instala Octave en

C:\Archivos de programa\GNU Octave 2.1.50\opt\octave\doc\octave_toc.html

pudiendo acceder en este archivo a una tabla que nos brinda acceso a las diferentes funciones que necesitemos.

Siempre es recomendable comentar el código. En Octave los comentarios empiezan con un signo **%** o **#**. Todo el texto que siga a continuación hasta el final de línea es considerado como un comentario y no es evaluado por Octave. Esto se muestra en el siguiente ejemplo:

```
>> manzanas=4              % Cantidad de manzanas
manzanas = 4
```

Si ocurre algún tipo de problema, se puede interrumpir la tarea que está realizando Octave con **Control-C**.

Para salir de Octave bastará con escribir **exit** en el prompt de Octave y presionar **ENTER**.

A.2. Sintaxis de Octave.

A.2.1 Tipos de datos

Octave ofrece soporte para datos predefinidos que incluyen escalares (reales y complejos), vectores, matrices, cadenas de caracteres y estructuras. También es posible la definición de nuevos datos por parte del usuario, programados en algún lenguaje que produzca código de máquina (Fortran, C, C++, ...), aunque esta posibilidad aun no ha sido suficientemente documentada. Por este motivo en esta sección nos centraremos únicamente en los datos predefinidos (Built-in).

A.2.1.1. Objetos numéricos.

Los objetos numéricos predefinidos de Octave son los escalares (reales y complejos) y las matrices. Todos los datos numéricos son almacenados como números de doble precisión, lo que significa que en sistemas que usan el formato en punto flotante se pueden representar números en el rango 2.2e-308 a 1.7e+308 y la precisión aproximada es de 2.2e-16.

A.2.1.1.1. Escalares.

Los números escalares se pueden especificar en formato decimal, hexadecimal (precedidos por el prefijo 0x). Los números en formato hexadecimal sólo pueden ser enteros. Los números decimales además permiten usar notación científica, pudiéndose utilizar los símbolos e, E, d ó D seguidos por un número n lo cual quiere decir "multiplicado por 10 elevado a la potencia n". Algunos ejemplos:

```
>> 105
ans = 105
>> 1.05e2
ans = 105
```

Nota. ans (answer) es una variable que se crea automáticamente para guardar los resultados de expresiones que no son asignadas explícitamente a

ninguna variable, como sucede en el caso anterior. Todos ellos representan el número 105.

Para representar el número complejo 3+4i se escribe de la siguiente forma

```
>> 3+4i
ans = 3 + 4i
>> 3.0+4.0i
ans = 3 + 4i
```

La única condición es que no haya un espacio entre el número y la **i** que representa el valor de la raíz cuadrada de -1. Tanto **i** y **j** y sus correspondientes mayúsculas representan la unidad imaginaria, y debe ser el último dígito de la cantidad imaginaria y sin espacios con el resto del número.

A.2.1.1.2. Rangos numéricos.

Un rango es una forma más cómoda de escribir un vector con elementos equiespaciados. Un rango se define como una base o primer valor del rango, un valor opcional de incremento entre elementos y un valor máximo que el rango no superará. Estos tres elementos se separan por el símbolo **:** (dos puntos). El incremento puede ser positivo o negativo, y en caso de omitirse se asumirá valor 1.

Escribiendo **1:5** será equivalente a la fila **[1,2,3,4,5]**. Se escribe de la siguiente forma

```
>> 1:5
ans =
     1  2  3  4  5
```

Otro ejemplo de rango es:

```
>> 7:-4:-18
ans =
     7   3  -1  -5  -9 -13 -17
```

A.2.1.1.3. Matrices.

A.2.1.1.3.1. ¿Cómo crear una matriz?

Los corchetes (**[y]**) se usan para definir vectores y matrices.

Si se trata de un vector fila se introducen los elementos separados por comas o por un espacio; si es un vector columna se introducen los elementos

separados por punto y coma. En el caso de las matrices, los elementos se introducen por filas. Esto se muestra en los siguientes ejemplos.

Vector fila:

```
>> m=[2,3,4,5]
m =
     2     3     4     5
```

ó

```
>> m=[2 3 4 5]
m =
     2     3     4     5
```

Vector Columna:

```
>> z=[1;2;3]
z  =
1
2
3
```

Matriz:

```
>> a=[1 2 3;4 5 6;7 8 9]
a =
     1     2     3
     4     5     6
     7     8     9
```

A.2.1.1.3.2 Tamaño de los objetos

Hay funciones como:

- **columns()** que devuelve el número de columnas que tiene una matriz.
- **rows()** que devuelve el número de filas que tiene una matriz.
- **length()** que retorna el número máximo de filas o de columnas (su uso está orientado a vectores que sean horizontales o verticales).
- **size()** que retorna un vector de dos componentes con el número de filas y de columnas.

Por ejemplo, ingresemos una matriz con elementos aleatorios y veamos cómo funcionan las expresiones mencionadas anteriormente.

```
>> a=rand(3,5)
a =
    0.22075  0.64006  0.12522  0.97024  0.56813
```

```
    0.59505  0.05093  0.96664  0.89523  0.69025
    0.89835  0.38869  0.47006  0.33910  0.77540

>> columns(a)          % Número de columnas
ans =
     5

>> rows(a)             % Número de filas
ans =
     3

>> length(a)           % Número máximo de filas o de columnas
ans =
     5

>> size(a)             % Número de filas y columnas
ans =
     3     5
```

A.2.1.1.3.3. Funciones que generan matrices.

Existen funciones que se utilizan muy a menudo para generar matrices de ciertas características. Se muestran dichas funciones y ejemplos ilustrativos acerca del uso de cada una de ellas.

- **zeros(n).** Crea una matriz cuadrada de dimensión n (con n filas y n columnas) con todos sus elementos ceros.
- **zeros(n,m).** Crea una matriz de n filas y m columnas con todos sus elementos ceros.
- **ones.** Genera una matriz con todos los elementos iguales a uno. Tiene las mismas formas de invocación que **zeros.**
- **eye.** Devuelve una matriz con sus elementos iguales a cero salvo en la diagonal principal, o primera diagonal. Tiene las mismas formas de invocación que **zeros.**
- **rand.** Devuelve una matriz con elementos aleatorios uniformemente distribuidos entre cero y uno. Tiene las mismas formas de invocación que **zeros.**
- **randn.** Devuelve una matriz de elementos aleatorios con distribución normal (gaussiana de media 0 y varianza 1). Tiene las mismas formas de invocación que **zeros**.
- **linspace(inicial,final,n).** Devuelve un vector de elementos equiespaciados a partir del valor inicial y hasta el final. Habrá tantos elementos como indique **n** ó **100** si este parámetro no se especifica. A

diferencia de los rangos, en que se conoce la separación de los elementos pero no su número, en este caso se conoce el número de elementos pero no su separación.

A.2.1.1.3.4. Indexación.

Cuando tenemos una variable matriz representada por su identificador es posible especificar submatrices de ésta para operar con ellas. Esto se realiza mediante las expresiones de indexación que consisten en:

- ***NombreVariable(indice_filas,indice_columnas)*** para el caso de matrices.
- ***NombreVariable(indice_elemento)*** para el caso de vectores.

Ingresemos una matriz para trabajar sobre sus índices.

```
>> A=[1,2,3;4,5,6;7,8,9]
A =
     1     2     3
     4     5     6
     7     8     9
```

Los posibles tipos de índices son:

Escalar. Entero que selecciona sólo la fila / columna indicada. Las filas / columnas se enumeran comenzando por el número **1**.

La expresión **A(3,1)** representa el elemento de la tercera fila, primera columna. En la matriz anterior su valor es **7**. Así, entonces, se escribe la orden:

```
>> A(3,1)
ans = 7
```

Los dos puntos (**:**) seleccionan todas las filas / columnas. Por ejemplo,

```
>> A(2,:)       % Representa toda la segunda fila
ans =
 4     5     6

>> A(:,:)       % Representa a todas las filas y columnas, es decir, a toda la matriz
ans =
     1     2     3
     4     5     6
     7     8     9
```

Vector. Se selecciona cada una de las filas / columnas indicadas por los elementos del vector y en el orden indicado por éste. Por ejemplo,

```
>> A(2,[3 1])          % Representa la fila 2 columnas 3 y 1
ans =
6          4
```

A.2.1.2. Cadenas de caracteres o strings.

Las cadenas de caracteres o strings, como se les denomina en Octave, son secuencias de caracteres encerrados entre comillas simples (**'**) o dobles (**"**).

Como la comilla simple (**'**) es también el operador de transposición, se recomienda utilizar comillas dobles (**"**).

Algunos caracteres no pueden ser incluidos en forma literal en un string y es necesario introducir secuencias de caracteres equivalentes (secuencias de escape). La siguiente tabla muestra algunas de las secuencias de escape:

Secuencia	Descripción
\\	Contra-barra (backslash): \.
\"	Comillas dobles ". Es necesario si se utilizan las comillas dobles para limitar el string.
\'	Comillas simples '. Es necesario si se utilizan las comillas simples para limitar el string.
\a	Representa el caracter “alert” (normalmente suena el pitido de terminal).
\b	Representa el caracter de borrado (backspace o control-h).
\f	Representa el caracter “formfeed”(nueva página).
\n	Representa el caracter “nueva línea”.

Las cadenas de caracteres pueden concatenarse utilizando la notación para definir matrices. Por ejemplo,

```
>> ["Bien","venidos"," al"," curso"]
ans = Bienvenidos al curso
```

Para más información, desde el prompt de Octave teclear **help-i strings**.

A.2.2. Variables.

Las variables nos permiten dar nombres a los valores que nos interesan guardar y poder referirnos a ellos más tarde. Ya se han visto

anteriormente muchos ejemplos de variables. El nombre de una variable debe ser una secuencia de letras, dígitos y guión bajo, pero no está permitido que empiecen con un dígito o con doble caracter de subrayado. No hay ninguna restricción respecto a la longitud de los nombres de variables. Éstos no deben superar los 30 caracteres. Las mayúsculas y minúsculas son consideradas como diferentes caracteres. Por ejemplo, son nombres de variables válidos: **X**, **x15**, **_quien_es**, **altura.**

El nombre de una variable es una expresión válida. Representa el valor almacenado en la variable. Una variable se define cuando se le asigna valor por primera vez, y no hay ningún problema en asignarle posteriormente distintos valores, aunque sean de otros tipos (por ejemplo, entero y string). Se da valor a una variable utilizando el operador igual (=), y si se trata de variables numéricas con los operadores de incremento.

A.2.3. Operadores.

Los operadores permiten construir sentencias más complejas. Es posible concatenar operaciones. Su precedencia es semejante a la de otros lenguajes.

A.2.3.1. Aritméticos.

Los operadores aritméticos son elementos que permiten construir expresiones más complejas partiendo de expresiones sencillas o atómicas (indivisibles). Operan sobre escalares y matrices.

Operación	Sintaxis	Descripción
Adición	**x + y**	Si son matrices el número de filas y columnas debe coincidir. Si un operando es escalar es sumado a cada elemento del otro operando.
Adición elemento a elemento	**x .+ y**	Este operador suma dos matrices elemento a elemento. Es equivalente a x + y.
Sustracción	**x - y**	Resta. El resultado es equivalente a x + (-y), donde -y representa el opuesto de y.
Sustracción elemento a elemento	**x .-y**	Resta elemento a elemento. Esta operación equivale a -.
Multiplicación de matrices	**x * y**	El número de columnas de x debe coincidir con el número de filas de y.
Multiplicación elemento a elemento	**x .* y**	Si ambos operandos son matrices deben coincidir en dimensión.
División a derecha	**x / y**	Este concepto es equivalente a efectuar el producto del inverso de y transpuesta por x transpuesta, todo transpuesto.

División a derecha elemento a elemento	**x ./ y**	División elemento a elemento, de los elementos de x divididos por los elementos de y.
División a izquierda	**x \ y**	El concepto es equivalente a multiplicar el inverso de x por y.
División a izquierda elemento a elemento	**x .\ y**	Cada elemento de y es dividido por el correspondiente de x.
Potencia	**x ^y** **x **y**	Si x e y son escalares devuelve x elevado a la potencia y. Ambos operadores no pueden ser matrices, y si alguno es matriz, deberá ser cuadrada.
Potencia elemento a elemento	**x .^y** **x .**y**	Si los dos operandos son matrices las dimensiones deben coincidir.
Negación	**-x**	Negación. Se obtiene el escalar opuesto o la matriz de igual di-mensión y cuyos elementos son los opuestos de la matriz original.
Suma unitaria	**+x**	Este operador no tiene efecto sobre el operando.
Conjugado complejo y transpuesta	**x'**	Para argumentos reales este operador es lo mismo que el opera-dor transpuesto. Para operandos complejos calcula el conjugado.
Transpuesta	**x.'**	Transposición de los elementos de x. No se conjugan los elementos.

Los siguientes ejemplos ilustran cómo se usan algunos de estos operadores.

```
>> c=[1 2;3 4];
>> d=[5 6;7 8];
>> c*d
ans =
      19     22
      43     50
>> c.*d
ans =
       5     12
      21     32
>>x = 121;
>> x**2
       ans = 14641
>> x^2
ans = 14641

>> z=[1;2;3]
z =
    1
| 456
```

```
    2
    3

>> z=[1;2;3]'
z =
    1     2     3

>> z=-[1;2;3]
z =
    -1
    -2
    -3
```

A.2.3.2. Incremento y decremento.

Los operadores de incremento, incrementan o decrementan el valor de una variable en 1. El operador de incremento se escribe como dos signos más seguidos sin espacio entre ellos (++), y el de decremento como dos signos menos seguidos sin espacio entre ellos (--). La forma de utilizarlos será antes o después de variables que ya se hayan definido, por ejemplo, **entradas++**, **y--**, **--altura**, **++indice**. Así, **++x** puede escribirse también como **x=x+1**, que se lee como **x se vuelve el valor que tiene la variable x más uno.**

En los siguientes ejemplos se muestra la diferencia entre utilizarlos antes o después de la variable:

```
>> x=2;
>> ++x        % Devuelve el valor original de x incrementado en 1
ans = 3
>> x
x = 3

>> x=2;
>> x++       % Devuelve el valor original de x e incrementa en 1 el valor de x
ans = 2
>> x
x = 3
```

Para matrices y vectores los operadores operan a nivel de elemento.

A.2.3.3. De comparación.

Los operadores de comparación comparan los valores numéricos con respecto a la igualdad. Todos los operadores de comparación devuelven un valor 1 si la operación es verdadera ó 0 si es falsa.

Operador	Descripción
x < y	Es verdadero si x es menor que y.
x <= y	Es verdadero si x es menor o igual que y.
x == y	Es verdadero si x es igual a y.
x >= y	Es verdadero si x es mayor o igual que y.
x > y	Es verdadero si x es mayor que y.
x != y	Es verdadero si x es distinto que y.
x ~= y	Es verdadero si x es distinto que y.
x <> y	Es verdadero si x es distinto que y.

Para matrices, los operadores trabajan elemento por elemento. Así, por ejemplo,

```
>> [1 2;3 4]==[1 3;2 4]
ans =
     1     0
     0     1
```

Si un operador fuera un escalar y el otro una matriz, el escalar se comparará con cada elemento de la matriz y el resultado tiene las mismas dimensiones que la matriz.

A.2.3.4. Lógicos o Booleanos.

El resultado de un operador booleano es una matriz de dimensión equivalente a los operandos, donde cada elemento es el resultado de aplicar el operador booleano a los elementos correspondientes. Únicamente hay tres operadores booleanos: "o" (|), "y" (**&**) y "no" (**!**), junto con paréntesis que controlan el orden de las operaciones lógicas. Realizan la operación booleana elemento a elemento, suponiendo 0 como falso y distinto de 0 como verdad. El resultado devuelto es 1 para verdad y 0 para falso. Los operadores booleanos se pueden emplear en las mismas situaciones que los operadores de comparación.

Operador	Descripción
x & y	Es verdadero sólo si x e y son verdaderos.
x | y	Es verdadero si x o y son verdaderos. Es falso si ambos elementos son falsos.
! x	Es verdadero si x es falso. Es falso si x es es verdadero.

El siguiente ejemplo muestra cómo se utlizan estos operadores:

```
>> x=2;
>> z=3;
>> y=4;
>> x>z | y>z
ans = 1
>> x>z | y<z
ans = 0
>> x>z & y<z
ans = 0
>> !(x>y)
ans = 1
```

A.2.3.5. Lógicos “short-circuit”.

Los operadores booleanos de "corto circuito" son semejantes a los operadores booleanos, con la diferencia de que si después de evaluar el primer operando ya es suficiente para obtener el resultado, no se comprueba el segundo operando.

Operador	Descripción
exp1&&exp2	Se evalúa **exp1**. Si el resultado tiene algún elemento falso (0), el operador devolverá 0 sin evaluar **exp2**. Si todos los elementos de **exp1** son verdad se pasa a evaluar **exp2**. Si el resultado de **exp2** tiene todos los elementos verdad el resultado será verdad (1), caso contrario será falso (0).
exp1 || exp2	Se evalúa **exp1**. Si el resultado tiene todos los elementos verdad (1), el operador devolverá 1 sin evaluar **exp2**. Si algún elemento de **exp1** es falso se pasa a evaluar **exp2**. Si el resultado de **exp2** tiene todos los elementos verdad el resultado será verdad (1), caso contrario será falso (0).

La diferencia entre los operadores binarios normales y los de corto circuito se verá mejor con un ejemplo:

```
>> a=0;b=0;a & b++,b
ans = 0
b = 1
>> a=0;b=0;a && b++,b
ans = 0
b = 0
```

En el segundo caso, **b** será incrementado sólo si **a** es verdadero. En el primer caso, **b** siempre será incrementado.

A.2.3.6. De asignación.

El signo igual (=) es el operador de asignación. Después de una asignación, una variable cambia de valor y de tipo para acomodarse al del nuevo valor. El operador asignación es la única manera de poder almacenar valores. En el lado derecho puede aparecer cualquier expresión de las descritas anteriormente o funciones que devuelvan un valor. En el lado izquierdo podemos tener variables, elementos de una matriz o vector, o listas de valores de retorno. Por ejemplo, para asignar un valor **1** a la variable **z** se utiliza la siguiente expresión

```
>> z=1
z = 1
```

Es importante notar que las variables pueden cambiar de tipo durante la ejecución de un programa; en un momento pueden ser un valor numérico y luego ser una matriz. Asignar un escalar a un conjunto de índices de una matriz hará que los elementos de la matriz se vuelvan el valor escalar. Así, por ejemplo, **a(:, 2) = 5 ENTER**, hará que todos los elementos de la segunda columna de la matriz **a** sean **5**. Esto es, definimos la matriz **a** y luego asignamos el valor **5** a toda la segunda columna:

```
 >> a=[1 2 3; 3 4 5; 6 3 1]
a =
     1     2     3
     3     4     5
     6     3     1

>> a(:,2)=5
a =
     1     5     3
     3     5     5
     6     5     1
```

A.2.3.7. Precedencia de los operadores.

La precedencia de los operadores indica cuales son las operaciones que se realizan primero al evaluar una expresión. Por ejemplo, '**-x^2**' se reduce a '**-(x^2)**', ya que el operador '^' tiene precedencia sobre el '**-**'.

- Operadores de fin de sentencia '**;**', '**,**'
- Operador de asignación '='

- Operadores lógicos cortocircuitados '||', '**&&**'
- Operadores lógicos '|', '**&**'
- Operadores de relación '<', '<=', '==', '>=', '>', '**!**', '~=', '<>'
- Operador "dos puntos" (rango) '**:**'
- Suma y resta '+', '-'
- Multiplicación y división '*', '/', '\', '.\', '**.***', '**./**'
- Transpuesta '**'**', '**.'**'
- Operadores unarios '+', '**-**', '++', '**--**', '**!**', '~'
- Exponenciación '^', '**', '**.^**', '**.****'

A.2.4. Funciones.

A.2.4.1. Matemáticas.

expm(A)	Exponencial de una matriz cuadrada (por serie de Taylor).
logm(A)	Logaritmo de una matriz cuadrada.
sqrtm(A)	Raíz cuadrada de una matriz cuadrada.
arg(X)	Argumento del número complejo.
abs(X)	Módulos; si son reales es equivalente al valor absoluto.
conj(X)	Complejo conjugado; parte imaginaria cambiada de signo.
imag(X)	Parte imaginaria como número real.
real(X)	Sólo parte real de los elementos.
rem(X,Y)	Resto de la división de matrices reales.
sign(X)	Signo de los elementos: 1 si es positivo; -1 si es negativo; 0 si es 0.
exp(X)	Exponencial.
sum(X)	Suma de todos los elementos de la columna o vector.
prod(X)	Producto de todos los elementos de la columna o vector.
log(X), log10(X), log2(X)	Logaritmos.
sqrt(X)	Raíz cuadrada.
sin, cos, tan, sec, csc, cot	Trigonométricas ordinarias.
asin, acos, atan, asec, acsc, acot	Trigonométricas inversas.
sinh, cosh, tanh, sech, csch, coth	Trigonométricas hiperbólicas.
asinh, acosh, atanh, asech, acsch, acoth	Trigonométricas hiperbólicas inversas.

A.2.4.2. Generales.

det(A)	Determinante de una matriz cuadrada.
trace(A)	Traza de la matriz: suma de los elementos de su diagonal principal.
inv(A)	Calcula la inversa de una matriz cuadrada no singular ($\|A\| \neq 0$).
rank(A)	Rango de una matriz.
eig(A)	Calcula los autovalores de una matriz cuadrada.

A.2.5. Expresiones de control de flujo.

Las expresiones de control de flujo influyen en la forma en que el código es ejecutado, y dividen el código en bloques. Los bloques empiezan con una palabra reservada, por ejemplo, **while** y terminan con la palabra reservada **endwhile**, formada por el prefijo **end** y la misma palabra que empezó el bucle. También es posible terminar un bloque con la palabra **end**. El código que se escribe entre dichas palabras se denomina el cuerpo de la expresión de control.

A.2.5.1. La sentencia if.

La sentencia **if** (si condicional) es la sentencia que permite tomar una decisión en Octave. Existen tres formas básicas; la más sencilla es:

if (<condición>)
  {cuerpo-entonces}
endif

donde *<condición>* es una expresión que controla lo que se va a hacer. Si *<condición>* es verdadera, las sentencias de *{cuerpo-entonces}* se ejecutarán. Si *<condición>* es falsa no se ejecutarán las sentencias de *{cuerpo-entonces}*.

Por ejemplo, en el siguiente fragmento de código, si **x** es menor que **2**, se sumará uno a **x**.

```
if (x< 2)
  x++;
endif
```

La segunda forma del **if** es

if (<condición>)
  {cuerpo-entonces}
else

{cuerpo-no-entonces}
endif

Si <*condición*> es verdadera se ejecutan las sentencias de *{cuerpo-entonces}*; si <*condición*> es falsa se ejecutan las sentencias de *{cuerpo-no-entonces}*.

Así, por ejemplo,

```
if (rem (x, 2) == 0)
  x++;
else
  x--;
endif
```

Ese fragmento hace que si el resto de dividir **x** por **2** (que se escribe **rem(x,2)**) es igual a **0** (**==0**) se sumará **1** a **x** (**x++**) y sino se le restará **1** a **x** (**x--**).

La tercera forma del **if** es la siguiente

```
if (<condición>)
  {cuerpo-entonces}
elseif (<condición>)
  {cuerpo-entonces-elseif}
elseif (<condición>)
  {cuerpo-entonces-elseif}
elseif (<condición>)
  {cuerpo-entonces-elseif}
else
  {cuerpo-no-entonces}
endif
```

donde cada <*condición*> se chequea, y cuando una <*condición*> es verdadera se ejecuta su correspondiente *{cuerpo-entonces}*. Si ninguna es verdadera, el último *{cuerpo-no-entonces}* será ejecutado. Puede haber muchos cuerpos **elseif**.

Así, por ejemplo,

```
if (rem (x, 2) == 0)
  x++;
elseif (rem (x, 3) == 0)
  x = x+5;
else
  x--;
endif
```

En este caso, si **x** es divisible por **2** se le sumará **1** a su valor, si es divisible por **3** se le sumará **5** a su valor y si no se le restará **1**.

A.2.5.2. La sentencia while.

En programación un loop significa que una parte de un programa se ejecutará más de una vez en forma sucesiva. La sentencia **while** es la forma más sencilla de hacer loops en Octave. Lo que hace un **while** es repetir un grupo de sentencias, mientras que una condición determinada sea verdadera.

while (condición)
  cuerpo
endwhile

Lo primero que sucede es la verificación de la *condición*. Si la *condición* es verdadera, se ejecuta el *cuerpo*. Una vez ejecutado el *cuerpo*, se vuelve a verificar la *condición*, y mientras siga siendo verdadera se vuelve a ejecutar el *cuerpo*, hasta que la *condición* no es verdadera. Si la *condición* es inicialmente falsa, el *cuerpo* no se ejecuta nunca.

El siguiente bloque de código genera números aleatorios mientras que el usuario ingrese 1, cuando ingrese 0 se detiene.

```
a=input('Ingrese 1 para generar un número aleatorio, 0 para detener el proceso');
    while a==1
          rand
          a=input('Ingrese 1 para continuar, 0 para detener el proceso');
    endwhile
```

A.2.5.3. La sentencia for.

El **for** es más conveniente cuando queremos contar la cantidad de veces que iteramos en el loop. La estructura de un **for** es como sigue

for nombre-variable = expresión
  cuerpo
endfor

La parte *nombre-variable* puede tener varias formas. La forma más común es el nombre de una variable simple. El operador de asignación (=) que figura en el **for** trabaja como sigue. En vez de asignar la *expresión* a *nombre-variable* de una vez, trabaja por columnas. Si la *expresión* es un rango, un vector o un escalar, el valor de *nombre-variable* será asignado cada vez que el cuerpo se ejecute. Esto puede verse en el siguiente código

que genera 10 números aleatorios. El rango **1:10** se va asignando a la variable **i** hasta que no quedan valores.

```
for i = 1:10
  r(i) =rand;
endfor
```

A.2.5.4. La sentencia input.

La sentencia **input** permite el ingreso de un valor numérico, como puede verse en el siguiente código

```
>> valor = input("Ingrese un valor ");
Ingrese un valor
```

Si se agrega el argumento "**s**", el valor ingresado se considera una cadena de caracteres.

A.2.5.5. La sentencia break.

La sentencia **break** hace que la ejecución "salte" fuera del ciclo **while** o **for** que la contiene.

En el siguiente ejemplo, el bucle terminará cuando la variable **var** valga **3**.

```
a = input("Ingrese un valor  ");
for var=1:10
a
        if(a==3)
           break
        endif
endfor
```

A.3. Archivos de funciones y de scripts.

A.3.1. Archivos de funciones.

Los programas complicados en Octave pueden ser simplificados definiendo funciones. Las funciones pueden ser definidas en la línea de comando o en archivos externos y pueden ser llamadas como funciones predefinidas.

A.3.1.1. Definiendo funciones.

La forma más simple de definición de una función es la siguiente

function nombre
  cuerpo
endfunction

Un *nombre* válido de función debe obedecer a las mismas reglas que un nombre válido de una variable. El *cuerpo* de la función consiste en un conjunto de sentencias Octave. En el *cuerpo* de la función se define que es lo que la función debe hacer. Normalmente, será necesario pasarle alguna información a la función que definimos. La estructura general para pasar parámetros a una función es

function nombre (arg-list)
  cuerpo
endfunction

donde *arg-list* es una lista de argumentos separados por comas. Cuando la función es llamada, los nombres de los argumentos son utilizados para contener los valores de los argumentos durante esa llamada. La función termina con la palabra *endfunction.*

Por ejemplo, esta sería una función que implementa **$y=3x^2-4x$**. Como la queremos guardar en un archivo para usarla varias veces, se la escribe en el Bloc de notas. Para ello, se deben seguir los siguientes pasos:

Ir a: **Inicio → Programas → Accesorios → Bloc de notas**, y escribir las sentencias:

```
function y(x)
  y=3*x^2-4*x
endfunction
```

Luego, se guarda en

C:\Archivos de programa\GNU Octave 2.1.50\octave_files\ f.m

Para invocar a la funcion y calcularla en un determinado valor, por ejemplo, **3.2654**, escribiremos

```
>> f(3.2654)
y = 18.927
```

Debemos recordar que el nombre con el cual guardamos la función, será luego con el que deberemos invocarla.

Por ejemplo, si queremos implementar una función que obtenga los primeros **n** múltiplos de un número entero **x**, podemos definir la siguiente función

```
function multiplos(x,n)
  for i=1:n
       x*i
  endfor
endfunction
```

Si a esta función la guardamos con el nombre **multiplo.m**, deberemos invocarla con dos argumentos:

```
>> multiplo(13,2)
ans = 13
ans = 26
```

A.3.1.2. Devolución de valores desde una función.

En la mayoría de los casos también se desea que alguna información regrese desde las funciones que se definan. La forma general de hacer esto es

```
function ret-val = name (arg-list)
  cuerpo
endfunction
```

donde *ret-val* es el nombre de la variable que contendrá el valor que la función retornará. Esta variable debe estar definida antes del fin del cuerpo de la función.

Las variables usadas en el cuerpo de la función son locales a la función. Las variables nombradas en la *arg-list* y *ret-val* también son locales a la función.

Por ejemplo, esta sería una función que calcula el valor promedio de una serie de valores almacenados en un vector

```
function retval = pro (v)
  retval = sum (v) / length (v);
endfunction
```

Este archivo lo guardamos con el nombre **pro.m**. Lo invocamos así

```
>> vec=rand(1,5)
vec =
      0.92892  0.18633  0.02686  0.54460  0.96861
```

```
>> pro(vec)
ans = 0.53106
```

A.3.1.3. Retorno de múltiples valores desde una función.

A diferencia de muchos otros lenguajes, Octave permite que una función retorne más de un valor. La sintaxis para definir funciones que devuelvan múltiples valores será

```
function [ret-list] = nombre (arg-list)
  cuerpo
endfunction
```

donde *ret-list* será una lista de variables separadas por comas que guardarán los valores devueltos por la función.

Por ejemplo, definimos una función donde se guardarán el elemento mayor de un vector y el índice correspondiente

```
function [max, idx] = vmax (v)
  idx = 1;
  max = v (idx);
  for i = 2:length (v)
    if (v (i) > max)
      max = v (i);
      idx = i;
    endif
  endfor
endfunction
```

A esta función la guardamos en un archivo con el nombre **mayor.m**. Luego, la invocamos desde Octave de la siguiente forma

```
>> v=rand(1,7)
v =
   0.69649  0.10844  0.45652  0.25819  0.48350  0.82244  0.98909
>> [elementomayor,indice]=mayor(v,i)
elementomayor = 0.98909
indice = 7
```

A.3.1.4. Retornando desde una función.

El cuerpo de una función puede contener una sentencia **return**. Esta sentencia hace que la ejecución vuelva al resto del programa. La sintaxis es

return

A.3.1.5. Archivos de funciones (.m).

Excepto para programas muy sencillos, no es práctico tener que definir todas las funciones que necesitemos cada vez que las debamos utilizar. En lugar de ello, la idea es guardar las funciones en un archivo que sea fácil de utilizar, ya sea para agregar funciones nuevas o modificar las que ya hemos creado y luego utilizarlas desde Octave.

Octave no requiere que se carguen las funciones antes de usarlas. Simplemente, es necesario que se guarden en un archivo con extensión "**.m**" en algún sitio donde Octave pueda encontrarlas, y que el nombre del archivo coincida con el de la función.

Cuando Octave encuentra un identificador que no ha sido definido, primero verifica que no sea una variable o función que ya haya sido usada. Si no la encuentra allí, busca en la lista de directorios que se encuentra almacenada en la variable predefinida **LOADPATH**, buscando archivos "**.m**" que tengan el mismo nombre que el identificador. Así, por ejemplo, si Octave encuentra el identificador **promedio** en un programa y no es una variable, buscará por el archivo **promedio.m** en la lista de directorios definidos en **LOADPATH**. Una vez que Octave encuentra el archivo, el contenido del mismo es leído. Si se define sólo una función dentro de ese archivo, esta función se ejecuta. El lugar por defecto en donde Octave espera que se encuentren los archivos es **<directorio de instalación>/octave_files**, entonces, por ejemplo, si instalamos Octave con las opciones por defecto este lugar sería **C:/Archivos de programas/GNU Octave 2.1.50/octave_files**.

A.3.2. Archivos de scripts (.m).

Los scritps son simples archivos de texto en el que se introducen las sentencias de la misma manera que se introducen en la línea de comandos. Su principal utilidad es definir un entorno de trabajo en cualquier momento, definiendo una serie de variables y/o funciones. Cada vez que queramos recuperar ese estado, deberemos volver a invocar el script por su nombre (sin la extensión). La única restricción es que la primera sentencia no debe ser la definición de una función. En caso contrario, Octave pensará que se trata de un fichero de función y los resultados no serán los esperados.

A.4. Polinomios.

A.4.1. ¿Cómo introducir un polinomio?.

Los polinomios se representan mediante un vector con los coeficientes en orden descendente. Se deben incluir los términos con coeficientes nulos.

Veamos algunos ejemplos ilustrativos.

```
>> p=[1 0 2 0 3]          % Polinomio p = x^4+2*x^2+3
p =
     1     0     2     0     3

>> q= [2 1 0]             % Polinomio q = 2*x^2+x
q =
     2     1     0
```

A.4.2. Evaluación.

Un polinomio se puede evaluar en un punto utilizando la función **polyval(p,x)**, que evalúa el polinomio **p** para el valor de **x**. Si **x** es un vector, el polinomio **p** se evalúa para cada elemento de **x**.

Los siguientes ejemplos muestran cómo se usa esta función.

```
>>polyval(p,-1)          % Evaluación del polinomio p = x^4+2*x^2+3
en x = -1
ans =
     6
```

lo que nos dice que **p** evaluado en **–1** es igual a **6**.

```
>> polyval(q,[1 –7])     % Evaluación del polinomio q = 2*x^2+x en x =
[1 –7]
ans =
     3    91
```

lo que nos dice que **q** evaluado en **1** y en **–7** es igual a **3** y **91**, respectivamente.

A.4.3. Raíces.

Las raíces de un polinomio se encuentran utilizando la función **roots(p)**.

Veamos un ejemplo ilustrativo.

```
>> roots (p)              % Raíces del polinomio p = x^4+2*x^2+3
ans =
      -0.60500 + 1.16877i
      -0.60500 - 1.16877i
       0.60500 + 1.16877i
       0.60500 - 1.16877i
```

A.4.4. Operaciones sobre polinomios.

Ya sabemos cómo introducir polinomios. Ahora vamos a sumar y restar polinomios.

```
>> p+[0 0 q]              % Suma de los polinomios p = x^4+2*x^2+3  y  q
= 2*x^2+x. Se
                          escriben así porque son de distinto grado
ans =
      1      0      4      1      3
```

lo que nos dice que **p+q = x^4+4x^2+x+3**.

```
>> p-[0 0 q]          % Resta de los polinomios  p = x^4+2*x^2+3  y  q =
                      2*x^2+x. Se escriben así porque son de distinto grado
ans =
      1      0      0     -1      3
```

lo que nos dice que **p-q = x^4-x+3**.

A continuación, se muestran otras funciones que realizan operaciones sobre polinomios y ejemplos ilustrativos sobre su uso:

- **conv(p,q).** Multiplica los dos polinomios **p** y **q**.

```
>>pro=conv(p,q)        % Producto de los polinomios p = x^4+2*x^2+3  y
                       q = 2*x^2+x
pro =
      2      1      4      2      6      3      0
```

lo que nos dice que **p.q = $2x^6+x^5+4x^4+2x^3+6x^2+3x$**.

- **deconv (p,q).** Divide el polinomio **p** por el polinomio **q**.

```
>>deconv(pro,p)    % Divide el polinomio pro 2*x^6+x^5+4*x^4+2*x^3+6*x^2+3*x
                    por el polinomio p = x^4+2*x^2+3; obviamente el
                    resultado es q
ans =
     2     1     0
```

lo que nos dice que **pro/p = $2x^2+x$**.

• **poly(v).** Siendo **v** un vector, devuelve un polinomio (con coeficiente principal la unidad) cuyas raíces son los elementos de **v**.

```
>> poly([1 0])        % Polinomio mónico que tiene por raíces a los números 1 y 0
ans =
     1    -1     0
```

la respuesta corresponde al polinomio **$x^2-x = x\,(x-1)$**.

• **polyder(p).** Calcula la derivada del polinomio **p.**

```
>> polyder(p)                % Derivada del polinomio p
ans =
     4     0     4     0
```

que corresponde al polinomio **$4x^3+4x$**, y que es justamente la derivada del polinomio **$p=x^4+2x^2+3$**.

A.5. Sistemas de ecuaciones lineales y no lineales.

A.5.1. Sistemas de ecuaciones lineales.

La solución de un sistema de ecuaciones lineales del tipo **Ax = b** se puede hacer de varias maneras:

1. Usando el operador "división por la izquierda", \ (método de eliminación de Gauss):
   >> A\b
2. Usando el comando **rref** sobre la matriz ampliada (método de Gauss-Jordan):
   >> rref([A b])
3. Invirtiendo la matriz A (método de la inversa):
   >> inv(A)*b

Para matrices de orden elevado la mejor forma es la dada en 1.

Consideremos, por ejemplo, el siguiente sistema lineal

$$\begin{aligned} x_1 - 2x_2 + 3x_3 &= 1 \\ 4x_1 + x_2 - 2x_3 &= -1 \\ 2x_1 - x_2 + 4x_3 &= 2 \end{aligned}$$

Para hallar la solución de este sistema, las instrucciones en Octave son:

```
>> A=[1 -2 3;4 1 -2;2 -1 4];   % Introducimos la matriz cuadrada de coeficientes A
>> b=[1;-1;2];                 % Introducimos la matriz columna de términos
                               independientes b
>> x=A\b                       % Vector solución según 1
x =
   -0.04167
    0.41667
    0.62500

>> x=rref([A b])               % Vector solución según 2
x =
   1.00000   0.00000   0.00000  -0.04167
   0.00000   1.00000   0.00000   0.41667
   0.00000   0.00000   1.00000   0.62500

>> x=inv(A)*b                  % Vector solución según 3
x =
   -0.04167
    0.41667
    0.62500
```

Si la matriz de coeficientes es singular, Octave emite un mensaje de advertencia y calculará una solución en el sentido de norma mínima.

Por lo tanto, si queremos resolver un sistema de n ecuaciones con n incógnitas, lo primero que deberíamos hacer es verificar que este sistema sea compatible determinado (es decir que la matriz de coeficientes del sistema sea no singular). Esto se hace simplemente escribiendo el comando **det(A)**. Sólo en el caso en que arroje un resultado distinto de cero, el sistema se podrá resolver.

```
>> A=[1 1 2;3 5 8;13 21 34];b=[.52;.83;.04];  % Introducimos la matriz cuadrada A
                                                 y la matriz columna b

>> A\b                                        % Solución del sistema según 1
warning: matrix singular to machine precision, rcond = 0
```

```
warning: attempting to find minimum norm solution
ans =
       0.67648
      -0.57352
       0.10296

>> det(A)                          % Evaluamos el determinante de la matriz A
ans = 0
```

Por ello es que resulta conveniente que primero siempre se evalúe el determinante de la matriz del sistema antes de resolverlo:

```
>> A=[1 2 3;4 5 6;7 1 4];b=[1;0;-1];    % Introducimos la matriz cuadrada A y la
                                          matriz columna b
>> det (A)                              % Evaluamos el determinante de la matriz A
ans = -27.000

>> % Distinto de cero: estupendo! Se puede resolver

>> A\b                                 % Solución del sistema según 1
ans =
      -0.62963
      -0.74074
       1.03704

>> rref([A b])                         % Solución del sistema según 2
ans =
       1.00000   0.00000   0.00000  -0.62963
       0.00000   1.00000   0.00000  -0.74074
       0.00000   0.00000   1.00000   1.03704

>> inv(A)*b                            % Solución del sistema según 3
ans =
      -0.62963
      -0.74074
       1.03704
```

A.5.2. Sistemas de ecuaciones no lineales.

Octave puede resolver sistemas de ecuaciones no lineales de la forma F(x)=0 usando la función **fsolve**. La forma es la siguiente:

[x,info]=fsolve (*funcion,val*)

donde *funcion* es el nombre de una función de la forma F(x) y *val* es un valor inicial.

Por ejemplo, para resolver el siguiente sistema de ecuaciones no lineales de la forma F(x)=0:

$$-2x^2 + 3xy + 4\,\text{sen}(y) = 6$$
$$3x^2 - 2xy^2 + 3\cos(x) = -4$$

primero necesitamos escribir una función para calcular el valor de la función dada. Es decir,

```
>> function y=f(x)
y(1)=-2*x(1)^2+3*x(1)*x(2)+4*sin(x(2))-6;
y(2)=3*x(1)^2-2*x(1)*x(2)^2+3*cos(x(1))+4;
endfunction
```

Luego, llamamos **fsolve** con una condición inicial especificada para encontrar las raíces del sistema de ecuaciones. En este caso, dada la función f definida antes, escribimos

```
>> [x,info]=fsolve("f",[1;2])          % Solución para el valor inicial [1;2]
x =
   0.57983
   2.54621

info = 1
```

y obtenemos así la solución del sistema dado. Un valor de **info = 1** indica que la solución es convergente.

Resolvamos el siguiente sistema de ecuaciones no lineales

$$4x^3 - 27xy^2 + 25 = 0$$
$$4x^2 - 3xy^3 - 1 = 0$$

ingresando distintos valores iniciales.

Primero escribimos la función:

```
>> function y=f(x)
y(1)=4*x(1)^3-27*x(1)*x(2)^2+25;
y(2)=4*x(1)^2-3*x(1)*x(2)^3-1;
endfunction
```

Luego, usamos **fsolve** para hallar las raíces del sistema de ecuaciones.

```
>> [x,info]=fsolve("f",[1;1])        % Solución para el valor inicial [1;1]
x =
   1.04429
   1.02382

info = 1
```

La solución es convergente pues **info = 1**.

```
>> [x,info]=fsolve("f",[2;2])        % Solución para el valor inicial [2;2]
x =
   1.04429
   1.02382

info = 1
```

La solución es convergente pues **info = 1**.

```
>> [x,info]=fsolve("f",[0;0])        % Solución para el valor inicial [0;0]
x =
    0
    0

info = 3
```

La solución no es convergente pues **info** no es igual a **1**.

A.6. Gráficas.

Para la realización de las gráficas Octave invoca al paquete Gnuplot.

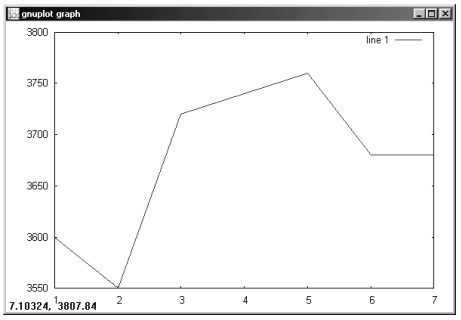
A.6.1 Gráficas en dos dimensiones.

En Octave podemos graficar utilizando la función **plot**. Esta crea una gráfica bidimensional a partir de vectores y/o columnas de matrices con escalas lineales sobre ambos ejes.

Se pueden representar los elementos de un vector frente a los de otro vector. Por ello, las longitudes de ambos han de coincidir, como se muestra en el siguiente ejemplo.

Ejemplo 1

```
>> dias=[1 2 3 4 5 6 7];          % Se definen los elementos de los vectores dias y bolsa
>> bolsa=[3600 3550 3720 3740 3760 3680 3680]; % Las longitudes de ambos vectores
                                                  deben coincidir
>> plot(dias,bolsa)          % Se grafica un conjunto de puntos considerando que dias
                               corresponde al eje horizontal y bolsa al eje vertical.
```



También podemos definir un vector **x**, con su valor inicial, el incremento y el valor final. Luego aplicamos a este vector la función deseada. Esto se muestra en el siguiente ejemplo.

Ejemplo 2

```
>> x=0.5:0.1:15;              % Se introduce el vector x
>> plot(x,sin(x)+log(x));     % Se grafica la función aplicada al vector x
```

Podemos graficar varias funciones en la misma ventana, simplemente colocando la variable con respecto a la cual obtendremos la gráfica y además separarlas por coma, por ejemplo

```
>>plot(x,log(x)+sin(x),x,sin(x),x,abs(x))
```

En este caso, en el cual graficaremos varias funciones juntas, es necesario poder distinguirlas. Para ello, podremos cambiar el tipo de línea o trazo y color de cada una, colocando después de la función y separado por coma, el símbolo y eventualmente el número que nos indicará el color.

Los posibles tipos de líneas del comando **plot** son los siguientes:

'-' Segmento uniendo los datos; formato por defecto.
'.' Puntos pequeños en cada dato.
'@' Puntos en cada dato.
'-@' Líneas uniendo cada dato con punto en cada dato.
'^' Gráfica estilo impulso; líneas de 0 al punto.
'L' Gráfica estilo escalera.
'n' Donde **n** es un dígito de 1 a 6; indica color. Algunos colores se pueden especificar por su inicial inglesa: r, g, w, etc.
'nm' Donde **n** y **m** son dígitos de 1 a 6; n indica color y m estilo de punto. Los estilos de puntos se pueden indicar por un símbolo: *, + o, etc.

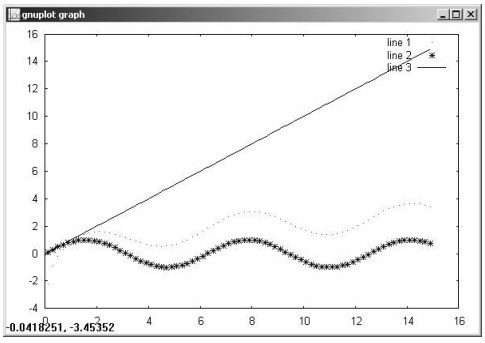
La relación entre el número y el color, y el estilo de línea es la siguiente:

Número	Color	Letra color	Tipo de punto	Símbolo punto
1	rojo	r	Círculo	o
2	verde	g	Cruces	+
3	azul	b	Cuadrado	
4	magenta	m	Aspa	x

En el siguiente ejemplo se grafican varias funciones juntas distinguiéndolas por el tipo de línea y color.

Ejemplo 3

>> plot(x,log(x)+sin(x),'.1',x,sin(x),'*3',x,abs(x),'-5')

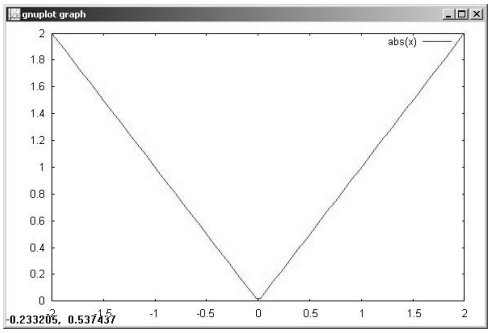


También por medio de la función **fplot** podemos obtener la gráfica de una función determinada, a la que se debe indicar el intervalo en el cual se quiere obtener la gráfica. La sintaxis es la siguiente

fplot(‘funcion’, [a,b])

Ejemplo 4

>> fplot('abs(x)',[-2,2])



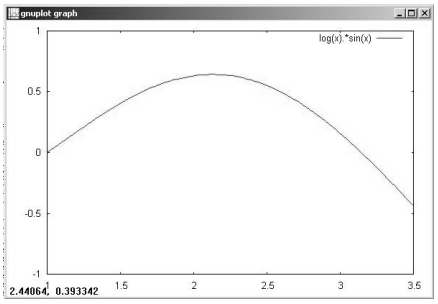
Por medio de la función **fplot** podemos indicar el intervalo que debe considerar en el eje y. Además, podemos seleccionar la cantidad de puntos que queremos que tome para trazar la gráfica de la función. Su sintaxis es

fplot(‘funcion’,[infx,supx,infy,supy],[cantidad de puntos])

y su uso se muestra en el siguiente ejemplo.

Ejemplo 5

>> fplot('log(x).*sin(x)',[1,3.5,-1,1])



A.6.2. Funciones para el control de la gráfica.

Existe un conjunto de funciones por medio de las cuales podemos modificar el aspecto de las gráficas realizadas, añadirles títulos, carteles, etc.

• **hold on.** Mantiene la gráfica actual en la ventana; los sucesivos se añadirán al actual.

• **hold off.** Desactiva la permanencia de la gráfica; los siguientes borrarán la ventana. Este es el estado por defecto.

• **clearplot.** Borra la ventana de gráficas actual.

• **closeplot.** Cierra la ventana de gráficas y el Gnuplot asociado.

• **replot.** Actualiza la gráfica.

A.6.3. Etiquetas en la gráfica.

Existen además otras funciones orientadas a añadir títulos a la gráfica, a cada uno de los ejes, a dibujar una cuadrícula auxiliar, a introducir texto, etc. Estas funciones son las siguientes:

• **title('título').** Añade un título a la gráfica.

• **xlabel('tal').** Añade una etiqueta al eje de abscisas.

• **ylabel('cual').** Añade una etiqueta al eje de ordenadas.

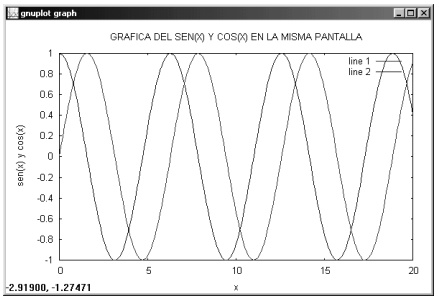
• **text(x,y,'texto').** Introduce 'texto' en el lugar especificado por las coordenadas x e y.

• **grid on y grid off.** Activan y desactivan, respectivamente, la grilla en la ventana que contiene a la gráfica.

El siguiente ejemplo muestra cómo se usan algunas de estas funciones.

Ejemplo 6

```
>> % Representa un seno frente a un coseno en la misma gráfica, escogiendo colores
>> x=0:0.1:20; y=sin(x); z=cos(x);
>> xlabel('x')
>> ylabel('sin(x) y cos(x)')
>> title('GRAFICA DEL SEN(X) Y  COS(X) EN LA MISMA PANTALLA')
>> plot(x,y,'r',x,z,'b')
```



A.6.4. Múltiples gráficas.

Por medio de la sentencia **subplot** podemos dividir la ventana en la cual aparecerá la gráfica en varias sub ventanas. De esta forma, luego podemos seleccionar en cual de ella queremos que aparezca nuestra gráfica. Nos permite obtener así en una misma ventana varias gráficas, facilitando luego trabajos de comparación y estudio. La sintaxis de la función es

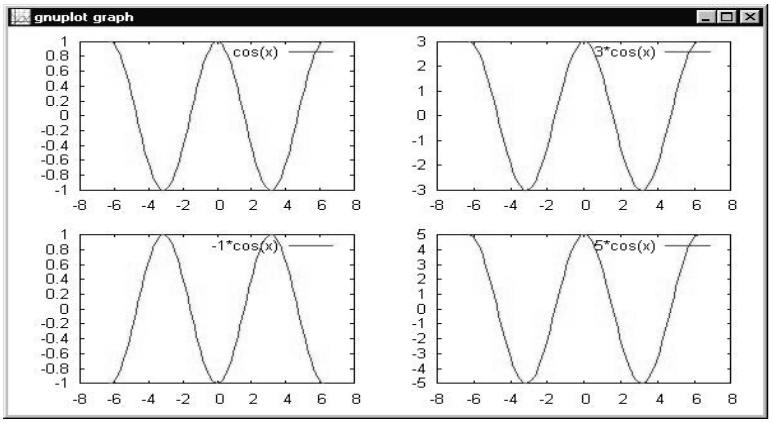
subplot(a,b,i)

donde **a** y **b** indican las columnas y las filas, respectivamente, en la que se dividirá la ventana, definiendo **axb** nuevas sub ventanas. El índice **i** nos indica en cual de todas las sub ventanas aparecerá la próxima gráfica a realizar. El índice va de **1** hasta **axb**.

El siguiente ejemplo ilustran cómo se usa esta función.

Ejemplo 7

```
>> subplot(2,2,1)
>> fplot('cos(x)',[-2*pi,2*pi])
>> subplot(2,2,2)
>> fplot('3*cos(x)',[-2*pi,2*pi])
>> subplot(2,2,3)
>> fplot('-1*cos(x)',[-2*pi,2*pi])
>> subplot(2,2,4)
>> fplot('5*cos(x)',[-2*pi,2*pi])
```



A.6.5. Gráficas tridimensionales.

Para realizar gráficas en tres dimensiones tenemos diferentes variantes según el tipo de gráfica a realizar. Ellas son:

- **plot3**
- **meshgrid**
- **mesh**
- **surf**

A.6.5.1. Gráficas de línea.

Extendemos la orden **plot** de 2-D a 3-D con **plot3**. El formato es el mismo, excepto que los datos están en tripletes. Es decir, para utilizar esta función debemos definir los rangos para **x**, para **y**, y luego indicar la función que queremos graficar.

A.6.5.2. Gráficas de malla y de superficie.

A.6.5.2.1. De malla.

Dados dos vectores **x**, **y**, la función **meshgrid** crea una matriz cuyas filas son copias del vector **x** y una matriz cuyas columnas son copias del vector **y**. Luego en un tercer vector **z** se le asigna la función aplicada a **x**, **y**. Una vez dada esta orden, la gráfica de malla se genera mediante la función **mesh** obteniéndose así la gráfica tridimensional.

A.6.5.2.2. De superficie.

Es como la gráfica de malla, excepto que se rellenan los espacios entre líneas. Las gráficas de este tipo se generan usando la función **surf** (con los mismos argumentos que la función **mesh**).

A.7. Funciones de entrada / salida.

A.7.1. Por terminal.

Algunas de estas funciones son:

- **disp(x).** Presenta el contenido del parámetro sin indicar su nombre.

Las siguientes órdenes muestran cómo se usa esta función:

```
>> x=25;
>> disp(x)                % Presenta el contenido de x
25
```

- **format** ***opción*****.** Permite controlar la forma en que se realiza la presentación en pantalla de los números. Las principales opciones son:

short. Trata de representar 3 cifras significativas en un máximo de 10 caracteres.

long. Trata de representar 15 cifras significativas en un máximo de 24 caracteres.

Las siguientes órdenes ilustran cómo se usan estas funciones:

```
>> 1/3                  % Escribe en formato estándar que es el de 5 cifras decimales
ans = 0.33333
>> format short         % Escribe con 3 cifras decimales
>> 1/3
ans = 0.333
>> format               % Vuelve al formato estándar que es el de 5 cifras decimales
>> 1/3
ans = 0.33333
>> format long          % Escribe con 15 cifras decimales
>> 1/3
ans = 0.333333333333333
>> format               % Vuelve al formato estándar que es el de 5 cifras decimales
>> 1/3
ans = 0.33333
```

A.7.2. Por archivo.

Estas funciones permiten salvar y recuperar variables de un archivo:

- **save** <*nombre_de_archivo v1 v2 ...*>. Salva las variables indicadas o todas las del espacio de trabajo actual si no se indica ninguna en ese archivo.
- **load** <*nombre_de_archivo v1 v2 ...*>. Carga las variables especificadas del archivo o todas si no se especifican variables.

Esto puede verse en el siguiente ejemplo:

```
>> A=[1 2 3;2 3 1;4 5 6];
>> b=[1;2;3];
>> x=A\b
x =
    0.22222
    0.55556
   -0.11111
```

```
>> save prueba     % En el archivo prueba se graban todas las variables anteriores
                   con sus valores
```

Después, con un **clear**, o simplemente cuando se necesiten, se pueden volver a cargarlas en el espacio de trabajo y usarlas.

En este caso, escribimos las siguientes órdenes:

```
>> clear
>> load prueba     % Volvemos a tener en memoria las variables que habíamos
                   empleado antes
>> A
A =
      1      2      3
      2      3      1
      4      5      6

>> b
b =
      1
      2
      3

>> x
x =
    0.22222
    0.55556
   -0.11111
```

Saliendo y volviendo a entrar a Octave:

```
>> load prueba     % Volvemos a tener en memoria las variables que habíamos
                   empleado antes
>> A
A =
      1      2      3
      2      3      1
      4      5      6

>> b
b =
      1
      2
      3
```

```
>> x
x =
     0.22222
     0.55556
    -0.11111
```

A.8. Otras funciones de interés.

A.8.1. Generales del entorno.

exit y **help** son funciones generales del entorno. Otra es:

- **eval(string).** Ejecuta el comando representado por **string** en el espacio de trabajo actual.

La siguiente orden ilustra cómo se usa:

```
>> eval('2+3')
ans = 5
```

A.8.2. Control del tiempo.

Se muestran las principales funciones de este grupo y ejemplos.

- **tic toc.** Permiten cronometrar la duración de un proceso.

tic pone a cero el cronómetro, mientras que **toc** devuelve el número de segundos que han transcurrido.

```
>> tic                          % Cronómetro en cero
>> A=[1 2 3;2 3 1;4 5 6];       % Ingreso de la matriz cuadrada de coeficientes A
>> b=[1;2;3];                   % Ingreso   los coeficientes de la columna b
>> x=A\b                        % Solución del sistema lineal ingresado
x =
     0.22222
     0.55556
    -0.11111

>> toc                          % Tiempo (en segundos) transcurrido
ans = 16.742
```

• **clock.** Devuelve un vector de 6 componentes con: año, mes, día, hora, minuto y segundos actuales.

```
>> clock
ans =
    1.0e+03  *
    2.00600  0.00900  0.02200  0.01000  0.00200  0.03075
```

• **pause(segundos).** Suspende la ejecución del programa el número de segundos indicado. Si no se indican segundos se espera hasta que se pulse una tecla.

A.9. Programas.

Una forma eficiente de construir programas es crear nuevas funciones que se almacenan como archivos cuya extensión es **.m**. Estos programas nos permiten especificar los datos que deben introducirse y los resultados que deben mostrarse y pueden ser llamados como subprogramas desde otros programas.

A modo de ejemplo, a continuación se muestran los programas básicos que implementan algunos de los métodos numéricos que se analizan en este libro.

A.9.1. Programa para implementar el método de iteración de punto fijo.

Para poder ejecutar el programa que implementa el método de iteración de punto fijo, se deben definir previamente las funciones que se utilizan en el método (ecuación a resolver y función generadora de las iteraciones) y que son guardadas en archivos **.m**. Luego, estos archivos serán invocados desde el programa principal.

El archivo creado para guardar la ecuación, $f(x) = 0$, que se desea resolver en este caso particular es:

```
% Ecuación a resolver
function f=y(x);
f = x-exp(-x);
```

El archivo creado para guardar la ecuación que genera las iteraciones, $x = g(x)$, para resolver la ecuación anterior es:

```
% Ecuación generadora de las iteraciones
function g=y(x);
g = exp(-x);
```

Programa principal

```
disp('   ');
disp('==== Método de iteración de punto fijo ==========\n');
%Ingresar los valores iniciales para aplicar el método de iteración de punto
fijo
error=input('Ingresar el valor del error ----->  ');
vi=input('Ingresar el valor inicial ----->  ');
iteracion=0; e=abs(vi-g(vi));
while((abs(f(vi))>error)&(e>error))
  t=vi; vi=g(vi); e=abs(vi-t);
 if iteracion > 100  % Luego  de  100 iteraciones, se detiene la ejecución
            break;
      end
  iteracion=iteracion+1;
end
disp('   ');
% Muestra los resultados obtenidos
if iteracion < 100
  disp('La raíz aproximada de la función es:  ');
  raiz=vi
  disp(' ');
  disp('El valor de la función en la raíz aproximada es: ');
  valor_de_la_funcion=f(vi)
  disp(' ');
  fprintf('Las iteraciones necesarias fueron: %d\n',iteracion);
  disp(' ');
else
     fprintf('El método no converge luego de 100 iteraciones. %\n');
  end
```

A.9.2. Programa para implementar el método de eliminación de Gauss.

```
disp('   ');
disp('==== Método de eliminación de Gauss ==========\n');
% Ingresar la dimensión de la matriz asociada al sistema a resolver
n=input('Ingresar la dimensión de la matriz asociada al sistema a resolver: ');
% Ingresar los valores de la matriz ampliada (A|B)
for i=1:n

  for j=1:n+1
```

```
      fprintf('Ingresar el elemento A (%d,%d) ---> ',i,j');
      A(i,j)=input(' ');
  end
end
% Muestra la matriz ampliada (A|B)
disp('   ');
disp('La matriz ampliada (A|B) es:  ');A
% Triangula la matriz ampliada
for i=1:n-1
  for k=i+1:n
     m=A(k,i)/A(i,i);
     for j=1:n+1
       A(k,j)=A(k,j)- m * A(i,j);
     end
  end
end
% Calcula las soluciones del sistema
for k=n:-1:1
  if k==n
     s(k)=A(k,k+1)/A(k,k);
  else
    suma=0;
    for j=n:-1:k+1
      suma=suma+A(k,j)* s(j);
    end
    s(k)=(A(k,n+1)-suma)/A(k,k);
  end
end
% Muestra la matriz triangular superior
disp('   ');
disp('La matriz triangular superior es:  ');A
% Muestra las soluciones
for i=1:n
 fprintf('La solución x%d es:  %f\n',i,s(i));
end
disp('   ');
```

A.9.3. Programa para implementar la regla del punto medio de segmentos múltiples.

Para poder ejecutar el programa que implementa la regla del punto medio de segmentos múltiples, se debe definir previamente la función a integrar que se guarda en un archivo **.m**. Luego, este archivo será invocado

desde el programa principal. En este caso particular, la función a integrar es $f(x) = x\text{-}exp(\text{-}x)$.

El archivo creado para guardar la función, $y = f(x)$, a integrar es:

```
% Ecuación a resolver
function f=y(x);
f = x-exp(-x);
```

Programa principal

```
disp('   ');
disp('=Regla del punto medio de segmentos múltiples=\n');
% Se ingresan los límites de integración y la cantidad de segmentos
a=input('Ingresar el límite inferior de la integral ---> ');
b=input('Ingresar el límite superior de la integral ---> ');
n=input('Ingresar el número de segmentos ---> ');
h=(b-a)/n;
% Se inicializan los vectores x y f
x=zeros(n,1); f=zeros(n,1);
% Se calculan los valores de las abscisas y de la función en esos puntos
for i=1:n
  x(i,1)=a+i*h-h/2; f(i,1)=funcion(x(i,1));
end
disp('   ');
disp('Las abscisas (puntos medios) utilizadas son: ');x
disp('Los valores de la función en esos puntos son: ');f
disp('   ');
% Se calcula el valor aproximado de la integral
suma=0;
for i=1:n
   suma=suma+f(i,1);
end
% Se muestra el valor aproximado de la integral
disp('El valor aproximado de la integral es: ');
integral=h*suma
disp('   ');
```

EJERCICIOS PROPUESTOS .....

Alcanzado este punto y luego de realizar los ejercicios que se listan a continuación, utilizando el lenguaje Octave, el lector debería ser capaz de crear programas que implementen los métodos numéricos desarrollados en este libro.

1. i) Ingresar en la línea de comandos la sentencia:

help format

y ver, principalmente, **format long** y **format short**. Luego ingresar la sentencia:

format long

ii) A continuación, realizar los siguientes apartados ingresando la operación correspondiente en la línea de comandos:

a) $\frac{423^2 \ 0.00006}{328}$-325 b) $\frac{log(1238)\ \sqrt[3]{\frac{3}{5}}}{e^3}$ c) 124/0

d) % 425/386 e) 16/∞ f) $3\ \sqrt{-64}+5$

iii) ¿Qué sucede en el punto d)?

iv) Ingresar en la línea de comandos la sentencia:

format short

v) Asignarle a la variable **x** la operación dada en a), a la variable **y** la dada en b) y a la variable **z** la dada en c). (Utilizar las flechas del teclado para obtener las sentencias ingresadas anteriormente).

vi) Ingresar en la línea de comandos la sentencia:

x,y,z

vii) Ingresar en la línea de comandos la sentencia:

x;y;z

viii) ¿Qué diferencias se pueden establecer entre las últimas dos entradas?

ix) Ingresar en la línea de comandos las sentencias:

a) **x >y** b) **y==x** c) **z<y**

x) ¿Qué respuestas obtiene?

xi) Ingresar en la línea de comandos la función:

clear

xii) Ingresar en la línea de comandos la sentencia:

x

xiii) ¿Qué conclusión puede extraer?

2. Resolver los siguientes problemas utilizando Octave como una herramienta para el cálculo de las operaciones necesarias. Cada línea de comando que se utilice debe estar comentada (**%** o #). Luego, copiar la sesión y pegarla en un archivo de texto. Guardar a este archivo con el nombre problema.

Se lanza un proyectil cuya trayectoria está dada por la función:

$$y = -x^2 + 4.25x$$

Indicar en qué momento x alcanza la mayor altura y de cuánto es la misma.

3. a) Asignarle a la variable **A** la siguiente matriz:

$$\begin{pmatrix} 2 & 24 & 4 & 10 \\ 12 & -4 & 5 & 8 \\ -6 & 71 & 27 & 0 \\ 4 & 2 & 26 & 1 \end{pmatrix}$$

b) Calcular el determinante de **A**.

c) Obtener el producto de la diagonal principal de **A** (realizar el producto elemento a elemento).
d) Obtener una matriz de nombre **B** de 4 filas y 5 columnas cuyos elementos sean todos 1.
e) Mostrar la columna 3 de la matriz **A**.
f) Utilizando el operador \ obtener una matriz **D** que sea la inversa de **A**.
g) Comprobar el resultado obtenido realizando la operación: **A*D**.
h) Obtener la inversa de **A** con la función **inv**.

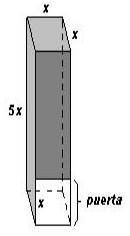
4. a) Ingresar la matriz asociada al sistema de ecuaciones lineales que resuelve el siguiente problema. Luego, hallar la solución correspondiente usando el método de Cramer.

Tres grifos han llenado un depósito de 11 m^3; el grifo P está abierto durante 1 hora, el grifo Q durante 2 horas y el grifo R durante 2 horas. Seguidamente llenan otro depósito de 16 m^3; el grifo P vierte agua durante 4 horas, el grifo Q 3 horas y el grifo R 2 horas. Finalmente, llenan otro depósito de 10 m^3, utilizando el grifo P 3 horas, el grifo Q 2 horas y el grifo R 1 hora. ¿Cuántos litros vierte por hora cada grifo?

b) Se pide ahora, resolver el sistema anterior de la forma Ax = B, multiplicando por la inversa en forma apropiada.

5. a) Ingresar el polinomio $p1(x) = -3x^4 + 6x^3 + 12x + 15$ como un vector, considerando cada coeficiente del polinomio completo como los elementos del vector.
b) Evaluar este polinomio para $x = 8$ utilizando la función **polyval**.
c) Calcular las raíces del polinomio $p1(x)$ utilizando la función **roots**.

6. a) Resolver el siguiente problema:



Las empresas de subtes desean construir ascensores para permitir el acceso de personas con discapacidad. Para cada ascensor se debe excavar un hueco desde la vereda con forma de prisma con base cuadrada de x metros de lado y profundidad igual al quíntuplo de la base. La base y las paredes del hueco deben estar totalmente cubiertas con cemento y deben llevar además, a lo largo de sus cuatro aristas verticales, potentes vigas de hierro para evitar su distorsión. Una superficie cuadrada igual a la base debe quedar libre para la puerta.

Para presupuestar la obra se consideran los siguientes gastos:

Máquina excavadora \$ 60 el m^3, cemento \$ 100 el m^2 y vigas de hierro \$ 30 el metro. Calcular las dimensiones del hueco del ascensor si su costo fue de \$27900.

b) Graficar las tres funciones polinómicas resultantes para corroborar los resultados obtenidos. Utilizar la función **plot**.

7. Utilizando la función **plot** graficar las siguientes funciones con las características que se indican:

a) $y = (sen(x))^2$ en el intervalo [-6, 6]
b) $y = abs(x)$ en el intervalo [-5, 5]

8. Utilizando la función **plot** hallar la raíz real exacta de la siguiente ecuación:

$$3x^3 + 12x - 8x^2 - 32 = 0$$

Utilizar la función **grid on** para facilitar la visualización del gráfico.

9. Para x variando desde -0.5 hasta 15 con paso 0.1, representar las dos funciones siguientes en un mismo grafico y hallar los intervalos en donde ambas se intersecan:

$y_1 = 2x - 1$ $y_2 = (1/5)\ x\ \hat{}\ 2$

10. Utilizando la función **plot** graficar la función $y = x^3 - 5x^2 + 6x$ en el intervalo [0, 3]. Luego colocar las siguientes etiquetas en la grafica:

a) Título: "Grafica de la función polinómica"
b) Eje x: x
c) Eje y: y
d) Colocar un texto que indique el punto mínimo de la gráfica en este intervalo.

11. Utilizando la función **plot** representar en un mismo gráfico las siguientes funciones variando los colores y tipo de trazo:

$y = sen(x)$ $y = 3sen(x)$ $y = sen(3x)$

12. Graficar en los mismos ejes coordenados las funciones dadas en el ejercicio anterior utilizando la función **fplot**.

13. Utilizando la función **subplot** graficar en una misma ventana las tres funciones dadas en el ejercicio 11.

14. Utilizando la función **fplot** graficar las siguientes funciones indicando los intervalos adecuados para x e y que permitan visualizar las gráficas:

$$y = cos(x)/x \qquad y = (tg(x)/x)^2$$

15. Utilizando la función **fsolve** resolver las siguientes ecuaciones no lineales:

a) $y = log(x) - 3x^2 + 2$
b) $y = sen(x) - x^2 - 4$

16. Resolver los siguientes sistemas de ecuaciones lineales de tres formas distintas utilizando el operador \, la función **rref** y la función **inv( )**:

a) $\begin{cases} 3x + 5y - 6z = -7 \\ x + y + z = 6 \\ 7x - y - z = 0 \end{cases}$ b) $\begin{cases} 3x + 5y = 81 \\ 6x + 10y = -23 \end{cases}$

17. Crear una función por medio de un archivo **.m** que contenga a las funciones no lineales del siguiente sistema y que usando la función **fsolve** se pueda resolver:

$$\begin{cases} 3x^3 + 4xy - 3yz = 4 \\ 2x^2 - y^2 - z = 0 \\ \frac{3}{x} - 4y - 6xz = -7 \end{cases}$$

18. Ejecutar los siguientes comandos para obtener una parábola en el espacio:

```
>> x=-2:0.1:2;
>> plot3(x,x,x.^2);
>> xlabel('x');
>> ylabel('y');
>> zlabel('z');
>> replot
```

Una vez realizado el gráfico haciendo clic sobre él con el mouse se podrá mover, obteniendo así diferentes ángulos desde los cuales se observará dicho grafico.

19. Con las funciones **meshgrid** y **mesh** graficar:

a) $z = x^3 - xy$
b) $z = log(x^2) - y$

20. Realizar los siguientes programas que permitan:

a) Ingresar un número natural y luego muestre todos los números pares menores que él.
b) Calcular las raíces de una ecuación de segundo grado ingresando los valores de a, b y c. Además, el programa deberá indicar las características de las raíces.
c) Calcular el valor de la hipotenusa de un triángulo ingresando los valores de sus catetos.
d) Ingresar un número e indique si es primo.
e) Generar números aleatorios y que el usuario decida cuando terminar ingresando su opción desde el techado.
f) Ingresar una matriz y luego muestre:
- El mayor de sus elementos.
- El menor de sus elementos.
- El promedio de sus elementos.

g) Ingresar dos matrices y calcular el producto.

……

Bibliografía

Atkinson, K., 1988, *An Introduction to Numerical Analysis*, 2ª Ed., John Wiley & Sons, N. Y.

Borrelli Nogueras, G., 2005, Manual: *Introducción Informal a MATLAB y Octave*, Calella, España. Disponible en:

http://torroja.dmt.upm.es/~guillem/matlab/index.html

Burden, R. y Faires, J., 2002, *Análisis Numérico*, International Thomson Editores, México. (Versión en inglés, 2001: *Numerical Analysis*, 7ª Ed., Brooks/Cole)

Cohen, A., Cutts, J., Fielder, R., Jones, D., Ribbans, J. y Stuart, E., 1977, *Análisis Numérico*, Editorial Reverté, S. A., España. (Versión en inglés, 1976: *Numerical Analysis*, Mc Graw – Hill Book Company (UK) Limited, England)

Conte, S. y Boor, C., 1980, *Elementary Numerical Analysis and Algorithmic Approach*, 3ª Ed., Mc Graw - Hill, N. Y.

Chapra, S. y Canale, R., 1988, *Métodos Numéricos para Ingenieros*. *Con Aplicaciones en Computadoras Personales*, Mc Graw - Hill / Interamericana de México, S. A. de C. V., México. (Versión en inglés, 1985: *Numerical Methods for Engineers with Personal Computer Applications*, 1ª Ed., Mc Graw - Hill, N. Y.)

Chapra, S. y Canale, R., 2003, *Métodos Numéricos para Ingenieros. Con Programas de Aplicación*, Mc Graw - Hill / Interamericana de España, S. A. U., España. (Versión en inglés, 2004: *Numerical Methods for Engineers with Software*, 4ª. Ed., Mc Graw - Hill, N. Y.)

Chapra, S. y Canale, R., 2005, *Numerical Methods for Engineers with Engineering Subscription Card*, 5ª Ed., Mc Graw - Hill / Interamericana de España, S. A. U., España.

Chapra, S. y Canale, R., 2007, *Applied Numerical Methods with MATLAB*, 2ª Ed., Mc Graw - Hill / Interamericana de España, S. A. U., España.

Chapra, S. y Canale, R., 2007, *Métodos Numéricos para Ingenieros*, 5ª Ed., Mc Graw - Hill / Interamericana de España, S. A. U., España.

Dahlquist, G. y Björck, A., 1974, *Numerical Methods*, Prentice Hall, Englewood Cliffs, N. J.

Demidovich, B. y Maron, I., 1993, *Cálculo Numérico Fundamental*, 4ª Ed., Editorial Paraninfo, S. A., España.

Demidovich, B., Maron, I. y Schuwalowa, E., 1980, *Métodos Numéricos de Análisis*, Editorial Paraninfo, S. A., España.

Eaton, J. W., 1997, Manual y software: *Octave: (octave). Interactive Language for Numerical Computations*, Versión 2.1.x. Disponible en:

http://www.gnu.org/software/octave/doc/interpreter/index.html

Eaton, J. W., 2002, *GNU Octave Manual. A high-level Interactive Language for Numerical Computations*, Publisher: Network Theory Ltd., Free License: GNU General Public License, Edition 3 for Octave version 2.0.13.

Forsythe, G., Malcom, M. y Moler, C., 1977, *Computer Methods for Mathematical Computation*, Prentice Hall, Englewood Cliffs, N. J.

Forsythe, G. y Moler, C., 1973, *Solución Mediante Computadoras de Sistemas Algebraicos Lineales*, Eudeba. (Versión en inglés, 1967: *Computer Solution of Linear Algebraic Systems*, Prentice Hall, Englewood Cliffs, N. J.)

García Rojo, J., 2003, Manual: *Herramientas en GNU/Linux para Estudiantes Universitarios. GNU/Octave: Cálculo Numérico por ordenador*, Free Software Foundation, Inc. 59 Temple Place, Boston, USA. Disponible en:

http://softwarelibre.unsa.edu.ar/docs/index.html

Gerald, C. y Wheatley, P., 2000, *Análisis Numérico con Aplicaciones*, 6ª Ed., Pearson Educación, México. (Versión en inglés, 1999: *Applied Numerical Analysis*, Sixth Ed., Addison Wesley)

Golubitsky, M. y Dellnitz, M., 2001, *Álgebra Lineal y Ecuaciones Diferenciales, con uso de MATLAB*, International Thomson Editores. (Versión en inglés, 1999: *Linear Algebra and Differential Equations using MATLAB*, Brooks/Cole Publishing Company)

González, H., 2002, *Análisis Numérico. Primer Curso*, 1ª Ed., Editorial Nueva Librería S. R. L., Buenos Aires, Argentina.

Gordon, J., 1985, *Algoritmos Numéricos*, La Plata, Buenos Aires, Argentina.

Hamilton Castro, A. F., 2004, *Introducción al Octave*, Grupo de Computadoras y Control, Departamento de Física, Electrónica y Sistemas de la ULL, España. Disponible en:

http://cyc.dfis.ull.es/asignaturas/Curso2004-2005/octave/ApuntesOctave/ApuntesOctave.html

Henrici, P., 1972, *Elementos de Análisis Numérico*, Editorial Trillas, México. (Versión en inglés, 1964: *Elements of Numerical Analysis*, John Wiley & Sons, N. Y.)

Henrici, P., 1982, *Essentials of Numerical Analysis with Pocket Calculator Demonstrations*, John Wiley & Sons, N. Y.

Isaacson, E. y Keller, H., 1966, *Analysis of Numerical Methods*, John Wiley & Sons, N. Y.

Kincaid, D. y Cheney, W., 1994, *Análisis Numérico. Las Matemáticas del Cálculo Científico*, Addison - Wesley Iberoamericana, S. A., E. U. A. (Versión en inglés, 1991: *Numerical Analysis: Mathematics of Scientific Computing*, Brooks/Cole Publishing Company, E. U. A.)

Long, P. J., 2005, *Introduction to Octave*, Department of Engineering, University of Cambridge. Disponible en:

www.mdp.eng.cam.ac.uk/CD/engapps/octave/octavetut.pdf

Mathews, J. y Fink, K., 2000, *Métodos Numéricos con MATLAB*, 3ª Ed., Prentice Hall, España. (Versión en inglés, 1999: *Numerical Methods using MATLAB*, Prentice Hall)

Medrano, C., Valiente, J. M., Plaza, L. y Ramos, P., 2006, *Evaluación de Herramientas de Software Libre para Cálculo Numérico*. Disponible en: http://www.euitt.upm.es/taee06/papers/S1/p46.pdf

Nakamura, S., 1992, *Métodos Numéricos Aplicados con Software*, 1ª Ed., Prentice Hall Hispanoamericana, S. A., México. (Versión en inglés, 1991: *Applied Numerical Methods with Software*, Prentice Hall)

Nakamura, S., 1997, *Análisis Numérico y Visualización Gráfica con MATLAB*, Pearson Educación, México. (Versión en inglés, 1996: *Numerical Analysis and Graphic Visualization with MATLAB*, Prentice Hall)

Quintela Estévez, P., 2000, *Matemáticas en Ingeniería con MATLAB*, Servicio de Publicaciones de la Universidad de Santiago de Compostela, España.

Ralston, A. y Rabinowitz, P., 1978, *A First Course in Numerical Analysis*, 2ª Ed., Mc Graw - Hill, N. Y. (Versión en español, 1986: *Introducción al Análisis Numérico*, Editorial Limusa, S. A. de C. V., México)

Sadosky, M., 1955, *Cálculo Numérico y Gráfico*, Librería del Colegio, Buenos Aires, Argentina.

Sánchez de la Rosa, L. L., Manual: *MATLAB / Octave*. Disponible en:

http://nereida.deioc.ull.es/~pcgull/ihiu01/cdrom/matlab/contenido/matlab.html

Sánchez, J. y Souto, A., 2005, *Problemas de Cálculo Numérico para Ingenieros con Aplicaciones MATLAB*, 1ª Ed., Mc Graw - Hill / Interamericana de España, S. A. U., España.

Scheid, F., 1988, *Análisis Numérico*, Mc Graw - Hill, N. Y. (Versión en inglés, 1988: *Numerical Analysis*, Mc Graw - Hill, N. Y.)

Scheid, F. y Di Costanzo, R., 1991, *Métodos Numéricos*, Mc Graw - Hill, N. Y. (Versión en inglés, 1988: *Numerical Analysis*, Mc Graw - Hill, N. Y.)