Programación Científica

Dr. en C. Luis Fernando Gutiérrez Marfileño

Universidad Autónoma de Aguascalientes
Centro de Ciencias Básicas
Depto. de Ciencias de la Computación
Academia de Inteligencia Artificial y Fundamentos Computacionales

23 de Enero de 2023

Índice general

Índice de figuras

Índice de tablas

Índice de algoritmos

Índice de definiciones

1

INTRODUCCIÓN A LA PROGRAMACIÓN CIENTÍFICA

"Comprender la importancia de los métodos aproximados y las implicaciones del uso de la computadora en soluciones aproximadas."

Objetivos particulares 1 y 2

1. Introducción a la Programación Científica

La **Programación Científica** es el área de las Ciencias Computacionales que se encarga de implementar una tercera herramienta para la solución de problemas que tienen que ver con la ciencia y la ingeniería, que es la **simulación computacional**[1] (las dos primeras son la **teoría** y la **experimentación**).

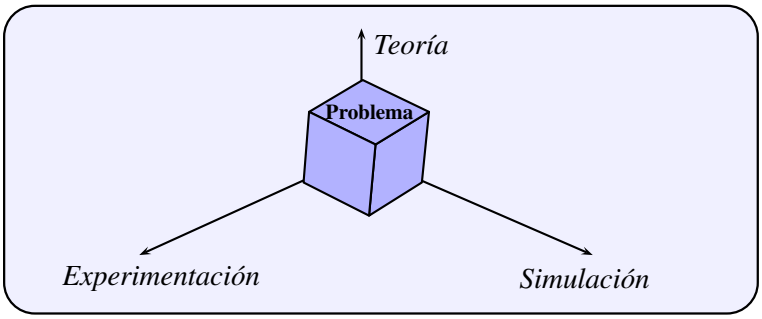

Figura 1.1: Enfoque computacional para la solución de problemas

[1] **Simulación computacional**. Es la representación y emulación de un sistema físico o proceso mediante una computadora.

El uso de las computadoras como un medio para simular muchos procesos requiere de un conocimiento a fondo de la implementación de modelos matemáticos utilizando lenguajes de computación.

DEFINICIÓN (1.1 ▷ Computación Científica)

Es la colección de herramientas, técnicas y teorías necesarias para resolver en una computadora modelos matemáticos de problemas en Ciencias e Ingeniería [Golub & Ortega, 1992].

HISTORIA

En 1945, IBM funda el Laboratorio de Computación Científica Watson en la Universidad de Columbia en la ciudad de Nueva York.

1: **Computación científica**

Actualmente el concepto de computación científica abarca todo el procedimiento, incluido el análisis del modelo matemático, el desarrollo y el análisis de un método numérico, la programación del algoritmo resultante y, finalmente, su ejecución en una computadora [Gustafsson, 2018].
En la vida cotidiana la **ingeniería** es el área que se encarga de aplicar principios científicos y matemáticos para desarrollar soluciones económicas a problemas técnicos [US Department of Labor].
Entre los principios de las matemáticas aplicadas, los modelos brindan una forma simplificada de representar fenómenos y resolver problemas asociados a ellos empleando el enfoque de sistemas.
Una **solución analítica** es una expresión matemática que proporciona toda la información sobre el comportamiento de un sistema, para cualquier valor de las variables y parámetros que intervienen en él, es una solución general (simbólica).
Por otro lado, una **solución numérica** expresa el comportamiento del sistema en función de números que se obtienen resolviendo las ecuaciones del sistema para valores concretos de sus variables y parámetros.
La parte cuya función es diseñar métodos para *aproximar* de forma eficiente las soluciones de problemas expresados matemáticamente se llama **Análisis Numérico** y su principal objetivo es encontrar soluciones a problemas complejos utilizando sólo operaciones aritméticas simples.
Se requiere de una secuencia de operaciones algebraicas y/o lógicas que producen la aproximación de la solución matemática al problema.
Así, el termino *solución numérica* se considera a menudo opuesto al de *solución analítica* de un problema, la diferencia entre ambas es sustancial.

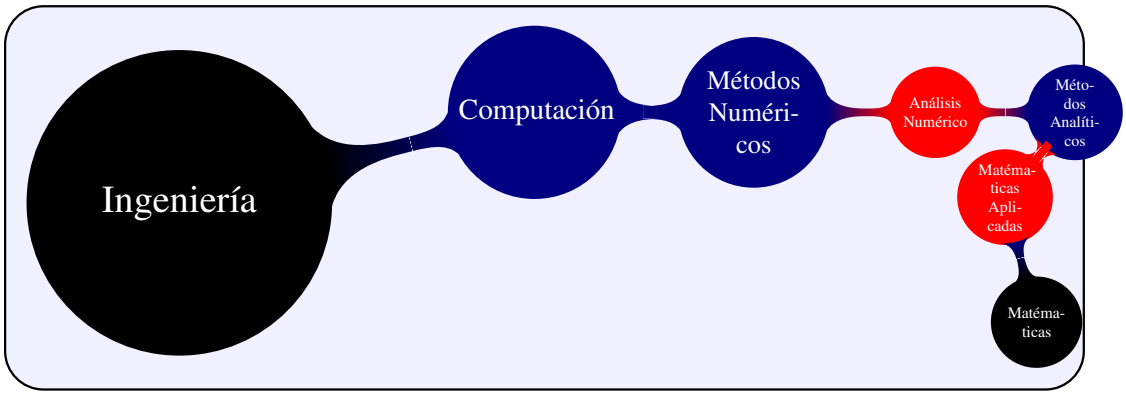

Figura 1.2: Métodos numéricos

Los **métodos numéricos** buscan números, mientras que los métodos analíticos buscan fórmulas matemáticas. Obviamente la solución analítica, al ser general, contiene todas las soluciones numéricas, mientras que, a partir de la solución numérica, es prácticamente imposible deducir la solución analítica.

2: **Métodos numéri-cos**

DEFINICIÓN **(1.2 ▷ Métodos Numéricos)**

Son técnicas mediante las cuales es posible formular problemas matemáticos de tal forma que puedan resolverse usando operaciones aritméticas [Chapra & Canale, 2008].

El objetivo de los métodos numéricos es reproducir lo más fielmente el comportamiento de fenómenos reales a través de números (esto se realiza empleando modelos matemáticos de dichos fenómenos), los cuales, expresados por medio de gráficos representan una visión de la realidad física, obviamente la solución numérica solo coincidirá con dicha realidad si:

- El modelo matemático incorpora todos los aspectos del mundo real
- El método numérico empleado puede resolver exactamente las ecuaciones de dicho modelo

En la práctica ninguna de estas condiciones se cumple por lo que la predicción numérica no coincidirá con el comportamiento del mundo real.
Por tanto se dice que la **solución numérica** *aproxima* la **solución real**.
Si se conoce la solución real del problema se puede comparar con la solución numérica y obtener el **error** de la predicción.

Los métodos numéricos pueden ser aplicados para resolver procedimientos matemáticos como:

- Facilitar operaciones con funciones
- Obtención de raíces
- Operaciones con matrices
- Ajuste de curvas
- Realizar interpolaciones
- Evaluar integrales

En la actualidad gran parte de la tecnología depende de la solución de modelos matemáticos, como ya se mencionó, éstas soluciones pueden obtenerse de manera analítica pero en muchos casos estas pueden no existir o ser demasiado complejas, por lo que se recurre a los métodos numéricos para aproximar dichas soluciones dentro de márgenes definidos de tolerancia.
Algunos ejemplos de las áreas en las que dichos métodos se aplican son:

- Ingeniería Industrial
- Ingeniería Química
- Ingeniería Civil
- Ingeniería Mecánica
- Ingeniería Eléctrica, etc...

1.1. Origen de los Métodos Numéricos

El enfoque mediante métodos numéricos es conocido desde hace miles de años, muchos de los grandes matemáticos del pasado trabajaron en el análisis numérico. Un ejemplo clásico de solución numérica fue la obtenida por Arquímedes (ver Fig.1.3) para obtener el valor aproximado del numero π a partir de la división de una circunferencia en polígonos obtenidos incrementando el número de lados y dividiendo el perímetro de cada polígono entre el radio del circulo, entre más aumente el número de lados de cada polígono, se obtiene una mayor precisión del valor de π [Rodríguez, 2006].
Arquímedes utilizó polígonos inscritos y circunscritos de 96 lados (ver Fig. 1.4) en una circunferencia y logró obtener un valor de π entre 3.14084 y 3.14285.

Figura 1.3: Arquímedes (Siracusa, 287a.C. - 212 a.C.)

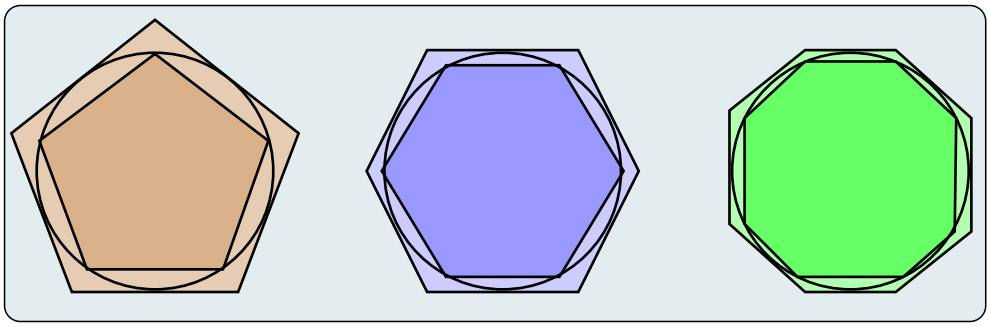

Figura 1.4: Método de aproximación del número π de Arquímedes

La técnica empleada por Arquímedes reúne las características de los métodos numéricos actuales:

- La solución numérica se obtiene dividiendo el dominio que se estudia (la circunferencia) mediante elementos (geométricos) sencillos (rectas), de los que se conocen todas sus propiedades (longitud).
- La solución numérica es aproximada y mejora (converge) al incrementar el número de divisiones del dominio.
- La solución numérica es la única alternativa ya que solución exacta del problema es desconocida (el valor exacto de π es inconmensurable).

La idea de que todo puede representarse por medio de números se ha ido fortaleciendo a lo largo del tiempo.
Pitágoras (ver Fig. 1.5), filósofo y matemático griego, fue un ferviente creyente de que el universo físico puede describirse de manera consistente en función de números.

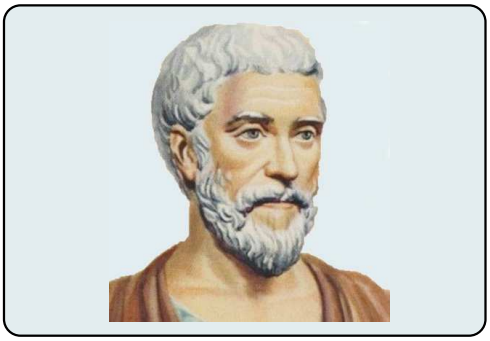

Figura 1.5: Pitágoras (Samos Jonia, 569a.C. - 475a.C.)

Y que, las conclusiones alcanzadas mediante razonamientos matemáticos son de mayor certeza que las obtenidas de cualquier otra forma.

Siglos más tarde Descartes (ver Fig 1.6) asumió y defendió la realidad objetiva de los métodos matemáticos.
Descartes creía que las ideas son entidades auto-existentes, extra-espaciales, extra-temporales, independientes de los hombres, intercambiables, perfectas y eternas, no creadas por la mente pero comprensibles y conocidas solo por medio de la razón a través de la dialéctica[2], no de los sentidos.

Figura 1.6: Rene Descartes (La Haye Francia, 1596 - 1650)

[2]**Dialéctica**. Teoría y técnica retórica de dialogar y discutir para descubrir la verdad mediante la exposición y confrontación de razonamientos y argumentaciones contrarios entre sí.

1.2. Ventajas y desventajas de los Métodos Numéricos

A continuación se mencionan algunas ventajas que se obtienen mediante el uso de los métodos numéricos:

- Permiten aproximar soluciones de ecuaciones no resolubles por otros métodos.
- Son más rápidos (si están desarrollados en software), en la mayoría de los casos.
- Una vez desarrollados y probados los algoritmos, tiene gran fiabilidad (dentro de los márgenes de error establecidos y con las condiciones iniciales adecuadas), y pueden manejar gran número de ecuaciones y variables sin errores de operación.

Algunas de las desventajas al utilizar métodos numéricos:

- No son 100 % precisos
- A menudo consumen mucha capacidad de proceso
- No todos los problemas se pueden resolver por métodos numéricos
- No avanzan hacia soluciones generales, teniendo que procesarse cada caso particular

1.3. Cómo elegir correctamente el Método Numérico

En general, éstos métodos se aplican cuando se necesita un valor numérico como solución a un problema matemático y, los procedimientos *exactos* o *analíticos* (manipulaciones algebraicas, teoría de ecuaciones diferenciales, métodos de integración, etc...) son incapaces de dar una respuesta. Debido a ello, son procedimientos de uso frecuente por físicos e ingenieros, y cuyo desarrollo se ha visto favorecido por la necesidad de éstos de obtener soluciones, aunque la precisión no sea completa.
Los problemas de esta disciplina se pueden dividir en dos grupos fundamentales:

- **Problemas de dimensión finita**: aquellos cuya respuesta son un conjunto finito de números, como las ecuaciones algebraicas, los determinantes, los problemas de valores propios, etc...
- **Problemas de dimensión infinita**: problemas en cuya solución o planteamiento intervienen elementos descritos por una cantidad infinita de números, como integración y derivación numéricas, cálculo de ecuaciones diferenciales, interpolación, etc...

Asimismo, existe una subclasificación de estos dos grandes apartados en tres categorías de problemas, atendiendo a su naturaleza o motivación para el empleo del cálculo numérico:

1. Problemas de tal complejidad que no poseen solución analítica.
2. Problemas en los cuales existe una solución analítica, pero ésta, por complejidad u otros motivos, no puede explotarse de forma sencilla en la práctica.
3. Problemas para los cuales existen métodos sencillos pero que, para elementos que se emplean en la práctica, requieren una cantidad de cálculos excesiva; mayor que la necesaria para un método numérico.

2. Representación computacional de los números

EL concepto de **numero** es difícil de definir y ha variado a lo largo del tiempo, por ejemplo para los antiguos griegos sólo los *naturales* (enteros positivos) podían considerarse como tales.
En la actualidad se aceptan las siguientes clasificaciones:

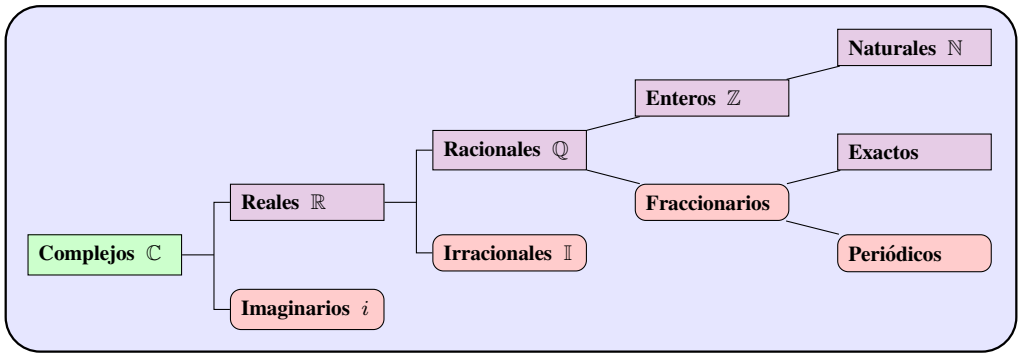

Figura 1.7: Tipos de números

Todos estos conjuntos tienen una capacidad infinita, lo cual es un concepto entendible por la mente humana (baste pensar en un número muy grande y agregarle un uno), pero los dispositivos mecánicos solo pueden representar y manejar conjuntos de números finitos y de precisión determinada.
Otro problema, en cualquier sistema numérico, es la representación de números grandes o muy pequeños, ya que requieren de mucho espacio, una forma de resolver esto es mediante la **notación científica**[3] , en la que se emplea la base del sistema y el exponente, lo que permite escalar el espacio requerido.
Las computadoras pueden representar las cantidades numéricas básicamente de dos formas, mediante números *enteros* y números *reales*.

2.1. Números Enteros

Los **números enteros** son una generalización de los números naturales, en este caso, un problema de representación adicional es la inclusión del signo, los símbolos empleados para el signo del número, para los negativos se antecede el (-) y para los positivos (aunque a veces suele omitirse) el (+).

3: **Números enteros**

[3] La **Notación científica** es una forma rápida y compacta de representar un número decimal mediante potencias de 10.

DEFINICIÓN (1.3 ▷ Enteros)

Esta formado por el conjunto de los números naturales[a], sus opuestos (negativos) y el cero, y se representa por:

$$\mathbb{Z} = \{..., -3, -2, -1, 0, +1, +2, +3, +...\}$$

[a] Los números naturales son aquellos que se emplean para contar cosas

Una forma de evitar este problema es representar la cantidad sin signo, aunque en algunos casos es poco práctico.
Como las computadoras solamente utilizan el sistema binario para la representación y manipulación de datos (números, texto, video, sonido), la representación de números enteros entre computadoras sería en binario directo, o si la información proviene de un ser humano habría que realizar la conversión de decimal a binario.

Números con signo

Para incluir el signo en números binarios en las computadoras se pueden emplear las siguientes formas:

- Notación signo magnitud (**SM**).
- Notación **Complemento a 1**.
- Notación **Complemento a 2**.

Formato SM

En el formato **SM** (ver Fig. 1.8), al conjunto de los bits que representa la magnitud del número se antepone (en la posición del bit más significativo **MSB**) el denominado *bit de signo*, que toma el valor de 0 para los números positivos, y 1 para los negativos.

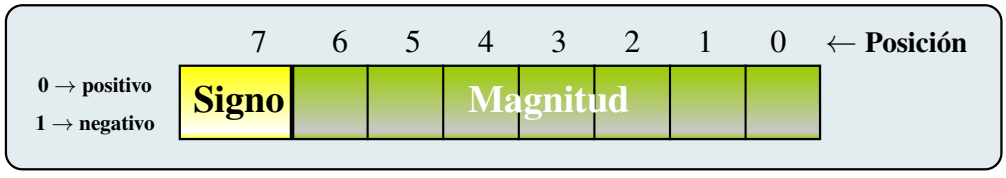

Figura 1.8: Formato **SM** para un *byte*

Formato Complemento a 1 (C1)

Los números positivos en formato de complemento a 1 se representan igual que los números positivos del formato **SM**. Sin embargo, los números negativos son el complemento a 1 del correspondiente número positivo. La operación de complemento implica cambiar cada bit por su opuesto (es decir, 1 por 0 y 0 por 1).

Formato Complemento a 2 (C2)

Los números positivos en formato de complemento a 2 se representan igual que los números positivos del formato **SM** y **C1**. Sin embargo, los números negativos son el complemento a 2 del correspondiente número positivo. La operación de complemento a 2 implica sumar un 1 al numero en **C1**.

Ejemplo (- 1)

Representar el número 2356_{10} en binario con signo en formato **SM** mediante 16 bits (como el signo es positivo, el MSB de los 16 es 0 y los siguientes a la derecha son 0´s hasta completar los 16 incluyendo el equivalente en binario).

$$2356_{10} = 0000100100110100_2$$

El -2356_{10} queda:

$$-2356_{10} = 1000100100110100_2$$

El -2356_{10} en complemento a 1 queda:

$$-2356_{10} = 1111011011001011_2$$

El -2356_{10} en complemento a 2 queda:

$$-2356_{10} = 1111011011001100_2$$

2.2. Números Reales

Los **números reales** son *todos los números* en la recta numérica.
Hay infinitamente muchos números reales así como hay infinitamente muchos números en cada uno de los conjuntos que los forman.

4: **Números reales**

DEFINICIÓN (1.4 ▷ Reales)

Esta formado por el conjunto de todos los números racionales[a], y todos los irracionales[b], y se representa por:

$$\mathbb{R} = \{..., -1, ..., -\frac{1}{2}, ..., -\frac{1}{3}, ..., 0, ..., +\frac{1}{3}, ..., +\frac{1}{2}, ..., +1, ...\}$$

[a] Los números racionales son aquellos números que pueden ser expresados como una relación exacta entre dos enteros.

[b] Los números irracionales son aquellos números que poseen infinitas cifras decimales no periódicas, que por lo tanto son fracciones irreductibles

En computación, los números reales pueden representarse de dos formas:

- Notación de **punto fijo**
- Notación de **punto flotante**

Formato de punto fijo

La representación de punto fijo consiste en que dado un espacio de n dígitos para almacenar un número N, se reservan s dígitos para el signo, M dígitos para almacenar la parte entera del número y f dígitos para almacenar la parte fraccional, respecto a cierta base β.
De esta forma el punto de la representación fraccional queda fijo en la *M-ésima* posición de la secuencia de dígitos (ver Fig. 1.9).

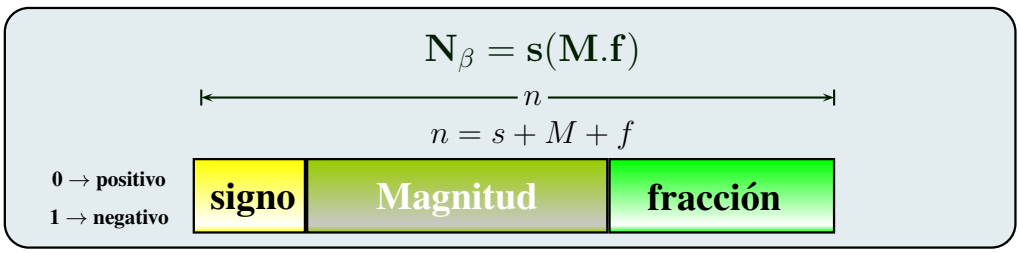

Figura 1.9: Formato de punto fijo para números reales

Ejemplo (- 2)

Suponga que $\beta = 10$, $N = 11$ y $f = 6$; entonces, se dispone de 6 dígitos para la parte fraccionaria y $N - f - 1 = 4$ dígitos para la parte entera. Para las siguientes cantidades determine el formato de punto fijo:
-30.412 : 1 0030 421000
0.0437 : 0 0000 043700

Para una longitud de palabra N y con f dígitos fijos para la parte fraccionaria, el rango de valores de los números reales que pueden representarse se encuentra dentro del intervalo $[-\beta^{N-k}, \beta^{N-k}]$.
Este rango es bastante limitado, los números muy grandes o las fracciones muy pequeñas no pueden representarse (a menos que se use una longitud de palabra N muy grande). Esta es una de las causas por las que la representación de punto fijo no es utilizada para implementar la representación de números reales. Sin embargo, con la convención $f = 0$, la representación de punto fijo puede ser útil para representar números enteros.
Considerando el sistema binario ($\beta = 2$), dada una palabra de N bits, existen 2^N combinaciones distintas que pueden generarse. Con ellas se quiere representar los enteros positivos, negativos y el cero.
Existen varias maneras de proceder, pero todas ellas tratan al bit más significativo (**MSB**) de la palabra como un bit de signo. Si dicho bit es 0, el entero es **positivo**, y si es 1, es **negativo**. Aquí también es aplicable la representación **SM**. En una palabra de N bits, mientras que el bit más significativo indica el signo del entero, los restantes $N - 1$ bits a la derecha representan la magnitud del entero.
Otra limitación de esta forma de representación tiene que ver con las operaciones aritméticas de suma y resta que requieren separar el bit que representa el signo de los restantes para llevar a cabo la operación en cuestión.
Además de que existen dos representaciones para el número 0, las cuales son:

$$000\cdots00, \qquad 1000\cdots00$$

Puesto que se tienen dos representaciones del cero, en la representación **SM** se representan la misma cantidad de enteros positivos que negativos.
El rango de enteros que puede representarse comprende entonces al intervalo

$$[-(2^{N-1}-1), 2^{N-1}-1]$$

Ejemplo (- 3)

Suponga que $\beta = 2$, $N = 32$ y $f = 16$; entonces, se dispone de $N - f - 1 = 15$ bits para la parte entera. Para las siguientes cantidades determine su representación en este formato de punto fijo:
Directamente en binario:
$26.32_{10} = 11010.01010001111010111000_2$
En punto fijo queda:
$0\,000000000011010010100011110101110000_2$
Para la siguiente cantidad:
$-125.42_{10} = 1000000001111101011010111000010 1_2$

Formato de punto flotante

Evidentemente, el mayor problema que tiene la representación de punto fijo es que limita el rango posible de números. Si se pudiese mover o *flotar* el punto libremente entre los n bits, se podría representar un rango mayor, para permitir trabajar tanto con números muy grandes como con números muy pequeños. La representación usada para esto se denomina representación de **punto flotante**.
Para lograr que el punto *flote* se debe codificar de alguna forma para cada número la posición actual del punto. Una representación decimal que permite esto es la representación de *notación científica*, la cual codifica un número como una multiplicación entre una *mantisa* con una *base* (10) elevada a un *exponente*. En esta representación, la mantisa representa el valor del número y, dado que multiplicar por una potencia de 10 en decimal es equivalente a mover el punto, el valor del exponente está indicando la posición del punto.
La representación de punto flotante antes descrita es una de muchas que se podría utilizar. Los parámetros relevantes para tal representación son: el número total de bits, el número de bits asignados a la mantisa, la normalización o no de la mantisa y el número de bits asignados al exponente.
En 1985 se definió el estándar IEEE-754 que especifica como deben representar las computadoras un número de punto flotante. El estándar define varias representaciones siendo dos las principales: punto flotante de precisión simple y punto flotante de precisión doble.

Formato de punto flotante de precisión simple

Este formato emplea 32 bits para la representación de un número en binario, 1 bit para el signo, 8 bits para el exponente y 23 bits para la fracción que debe estar *normalizada*[4] (es decir, se toman de forma que el bit más significativo, el bit más a la izquierda sea siempre 1) y cuyo rango es $\pm 1.18 \times 10^{-38} a \pm 3.4 \times 10^{38}$ ver Fig. 1.10.

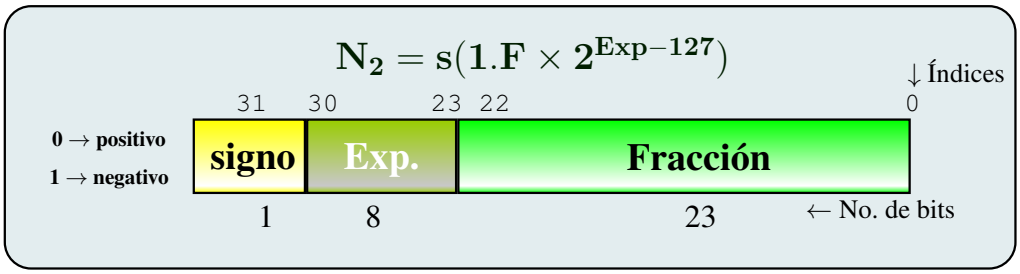

Figura 1.10: Formato de punto flotante precisión simple

Formato de punto flotante de precisión doble

En el formato de precisión doble, más bits están disponibles para el exponente y los decimales, como se muestra en la Fig. 1.11, 1 bit para el signo, 11 bits para el exponente y 52 bits para la fracción. Como resultado, pueden ser representados números tan pequeños como 10^{-308} y tan grandes como 10^{+308}.

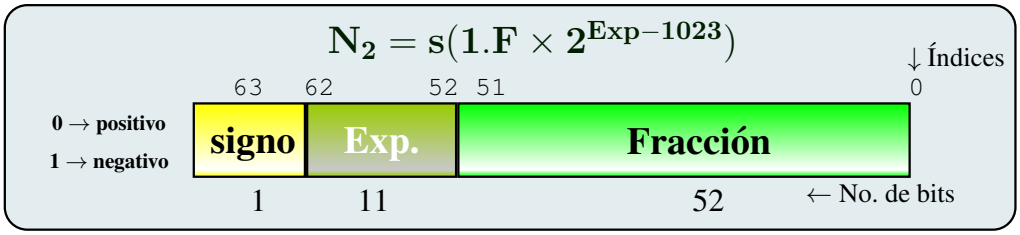

Figura 1.11: Formato de punto flotante precisión doble

[4] Un número que no va precedido por ceros se dice que está **normalizado**.

Ejemplo (- 4)

Para las siguientes cantidades determine su representación en formato de punto flotante (IEEE 754-1985):

a) 5777_{10}

El signo es positivo, por lo tanto:

$s = 0$

Se convierte a binario y se normaliza:

$5777_{10} = 1011010010001_2 = 1,011010010001 \times 2^{12}$

Con un desplazamiento, el exponente almacenado queda:

$E = 12 + 127 = 139_{10} = 10001011_2$

La fracción queda:

$F = 011010010001$

El número en formato de punto flotante precisión simple queda:

$N_{(IEEE-754)} = 01000101101101001000100000000000$

b) $N_{(IEEE-754)} = 11000000101001001110001000000000$

El signo es negativo ya que:

$s = 1$

El exponente almacenado queda:

$E = 10000001_2 = 129_{10} - 127 = 2_{10}$

La fracción es:

$F = 01001001110001$

La fracción normalizada queda:

$F = (-1)(1,01001001110001) \times 2^2 =$

Convirtiendo a decimal:

$= -101.001001110001_2 = -5.152587890625$

c) $N_{10} = 57.2310$

El signo es positivo, por lo tanto:

$s = 0$

Se convierte a binario y se normaliza:

$5777_{10} = 111001,00111_2 = 1.1100100111 \times 2^5$

El exponente almacenado queda:

$E = 5 + 127 = 132_{10} = 10000100_2$

La fracción normalizada queda:

$F = 1100100111$

El número en formato de punto flotante precisión simple queda:

$N_{(IEEE-754)} = 01000010011001001110000000000000$

Características importantes de la Norma de punto flotante

1. -126 es el menor exponente de un número normalizado en precisión simple.
2. El valor de un número cuyo exponente contiene todos 0's y la fracción todos 0's es $+0$ (sí $S = 0$) y -0 (sí $S = 1$).
3. El valor de un número cuyo exponente contiene todos 1's y la fracción todos 0's es $+\infty$ (sí $S = 0$) o $-\infty$ (sí $S = 1$).
4. Hay cantidades que pueden ser consideradas como no números (*NaNs Not a Number*) pero que pueden tener un signo y un significando, pero estos no tienen otro significado que el que puedan aportar en pruebas de diagnostico; el primer bit del significando es a menudo utilizado para distinguir los NaNs señalizados (*SnaN: Signalling Not a Number*) de los NaNs silenciosos (*QNaN: Quiet Not a Number*). Los primeros son utilizados cuando el dato ingresado al sistema no tiene el formato de un número válido especificado por la norma de punto flotante. Los QNaN especifican que la operación realizada con las cifras ingresadas produce un valor numérico no válido.
5. Los NaNs y los infinitos tienen todos los bits puestos a 1 en el campo Exp.

Problemas con la aritmética en punto flotante

Cuando se realizan operaciones aritméticas en punto flotante, pueden surgir problemas como resultado de estas operaciones. Estos pueden ser:

- **Desbordamiento del exponente**: un exponente positivo que excede el valor máximo representable. Este casos se puede representar como $+\infty$ ó $-\infty$.
- **Desbordamiento a cero del exponente**: un exponente negativo que exceda el valor máximo permitido. Esto quiere decir que el número es demasiado pequeño para ser representado, y puede ser aproximado por +0 ó -0.
- **Desbordamiento a cero de la mantisa**: debido a que en la operación de suma y resta los exponentes se deben igualar, pueden perderse dígitos por la parte derecha de la mantisa, durante el proceso de alineación de la misma.
- **Desbordamiento de la mantisa**: la suma de dos mantisas del mismo signo puede producir un acarreo debido al bit más significativo.

Otras consideraciones

Las operaciones aritméticas con infinito son tratadas como casos límite de la aritmética real, dándose la siguiente interpretación a los valores de infinitos:

$$-\infty < (cualquier\_numero\_finito) < +\infty$$

Algunas operaciones con infinitos son:

número + $(+\infty)$ =	$+\infty$
número - $(+\infty)$ =	$-\infty$
número + $(-\infty)$ =	$-\infty$
número - $(-\infty)$ =	$+\infty$
$(+\infty)$ + $(+\infty)$ =	$+\infty$
$(-\infty)$ + $(-\infty)$ =	$-\infty$
$(-\infty)$ - $(+\infty)$ =	$-\infty$
$(+\infty)$ - $(-\infty)$ =	+

Tabla 1.1: Operaciones con infinitos

3. Cifras significativas, exactitud y precisión

3.1. Análisis de errores

Como ya se estableció, los métodos numéricos ofrecen soluciones aproximadas muy cercanas a las soluciones exactas, la discrepancia entre una solución verdadera y una aproximada constituye lo que se denomina un *error*.
El análisis del error en un resultado numérico es esencial en cualquier cálculo (humano o computacional).
Los datos de entrada rara vez son exactos puesto que se basan en ensayos experimentales o bien son estimados y los métodos numéricos introducen errores de varios tipos, por ello brindan resultados aproximados.
Generalmente, los errores son costosos y en algunos casos fatales, por lo que es necesario tener en claro los siguientes conceptos:

Cifras significativas

La confiabilidad de un valor numérico está dada por sus **cifras significativas** que se definen como el número de dígitos, más un dígito estimado que se pueda usar con confianza.
Un cero puede ser significativo o no, dependiendo de su posición en un número dado. Los ceros que solamente sitúan la cifra decimal no son significativos. Los ceros al final de un número pueden ser significativos o no.

5: **Cifras significativas**

Exactitud

La **exactitud** se refiere a la aproximación de un número o de una medida al valor verdadero que se supone representa.

6: **Exactitud**

Precisión

La **precisión** se refiere al número de cifras significativas que representan una cantidad ó a la extensión en las lecturas repetidas de un instrumento que mide alguna propiedad física.

7: **Precisión**

Error absoluto

Sea x el valor exacto de una cantidad, y sea x^* su valor aproximado, se define el **error absoluto** como:

$$E_x = x - x^* \tag{1.1}$$

8: **Error absoluto**

Es decir, el error absoluto calcula la diferencia entre el valor exacto de una cantidad y su valor aproximado.

Error relativo

Para cuantificar la importancia del error respecto del valor exacto de una cantidad x se introduce el concepto de **error relativo**, que se define como:

9: **Error relativo**

$$r_x = \frac{E_x}{x} = \frac{x - x^*}{x} \tag{1.2}$$

Como puede observarse, el error relativo no esta definido para $x = 0$. La ecuación 1.2 muestra que el error relativo es una cantidad adimensional, que generalmente se expresa en porcentaje.

Es importante señalar que generalmente no se conoce el valor exacto de la cantidad x, en consecuencia, tampoco se puede conocer ni el error absoluto ni el error relativo cometido y hay que conformarse con calcular una cota del error.

Convergencia

Se entiende por **convergencia** de un método numérico la garantía de que, al realizar un *buen número* de iteraciones, las aproximaciones obtenidas terminan por acercarse cada vez más al verdadero valor buscado. En la medida en la que un método numérico requiera de un menor número de iteraciones que otro, para acercarse al valor deseado, se dice que tiene una mayor rapidez de convergencia.

10: **Convergencia**

Estabilidad

Se entiende por **estabilidad** de un método numérico el nivel de garantía de convergencia, y es que algunos métodos numéricos no siempre convergen y, por el contrario, divergen; esto es, se alejan cada vez más del resultado deseado. En la medida en la que un método numérico, ante una muy amplia gama de posibilidades de modelado matemático, es más seguro que converja que otro, se dice que tiene una mayor estabilidad. Es común encontrar métodos que convergen rápidamente, pero que son muy inestables y, en contraparte, modelos muy estables, pero de lenta convergencia.

11: **Estabilidad**

3.2. Clasificación de los errores

En el contexto de los métodos numéricos, se considera que el error total que contiene un número puede ser debido a los siguientes tipos de errores:

1. *Error inherente*. Se producen por la propia variabilidad de los fenómenos; al ser caracterizados a través de cantidades físicas, las mediciones conllevan incertidumbre, pues los instrumentos de medición ofrecen sólo una aproximación numérica del valor verdadero de la magnitud medida, pues se calibran para considerar solamente un determinado número de cifras significativas. Todas las magnitudes que se manejan en ingeniería son susceptibles a este tipo de errores.

2. *Error de redondeo*. Se producen al realizar operaciones aritméticas en las que el resultado produce una mantisa cuyo número de dígitos difiere significativamente del número de dígitos de la mantisa de alguno de los valores numéricos involucrados en la operación. Al manejar un determinado número de cifras significativas en los cálculos, el resultado tiene que ser redondeado de alguna manera, sobrestimando o subestimando el valor resultante verdadero.

3. *Error de truncamiento*. Los errores por truncamiento ocurren cuando un número, cuya parte fraccionaria está constituida por un número infinito de dígitos, requiere ser representado numéricamente en forma aproximada, utilizando un determinado número de cifras significativas.

3.3. Propagación del error

Cuando se cuantifica la propagación del error al efectuar operaciones se obtienen expresiones que relacionan el error del resultado obtenido con el error de los datos. Las consecuencias de la existencia de un error en los datos de un problema son más importantes de lo que aparentemente puede parecer. Desafortunadamente, estos errores se propagan y amplifican al realizar operaciones con dichos datos, hasta el punto de que puede suceder que el resultado carezca de significado.
Existen normas distintivas para reducir la propagación de los errores, como por ejemplo evitar restar números muy parecidos o evitar dividir por números muy pequeños comparados con el numerador, a continuación un ejemplo de la propagación del error en la suma.

Propagación del error en la suma

En ésta demostración se denotará por x e y los valores exactos de dos números y por x^* e y^* sus valores aproximados. Así mismo, los errores absolutos y relativos de estas cantidades se denotaran por E_x, E_y, r_x, r_y, respectivamente. Si se representa por $s = x + y$ al valor exacto de la suma y a $s^* = x^* + y^*$ su valor aproximado, entonces el error absoluto de la suma es:

$$E_s = s - s^* = (x + y) - (x^* + y^*) = E_x + E_y \tag{1.3}$$

La expresión anterior indica que el error absoluto de la suma es la suma de los errores absolutos de los sumandos. El error relativo vale:

$$r_s = \frac{E_s}{s} = \frac{E_x + E_y}{x + y} = \frac{x}{x + y} r_x + \frac{y}{x + y} r_y \tag{1.4}$$

donde se puede observar que el error relativo de la suma es la suma de los errores relativos de los datos multiplicados por unos factores que dependen de dichos datos. Esta dependencia se muestra gráficamente en la Figura 1.12

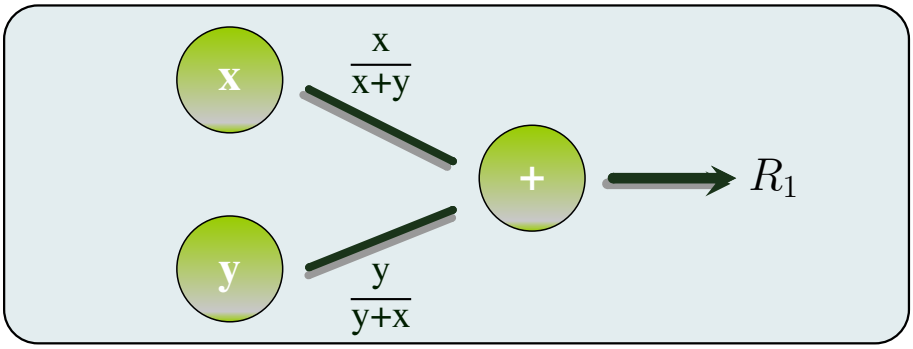

Figura 1.12: Propagación del error relativo en una suma

Propagación del error en una función

Sea $z = f(x)$ la imagen mediante la función f del valor exacto de un numero x y sea $z^* = f(x^*)$ la imagen de su valor aproximado. Entonces el error absoluto de la imagen es:

$$E_d = d - d^* = \frac{x}{y} - \frac{x^*}{y^*} = \frac{x}{y} - \frac{(x - E_x)}{(y - E_y)} \tag{1.5}$$

$$= \frac{yE_x - xE_y}{y(y - E_y)} \tag{1.6}$$

$$\approx \frac{yE_x - xE_y}{y^2} \tag{1.7}$$

En consecuencia el error relativo que se comete al evaluar la función f está determinado por la expresión:

$$r_z = \frac{E_z}{z} = \frac{f'(x)E_x - r_y}{f(x)} \approx x\frac{f'(x)}{f(x)}r_x \tag{1.8}$$

que gráficamente se muestra en la figura 1.13.

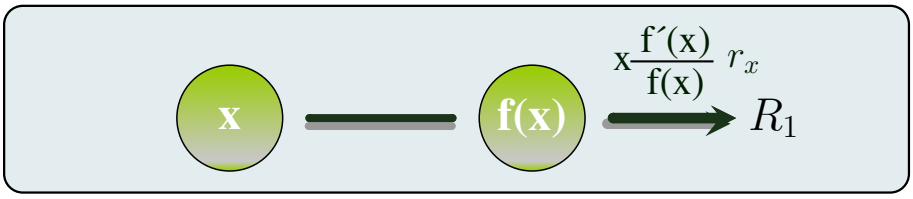

Figura 1.13: Propagación del error relativo al evaluar una función

4. Serie de Taylor y error

MAtemáticamente, las funciones reales más sencillas de evaluar son los polinomios[5], así que, cuando se tiene una función de otro tipo que es difícil de evaluar se busca un polinomio parecido el cual pueda sustituirla, donde desde luego, la solución obtenida es aproximada.

Las aproximaciones locales de una función se basan en la construcción de un polinomio que coincida con la función de partida y con algunas de sus derivadas en un único punto. En las proximidades de ese punto el polinomio toma valores muy parecidos a la función, pero no necesariamente iguales, y lejos de ese punto el polinomio no tiene por qué parecerse a la función.

Una **serie de Taylor** es una representación de una función como una suma infinita de términos que son calculados mediante los valores de las derivadas de esa función en un punto dado.

12: **Serie de Taylor**

El concepto de serie de Taylor fue introducido por el matemático inglés Brook Taylor en 1715 (ver Fig. 1.14).

DEFINICIÓN (1.5 ▷ Serie de Taylor)

La serie de Taylor de una función $f(x)$ de valor real o complejo que es infinitamente diferenciable en la vecindad de un número real o complejo a es igual a la serie de potencias:

$$f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots, \quad (1.9)$$

cuya forma compacta es:

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \quad (1.10)$$

donde $n!$ denota el factorial de n y $f^{(n)}(a)$ denota la $n-esima$ derivada de f evaluada en el punto a. La derivada de orden cero de f se define por sí misma y $(x-a)^0$ y $0!$ ambos son definidos como 1.

[5] **Polinomio**: es una expresión matemática constituida por un conjunto finito de variables (desconocidas) y constantes (coeficientes), utilizando únicamente las operaciones aritméticas de suma, resta y multiplicación, así como también exponentes enteros positivos que tiene la forma $P(x) = a_n x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + ... + a_1 x^1 + a_0$.

Si la serie de Taylor está centrada en 0, entonces esta se conoce como serie de McLaurin, nombrada así en honor al matemático escoces Colin Mc Laurin quien hizo gran uso de este caso especial de las series de Taylor durante el siglo XVIII.

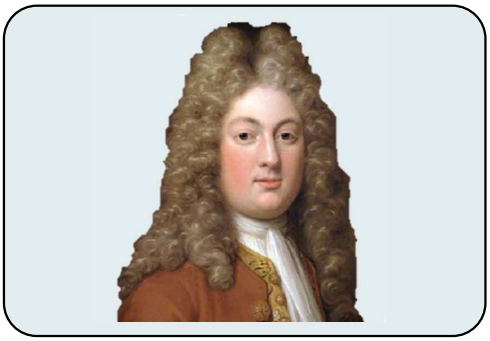

Figura 1.14: Brook Taylor (Inglaterra, 1685 - 1731)

Es una practica común aproximar una función usando un número finito de términos de su serie de Taylor. El teorema de Taylor da estimaciones cuantitativas del error en esta aproximación. Cualquier número finito de términos iniciales de la serie de Taylor de una función es llamado **polinomio de Taylor** (ver ec.1.11).

$$P_n(x) = f(x_0)+\frac{f'(x_0)}{1!}(x-x_0)+\frac{f''(x_0)}{2!}(x-x_0)^2+\frac{f'''(x_0)}{3!}(x-x_0)^3+\cdots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n \tag{1.11}$$

La serie de una función es el limite del polinomio de Taylor de esa función siempre y cuando exista. Una función puede no ser igual a su serie de Taylor incluso si la serie de Taylor converge en cada punto. Una función que es igual a su serie de Taylor en un intervalo abierto (o un disco en un plano complejo) se conoce como una función analítica.
Así que podemos decir que cualquier función que sea n-derivable en un punto x_0, puede aproximarse mediante la ec. 1.12:

$$f(x) \approx f(x_0)+\frac{f'(x_0)}{1!}(x-x_0)+\frac{f''(x_0)}{2!}(x-x_0)^2+\frac{f'''(x_0)}{3!}(x-x_0)^3+\cdots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n \tag{1.12}$$

4.1. Cálculo de la serie de Taylor

Existen varios métodos para el cálculo de series de Taylor de un gran número de funciones. Se puede intentar usar la serie de Taylor tal cual y generalizar la forma de los coeficientes, otra puede usar manipulaciones tales como sustitución, multiplicación o división, suma o resta de la serie estándar de Taylor para construir la serie de Taylor de una función en virtud de que la serie de Taylor es una serie de potencias. En algunos casos, también se puede derivar la serie de Taylor por la aplicación repetida de la integración por partes. Particularmente conveniente es el uso de sistemas de álgebra computacional para calcular series de Taylor.

Ejemplo (- 5)

Calcular la serie de Taylor para la función $f(x) = seno(x)$ en el punto $x_0 = 0$ hasta el 7o grado.

Primero se va a calcular el valor en el punto 0 de la función seno y las derivadas sucesivas.

$f(x) = seno(x) \Longrightarrow f(0) = seno(0) = 0$
$f'(x) = coseno(x) \Longrightarrow f'(0) = coseno(0) = 1$
$f''(x) = -seno(x) \Longrightarrow f''(0) = -seno(0) = 0$
$f'''(x) = -coseno(x) \Longrightarrow f'''(0) = -coseno(0) = -1$
$f^{(4)}(x) = f(x) = seno(x) \Longrightarrow f^{(4)}(0) = f(0) = seno(0) = 0$
$f^{(5)}(x) = f'(x) = coseno(x) \Longrightarrow f^{(5)}(0) = f'(0) = coseno(0) = 1$
$f^{(6)}(x) = f''(x) = -seno(x) \Longrightarrow f^{(6)}(0) = f''(0) = -seno(0) = 0$

Aplicando la fórmula de aproximación de Taylor se obtiene:

$$seno(x) \approx 0 + \frac{1}{1!}x + \frac{0}{2!}x^2 - \frac{1}{3!}x^3 + \frac{0}{4!}x^4 + \frac{1}{5!}x^5 + \frac{0}{6!}x^6$$

Realizando operaciones

$$seno(x) \approx x - \frac{x^3}{6} + \frac{x^5}{120}$$

Dando a la expresión el formato de serie de potencias queda:

$$seno(x) \approx \sum_{k=0}^{6} \frac{(-1)^k}{(2k+1)!}x^{2k+1}$$

Ejemplo (- 6)

Calcular la serie de Taylor para la función $f(x) = coseno(x)$ en el punto $x_0 = 0$ hasta el 6o grado.

Primero se va a calcular el valor en el punto 0 de la función coseno y las derivadas sucesivas.

$f(x) = coseno(x) \Longrightarrow f(0) = coseno(0) = 1$
$f'(x) = -seno(x) \Longrightarrow f'(0) = -seno(0) = 0$
$f''(x) = -coseno(x) \Longrightarrow f''(0) = -coseno(0) = -1$
$f'''(x) = seno(x) \Longrightarrow f'''(0) = seno(0) = 0$
$f^{(4)}(x) = f(x) = coseno(x) \Longrightarrow f^{(4)}(0) = f(0) = coseno(0) = 1$
$f^{(5)}(x) = f'(x) = -seno(x) \Longrightarrow f^{(5)}(0) = f'(0) = -seno(0) = 0$
Aplicando la fórmula de aproximación de Taylor se obtiene:

$$coseno(x) \approx 1 + \frac{0}{1!}x - \frac{1}{2!}x^2 + \frac{0}{3!}x^3 + \frac{1}{4!}x^4 + \frac{0}{5!}x^5$$

Realizando operaciones

$$coseno(x) \approx 1 - \frac{x^2}{2} + \frac{x^4}{24}$$

Dando a la expresión el formato de serie de potencias queda:

$$coseno(x) \approx \sum_{k=0}^{5} \frac{(-1)^k}{(2k)!}x^{2k}$$

Ejemplo (- 7)

Calcular la serie de Taylor para la función $f(x) = e^{(x)}$ en el punto $x_0 = 0$ hasta el 4o grado.

Primero se va a calcular el valor en el punto 0 de la función exponencial y las derivadas sucesivas.

$f(x) = e^{(x)} \Longrightarrow f(0) = e^{(0)} = 1$

$f'(x) = e^{(x)} \Longrightarrow f'(0) = e^{(0)} = 1$

$f''(x) = e^{(x)} \Longrightarrow f''(0) = e^{(0)} = 1$

$f'''(x) = e^{(x)} \Longrightarrow f'''(0) = e^{(0)} = 1$

Aplicando la fórmula de aproximación de Taylor se obtiene:

$e^{(x)} \approx 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!}$

Realizando operaciones

$e^{(x)} \approx 1 + x + \frac{x^2}{2} + \frac{x^3}{6}$

Dando a la expresión el formato de serie de potencias queda:

$e^{(x)} \approx \sum_{k=0}^{3} \frac{x^k}{k!}$

Ahora se van a programar series de Taylor, en este caso para la función exponencial (e).
El número exponencial (e) es un número irracional denominado más formalmente **Número de Euler** (o Constante de Napier).
Su obtención fue definida en 1748 por el matemático Leonard Euler (el mismo matemático que mejoró la aproximación del número π), y consta de la obtención del número de acuerdo la siguiente serie:

$$e = 1 + \frac{1}{1} + \frac{1}{1 * 2} + \frac{1}{1 * 2 * 3} + ...$$

Como se muestra en el formato de serie del ejemplo 7 el objetivo es realizar la suma de los términos para cada iteración n, en donde n tiende al infinito.

Evidentemente computacionalmente es imposible obtener la convergencia de una operación que tiende al infinito, sin embargo con un suficiente número de iteraciones se obtiene una buena aproximación, que es el objetivo de la serie de Taylor. El pseudocódigo se muestra a continuación:

Algorithm 1: SERIE DE TAYLOR DE e^x

Entradas: int Val_x, int Num_iter
Salidas: $suma$

```
1  INICIO
       Leer: Val_x,Num_iter;
2      float suma ← 0;
3      int contador ← 1;
4      int term ← 1;
5      Mientras Val_x ≤contador hacer
6          suma = suma + term;
7          term = term * Val_x/contador;
8          contador = contador + 1;
9      fin
       Escribir: suma;
10 FIN
```

Y el diagrama de flujo:

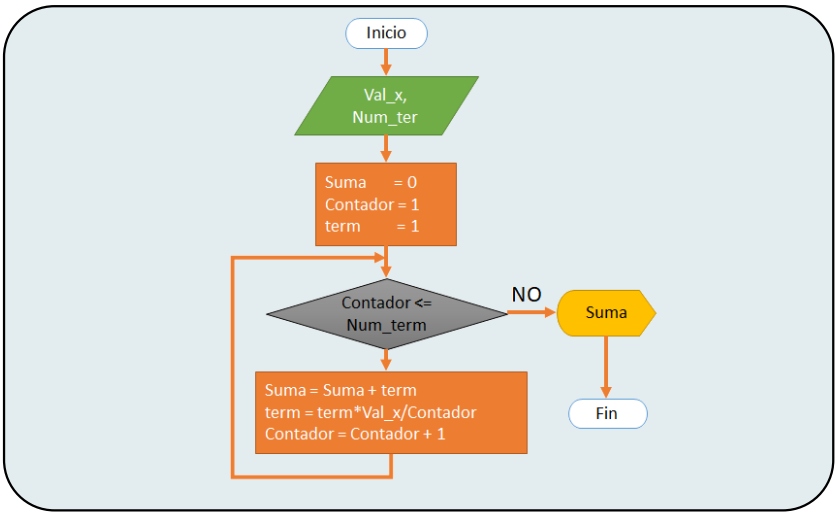

Figura 1.15: Diagrama de flujo de Serie de Taylor

4.2. Error por truncamiento

La serie de la ec.1.12 se puede truncar en cualquier punto, así el subíndice n indica que se han incluido en la aproximación los primeros $(n+1)$ términos de la serie. La siguiente expresión se conoce como expansión en serie de Taylor con residuo, y es idéntica a la ec.1.12, excepto porque se ha agregado el término R_n, que sintetiza los términos de la serie que se han despreciado y se conoce con el nombre de residuo de la aproximación al n-ésimo orden (ec.1.13).

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \frac{f'''(x_0)}{3!}(x-x_0)^3 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + R_n \quad (1.13)$$

Donde R_n es el error por truncamiento cuando se aproxima el valor de una función $f(x-x_0)$, considerando solamente los $(n+1)$ primeros términos de la expansión en serie de Taylor correspondiente a la función. Así, mientras mayor sea el valor de n (es decir, el número de términos de la serie, considerados al aproximar el valor de la función), menor será el residuo y mejor la aproximación al valor de la función.

Ejemplo (- 8)

Calcular el error por truncamiento asociado a la serie de Taylor de la función definida en el ejemplo de la página 25, en el punto $x_0 = 45°$.

Primero se va a convertir el valor de $45°$ a radianes.

$45° = \frac{45° \times \pi rad}{180°} = 0.785398163 rad$

Mediante calculadora el valor considerado exacto es:

$sen(0.785398163 rad) = 0.707106781$

Mediante la serie de Taylor el valor aproximado calculado es:

$$seno(0.785398163 rad) \approx 0.785398163 - \frac{(0.785398163)^3}{6} + \frac{(0.785398163)^5}{120}$$

$$\approx 0.787888558$$

De la expansión de serie de Taylor con residuo (ecuación 1.13), el error por truncamiento queda:

$R_7 = f(x) - P_7(x) = 0.707106781 - 0.787888558 = -0.080781777$

Ejemplo (- 9)

Calcular el error por truncamiento asociado a la serie de Taylor de la función definida en el ejemplo de la página 26, en el punto $x_0 = 5$.

Mediante calculadora el valor considerado exacto es:

$e^{(5)} = 148.413159103$

Mediante la serie de Taylor el valor aproximado calculado es:

$$e^{(5)} \approx 1 + (5) + \frac{(5)^2}{2} + \frac{(5)^3}{6}$$

$$\approx 39.333333333$$

De la expansión de serie de Taylor con residuo (ecuación 1.13), el error por truncamiento queda:

$$R_4 = f(x) - P_4(x) = 148.413159103 - 39.333333333 = 109.07982577$$

Ejemplo (- 10)

Calcular el error por truncamiento asociado a la serie de Taylor de la función definida en el ejemplo de la página 27, en el punto $x_0 = 60°$.

Primero se va a convertir el valor de $60°$ a radianes.

$60° = \frac{60° \times \pi rad}{180°} = 1.047197551 rad$

Mediante calculadora el valor considerado exacto es:

$cos(1.047197551 rad) = 0.5$

Mediante la serie de Taylor el valor aproximado calculado es:

$$coseno(1.047197551 rad) \approx 1 - \frac{(1.047197551)^2}{2} + \frac{(1.047197551)^4}{24}$$

$$\approx 0.501796202$$

De la expansión de serie de Taylor con residuo (ecuación 1.13), el error por truncamiento queda:

$R_6 = f(x) - P_6(x) = 0.5 - 0.501796202 = -0.001796202$

5. Ejemplo de aproximaciones sucesivas (Método de Punto Fijo)

EN ocasiones en el ámbito de la ingeniería es necesario resolver ecuaciones no lineales que no tienen solución analítica o que es muy complicado hallarlas, para estos casos, deben utilizarse métodos de solución numérica de ecuaciones.
Dada una ecuación de una variable independiente x,

$$f(x) = 0 \tag{1.14}$$

El objeto del cálculo de las raíces de una ecuación es determinar los valores de x para los que se cumple la ecuación 1.14.
Encontrar una solución (ó una raíz real) de una ecuación, es hallar el valor de la variable independiente x, que anule el valor de la función $f(x)$, que se exprese en términos de la variable citada. Es decir, si la función se desarrolla en el plano cartesiano xy, la solución real de esa función es el valor de x que corresponda a la intersección del eje de las abscisas con la curva definida por la función $f(x)$, como se muestra en la Fig. 1.16.
Si la curva no corta al eje x, entonces, la ecuación no tiene una solución real, pero puede tener raíces imaginarias (que no serán tratadas aquí).

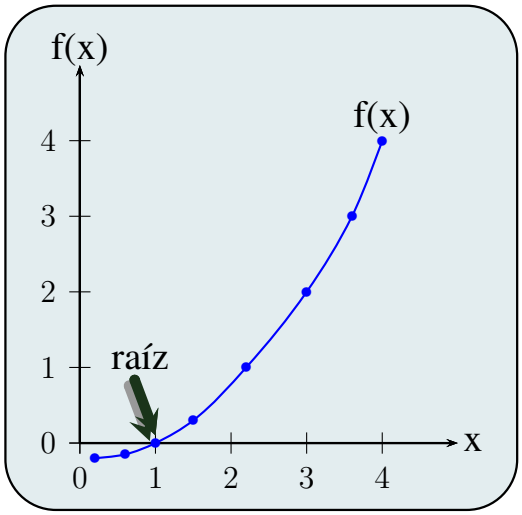

Figura 1.16: Concepto gráfico de raíz

La determinación de las raíces (ó ceros) de una ecuación es uno de los problemas más antiguos en matemáticas y se han realizado un gran número de esfuerzos en este sentido.
Su importancia radica en que si se pueden determinar las raíces de una ecuación también se pueden determinar máximos y mínimos, valores propios de matrices, resolver sistemas de ecuaciones lineales y diferenciales, etc..

La determinación de las soluciones de la ecuación 1.14 puede llegar a ser un problema muy difícil.
Si $f(x)$ es una función polinómica de grado 1 ó 2, se conocen expresiones simples que permiten determinar sus raíces.
Por ejemplo, si f es un polinomio de primer grado:

$$f(x) = ax + b \tag{1.15}$$

Entonces, hay un solo cero (real) en $-b/a$.
O, sí f es un polinomio de segundo grado:

$$f(x) = ax^2 + bx + c \tag{1.16}$$

Entonces, el número de ceros (reales) depende de cuanto valga el discriminante $\Delta = b^2 - 4ac$; para $\Delta > 0$, la funcion f tiene dos ceros $x = (-b \pm sqrt\Delta)/2a$.
Para polinomios de grado 3 ó 4 es necesario emplear métodos complejos y laboriosos, sin embargo, si $f(x)$ es de grado mayor de cuatro o bien no es polinómica, no hay ninguna fórmula conocida que permita saber a priori cuantos ceros tiene la función: ¿varios, uno, ninguno? (excepto en casos muy particulares).
Sin embargo, existen una serie de reglas que pueden ayudar a determinar las raíces de una ecuación:

- El teorema de Bolzano, que establece que si una función continua, $f(x)$, toma en los extremos del intervalo [a,b] valores de signo opuesto, entonces la función admite, al menos, una raíz en dicho intervalo.
- En el caso en que $f(x)$ sea una función algebraica (polinómica) de grado n y coeficientes reales, podemos afirmar que tendrá n raíces reales o complejas.
- La propiedad más importante que verifican las raíces racionales de una ecuación algebraica establece que si p/q es una raíz racional de la ecuación de coeficientes enteros:

  $$f(x) = a_0 + a_1x + a_2x^2 + ... + a_nx^n = 0 \quad (a_i \epsilon Z) \tag{1.17}$$

  entonces el denominador q divide al coeficientes a_n y el numerador p divide al término independiente a_0.

Para resolver este tipo de problemas se puede utilizar técnicas numéricas iterativas: a partir de una aproximación inicial x^0 a un cero x^* de f, se construye iterativamente una sucesión de aproximaciones $\{x^k\}$.
El superíndice k es el contador de iteraciones: en la primera iteración, se calcula x^1; en la segunda, x^2, y así sucesivamente.
El proceso iterativo se detiene cuando, para un cierto valor de k, el valor x^k es una aproximación suficientemente buena a x^*.

Como puede verse, para obtener numéricamente un cero de f hay que responder a las siguientes tres preguntas:

1. ¿Cómo se elige la aproximación inicial x^0?
2. ¿Cómo se construye la sucesión $\{x^k\}$ de aproximaciones?
3. ¿Cómo se decide si x^k es una aproximación suficientemente buena a x^*?

El método de punto fijo consiste en una forma iterativa de resolver una ecuación de la forma $f(x) = x$.
En este método se debe elegir una aproximación inicial x_0 y realizar la iteración

$$x_{k+1} = f(x_k) \tag{1.18}$$

Hasta que la diferencia $|x_{k+1} - x_k|$ sea muy cercana a cero, para lo cual se establece una tolerancia (ε) a criterio del usuario.

Algorithm 2: MÉTODO DE PUNTO FIJO

Entradas: aproximacion inicial x_0, tolerancia ε
Salidas: valor x tal que $f(x) = x$

```
INICIO
    s_w = 1;
    x_1 = x_0;
    Mientras s_w == 1 hacer
        x_2 = f(x_1);
        if abs(x_2 - x_1) ≤ ε then
            x = x_2;
            s_w = 0;
    x_1 = x_2;
FIN
```

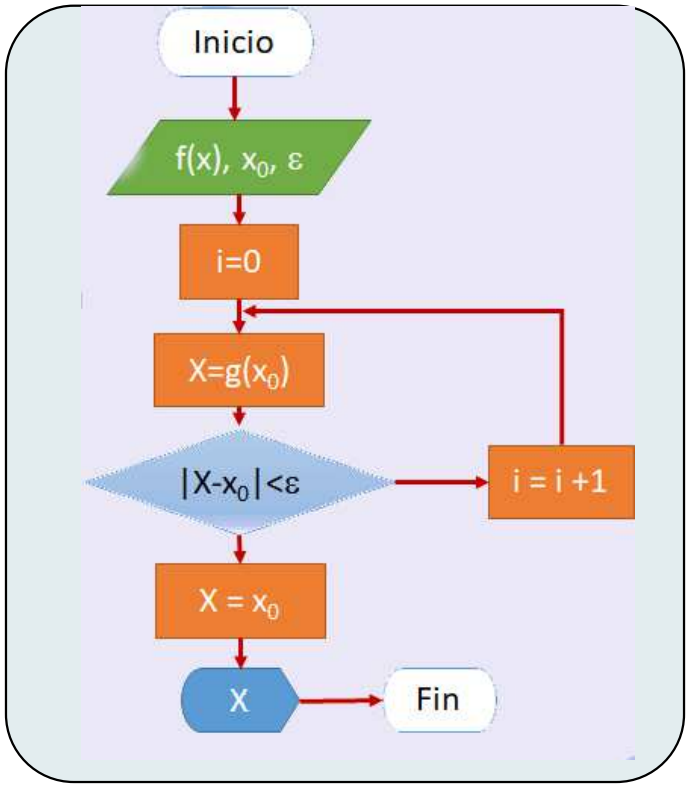

Figura 1.17: Diagrama de flujo de Punto Fijo

Ejemplo (- 11)

Determinar, aplicando el método de aproximaciones sucesivas de punto fijo, una de las raíces de la ecuación:

$$f(x) = x^2 - 4x + 2 \tag{1.19}$$

Para encontrar las raíces:

$$f(x) = 0 \qquad \therefore \qquad x^2 - 4x + 2 = 0 \tag{1.20}$$

Para cambiar la ecuación a la forma, $x = g(x)$, se despeja x de la ecuación original 1.20 :

$$x = \frac{x^2 + 2}{4} \tag{1.21}$$

Por lo tanto, la ecuación de recurrencia o iteración es:

$$x_{i+1} = \frac{x_i^2 + 2}{4} \tag{1.22}$$

El valor inicial propuesto para x_0 es 1 (considerando un valor cercano a la solución) y la tolerancia de error (ε) se establece en 0.0001 (este valor dependerá de cada caso en particular).
El el diagrama de flujo del algoritmo empleado se muestra en la figura 1.15:

```c
/*****************************************************************
*                                UAA                             *
*  Autor:                                                        *
*  Carrera:                                                      *
*  Fecha:                                Versión: 1.0            *
*****************************************************************
*  Descripción del Programa:                                     *
*  Programa de Metodos Numericos: Aproximaciones Sucesivas o de  *
*  punto fijo para la funcion:                                   *
                              x^2 - 4x + 2                      *
*****************************************************************/
#include<stdio.h>
#include<stdlib.h>
#include<math.h>

int main()
{
    float x1,x2,x,y,Epsilon;
    int Iteracion,Contador;
    printf("================================================\n");
    printf("Este programa calcula una raiz del polinomio: \n");
    printf("               x^2 - 4x + 2                     \n");
    printf("================================================\n\n");
    printf("Dame el valor de x: ");
    scanf("%f",&x);
    printf("Dame el valor de Epsilon ");
    scanf("%f",&Epsilon);
    printf("Dame el valor del Contador: ");
    scanf("%d",&Contador);
    x1=(pow(x,2)+2)/4;
    Iteracion=1;
    printf("\n=====================================================\n");
    printf("No.it.\tx0\t\tGx\t\tFx \n");
    printf("=====================================================\n");
    while(fabs(x1-x)>=Epsilon && Iteracion<=Contador)
         {
           x=x1;
           y=pow(x,2)-4*x+2;
           x1=(pow(x,2)+2)/4;
           printf("%d\t%f\t%f\t%f \n",Iteracion,x,x1,y);
           Iteracion=Iteracion++;
           }
    printf("=====================================================\n\n");
    if(fabs(x1-x))
    printf("============================\nLa raiz es igual a %f\n",x1);
    else
    printf("============================\nNo alcanza con el numero de Iteraciones\n");
    printf("============================\n\n");
    system("Pause");
    return 0;
}
```

Figura 1.18: Programa de Punto Fijo

Dentro del rango de valores de 10 a -10 para x, la función $f(x)$ toma los siguientes valores:

x	10	9	8	7	6	5	4	3	2	1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
f(x)	62	47	34	23	14	7	2	-1	-2	-1	2	7	14	23	34	47	62	79	98	119	142

Tabla 1.2: Valores de $f(x)$

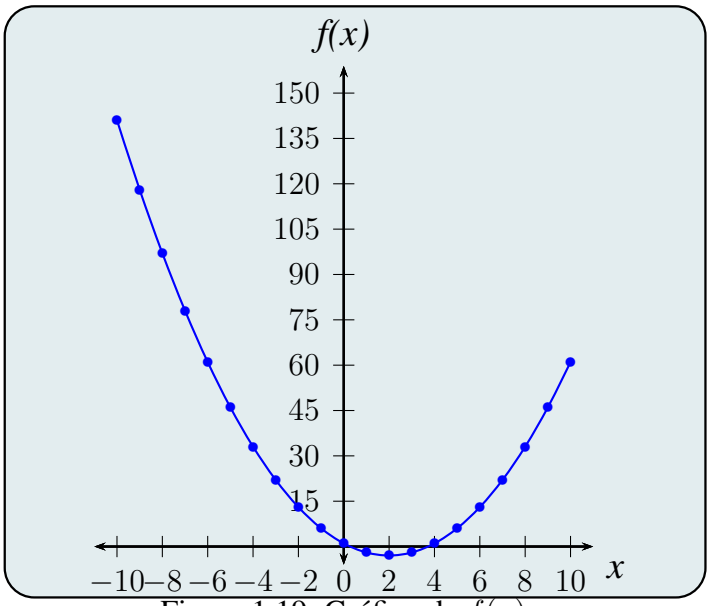

Figura 1.19: Gráfico de $f(x)$

Ejemplo (- 12)

Obtener una raíz, por el método de aproximaciones sucesivas de punto fijo, de la ecuación:

$$f(x) = e^{-x} - x \tag{1.23}$$

Se despeja x de la ecuación original:

$$x = e^{-x} \tag{1.24}$$

El valor inicial propuesto para x_0 es 0 y la tolerancia de error se establece en 0.0001.
Por lo tanto la ecuación de recurrencia o iteración es:

$$x_{i+1} = e^{x_i} \tag{1.25}$$

Dentro del rango de valores de 5 a -5 para x, la función $f(x)$ toma los siguientes valores:

x	5	4	3	2	1	0	-1	-2	-3	-4	-5
f(x)	-4.9	-3.9	-2.9	-1.8	-0.6	1	3.7	9.3	23	58.5	153.4

Tabla 1.3: Valores de $f(x)$

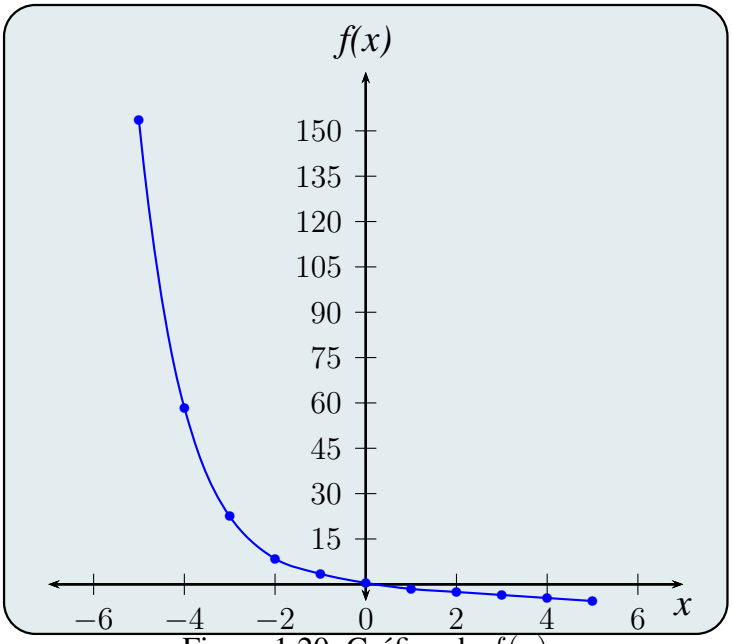

Figura 1.20: Gráfico de $f(x)$

2

RAÍCES DE ECUACIONES

"Conocer distintos métodos que existen para encontrar raíces de ecuaciones mediante la programación, identificar las ventajas y desventajas de dichos métodos, implementar programas que desarrollen los métodos discutidos."

Objetivos particulares 1, 2 y 3

1. Introducción

Un problema básico del cálculo numérico es la búsqueda de la solución de una ecuación. Lo que significa encontrar los valores de la variable x que satisfacen la ecuación $f(x) = ax^k + ...$, para una función f dada.
Se va a suponer que tanto x como $f(x)$ son reales (aunque algunos de los algoritmos que se verán son válidos para funciones complejas analíticas de variable compleja), es decir:

$$f : \Omega \subset \mathbb{R} \longrightarrow \mathbb{R}$$

Las ecuaciones pueden ser algebraicas (la función f es un polinomio), por ejemplo:

$$x^2 + 5x - 4 = 0$$

o bien **trascendentes** (aquellas que están constituidas por funciones exponenciales, trigonométricas, logarítmicas), etc., por ejemplo:

$$e^{-x} - x; \qquad sen(x); \qquad ln(x^2) - 1$$

Solamente en casos muy simples, de ecuaciones algebraicas, existen fórmulas que permiten resolverlas en términos de sus coeficientes, para el resto de las ecuaciones se utilizan métodos aproximados que permiten mejorar la solución por simple repetición del mismo método hasta adquirir el grado de aproximación requerido.
No existe un método universal de solución de sistemas de ecuaciones no lineales.
Los métodos para calcular una raíz real de una ecuación involucran dos pasos, en primer lugar la determinación del intervalo de búsqueda (es decir el intervalo al que la raíz pertenece), siempre que la ecuación esté vinculada a un sistema físico y en segundo lugar la selección y aplicación de un método numérico apropiado para determinar la raíz con la exactitud adecuada.
Estos métodos se clasifican en dos categorías:

- Métodos **cerrados**
- Métodos **abiertos**

Los métodos cerrados (como el de la bisección y el de la falsa posición), son aquellos que usan intervalos, se caracterizan por ser siempre convergentes pero la velocidad de convergencia es lenta.
Los métodos abiertos (como el de Newton-Raphson), requieren información únicamente de un punto, o de dos pero que no necesariamente encierran a la raíz, la convergencia es más rápida (aunque algunas veces divergen).

2. Métodos cerrados

2.1. Método gráfico

A Diferencia de las ecuaciones lineales, las cuales representan ecuaciones de líneas rectas, las ecuaciones no lineales representan ecuaciones de curvas (círculos, parábolas, hipérbolas, elipses, etc.), en esta unidad nuevamente el objetivo es determinar para funciones de este tipo si la curva cruza en algún punto el eje de los reales (es decir si existe al menos una raíz).
Así, se tiene un sistema de ecuaciones no lineales, la curva de la función y la recta que representa al eje de los números reales, para resolverlo mediante el método gráfico se requiere:

- Definir un intervalo de valores [a,b] para x.
- Calcular el valor de la función $f(x)$ para cada punto del intervalo.
- Y luego graficarla para verificar que $f(x)$ toma, dentro de dicho intervalo, valores de signo opuesto (Teorema de Bolzano), lo que confirmaría la existencia de al menos una raíz.

El resultado se obtiene mediante la interpolación directa entre los valores de x que corresponden a los de la función que presentan signos opuesto.

Ejemplo (- 13)

Para la función:

$$f(x) = 4x^2 - 10$$

Determinar si existen ceros y su ubicación en el intervalo [-3,4] empleando el método gráfico.
Calculando el valor de la función:

x	-3	-2	-1	0	1	2	3	4
f(x)	26	6	-6	-10	-6	6	26	54

Tabla 2.1: Tabulación de *f(x)*

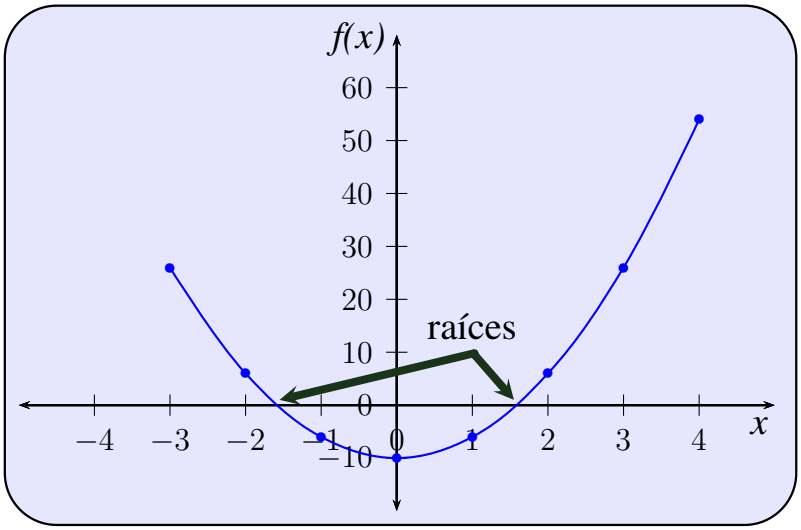

Figura 2.1: Gráfica de *f(x)*

Interpolando directamente:

$$Z_1 = \frac{-2-1}{2} = -1.5$$

$$Z_2 = \frac{2+1}{2} = 1.5$$

2.2. Método de bisecciones sucesivas

El método de la bisección, conocido también como *de corte binario*, *de partición en dos intervalos iguales*, *de búsqueda binaria* o *de Bolzano* es un método cerrado que se basa en los siguientes teoremas.

Teorema del valor intermedio:
Si $f \in [a, b]$ y k es un número cualquiera comprendido entre $f(a)$ y $f(b)$ entonces existe un punto c en el intervalo (a, b) tal que $f(c) = k$

Teorema de Bolzano:
Sea f una función contínua en el intervalo $[a, b]$, con $f(a)f(b) < 0$ entonces existe al menos un punto $c \in [a, b]$ tal que $f(c) = 0$

Así pues, si se tiene una función $f(x)$ continua en el intervalo $[x_i, x_s]$, con $f(x_i)$ y $f(x_s)$ de signos opuestos, por el teorema anterior, existe un valor x^* incluido en el intervalo (x_i, x_s) tal que $f(x^*) = 0$.
El método requiere de dividir el intervalo a la mitad y localizar la mitad que contiene a la raíz.
El proceso se repite y su aproximación mejora a medida que los subintervalos se dividen en intervalos más y más pequeños; la primera aproximación a la raíz, se determina como:

$$x_M = \frac{(x_i + x_s)}{2} \tag{2.1}$$

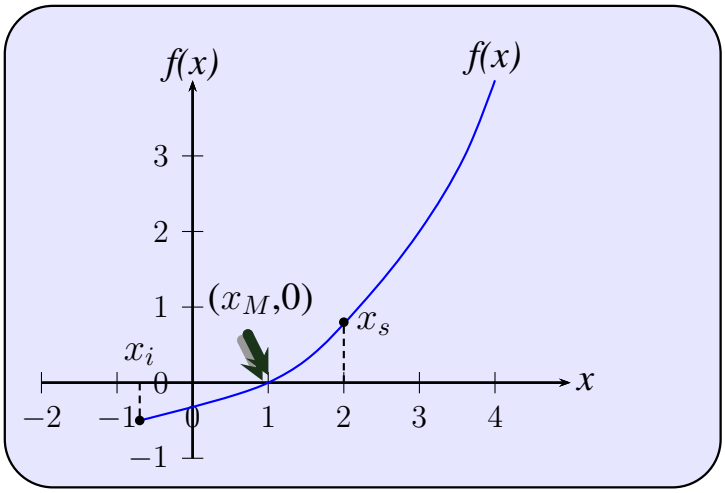

Figura 2.2: Esquema del método de la Bisección

Para determinar en qué subintervalo está situada la raíz, hay que considerar lo siguiente:

- Si $f(x_M) = 0$, entonces la raíz es igual a x_M.
- Si $f(x_i) * f(x_M) < 0$, la raíz está en el primer subintervalo (x_i, x_M)
- Si $f(x_i) * f(x_M) > 0$, la raíz está en el segundo subintervalo (x_M, x_s).

Se calcula una nueva aproximación a la raíz en el nuevo subintervalo y se continúa con las iteraciones hasta que se alcanza el margen de error fijado de antemano (ε). Una de las ventajas de este método es que siempre es convergente.
Las desventajas son que converge muy lentamente y que, si existe más de una raíz en el intervalo, el método solo permite encontrar una de ellas.

Algorithm 3: MÉTODO DE BISECCIÓN

Entradas: límite inferior x_i, límite superior x_s, raíz aproximada x_M, tolerancia ε. número máximo de iteraciones N
Salidas: valor final aproximado de la raíz x_f

```
1  INICIO
2      i = 1;
3      Mientras i ≤ N hacer
4          x_M = (x_i + x_s)/2;
5          if |(x_M − x_f)| ≤ ε o f(x_M = 0) then
6              x_f = x_M;
7          i = i + 1;
8          if f(x_i) * f(x_m) > 0 then
9              x_i = x_M
10         else
11             x_s = x_M
12     "Procedimiento completado sin éxito después de N iteraciones";
13 FIN
```

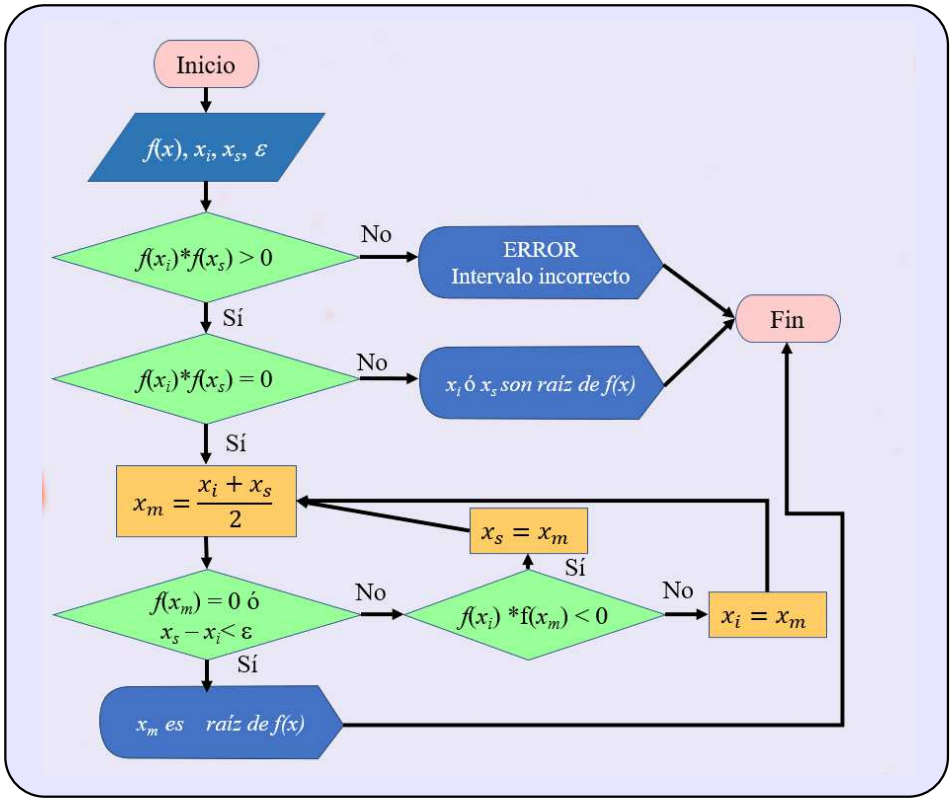

Figura 2.3: Diagrama de flujo del método de la Bisección

Ejemplo (- 14)

Dada la función:

$$x^3 + x^2 - 3x - 3 = 0$$

Obtener una de sus raíces por el método de bisección, considerando un margen de error de: $\varepsilon = 0.01$

1a Iteración

Considerando el intervalo $[1, 2]$:

$x_{M1} = \frac{(1+2)}{2} = 1.5$

Ahora se calcula $f(x_i)$*$f(x_M)$:

$f(x_{i1}) = f(1) = (1)^3 + (1)^2 - 3(1) - 3 = -4$

$f(x_{M1}) = f(1.5) = (1.5)^3 + (1.5)^2 - 3(1.5) - 3 = -1.875$

$$f(x_{i1}) * f(x_{M1}) = (-4) * (-1.875) = 7.5$$

Como $f(x_i)$*$f(x_M) > 0$, la raiz se encuentra en el 2o subintervalo.

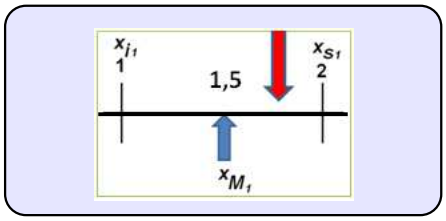

Figura 2.4: Intervalo 1a iteración

2a Iteración

Considerando el intervalo $[1.5, 2]$:

$x_{M2} = \frac{(1.5+2)}{2} = 1.75$

Ahora se calcula $f(x_i)$*$f(x_M)$:

$f(x_{i2}) = f(1.5) = (1.5)^3 + (1.5)^2 - 3(1.5) - 3 = -1.875$

$f(x_{M2}) = f(1.75) = (1.75)^3 + (1.75)^2 - 3(1.5) - 3 = 0.171875$

$$f(x_{i2}) * f(x_{M2}) = (-1.875) * (0.171875) = -0.322265625$$

Como $f(x_i)$*$f(x_M) < 0$, la raiz se encuentra en el 1er subintervalo.

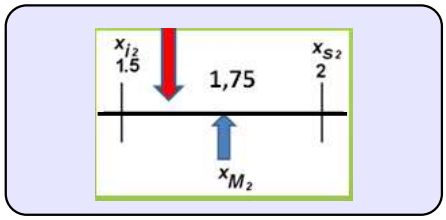

Figura 2.5: Intervalo 2a iteración

3a Iteración

Considerando el intervalo $[1.5, 1.75]$:

$$x_{M3} = \frac{(1.5 + 1.75)}{2} = 1.625$$

Ahora se calcula $f(x_i)$*$f(x_M)$:

$f(x_{i3}) = f(1.5) = (1.5)^3 + (1.5)^2 - 3(1.5) - 3 = -1.875$

$f(x_{M3}) = f(1.625) = (1.625)^3 + (1.625)^2 - 3(1.625) - 3 = -0.943359375$

$f(x_{i3}) * f(x_{M3}) = (-1.875) * (-0.943359375) = 1.768798828125$

Como $f(x_i)$*$f(x_M) > 0$, la raiz se encuentra en el 2o subintervalo.

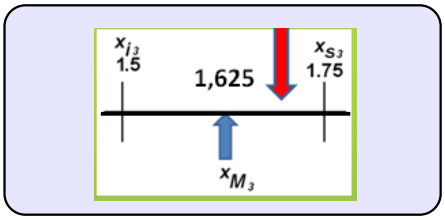

Figura 2.6: Intervalo 3a iteración

4a Iteración

Considerando el intervalo $[1.625, 1.75]$:

$$x_{M4} = \frac{(1.625 + 1.75)}{2} = 1.6875$$

Ahora se calcula $f(x_i)$*$f(x_M)$:

$f(x_{i4}) = f(1.625) = (1.625)^3 + (1.625)^2 - 3(1.625) - 3 = -0.943359375$

$f(x_{M4}) = f(1.6875) = (1.6875)^3 + (1.6875)^2 - 3(1.6875) - 3 = -0.409423828125$

$f(x_{i4}) * f(x_{M4}) = (-0.943359375) * (-0.409423828125) = 0.386233806610107421875$

Como $f(x_i)$*$f(x_M) > 0$, la raiz se encuentra en el 2o subintervalo.

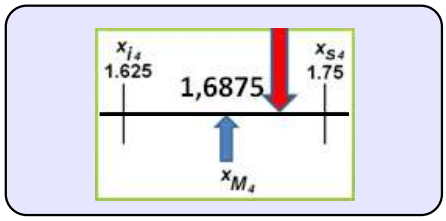

Figura 2.7: Intervalo 4a iteración

5a Iteración

Considerando el intervalo $[1.6875, 1.75]$:

$$x_{M5} = \frac{(1.6875 + 1.75)}{2} = 1.71875$$

Ahora se calcula $f(x_i)$*$f(x_M)$:

$f(x_{i5}) = f(1.6875) = (1.6875)^3 + (1.6875)^2 - 3(1.6875) - 3 = -0.409423828125$

$f(x_{M5}) = f(1.71875) = (1.71875)^3 + (1.71875)^2 - 3(1.71875) - 3 = -0.124786376953125$

$f(x_{i5})*f(x_{M5}) = (-0.09423828125)*(-0.124786376953125) = 0.0511905161499977111181640625$

Como $f(x_i)*f(x_M) > 0$, la raiz se encuentra en el 2o subintervalo.

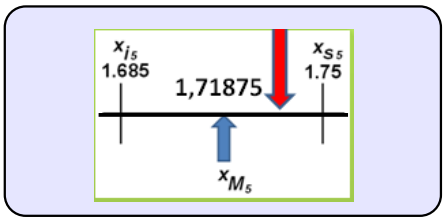

Figura 2.8: Intervalo 5a iteración

6a Iteración

Considerando el intervalo $[1.71875, 1.75]$:

$$x_{M6} = \frac{(1.71875 + 1.75)}{2} = 1.734375$$

Ahora se calcula $f(x_i)*f(x_M)$:

$f(x_{i6}) = f(1.71875) = (1.71875)^3+(1.71785)^2-3(1.71875)-3 = -0.124786376953125$

$f(x_{M6}) = f(1.734375) = (1.734375)^3+(1.734375)^2-3(1.734375)-3 = 0.022029876708984375$

$f(x_{i6})*f(x_{M6}) = (-0.124786376953125)*(0.022029876708984375) = -0.00274902849923819303512573242 1875$

Como $f(x_i)*f(x_M) < 0$, la raiz se encuentra en el 1er subintervalo.

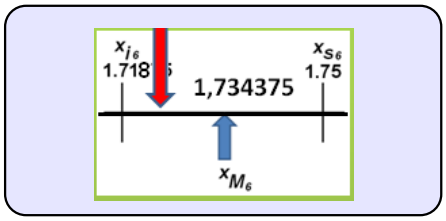

Figura 2.9: Intervalo 6a iteración

7a Iteración

Considerando el intervalo $[1.71875, 1.734375]$:

$$x_{M7} = \frac{(1.71875+1.734375)}{2} = 1.7265625$$

Como:

$|x_{M6} - x_{M7}| = |(1.734375 - 1.725625)| = 0.0078125$

es menor que ($\varepsilon = 0.01$) tomamos del valor de x_{M6} como:

La raiz es igual a 1.734375

.
.
.
.
.

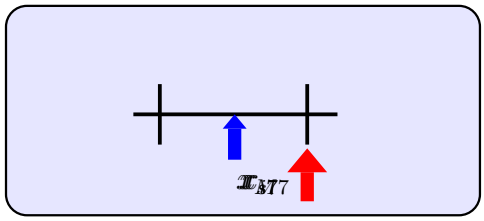

Figura 2.10: Intervalo 7a iteración

2.3. Método de la falsa posición (interpolación lineal inversa)

Este método se basa en el siguiente teorema:

Teorema de la Regla Falsa:
Sea $f(x)$ un polinomio de coeficientes reales, con grado $f(x) \geq 2$, y sean a y b numeros reales ($a < b$) tales que:

1. $f(a) * f(b) < 0$
2. $f''(x)$ no tiene raices en $[a, b]$

Si β_1 es el extremo del intervalo $[a, b]$, tal que $f(\beta_1) * f''(\beta_1) < 0$ (es decir, $\beta_1 = a$ si $f(a) * f''(a) < 0$ ó $\beta_1 = b$ si $f(b) * f''(b) < 0$) y a_1 es el extremo del intervalo $[a, b]$ tal que $f(a_1) * f''(a_1) > 0$ (es decir, $a_1 = a$ si $f(a) * f''(a) > 0$ ó $a_1 = b$ si $f(b) * f''(b) > 0$), entonces la sucesión β_n, donde β_1 es como ya se dijo, y

$$\beta_1 = a_1 - \frac{f(a_1)}{f(b_n) - f(a_1)}(\beta_1 - \alpha_1) \tag{2.2}$$

para n = 1, 2, 3, ..., converge a la única raíz ζ de $f(x)$ en $[a, b]$.

El método es similar al de la bisección salvo que la siguiente iteración se toma en la intersección de una recta entre el par de valores de x y el eje de las abscisas en lugar de tomar el punto medio. El reemplazo de la curva por una línea recta da una *posición falsa* de la raíz, de aquí el nombre del método también llamado de la regla falsa.
Para aplicarlo se eligen los extremos x_i y x_s del intervalo entre los que se encuentra la raíz, verificando que se cumpla que $f(x_i) * f(x_s) < 0$.
De acuerdo con la Figura **??**, por semejanza de triángulos, se tiene la siguiente igualdad:

$$\frac{f(x_i)}{x_M - x_i} = \frac{f(x_s)}{x_M - x_s} \tag{2.3}$$

Y despejando de la ecuación 2.3 el valor de x_M, que es una aproximación de la raíz, se obtiene la siguiente fórmula de iteración o recurrencia:

$$x_M = x_s - f(x_s)\frac{x_s - x_i}{f(x_s) - f(x_i)} \tag{2.4}$$

El valor de x_M, calculado con la ecuación 2.4, se reemplaza a uno de los dos valores, x_i o x_s que produzca un valor de la función que tenga el mismo signo de $f(x_M)$. De esta manera los valores x_i y x_s siempre encierran a la raíz.

- Si $f(x_M) = 0$ el proceso termina.
- Si $f(x_M)$ tiene el mismo signo de $f(x_i)$, el próximo paso es elegir $x_i = x_M$ y $x_s = x_s$.
- Si $f(x_M$, tiene el mismo signo de $f(x_s)$, el próximo paso es elegir $x_i = x_i$ y $x_s = x_M$.

El proceso iterativo continúa hasta alcanzar el margen de error.
La Figura 2.11 muestra gráficamente un esquema del método.

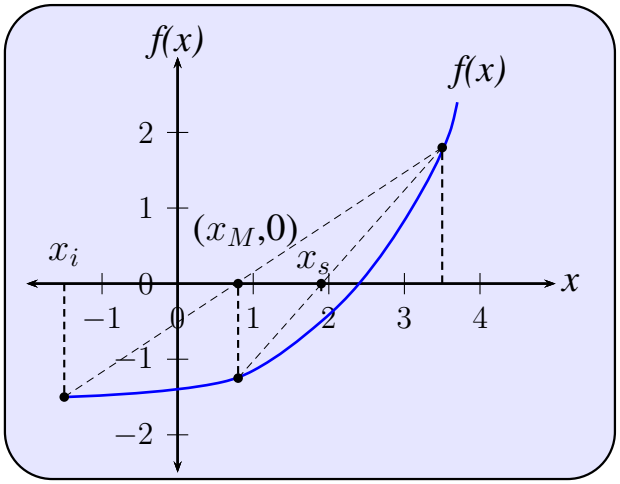

Figura 2.11: Esquema del método de la Falsa Posición

Una de la ventajas de este método es que siempre es convergente, y lo hace más rápidamente que el método de la bisección.
Una desventaja es que si existe más de una raíz en el intervalo, el método permite encontrar sólo una de ellas.

Algorithm 4: MÉTODO DE FALSA POSICIÓN

Entradas: límite inferior x_i, límite superior x_s, raíz aproximada x_M, valor de la función en x_i $f(x_i)$, valor de la función en x_s $f(x_s)$, tolerancia ε. número máximo de iteraciones N

Salidas: valor final aproximado de la raíz x_f

```
INICIO
    i = 1;
    Mientras i ≤ N hacer
        x_M = x_s − f(x_s) ∗ (x_s − x_i)/(f(x_s) − f(x_i));
        if |(x_M − x_f)| ≤ ε o f(x_M) = 0 then
            x_f = x_M;
        i = i + 1;
        if f(x_i) ∗ f(x_M) > 0 then
            x_i = x_M
        else
            x_s = x_M
    "Procedimiento completado sin éxito después de N iteraciones";
FIN
```

Ejemplo (- 15)

Determinar, aplicando el método de la falsa posición, una de las raíces de la función:

$$x^3 + x^2 - 3x - 3 = 0$$

considerando que la función cambia de signo en el intervalo (1,2) y un margen de error de $\varepsilon = 0.01$.

Se iniciarán los cálculos con los valores iniciales $x_i = 1$ y $x_s = 2$

1a Iteración

Para: $x_{i1} = 1$ tenemos que: $f(x_{i1}) = (1)^3 + (1)^2 - 3(1) - 3 = -4$

Para: $x_{s1} = 2$ tenemos que: $f(x_{s1}) = (2)^3 + (2)^2 - 3(2) - 3 = 3$

$$x_{M1} = x_{s1} - f(x_{s1})\frac{x_{s1} - x_{i1}}{f(x_{s1}) - f(x_{i1})} = 2 - 3\frac{2-1}{3-(-4)} = 1.571428571428571428571428571428571$$

$f(x_{M1}) = f(_{1.571428571428571428571428571428571}) = (1.571428571428571428571428571428571)^3 +$
$(1.571428571428571428571428571428571)^2 - 3(1.571428571428571428571428571428571) - 3 =$

$-1.364431486880466472303206997088 9$

$$f(x_{i1}) * f(x_{M1}) = (-4) * (-1.875) = 7.5$$

Como $f(x_i)$*$f(x_M) > 0$, la raiz se encuentra en el 2o subintervalo.

2a Iteración

Como $f(x_{M1}) = -1,36449$ tiene el mismo signo que $f(x_{i1}) = -4$, x_{M1} se convierte en el límite superior de la siguiente iteración
Para: $x_{i2} = 1$ tenemos que: $f(x_{i2}) = (1)^3 + (1)^2 - 3(1) - 3 = -4$
Para: $x_{s2} = 1.571428571428571428571428571428 6$

tenemos que:

$f(x_{s2}) = (1.571428571428571428571428571428 6)^3$
$+(1.571428571428571428571428571428 6)^2$
$-3(1.571428571428571428571428571428 6)$
$-3 = -1.364431486880466472303206997088 9$
$f(x_{i2}) * f(x_{M2}) = (-4) * (-1.364431486880466472303206997088 9) =$
$5.457725947521865889212827988 3555$

Como $f(x_i)$*$f(x_M) > 0$, la raiz se encuentra en el 2o subintervalo.

3. Métodos abiertos

3.1. Método de Newton-Raphson

Es el año de 1669 en la Inglaterra gobernada por el rey Carlos II y en la Universidad de Cambridge Issac Newton físico, filosofo y matemático inglés estando a cargo de la cátedra Lucasiana genera la primera referencia escrita sobre su método para encontrar raíces mediante una carta a sus colegas Barrow y Collins titulada *De analysis per aequations numero terminorum infinitas*.
El así denominado Método de Newton-Raphson es un método muy poderoso para resolver ecuaciones de la forma:

$$f(x) = 0$$

Una primera aproximación al método es partir del método de la falsa posición, y en vez de trazar una cuerda entre los dos extremos del intervalo, se traza una tangente. El punto donde esta tangente corta al eje x representa una aproximación mejorada de la raíz.
Suponiendo que para el mismo intervalo $[a, b]$ se traza la tangente que pasa por $f(b)$. En consecuencia se tiene que:

$$t(x) = f'(b)(x - b) + f(b) \tag{2.5}$$

Cuando $f(x) = 0$ también se cumple que $t(x) = 0$, entonces se busca una x_1 tal que $y(x_1) = 0$ para ir aproximando la raíz. Así se obtiene:

$$t(x_1) = 0 = f'(b)(x_1 - b) + f(b) \tag{2.6}$$

$$x_1 = b - \frac{f(b)}{f'(b)} \tag{2.7}$$

Generalizando se tiene:

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \tag{2.8}$$

La Figura 2.12 muestra gráficamente un esquema del método.

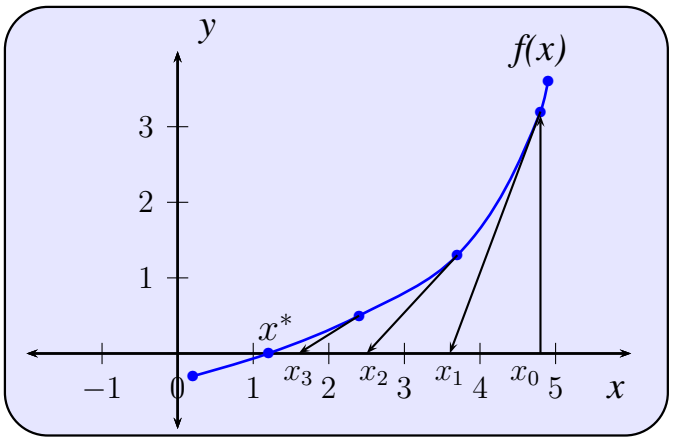

Figura 2.12: Esquema del método de Newton-Raphson

Otra forma de deducirlo es a través de la expansión de $f(x)$ de una serie de Taylor alrededor de x_0, se tiene:

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x - x_0)^2}{2!}f''(x_0) + ...+ \tag{2.9}$$

Si $f(x) = 0$, entonces, x es una raíz y el lado derecho de ecuación 2.9 constituye una ecuación para obtener esa raíz, pero esta ecuación es un polinomio de grado infinito.
Sin embargo, un valor aproximado de la raíz x puede ser obtenido, tomando solamente los dos primeros términos de la serie anterior, queda:

$$0 = x_0 - \frac{f(x_0)}{f'(x_0)} \tag{2.10}$$

de donde, al resolver para x, se tiene:

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \tag{2.11}$$

Como se ve, las ecuaciones 2.8 y 2.11 muestran la misma expresión denominada Fórmula de Newton - Raphson; una ventaja de este método es su rápida convergencia y algunas desventajas son que no siempre converge, depende de la función, además no es conveniente en caso de raíces múltiples y puede alejarse del área de interés si la pendiente es cercana a cero (lo que muestra una tendencia a caer en un máximo o en un mínimo de una función).

Algorithm 5: MÉTODO DE NEWTON - RAPHSON

Entradas: valor inicial de la de la raíz aproximada x_0, raíz aproximada x_M, valor de la función en x_M $f(x_M)$, valor de la 1a derivada de la función en x_M $f'(x_M)$, tolerancia ε. número de iteraciones N

Salidas: valor final aproximado de la raíz x_f

```
INICIO
    i = 1;
    Mientras i ≤ N hacer
        x_M = x_0 − f(x_0)/f'(x_0);
        if |(x_M − x_0)| ≤ ε o f(x_M = 0) then
            x_f = x_M;
        i = i + 1;
        x_0 = x_M
    "Procedimiento completado sin éxito después de N iteraciones";
FIN
```

Ejemplo (- 16)

Determinar, aplicando el método de Newton-Raphson, una de las raíces de la función:

$$x^2 - 4x + 2 = 0$$

Como $f(x) = x^2 - 4x + 2$, su primera derivada es: $f'(x) = 2x - 4$, de acuerdo a la fórmula de recurrencia (ec. 2.8):

1a Iteración

Considerando un valor de $x_0 = 1$ y un margen de error de $\varepsilon = 0.01$.

$$x_{M1} = 1 - \frac{(1)^2 - 4(1) + 2}{2(1) - 4} = 0.5$$

Como $|0.5 - 1| = 0.5 > \varepsilon(0.01)$ pasamos a:

2a Iteración

$$x_{M2} = 0.5 - \frac{(0.5)^2 - 4(0.5) + 2}{2(0.5) - 4} = 0.583333333$$

Como $|0.583333333 - 0.5| = 0.083333333 > \varepsilon(0.01)$ pasamos a:

3a Iteración

$$x_{M3} = 0.583333333 - \frac{(0.583333333)^2 - 4(0.583333333) + 2}{2(0.583333333) - 4} = 0.585784314$$

Como $|0.585784314 - 0.583333333| = 0.002450981 < \varepsilon(0.01)$
El valor aproximado de la raiz es igual a:

$x_f = 0.585784314$

3.2. Prueba de convergencia de Newton-Raphson

Como se puede observar si no se elige un x_0 lo suficientemente cerca, el método puede no converger. Para esto se tiene el siguiente teorema.

Teorema:
Sea $f \in C^2[a;b]$; si $\bar{x} \in [a;b]$ es tal que $f(\bar{x}) = 0$ y $f'(\bar{x}) = 0$, entonces existe un un $\delta > 0$ tal que el método de Newton-Raphson genera una sucesión $\{x_n\}_{n=1}^{\infty}$ que converge a $\bar{x}$ para cualquier aproximación inicial $x_0 \in [\bar{x} - \delta; \bar{x} + \delta]$.

Demostración
La demostración se basa en analizar el método de Newton-Raphson como si fuera el método de las aproximaciones sucesivas, considerando que
$x_n = g(x_{n-1})$, y $n \geq 1$, y que

$$g(x) = x - \frac{f(x)}{f_0(x)}$$

Entonces, sea k un número cualquiera en (0;1). En primer lugar debemos encontrar un intervalo $[\bar{x} - \delta; \bar{x} + \delta]$ que g «mapee» en sí mismo y en el que $|g'(x)| \leq k$ para toda $x \in [\bar{x} - \delta; \bar{x} + \delta]$.
Como $f'(\bar{x}) \neq 0$ y f' es continua, existe $\delta_1 > 0$ tal que $f'(x) \neq 0$ para $x \in [\bar{x} - \delta; \bar{x} + \delta] \subset [a;b]$.
Por lo tanto, g está definida y es continua en $[\bar{x} - \delta; \bar{x} + \delta]$. Por otro lado tenemos que

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

para $x \in [\bar{x} - \delta_1; \bar{x} + \delta_1]$ y como $f \in C^2[a;b]$, tendremos que $g \in C^1[a;b]$.
Como hemos supuesto que $f(\bar{x}) = 0$, entonces

$$g'(\bar{x}) = \frac{f(\bar{x})f''(\bar{x})}{[f'(\bar{x})]^2} = 0$$

.

Además g' es continua y k es tal que $0 < k < 1$, entonces existe un δ, tal que $0 < \delta < \delta_1$, y

$$g'(x) \leq k \; para \; toda \; x \in [\bar{x} - \delta; \bar{x} + \delta]$$

.

Nos falta todavía demostrar que $g : [\bar{x} - \delta; \bar{x} + \delta] \rightarrow [\bar{x} - \delta; \bar{x} + \delta]$. Si $x \in [\bar{x} - \delta; \bar{x} + \delta]$. El teorema del valor medio implica que existe un número ξ entre x y $\bar{x}$ para el que se cumple

$$|g(x) - g(\bar{x})| = |g'(\xi)||x - \bar{x}|$$

Por lo tanto, se cumple que

$$|g(x) - \bar{x}| = |g(x) - g(\bar{x})| = |g'(\xi)||x - \bar{x}| \leq k|x - \bar{x}| < |x - \bar{x}|$$

Como $x \in [\bar{x}-\delta; \bar{x}+\delta]$, podemos deducir que $|x-\bar{x}| < \delta$ y que $|g(x)-\bar{x}| < \delta$. Este último resultado nos muestra que $g : [\bar{x} - \delta; \bar{x} + \delta] \to [\bar{x} - \delta; \bar{x} + \delta]$.

En consecuencia, la función $g(x) = x - f(x)/f'(x)$ satisface todas las hipótesis del teorema 3.5, de modo que la sucesión $\{x_n\}_{n=1}^{\infty}$ definida por

$x_n = g(x_{n-1}) = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}$, para $n \geq 1$,

converge a $\bar{x}$ para cualquier $x_0 \in [\bar{x} - \delta; \bar{x} + \delta]$.

Como vimos este método es una variante del método de las aproximaciones sucesivas. Si la función $f(x)$ no tiene derivada en el entorno $[a; b]$ no es posible aplicarlo, pero si resulta difícil calcularla o evaluarla, existe un método alternativo denominado *método de la secante*, el cual reemplaza $f'(x_{n-1})$ por su aproximación discreta, es decir,

$$f'(x_{n-1}) = \frac{f(x_{n-1}) - f(x_{n-2})}{x_{n-1} - x_{n-2}}.$$

Si reemplazamos esto último en la fórmula de Newton-Raphson tenemos

$$x_n = x_{n-1} - \frac{f(x_{n-1})(x_{n-1} - x_{n-2})}{f(x_{n-1}) - f(x_{n-2})}$$

,

que también podemos escribir como

$$x_n = \frac{f(x_{n-1})x_{n-2} - f(x_{n-2})x_{n-1}}{f(x_{n-1}) - f(x_{n-2})}$$

3.3. Condiciones de convergencia

1. **La existencia de al menos una Raíz.**

   Dado un cierto intervalo de trabajo $[a, b]$, dentro del mismo debe cumplirse que $f(a) * f(b) < 0$.

2. **Unicidad de la Raíz.**

   Dentro del intervalo de trabajo $[a, b]$, la derivada de $f(x)$ debe ser diferente de cero.

3. **Concavidad.**

   La gráfica de la función $f(x)$ dentro del intervalo de trabajo $[a, b]$, debe ser cóncava, hacia arriba o hacia abajo. Para ello debe verificarse que:

   $f''(x) <= 0$ ó $f''(x) >= 0$ para toda x que pertenezca a $[a, b]$

4. **Intersección de la Tangente a f(x), dentro de [a,b]**

   Se debe asegurar que la tangente a la curva en el EXTREMO del intervalo $[a, b]$ en el cual $f'(x)$ sea mínima, intersecte al eje x dentro del intervalo $[a, b]$.

   De esta manera se asegura que la sucesión de valores de x_i caigan dentro de $[a, b]$.

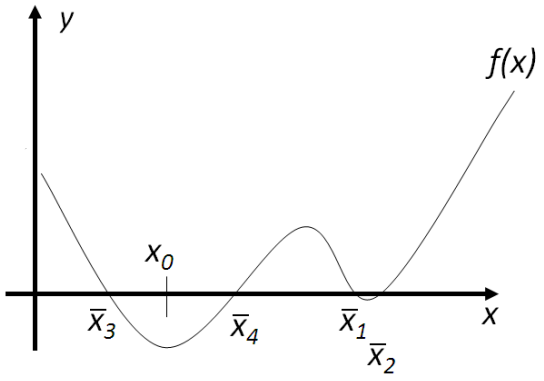

Figura 2.13: Raíces repetidas por pares y muy cercanas entre sí

3

APROXIMACIÓN DE SISTEMAS DE ECUACIONES

"Reconocer la importancia de resolver sistemas de ecuaciones utilizando la computadora e implementar programas que desarrollen los métodos discutidos."

Objetivos particulares 1 y 2

1. Definiciones y nomenclatura

Matrices

13: **Matriz**

DEFINICIÓN **(3.1 ▷ Matriz)**

Es una tabla o arreglo rectangular de números, los cuales se denominan *elementos de la matriz*; en este arreglo, las líneas horizontales se denominan **filas** y las líneas verticales se denominan **columnas**. A una matriz con m filas y n columnas se le denomina matriz $(m \times n)$, y m y n son sus **dimensiones**. La dimensión (u orden) de una matriz se da con el número de filas primero y el número de columnas después.

HISTORIA

James Joseph Sylvester (1814,1897), matemático inglés, fue el primero que empleó el término *matriz*, en el año 1850.

EL elemento de una matriz A que se encuentra en la i-ésima fila y la j-ésima columna se conoce como elemento i, j o elemento (i, j)-ésimo de A; esto se representa como Ai, j o $A[i, j]$.
Normalmente se escribe $A := a_{i,j}(m \times n)$ para definir una matriz $A(m \times n)$ con cada elemento en la matriz $A[i, j]$ llamado a_{ij} para todo $1 \leq i \leq m$ y $1 \leq j \leq n$, aunque esta convención del inicio de los índices i y j en 1 no es universal: algunos lenguajes de programación inician en cero, en cuyo caso se tiene $0 \leq i \leq m - 1$ y $0 \leq j \leq n - 1$.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix}$$

Matriz $(m \times n)$

Una matriz con una sola columna o una sola fila se conoce como $vector$, y se interpreta como un elemento del espacio euclídeo.
La **diagonal principal** de una matriz es la diagonal que va desde la esquina superior izquierda hasta la esquina inferior derecha, desde luego esto solo ocurre en aquellas matrices donde $m = n$.

Algunos tipos de matrices

A continuación una clasificacion basica de las matrices de acuerdo con los siguientes criterios:

Según su orden

Vector fila
Su orden es $[1 \times n]$ ya que consisten de una fila y n columnas

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \dots & a_{1n} \end{bmatrix}$$

Matriz $(1 \times n)$

Vector columna
Su orden es $[m \times 1]$ ya que consisten de m filas y una columna.

$$A = \begin{bmatrix} a_{11} \\ . \\ . \\ . \\ a_{m1} \end{bmatrix}$$

Matriz $(m \times 1)$

Matriz cuadrada
Su orden es $(m \times m)$, por lo que normalmente se conocen simplemente como de orden m.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & ... & a_{1m} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ a_{m1} & a_{m2} & a_{m3} & ... & a_{mm} \end{bmatrix}$$

Matriz cuadrada $(m \times m)$

Matriz rectangular
Su orden es $(m \times n)$ pero el numero de filas y el de columnas no coincide, es decir $m \neq n$.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & ... & a_{1n} \\ . & . & . & . & . \\ a_{m1} & a_{m2} & a_{m3} & ... & a_{mn} \end{bmatrix}$$

Matriz rectangular$(m \times n)$

Según su contenido

Matrices llenas pero no muy grandes
Por *llenas* se entiende que poseen pocos elementos nulos y por *no muy grandes* que el número de ecuaciones es de unos pocos miles a lo sumo. Estas matrices aparecen en problemas estadísticos, matemáticos, físicos y de ingeniería.
Para resolver este tipo de matrices generalmente, se requiere de **métodos directos**.

Matrices vacías y muy grandes
En oposición al caso anterior, *vacías* indica que hay pocos elementos no nulos y además están situados con una cierta regularidad. En la mayoría de estos casos el numero de ecuaciones supera los miles y puede llegar en ocasiones a los millones. Estas matrices son comunes en la resolución de ecuaciones diferenciales de problemas de ingeniería.
Para resolver este tipo de matrices generalmente, se requiere de **métodos indirectos** (o de aproximaciones sucesivas).

Según sus elementos

Matriz nula
Es aquella en la que todos sus elementos son nulos, se representa por $O_{m\times n}$.

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Matriz $(O_{2\times 3})$

Matriz escalonada
Es una matriz en la que, al principio de cada fila (columna) hay al menos un elemento nulo mas que en la fila (columna) anterior.

$$A = \begin{bmatrix} 3 & 0 & 5 \\ 0 & 4 & -1 \\ 0 & 0 & 5 \end{bmatrix}$$

Matriz escalonada por filas

$$A = \begin{bmatrix} 3 & 0 & 0 \\ 2 & 0 & 0 \\ 4 & 1 & 0 \\ -6 & 4 & -3 \end{bmatrix}$$

Matriz escalonada por columnas

Matriz triangular superior
Es una matriz cuadrada en la que todos los elementos que están por debajo de su diagonal principal son nulos.

$$A = \begin{bmatrix} 1 & 6 & 4 \\ 0 & 3 & 8 \\ 0 & 0 & 7 \end{bmatrix}$$

Matriz triangular superior

Matriz triangular inferior
Es una matriz cuadrada en la que todos los elementos que están por encima de su diagonal principal son nulos.

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ 4 & 5 & 7 \end{bmatrix}$$

Matriz triangular inferior

Matriz diagonal
Es una matriz cuadrada en la que todos los elementos que no están en su diagonal principal son nulos.

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 7 \end{bmatrix}$$

Matriz diagonal

Matriz escalar
Es una matriz cuadrada en la que todos los elementos que están en su diagonal principal son iguales.

$$A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

Matriz escalar

Matriz identidad (unidad)
Es una matriz cuadrada en la que todos los elementos que están en su diagonal principal son 1.

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Matriz identidad (unidad) I_3

2. Determinantes

FUe Carl Friedrich Gauss introdujo por primera vez el concepto de **determinante** el cual aparece en su obra *Disquisitiones arithmeticae* (1801) mientras discutía las formas cuadráticas; sin embargo el concepto presentado por Gauus no es el mismo que el concepto actual de determinante.

14: **Determinante**

HISTORIA

Carl Gustav Jacob Jacobi publicó tres tratados sobre determinantes en 1841, en los cuales da a conocer ampliamente la idea de determinante y presenta por primera vez la definición de determinante de manera algorítmica y, al no especificar las entradas en el determinante, sus resultados tuvieron una mayor aplicación.

DEFINICIÓN (**3.3 ▷ Determinante**)

Es una función la cual acepta como entrada una matriz de $m \times m$ y cuya salida puede ser un número real, cero o un número complejo llamado determinante de la matriz de entrada. Una forma para definir el valor del determinante es:

$$det(A) = \sum_{i1,i2,...in} \pm a_{i1}a_{i2}...a_{in}$$

Donde los términos $(i_1 i_2 ... i_n)$ se suman en todas las permutaciones, y su signo es positivo (+) si la permutación es par, y negativo (-), si es impar.

HISTORIA

Takakazu Seki, matemático japonés, fue el primero en estudiar los factores determinantes en 1683. Diez años más tarde, Leibniz, de forma independiente, utilizó determinantes para resolver ecuaciones simultáneas, aunque la versión de Seki es la más general.

Para la siguiente matriz A de (2×2):

$$A = \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} \tag{3.1}$$

Existe una forma sencilla para calcular su determinante mediante la siguiente expresión (fórmula de Leibniz):

$$det(A) = a_1 * b_2 - b_1 * a_2 \tag{3.2}$$

Para la siguiente matriz A de (3×3):

$$A = \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix} \tag{3.3}$$

El determinante de una matriz de (3×3) puede calcularse agregando las dos primeras columnas de A (que conformarían las columnas 4 y 5 de la matriz ampliada A') y luego sumando los productos de las 3 diagonales principales y restándoles los productos de las tres diagonales secundarias (regla de Sarrus):

$$A' = \begin{bmatrix} a_1 & b_1 & c_1 & a_1 & b_1 \\ a_2 & b_2 & c_2 & a_2 & b_2 \\ a_3 & b_3 & c_3 & a_3 & b_3 \end{bmatrix} \tag{3.4}$$

$$det(A) = a_1 * b_2 * c_3 + b_1 * c_2 * a_3 + c_1 * a_2 * b_3 - c_1 * b_2 * a_3 - a_1 * c_2 * b_3 - a_1 * c_2 * b_3 \tag{3.5}$$

Para la siguiente matriz A de (4×4):

$$A = \begin{bmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \\ a_4 & b_4 & c_4 & d_4 \end{bmatrix} \tag{3.6}$$

El determinante de una matriz de (4×4) puede calcularse mediante cofactores (regla de Laplace) para el cálculo de cofactores ver ecuación 3.8.

$$\begin{array}{rcl} det(A) & = & ((a_1b_2 - a_2b_1)c_3 + (a_3b_1 - a_1b_3)c_2 + (a_2b_3 - a_3b_2)c_1)d_4 \\ & + & ((a_2b_1 - a_1b_2)c_4 + (a_1b_4 - a_4b_1)c_2 + (a_4b_2 - a_2b_4)c_1)d_3 \\ & + & ((a_1b_3 - a_3b_1)c_4 + (a_4b_1 - a_1b_4)c_3 + (a_3b_4 - a_4b_3)c_1)d_2 \\ & + & ((a_3b_2 - a_2b_3)c_4 + (a_2b_4 - a_4b_2)c_3 + (a_4b_3 - a_3b_4)c_2d_1 \end{array} \tag{3.7}$$

Propiedades de los determinantes:

1. Si una matriz tiene una fila o columna con valores nulos, el determinante vale cero

2. Si una matriz tiene dos filas iguales o proporcionales, su determinante es nulo.

3. Si se permutan dos líneas paralelas de una matriz cuadrada su determinante cambia de signo

4. Si se multiplican todos los elementos de una línea de un determinante por un número, el determinante queda multiplicado por ese número

5. Si a una línea de una matriz se le suma otra línea multiplicada por un número, el determinante no cambia

6. El determinante de una matriz es igual al de su transpuesta

7. Si A tiene matriz inversa (A^{-1}), se verifica que

$$det(A^{-1}) = \frac{1}{det(A)}$$

Cofactores

Sea A una matriz cuadrada:
El *menor* del elemento a_{ij} se denota por M_{ij} y es el determinante de la matriz que queda después de borrar el renglón i y la columna j de A.
El *cofactor* de a_{ij} se denota como C_{ij} y está dado por

$$C_{ij} = (-1)^{i+j} M_{ij} \tag{3.8}$$

Observe que menor y el cofactor a veces difieren en el signo y a veces tienen el mismo signo

$$C_{ij} = \pm M_{ij} \tag{3.9}$$

Ejemplo (- 17)

Determinar el menor y el cofactor de a_{11} y a_{32} de la siguiente matriz A.

$$A = \begin{bmatrix} 1 & 0 & 3 \\ 4 & -1 & 2 \\ 0 & -2 & 1 \end{bmatrix}$$

$$M_{11} = \begin{bmatrix} . & . & . \\ . & -1 & 2 \\ . & -2 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ -2 & 1 \end{bmatrix} = (-1 \times 1) - (2 \times (-2)) = 3$$

$$C_{11} = (-1)^{1+1} M_{11} = (-1)^2(3) = 3$$

$$M_{32} = \begin{bmatrix} 1 & . & 3 \\ 4 & . & 2 \\ . & . & . \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix} = (1 \times 2) - (3 \times 4) = -10$$

$$C_{32} = (-1)^{3+2} M_{32} = (-1)^5(-10) = 10$$

El determinante de una matriz cuadrada es la suma de los productos de los elementos de la 1a fila por sus cofactores.

Si A es de 3×3, $|A| = a_{11}C_{11} + a_{12}C_{12} + a_{13}C_{13}$
Si A es de 4×4, $|A| = a_{11}C_{11} + a_{12}C_{12} + a_{13}C_{13} + a_{14}C_{14}$
Si A es de $n \times n$, $|A| = a_{11}C_{11} + a_{12}C_{12} + a_{13}C_{13} + ... + a_{1n}C_{1n}$
expansion de cofactores de $|A|$

El determinante de cualquier matriz cuadrada es la suma de los productos de los elementos de cualquier fila o columna, por sus cofactores.
Expansion a lo largo de la fila i:
$|A| = a_{i1}C_{i1} + a_{i2}C_{i2} + a_{i3}C_{i3} + ... + a_{in}C_{in}$
Expansion a lo largo de la columna j:
$|A| = a_{1j}C_{1j} + a_{2j}C_{2j} + a_{3j}C_{3j} + ... + a_{nj}C_{nj}$

3. Matriz inversa

LA matriz identidad (*pag.*66) juega un rol, en algebra matricial, similar al que juega el número 1 en el álgebra regular. En particular, si A es cualquier matriz $(m \times n)$, entonces el producto $A \cdot I_n = A$, y si B es cualquier matriz $(m \times n)$, entonces el producto $B \cdot I_n = B$.
En el álgebra regular cada número real $a \neq 0$ tiene un único *inverso multiplicativo*; esto significa que hay un único número real a^{-1} tal que el producto $a \cdot a^{-1} = 1$. Por ejemplo, el inverso multiplicativo de 5 es 1/5 (el cual se denota por 5^{-1}) por lo que: $5 \cdot 5^{-1} = 1$.
Aplicando estas cuestiones a las matrices cuadradas tenemos que:

¿Dada una matriz $A(n \times n)$, es posible encontrar una matriz $B(n \times n)$ tal que, $A \cdot B = I_n$?

A continuación algunas definiciones que se aplican a matrices que no son necesariamente cuadradas.

Propiedades de la matriz inversa

- La inversa de una matriz, si existe, es única.
- La inversa del producto de dos matrices es el producto de las inversas cambiando el orden:

  $(A \cdot B)^{-1} = B^{-1} \cdot A^{-1}$
- Si la matriz es invertible, también lo es su transpuesta, y el inverso de su transpuesta es la transpuesta de su inversa, es decir:

  $(A^T)^{-1} = (A^{-1})^T$
- Y, como ya se señaló:

  $(A^{-1})^{-1} = 1$
- Una matriz es invertible si y sólo si el determinante de A es distinto de cero. Además la inversa satisface la igualdad:

  $A^{-1} = \frac{1}{det(A)} adj(A^T)$

  donde $adj(A^T)$ es la transpuesta de la matriz de adjuntos de A.

4. Soluciones de sistemas de ecuaciones

LOs problemas de matemáticas aplicadas sobre ciencia e ingeniería en muchas casos pueden reducirse a **sistemas de ecuaciones algebraicas lineales** que podrían llegar a incorporar miles de ecuaciones.
Para resolver problemas que involucran sistemas de ecuaciones lineales pueden utilizarse métodos directos y métodos indirectos:

- Los **métodos directos** (o exactos) proporcionan soluciones exactas del problema después de un número de operaciones algebraicas básicas, no presentan errores por truncamiento, y son usados cuando la mayoría de los coeficientes de la matriz A son distintos de cero y las matrices no son demasiado grandes, suelen ser algoritmos complicados de implementar.

- Los **métodos indirectos** (o iterativos) encuentran una solución x para un problema dado por el límite de una secuencia de soluciones aproximadas x_k, tienen asociado un error de truncamiento y se usan preferiblemente para matrices grandes ($n \gg 1000$) cuando los coeficientes de A son la mayoría nulos (matrices dispersas), son algoritmos sencillos de implementar que requieren una aproximación inicial y que en general no tiene porqué converger por lo que requieren un análisis de convergencia previo.

  A diferencia de los métodos directos, en los cuales se debe terminar el proceso para tener la respuesta, en los métodos iterativos se puede suspender el proceso al termino de una iteración y se obtiene una aproximación a la solución.

A continuación, algunos conceptos previos:
Un *sistema de ecuaciones lineales* es un conjunto de ecuaciones que deben resolverse *simultáneamente*, que tienen la forma general:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 ... + a_{1n}x_n = & c_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 ... + a_{2n}x_n = & c_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 ... + a_{3n}x_n = & c_3 \\ \vdots & \vdots \\ a_{11}x_1 + a_{12}x_2 + a_{13}x_3 ... + a_{1n}x_n = & c_n \end{cases} \tag{3.10}$$

Aplicando la definición de producto de matrices, en este sistema de n ecuaciones algebraicas lineales con n incógnitas, puede escribirse en forma matricial:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13}... & a_{1n} \\ a_{21} & a_{22} & a_{23}... & a_{2n} \\ a_{31} & a_{32} & a_{33}... & a_{3n} \\ \vdots & & & \\ a_{n1} & a_{n2} & a_{n3}... & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_n \end{bmatrix} \tag{3.11}$$

Este sistema de ecuaciones se simboliza como $[A]_{n\times n}[x]_{n\times 1} = [c]_{n\times 1}$, en donde A es la matriz del sistema, x es el vector incógnita y c es el vector de términos independientes, que en forma sintética se simboliza como $Ax = c$.
La matriz formada por A, a la que se le agrega el vector de términos independientes como última columna, se denomina **matriz ampliada del sistema** y se representa por:

$$[A^a] = \begin{bmatrix} a_{11} & a_{12} & a_{13}... & a_{1n} & c_1 \\ a_{21} & a_{22} & a_{23}... & a_{2n} & c_2 \\ a_{31} & a_{32} & a_{33}... & a_{3n} & c_3 \\ \vdots & & & & \\ a_{n1} & a_{n2} & a_{n3}... & a_{nn} & c_n \end{bmatrix} \tag{3.12}$$

La **solución de un sistema de ecuaciones** es un conjunto de valores de las incógnitas que satisfacen simultáneamente a todas las ecuaciones del sistema.
De acuerdo con su solución, un sistema puede ser **compatible**, si admite solución; o **incompatible** si no admite solución.
Un sistema *compatible* puede ser **determinado**, si la solución es única; o **indeterminado**, si la solución no es única.
Las soluciones para un sistema compatible de la forma $Ax = c$ permanecen invariantes ante las siguientes operaciones elementales:

- Intercambio de dos filas o renglones cualesquiera.
- Multiplicación de una fila por un escalar no nulo.
- Suma a una fila de una combinación lineal no nula de otro renglón.

Métodos Directos

4.1. Regla de Cramer

Un sistema de ecuaciones se denomina **sistema de Cramer** si tiene tantas ecuaciones como incógnitas, en ese caso la matriz es una matriz cuadrada.
Un sistema de ecuaciones es *compatible determinado* si tiene solución única.
Un sistema de Cramer es compatible determinado si y sólo si $det(A) \neq 0$.
En ese caso, se define la matriz A_j como la que se obtiene a partir de A sustituyendo la columna j por el vector c, esto es, si c_j es la columna j de A,

$$A = (c_1, c_2, ..., c_n), \qquad c_j = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{bmatrix} \tag{3.13}$$

entonces la matriz A_j tiene la siguiente estructura

$$A_j = (c_1, c_2, ..., c_{j-1}, c_{j+1}, ..., c_n) \tag{3.14}$$

El determinante de A_j queda,

$$det(A_j) = det(c_1, c_2, ..., c_{j-1}, c_{j+1}, ..., c_n). \tag{3.15}$$

Entonces la solución del sistema viene dada por la así denominada *regla de Cramer*

$$x_1 = \frac{det(A_1)}{det(A)}, x_2 = \frac{det(A_2)}{det(A)}, ..., x_n = \frac{det(A_n)}{det(A)}. \tag{3.16}$$

La expresión general de la solución por la *regla de Cramer* es:

$$x_i = \frac{\begin{vmatrix} a_{11} & ... & a_{1,i-1} & c_1 & a_{1,i+1} & ... & a_{1n} \\ a_{21} & ... & a_{2,i-1} & c_2 & a_{2,i+1} & ... & a_{2n} \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{n1} & ... & a_{n,i-1} & c_n & a_{n,i+1} & ... & a_{nn} \end{vmatrix}}{det(A)} \tag{3.17}$$

Para un sistema de ecuaciones lineales del tipo:

$$x_1 = \frac{\begin{vmatrix} c & b \\ f & b \end{vmatrix}}{\begin{vmatrix} a & b \\ d & e \end{vmatrix}} \qquad x_2 = \frac{\begin{vmatrix} a & c \\ d & f \end{vmatrix}}{\begin{vmatrix} a & b \\ d & e \end{vmatrix}} \tag{3.18}$$

Donde ni x_1 ni x_2 pueden ser negativos:

Algorithm 6: REGLA DE CRAMER

Entradas: a, b, c, d, e, f
Salidas: x_1, x_2

INICIO
  $x_1 = ((c * e) - (f * b))/((a * e) - (d * b))$;
  $x_2 = ((a * f) - (d * c))/((a * e) - (d * b))$;
  if $x_1 < 0$ *o* $x_2 < 0$) **then**
    El sistema no tiene soluciones;
  else
    Desplegar x_1, x_2
FIN

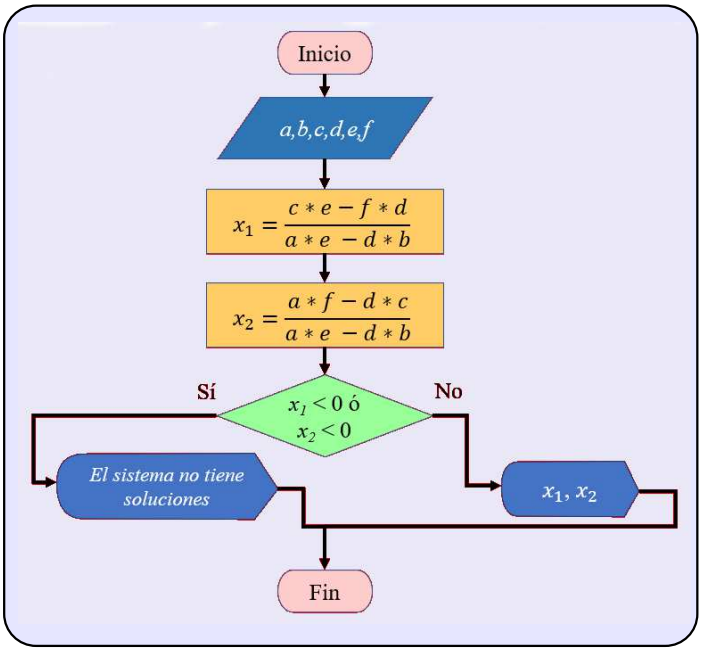

Figura 3.1: Diagrama de flujo de la regla de Cramer

Ejemplo (- 18)

Determinar, aplicando la regla de Cramer, los valores de las incógnitas del siguiente sistema de ecuaciones:

$$\begin{cases} 2x_1 + 3x_2 + x_3 = 3 \\ x_1 - x_2 + x_3 = 5 \\ x_2 + x_3 = -2 \end{cases}$$

La expresión matricial queda

$$\begin{bmatrix} 2 & 3 & 1 \\ 1 & -1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \\ -2 \end{bmatrix}$$

Calculando el determinante de la matriz de coeficientes queda

$$det(A) = \begin{vmatrix} 2 & 3 & 1 \\ 1 & -1 & 1 \\ 0 & 1 & 1 \end{vmatrix} = (-2) + 0 + 1 - 0 - 3 - 2 = -6$$

por lo tanto, el sistema es compatible determinado. Ahora se calcula

$$det(A_1) = \begin{vmatrix} 3 & 3 & 1 \\ 5 & -1 & 1 \\ -2 & 1 & 1 \end{vmatrix} = (-3) + (-6) + 5 + -2 - 15 - 3 = -24$$

$$det(A_2) = \begin{vmatrix} 2 & 3 & 1 \\ 1 & 5 & 1 \\ 0 & -2 & 1 \end{vmatrix} = 10 + 0 + (-2) - 0 - 3 - (-4) = 9$$

$$det(A_3) = \begin{vmatrix} 2 & 3 & 3 \\ 1 & -1 & 5 \\ 0 & 1 & -2 \end{vmatrix} = 4 + 0 + 3 - 0 - 10 - (-6) = 3$$

De donde se obtiene la solución del sistema

$$x_1 = \frac{det(A_1)}{det(A)} = \frac{-24}{-6} = 4, \quad x_2 = \frac{det(A_2)}{det(A)} = \frac{9}{-6} = -\frac{3}{2}, \quad x_3 = \frac{det(A_3)}{det(A)} = \frac{3}{-6} = -\frac{1}{2}$$

4.2. Método de Gauss

El procedimiento para resolver un sistema de ecuaciones lineales por medio del método de Gauss consta de dos pasos:

1. Eliminación hacia adelante de incógnitas.
2. Sustitución hacia atrás

Eliminación hacia adelante

En la eliminación hacia adelante, se reduce el conjunto de ecuaciones a un sistema triangular superior. El primer paso es multiplicar la primera ecuación (sistema de ecuaciones 3.10) por el cociente entre los coeficientes de la primera incógnita de la segunda y primera ecuación, $-\frac{a_{21}}{a_{11}}$, obteniéndose:

$$a_{21}x_1 + \frac{a_{21}}{a_{11}}a_{12}x_2 + ... + \frac{a_{21}}{a_{11}}a_{1n}x_n = \frac{a_{21}}{a_{11}}c_1 \tag{3.19}$$

Como el primer término de la primera ecuación modificada (3.19) es idéntico al primer término de la segunda ecuación del sistema, se elimina la primera incógnita restando la última ecuación de esta y se llega a:

$$(a_{22} - a_{21}\frac{a_{12}}{a_{11}})x_2 + ... + (a_{2n} - a_{21}\frac{a_{1n}}{a_{11}})x_n = c_2 - \frac{a_{21}}{a_{11}}c_1 \tag{3.20}$$

El procedimiento se repite con las ecuaciones restantes, en los pasos anteriores a la primera ecuación del sistema 3.10 se llama ecuación *pivote* y a_{11} se denomina *coeficiente* o $elemento\ pivote$.

$$Eliminacion_{adelante} \begin{bmatrix} a_{11} & a_{12} & a_{13} & \vdots & c_1 \\ a_{21} & a_{22} & a_{23} & \vdots & c_2 \\ a_{31} & a_{32} & a_{33} & \vdots & c_3 \end{bmatrix} \Rightarrow \begin{bmatrix} a_{11} & a_{12} & a_{13} & \vdots & c_1 \\ & a'_{22} & a'_{23} & \vdots & c'_2 \\ & & a''_{33} & \vdots & c''_3 \end{bmatrix} \tag{3.21}$$

Sustitución hacia atrás

De la primera ecuación del sistema 3.10 se despeja x_n:

$$x_n = \frac{c_n^{n-1}}{a_{nn}^{n-1}} \tag{3.22}$$

Este resultado se puede sustituir hacia atrás en la $(n-1)$iésima ecuación y despejar x_{n-1}, el procedimiento para despejar las incógnitas restantes se representa mediante la fórmula 3.23:

$$x_1 = \frac{c_i^{i-1} - \sum_{j=i+1}^{n} a_{ij}^{i-1}x_j}{a_{ii}^{i-1}} \text{ para } i = n-1, n-2, ..., 1 \tag{3.23}$$

$$Sustitucion_{atras} \begin{bmatrix} x_3 & = & c''_3 & / & a''_{33} \\ x_2 & = & (c'_2 - a'_{23}x_3) & / & a'_{22} \\ x_1 & = & (c_1 - a_{12}x_2 - a_{13}x_3) & / & a_{11} \end{bmatrix} \tag{3.24}$$

Algoritmo
Método de Gauss
Entradas: numero de ecuaciones n, elementos de la matriz ampliada $A'(1 \leq i \leq n, 1 \leq j \leq n+1)$: a_{ij}, indice del pivote: p, fila i: F_i.
Salidas:X (es decir $(x_1, x_2, ..., x_n)$)

Paso 1: Para $i = 1, ..., n-1$ seguir los Pasos 2 a 4. (Eliminación hacia adelante)
Paso 2: Sea p el menor entero con $i \leq p \leq n$ y $a_{pi} \neq 0$.
Si p no puede encontrarse, SALIDA ("No existe solución única")
PARAR
Paso 3: Si $p \neq i$ intercambiar la fila p por la fila i
Paso 4: Para $j = i+1, ..., n$ seguir los Pasos 5 a 6
Paso 5: Hacer $m_{ij} = \frac{a_{ij}}{a_{ii}}$
Paso 6: Realizar $F_j - m_{ij}F_i$ e intercambiarla por la fila F_i
Paso 7: Si $a_{nn} = 0$ entonces SALIDA ("No existe solucion unica")
PARAR
Paso 8: Hacer $x_n = \frac{c_n}{a_{nn}}$ (Sustitucion hacia atras)
Paso 9: Para $i = n-1, ..., 1$ tomar $x_i = \frac{1}{a_{ii}}(c_i - \sum_{j=i+1}^{n} a_{ij}x_j)$
Paso 10: SALIDA X (es decir $(x_1, x_2, ..., x_n)$)
PARAR

Ejemplo (- 19)

Determinar, aplicando el método de Gauss, los valores de las incógnitas del siguiente sistema de ecuaciones:

$$\begin{cases} x_1 + x_2 + 2x_3 = 3 \\ 3x_1 - x_2 + x_3 = 1 \\ 2x_1 + 3x_2 - 4x_3 = 8 \end{cases}$$

Eliminación hacia adelante

Como el coeficiente de la primera incógnita es 1, se multiplica la primera ecuación por 3 y se resta el resultado de la segunda ecuación, luego se multiplica por 2 la primera ecuación y se resta de la tercera de manera que el sistema queda reducido a:

$$\begin{cases} x_1 + x_2 + 2x_3 = 3 \\ -4x_2 - 5x_3 = -8 \\ x_2 - 8x_3 = 2 \end{cases}$$

Se procede ahora a eliminar la segunda incógnita de la tercera ecuación, para ello se divide la segunda ecuación por -4 y se multiplica por el coeficiente de la tercera ecuación que en este caso es 1, quedando la segunda como: $x_2 + \frac{5}{4}x_3 = 2$ y se resta este resultado de la tercera ecuación, el sistema queda:

$$\begin{cases} x_1 + x_2 + 2x_3 = 3 \\ -4x_2 - 5x_3 = -8 \\ -\frac{37}{4}x_3 = 0 \end{cases}$$

Sustitución hacia atrás

Se despeja x_3 de la tercera ecuación, en este caso: $x_3 = 0$, se reemplaza este valor en la segunda ecuación: $-4x_2 = -8$ por lo tanto $x_2 = 2$ y por último se reemplazan estos valores en la primera ecuación $x_1 + 2 = 3$ entonces $x_1 = 1$.

4.3. Método de Gauss-Jordan

Es diferente al método de eliminación gaussiana puesto que 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 identidad en lugar de una matriz triangular.
Además todos los renglones se normalizan al dividirlos entre su elemento pivote. De esta forma el paso de eliminación genera una matriz identidad en vez de una triangular.
Por consiguiente no es necesario emplear la sustitución hacia atrás para obtener la solución. A continuación se esquematiza el método.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & |c_1 \\ a_{21} & a_{22} & a_{23} & |c_2 \\ a_{31} & a_{32} & a_{33} & |c_3 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & 0 & |c_1^* \\ 0 & 1 & 0 & |c_2^* \\ 0 & 0 & 1 & |c_3^* \end{bmatrix} \Rightarrow \begin{array}{cccc} x_1 & & & = c_1^* \\ & x_2 & & = c_2^* \\ & & x_3 & = c_3^* \end{array} \tag{3.25}$$

Figura 3.2: Esquematizacion del metodo de Gauss-Jordan

Algorithm 7: MÉTODO DE GAUSS-JORDAN

Entradas: numero de ecuaciones: n, elementos de la matriz ampliada $A'(1 \leq i \leq n, 1 \leq j \leq n+1)$: a_{ij}, indice del pivote: p, fila i: F_i
Salidas: X (es decir $(x_1, x_2, ..., x_n)$)

```
INICIO
    A' ← [Ac];
    for i ← 1 to n do
        A(i,:) ← A(i,:)/A(i,i);
        for j ← 1 to n do
            if j ≠ i then
                A(j,:) ← A(j,:) − A(j,i)A(i,:)
    X ← A(:, n+1);
    return X
FIN
```

Ejemplo (- 20)

Se desea resolver el sistema anterior (ejemplo 19) utilizando este método, se escribe el sistema en forma matricial, se trabaja con la matriz ampliada (formada por la matriz de coeficientes a la que se le adiciona una última columna constituida por los términos independientes), luego se efectúan operaciones elementales en las filas hasta llegar a la matriz identidad quedando los valores de las incógnitas en la última columna de la matriz ampliada.

$$\begin{bmatrix} 1 & 1 & 2 & |3 \\ 3 & -1 & 1 & |1 \\ 2 & 3 & -4 & |8 \end{bmatrix} \begin{matrix} -3F_1+F_2 \\ -2F_1+F_2 \\ \longrightarrow \end{matrix} \begin{bmatrix} 1 & 1 & 2 & |3 \\ 0 & -4 & 5/4 & |2 \\ 0 & 1 & -8 & |2 \end{bmatrix} \begin{matrix} \\ -(1/4)F_{23} \\ \longrightarrow \end{matrix} \begin{bmatrix} 1 & 1 & 2 & |3 \\ 0 & 1 & -4 & |-5 \\ 0 & 1 & -8 & |2 \end{bmatrix} \begin{matrix} -F_2+F_1 \\ -F_2+F_3 \\ \longrightarrow \end{matrix}$$

$$\begin{bmatrix} 1 & 0 & 3/4 & |1 \\ 0 & 1 & 5/4 & |2 \\ 0 & 0 & -(37/4) & |0 \end{bmatrix} \begin{matrix} \\ -(4/37)F_3 \\ \longrightarrow \end{matrix} \begin{bmatrix} 1 & 0 & 3/4 & |1 \\ 0 & 1 & 5/4 & |2 \\ 0 & 0 & 1 & |0 \end{bmatrix} \begin{matrix} -(3/4)F_3+F_1 \\ -(5/4)F_3+F_2 \\ \longrightarrow \end{matrix} \begin{bmatrix} 1 & 0 & 0 & |1 \\ 0 & 1 & 0 & |2 \\ 0 & 0 & 1 & |0 \end{bmatrix}$$

Se llega al mismo resultado que con el método anterior, es decir $x_1 = 1$, $x_2 = 2$ y $x_3 = 0$, como puede verse, es un método que involucra alrededor de un 50 % de cálculos adicionales.

Métodos Indirectos

4.4. Método de Gauss Seidel

Los métodos indirectos (o iterativos) tienen la desventaja de que no se pueden aplicar, por lo menos de forma elemental, a cualquier sistema de ecuaciones lineales y tienen la ventaja de que son más eficaces para sistemas cuyo orden es superior a 1000.
Para éstos sistemas muy grandes los métodos iterativos frecuentemente ofrecen ventajas decisivas sobre los métodos directos en términos de velocidad y demandas sobre la capacidad de memoria de una computadora.
Otra ventaja importante de los métodos iterativos es que son por lo general estables, y que en realidad amortiguan los errores, debido al redondeo de errores de menor importancia, ya que el proceso es continúo.
Un método iterativo para resolver un sistema de ecuaciones lineales empieza con una aproximación inicial x_0 para una solución x y genera una secuencia de vectores $\{x^k\}_{k=0}^{\infty}$ que converge hacia x.
Partiendo del modelo de los sistemas de ecuaciones lineales (ec. 3.11) simbolizado por:

$$Ax = c \tag{3.26}$$

Se procede a descomponer la matriz de coeficientes A en:

$$A = D - E - F \tag{3.27}$$

Donde:
D es la diagonal principal de A.
E es estrictamente la matriz triangular inferior de A.
F es estrictamente la matriz triangular superior de A.

Gráficamente tenemos (fig.3.3):

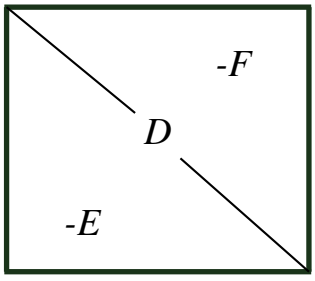

Figura 3.3: Particionamiento incial de la matriz A

Sustituyendo la ec.3.27 en la ec.3.26 tenemos:

$$(D - E - F)x = c \tag{3.28}$$

Despejando x de la ec.3.28 tenemos:

$$x = D^{-1}(E + F)x + D^{-1}c \tag{3.29}$$

Reescribiendo la ec.3.28 tenemos:

$$(D - F)x = Ex + c \tag{3.30}$$

Reemplazando la x de la derecha por $x^{(k)}$, que es la solución aproximada en el paso de iteración k, y la x de la izquierda por $x^{(k+1)}$, que es la solución aproximada en el paso de iteración $k + 1$, tenemos:

$$x^{(k+1)} = (D - F)Ex^{(k)} + (D - F)^{-1}c \tag{3.31}$$

Y éste es el llamado método de Gauss-Seidel.
Representando la ec.3.30 en forma matricial tenemos:

$$\begin{cases} a_{11}x_1^{(k+1)} & & & = c_1 - a_{12}x_2^{(k)} - ... - a_{1n}x_n^{(k)} \\ a_{21}x_1^{(k+1)} + a_{22}x_2^{(k+1)} & & & = c_2 - a_{23}x_3^{(k)} - ... - a_{2n}x_n^{(k)} \\ a_{31}x_1^{(k+1)} + a_{32}x_2^{(k+1)} + a_{33}x_3^{(k+1)} & & & = c_3 ... - a_{3n}x_n^{(k)} \\ \vdots & & & \vdots \\ a_{n1}x_1^{(k+1)} + a_{n2}x_2^{(k+1)} + ... + a_{nn}x_n^{(k+1)} & & & = c_n \end{cases} \tag{3.32}$$

Pero la notación matricial requiere calcular las inversas, por lo que es preferible una representación que resuelva en cada iteración, de manera secuencial cada elemento del vector solución, para ello se da a la expresión (3.31) el formato de serie simbolizado por:

$$x_i^{(k)} = \frac{-\sum_{(j=1)}^{(i-1)}(a_{ij}x_j^{(k))} - \sum_{(j=i+1)}^{(n)}(a_{ij}x_j^{(k-1)}) + b_i}{a_{ii}} \tag{3.33}$$

Donde:
$I = 1, 2, 3, ..., n$

Algorithm 8: MÉTODO DE GAUSS-SEIDEL

Entradas: numero de ecuaciones: n, elementos de la matriz ampliada $A'(1 \leq i \leq n, 1 \leq j \leq n+1)$: a_{ij}, indice del pivote: p, fila i: F_i

Salidas: X (es decir $(x_1, x_2, ..., x_n)$)

```
INICIO
    A' ← [Ac];
    for i ← 1 to n do
        A(i,:) ← A(i,:)/A(i,i);
        for j ← 1 to n do
            if j ≠ i then
                A(j,:) ← A(j,:) − A(j,i)A(i,:)
    X ← A(:, n+1);
    return X
FIN
```

Ejemplo (- 21)

Se tiene el siguiente sistema de ecuaciones lineales, resolverlo por el método de Gauss-Seidel, partiendo de ($x = 1$, $y = 2$) con una tolerancia al error (ε) de 0.001.

$$5x + 2y = 1$$

$$x - 4y = 0$$

Despejar la incógnita de la ecuación correspondiente:

$$\begin{array}{lll} 5x = 1 - 2y & \therefore \quad x = \frac{1}{5} - \frac{2}{5}y & \therefore \quad x = 0.2 - 0.4y \\ 4y = x & \therefore \quad y = \frac{1}{4}x & \therefore \quad y = 0.25x \end{array}$$

1a Iteración:
$x_1 = 0.2 - 0.4(2) = -0.6$
$y_1 = 0.25(-0.6) = -0.15$

2a Iteración:
$x_2 = 0.2 - 0.4(-0.15) = 0.26$
$y_2 = 0.25(0.26) = 0.065$

3a Iteración:
$x_3 = 0.2 - 0.4(0.065) = 0.174$
$y_3 = 0.25(0.174) = 0.0435$

4a Iteración:
$x_4 = 0.2 - 0.4(0.0435) = 0.1826$
$y_4 = 0.25(0.1826) = 0.04565$

5a Iteración:
$x_5 = 0.2 - 0.4(0.04565) = 0.18174$
$y_5 = 0.25(0.1874) = 0.045435$

Sustituyendo los valores (aproximados) de x_5 y y_5 en el sistema de ecuaciones tenemos:
$5(0.18174) + 2(0.045435) = 0.99957$
$(0.18174) - 4(0.045435) = 0$

La magnitud del error es:
$1 - 0.99957 = 0.00043 < \varepsilon(0.001)$

4 AJUSTE DE CURVAS

"Utilizar un método computacional para interpolar valores, a partir de un grupo de datos así como obtener una regresión a una función, a partir de un grupo de datos mediante la programación e implementar programas que desarrollen los métodos discutidos"

Objetivos particulares 1, 2 y 3

1. Introducción

UNa situación frecuente en ingeniería es la de tener una función f dada por un conjunto de puntos $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ donde falta la representación del modelo analítico.
La función f puede representar los datos de un experimento o el resultado de mediciones a gran escala de una magnitud física que no se puede convertir a una forma analítica sencilla.
Puede ser necesario evaluar la función f en algún punto x en el conjunto de datos $x_1, ..., x_n$, pero donde x es diferente de los valores tabulados, este proceso se denomina **ajuste de curvas**.

Tipos de ajustes

- **Ajuste por aproximación**: Se asume un modelo que se aproxima a lo observado.
- **Ajuste por interpolación**: Se asume un modelo que es exactamente lo observado.

En el ajuste por aproximación se puede tener relativa confianza en valores fuera del rango de la muestra y es robusto con respecto al ruido.
En el ajuste por interpolación se tiene relativa confianza solamente en valores dentro del rango de la muestra y es poco robusto con respecto al ruido.

Método de ajuste por aproximación

2. Método de Mínimos Cuadrados (aplicación regresión)

DAda la siguiente representación del conjunto de observaciones de un experimento se tiene:

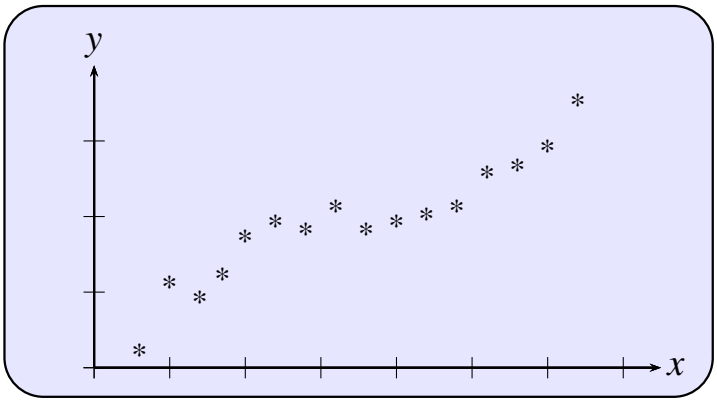

Figura 4.1: Imagen real con los datos originales

El ajuste de curvas por aproximación mediante el método de mínimos cuadrados es un proceso en el cual, dado un conjunto de N pares de puntos (x_n, y_n) (siendo x la variable independente y y la dependiente), se determina una función matemática $f(x)$ tal que la suma de los cuadrados de la diferencia entre la imagen real y la correspondiente obtenida mediante la función ajustada en cada punto sea mínima:

$$\varepsilon = min(\sum_{i}^{N}(y_i - f(x_i))^2) \tag{4.1}$$

Generalmente, se escoge una función genérica $f(x)$ en función de uno o más parámetros y se ajustan estos valores para que se minimice el error cuadrático, ε. La forma más típica de esta función ajustada es la de un polinomio de grado M; obteniéndose para $M = 1$ un ajuste lineal (o regresión lineal):

$$f(x) = a_0 + a_1 x \tag{4.2}$$

Como puede apreciarse la función describe una línea recta, en donde a_1 representa la pendiente de la recta y a_0 el punto de intersección sobre la ordenada.

En la regresión lineal existen dos tipos de relación:

- Sí $a_1 > 0$ hay relación lineal positiva (pendiente positiva).
- Sí $a_1 < 0$ hay relación lineal negativa (pendiente negativa).

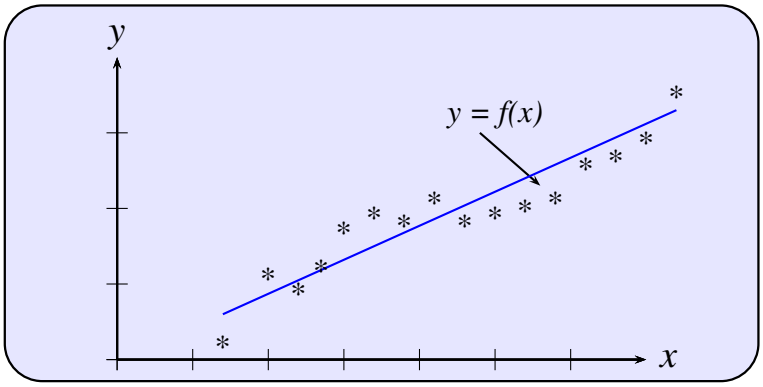

Figura 4.2: Datos originales y línea de ajuste de los datos

Entre los datos originales y las curvas aproximadas existe una diferencia de ordenadas que se conoce como desviación.

$$d = \hat{y} - y \tag{4.3}$$

Donde:
$d \Rightarrow$ desviación
$\hat{y} \Rightarrow$ ordenada del dato original
$y \Rightarrow$ ordenada de la curva de ajuste

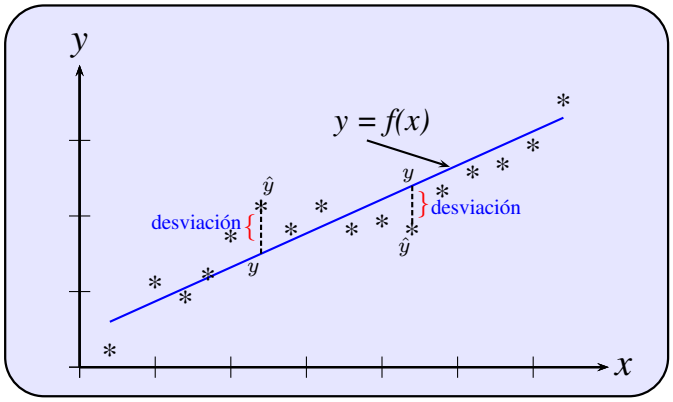

Figura 4.3: Desviaciones

La minimización de la sumatoria de las desviaciones cuadradas es el enfoque que generalmente se ha utilizado para encontrar las curvas o las funciones que mejor se ajustan a determinados datos.

Para calcular la pendiente (a_1) tenemos:

$$a_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \tag{4.4}$$

Para calcular la intersección (a_0) tenemos:

$$a_0 = \frac{\sum y - a_1(\sum x)}{n} \tag{4.5}$$

Donde:
n es el número de puntos de la imagen real
$\sum y = y_1 + y_2 + y_3 + ... + y_n$
$\sum x = x_1 + x_2 + x_3 + ... + x_n$
$\sum x^2 = x_1^2 + x_2^2 + x_3^2 + ... + x_n^2$
$\sum xy = x_1y_1 + x_2y_2 + x_3y_3 + ... + x_ny_n$
.

Algorithm 9: MÉTODO DE MÍNIMOS CUADRADOS

Entradas: numero de coordenadas: n, datos (x, y):
Salidas: a_0, a_1

INICIO
  $sum\_x, sum\_y, sum\_xy, sum\_x^2 = 0$;
  Mientras $i \leq (n-1)$ **hacer**
    $sum\_x \leftarrow sum\_x + x(i)$;
    $sum\_y \leftarrow sum\_y + y(i)$;
    $sum\_x^2 \leftarrow sum\_x + (x(i) * x(i))$;
    $sum\_xy \leftarrow sum\_xy + (x(i) * y(i))$;
    $i \leftarrow i + 1$;
  $Denominador \leftarrow sum\_x * sum\_y - n * sum\_x^2$;
  $a_1 \leftarrow (sum\_x * sum\_y - n * sum\_xy)/Denominador$;
  $a_0 \leftarrow (sum\_x * sum\_xy - sum\_x^2 * sum\_y)/Denominador$;
  Imprimir a_1, a_0
FIN

Ejemplo (- 22)

Calcular la función de regresión lineal por el método de mínimos cuadrados para el siguiente conjunto de datos:

y	x
7000	1
9000	2
5000	3
11000	4
10000	5
13000	6

Calcular las sumatorias:
$\sum y = 7000 + 9000 + 5000 + 11000 + 10000 + 13000 = 55000$
$\sum x = 1 + 2 + 3 + 4 + 5 + 6 = 21$
$\sum x^2 = (1)^2 + (2)^2 + (3)^2 + (4)^2(4)^2(6)^2 = 91$
$\sum xy = (1)(7000) + (2)(9000) + (3)(5000) + (4)(11000) + (5)(10000) + (6)(13000) = 212000$
Calcular la pendiente:
$a_1 = \frac{(6)(212000)-(21)(55000)}{(6)(91)-(21)^2} = 1114.285714$

Calcular la intersección:
$a_0 = \frac{(55000)-(1114.285714)(21)}{6} = 5266.666667$

La función queda:
$y = 5266.666667 + 1114.285714x$

3. Interpolación

LA interpolación consiste en determinar los parámetros de un modelo $y = f(x)$ que se ajuste mejor a los datos $(x_1, y_1), ..., (x_n, y_n)$ que están sujetos a errores aleatorios producidos por incertidumbres en las mediciones o, por un deficiente control de las condiciones en el que se realiza un experimento.
Este proceso sirve por ejemplo, para predecir una tendencia o comportamiento en el futuro (extrapolación) o en el pasado (interpolación) cuando se observan y se obtienen muestras u observaciones objetivas de un evento.
Así, si las muestras u observaciones son métricas (medibles con números reales) se puede buscar una función/curva que las relacione.

Métodos de ajuste por interpolación

3.1. Polinomio único de interpolación

El problema de la interpolación consiste en estimar el valor de una función en un punto a partir de valores conocidos en puntos cercanos.
En el caso de la **interpolación polinómica**, la función incógnita se sustituye por un polinomio que coincide con aquella en los puntos conocidos.
Se eligen los polinomios porque son fáciles de evaluar y por el hecho fundamental de que dados $n+1$ puntos de abscisa distinta, $(x_0, y_0), (x_1, y_1), ..., (x_n, y_n)$, existe exactamente un polinomio $P_n(x)$ de grado no superior a n, que pasa por dichos puntos, es decir, tal que:

$$P_n(x_i) = y_i, \qquad i = 0, 1, 2..., n \tag{4.6}$$

En la interpolación lineal, la función se sustituye por la recta que pasa por dos puntos. Tres datos se interpolan con un polinomio de segundo grado, gráficamente una parábola que pasa por esos tres puntos.
De acuerdo con lo anterior, se podría pensar que al aumentar el grado se obtiene mejor aproximación, pero esto es falso en general. La coincidencia del polinomio con muchos puntos de interpolación se consigue a costa de grandes oscilaciones en los intervalos entre nodos o puntos de interpolación dados.
Asumiendo un polinomio de la forma:

$$P_n(x_i) = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n \tag{4.7}$$

Al tener que cumplir con las restricciones (4.6), se generan $n + 1$ ecuaciones en $n + 1$ incógnitas; siendo éstas los coeficientes a_i's:

$$\begin{array}{rl} a_0 + a_1x_0 + a_2x_0^2 + a_3x_0^3 ... + a_nx_0^n = & y_0 \\ a_0 + a_1x_1 + a_2x_1^2 + a_3x_1^3 ... + a_nx_1^n = & y_1 \\ a_0 + a_1x_2 + a_2x_2^2 + a_3x_2^3 ... + a_nx_2^n = & y_2 \\ \vdots & \vdots \\ a_0 + a_1x_n + a_2x_n^2 + a_3x_n^3 ... + a_nx_n^n = & y_n \end{array} \tag{4.8}$$

y, en forma matricial:

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ 1 & x_2 & x_2^2 & \dots & x_2^n \\ & & \vdots & & \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{4.9}$$

Al resolver el sistema se encuentran los valores del vector **a** = $[a_0, a_1, a_2, ..., a_n]$.

Ejemplo (- 23)

Encontrar el polinomio de interpolación único para los valores:

$$(10, 0.1763), (20, 0.3640), (30, 0.5774)$$

e interpolar el valor para $x = 21$.

$$P_2(x) = a_0 + a_1x + a_2x^2$$

$$\begin{bmatrix} 1 & 10 & 100 \\ 1 & 20 & 400 \\ 1 & 30 & 900 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.1763 \\ 0.364 \\ 0.5774 \end{bmatrix} \Rightarrow \mathbf{a} = \begin{bmatrix} 0.0143 \\ 0.014915 \\ 0.0001285 \end{bmatrix}$$

$$P_2(x) = 0.0143 + 014915x + 0.0001285x^2$$

y evaluando para $x = 21$:

$$P(21) = 0.3841835$$

3.2. Método de interpolación de Lagrange

La obtención del polinomio de interpolación en forma normal requiere la resolución de un sistema de ecuaciones lineales, cuyo costo aritmético es del orden de n^3, siendo n el número de nodos.
Para reducir este costo se puede tomar una base del espacio de polinomios más adecuada, en la que sea más cómodo definir las condiciones de interpolación.
Esta base, formada por polinomios $L_{in}(x)$, $i = 0, ..., n$, dependientes de las abscisas $x_0, x_1, ..., x_n$, de los nodos considerados, proporcionará el polinomio de interpolación sin hacer un cálculo.
Sea $L_{in}(x)$ un polinomio de grado n, que se anule en todos los puntos x_j, $j = 0, 1, ..., n$, salvo en el *i-ésimo*, donde vale 1; es decir, tal que:

$L_i(x_j) = 0$ si $j \neq i$ y $L_i(x_i) = 1$

La existencia de este polinomio se deriva del resultado anterior, pero puede obtenerse directamente, sin necesidad de resolver un sistema, gracias a la siguiente fórmula debida a Lagrange:

$$L_{in}(x) = \frac{(x - x_0)...(x - x_{i-1})(x - x_{i+1})...(x - x_n)}{(x_i - x_0)...(x_i - x_{i-1})(x_i - x_{i+1})...(x_i - x_n)} \tag{4.10}$$

Es inmediato comprobar entonces que el polinomio:

$$P_n(x) = y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x) + ... + y_n L_n(x) \tag{4.11}$$

cumple las condiciones de la ec.4.6, lo que prueba directamente la existencia del polinomio de interpolación. La unicidad se puede garantizar utilizando el hecho de que un polinomio de grado n puede tener a lo sumo n raíces. Si dos polinomios de grado $\leq n$ interpolan $n+1$ puntos, su diferencia se anula en dichos puntos, por lo que sólo puede ser el polinomio idénticamente nulo.
Así pues existe un único polinomio $P_n(x)$ de menor grado o igual que n que verifica la ec.(4.6) a este polinomio se le denomina polinomio de interpolación de f en los nodos $\{x_0, x_1, ..., x_n\}$ y viene dado por:

$$P_n(x) = \sum_{n=0}^{n} f(x_i) L_i(x) \tag{4.12}$$

donde para cada $i \in (0, 1, ..., n)$

$$L_i(x) = \prod_{j=0 j \neq i}^{n} \frac{x - x_j}{x_i - x_j} \tag{4.13}$$

La ec.(4.12) se conoce como fórmula de Lagrange del polinomio de interpolación.

Algorithm 10: MÉTODO DE INTERPOLACIÓN DE LAGRANGE

Entradas: numero de coordenadas: n, datos (x, y) y el valor que se desea interpolar x_{int}:

Salidas: $f(x_{int})$

INICIO
- $f(x_{int}) \leftarrow 0$;
- $i \leftarrow 0$;
- **Mientras** $i \leq (n-1)$ **hacer**
  - $L \leftarrow 1$;
  - $j \leftarrow 0$;
  - **Mientras** $j \leq (n-1)$ **hacer**
    - **if** $i \neq j$ **then**
      - $L \leftarrow L * \frac{x_{int} - x(j)}{x(i) - x(j)}$;
    - $j \leftarrow j + 1$;
  - $f(x_{int}) \leftarrow f(x_{int}) + L * f(x(i))$;
  - $i \leftarrow i + 1$;
- *Imprimir* $f(x_{int})$

FIN

Ejemplo (- 24)

Obtener el polinomio de interpolación usando la fórmula de interpolación de Lagrange con la siguiente tabla de valores, y obtener el valor de interpolación en el punto $x = -4$.

y	x
-2	-5
-2	-7
-2	4

$L_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = -1/18(x+7)(x-4) = -1/18x^2-1/6x+14/9$
$L_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = 1/22(x+5)(x-4) = 1/22x^2+1/22x-10/11$
$L_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = 1/99(x+5)(x+7) = 1/99x2+4/33x+35/99$

El polinomio solución queda:
$P(x) = \sum_{i=0}^{2} y_i L_i(x) = -2L_0(x) - 2L_1(x) - 2L_2(x) = -2$

Para calcular el valor del polinomio de interpolación en un punto concreto, se sustituye la variable x de la fórmula por ese valor y se realizan las operaciones correspondientes:

$L_0(-4) = 4/3$, $L_1(-4) = -4/11$, $L_2(-4) = 1/33$ y por tanto:

$P(-4) = \sum_{i=0}^{2} y_i L_i(-4) = -2L_0(-4) - 2L_1(-4) - 2L_2(-4) = -2$

3.3. Polinomio de interpolación de Newton

El método de Lagrange tiene la dificultad de que al modificar el soporte hay que comenzar de nuevo, el siguiente método permite reutilizar el polinomio P_n para calcular el P_{n+k} consecuencia de ampliar en k puntos el soporte.
Dada una familia de $n+1$ puntos $\{(x_0, f(x_0)), (x_1, f(x_1)), ..., (x_n, f(x_n))\}$, todos con abscisas distintas, definimos la base de Newton asociada a dicho soporte, como:

$$\{1, (x-x_0), (x-x_0), (x-x_1), (x-x_0), (x-x_1), (x-x_2), ..., \prod_{i=0}^{n-1}(x-x_i)\} \quad (4.14)$$

Aunque no es necesario, se trabajará con el soporte ordenado de menor a mayor.

Dada una familia de $n+1$ puntos $\{(x_0, y_0), (x_1, y_1), ..., (x_n, y_n)\}$, todos con abscisas distintas, definimos el polinomio interpolador de Newton asociado a dicho soporte, como:

$$P_n(x) = A_0 + A_1(x-x_0) + A_2(x-x_0)+(x-x_1) + A_3(x-x_0),(x-x_1),(x-x_2)+...+A_n \prod_{i=0}^{(n-1)} (x-x_i) \quad (4.15)$$

siempre que $P_n(x_i) = y_i$ para $i = 0, 1, 2, ..., n$.
Es importante hacer notar que el polinomio que se acaba de definir, de grado n, es una combinación lineal de la base de Newton, lo que permitirá ampliarlo con facilidad; y además es el mismo que se obtiene por Lagrange para esos mismos datos (recordar que el polinomio interpolador asociado a un soporte es único).
Sólo hay que explicitar cómo calcular los coeficientes A_i.
Dada una familia de $n+1$ puntos $\{(x_0, y_0), (x_1, y_1), ..., (x_n, y_n)\}$, todos con abscisas distintas donde $f(x_i) = y_i$ definimos las DIFERENCIAS DIVIDIDAS de NEWTON asociadas a dichos datos, de la siguiente manera:

Diferencias divididas de orden 0:

$f[x_i] = f(x_i)$, para $i = 0, 1, 2, ..., n$.

Diferencias divididas de orden 1:

$f[x_i, x_{(i+1)}] = f[x_{(i+1)}] - f[x_i]x_{(i+1)} - x_i$, para $i = 0, 1, 2, ..., (n-1)$.

Y en general, diferencias divididas de orden $k < n$:

$$f[x_i, x_{(i+1)}, ..., x_{(i+k)}] = \frac{f[x_{(i+1)}, ..., x_{(i+k)}] - f[x_i, x_{(i+1)}, ..., x_{(i+k-1)}]}{x_{(i+k)} - x_i} \quad (4.16)$$

para $i = 0, 1, 2, ..., (n-k)$ con $1 \leq k$.

Algorithm 11: MÉTODO DE INTERPOLACIÓN DE NEWTON

Entradas: numero de coordenadas: n, datos (x, y) y el valor que se desea interpolar x_{int}:

Salidas: $f(x_{int})$

INICIO

`/* Tabla de diferencias */`

$m \leftarrow n - 1$;

$i \leftarrow 0$;

$f(x_{int}) \leftarrow f(x_0)$;

Mientras $i \leq (m - 1)$ **hacer**

  $T(i, 0) \leftarrow \frac{f(x(i+i)) - f(x(i))}{x(i+1) - x(i)}$;

  $i \leftarrow i + 1$;

$j \leftarrow 1$;

Mientras $j \leq m - 1$ **hacer**

  $i \leftarrow j$;

  Mientras $i \leq m - 1$ **hacer**

    $T(i, j) \leftarrow \frac{T(i,j-1) - T(i-1,j-1)}{x(i+1) - x(i-j)}$;

    $i \leftarrow i + 1$;

  $j \leftarrow j + 1$;

`/* Inicia función principal */`

$f(x_{int}) \leftarrow f(x(0))$;

$i \leftarrow 0$;

Mientras $i \leq n - 1$ **hacer**

  $p \leftarrow 1$;

  $j \leftarrow 0$;

  Mientras $j \leq i$ **hacer**

    $p \leftarrow p * (x_{int} - x(j))$;

    $j \leftarrow j + 1$;

  $f(x_{int}) \leftarrow f(x_{int}) + T(i, i) * p$;

  $i \leftarrow i + 1$;

Imprimir $f(x_{int})$

FIN

Ejemplo (- 25)

Calcule el polinomio de interpolación de Newton para los datos:
$\{(-2,4),(-1,1),(2,4),(3,9)\}$

El polinomio queda:
$p_3(x) = A_0 + A_1(x-x_0) + A_2(x-x_0)(x-x_1) + A_3(x-x_0)(x-x_1)(x-x_2)$

La tabla de diferencias divididas para obtener los coeficientes queda:

x_i	y_i	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_0, x_1, x_2, x_3]$
-2	4			
-1	1	$f[x_0, x_1] = \frac{1-4}{1-(-2)} = -3$		
2	4	$f[x_1, x_2] = \frac{4-1}{2-(-1)} = 1$	$f[x_0, x_1, x_2] = \frac{1-(-3)}{2-(-2)} = 1$	
3	9	$f[x_2, x_3] = \frac{9-4}{3-2} = 5$	$f[x_1, x_2, x_3] = \frac{5-1}{3-(-1)} = 1$	$f[x_0, x_1, x_2, x_3] = \frac{1-1}{3-(-1)} = 0$

$p_3(x) = 4 - 3(x+2) + (x+2)(x+1) + 0(x+2)(x+1)(x-2) = x^2$

5 INTEGRACIÓN

"Resolver integrales mediante técnicas numéricas e implementar programas que desarrollen los métodos discutidos"

Objetivos particulares 1 y 2

1. Cálculo de Áreas

UN **área** es la superficie comprendida entre ciertos límites [Diccionario Larousse, 2003]. Los límites definidos para las figuras geométricas determinan su área, ubicadas en el plano cartesiano, las funciones también definen áreas bajo la curva que generan, el eje x y dos rectas perpendiculares a éste ver figura 5.1.
La operación que realiza el cálculo de esta aproximación es la **integración**, y la integral que representa a la función de la figura 5.1 es:

$$I = \int_b^a f(x)dx \tag{5.1}$$

En forma más precisa, la integral I puede definirse a partir de *aproximaciones rectangulares* (superior e inferior). Para ello, se divide el intervalo $[a; b]$ en n subintervalos iguales, de longitud $h = \frac{b-a}{n}$, mediante los puntos $x_0 = a$, $x_1 = a + h$, $x_2 = a + 2h$, ..., $x_i = a + ih$, ...,$x_n = b$ (véase la figura 8.2). A continuación se construyen los rectángulos *superior* e *inferior* para cada subintervalo $[x_i; x_{i+1}]$.
Suponiendo que la función $f(x)$ es creciente en el intervalo $[a; b]$, como ocurre en la figura 8.2. En ese caso, la altura del rectángulo inferior es $f(x_i)$ (extremo izquierdo) y la altura del rectángulo superior es $f(x_{i+1})$ (extremo derecho).
La aproximación rectangular inferior se define como la suma de las áreas de todos los rectángulos inferiores

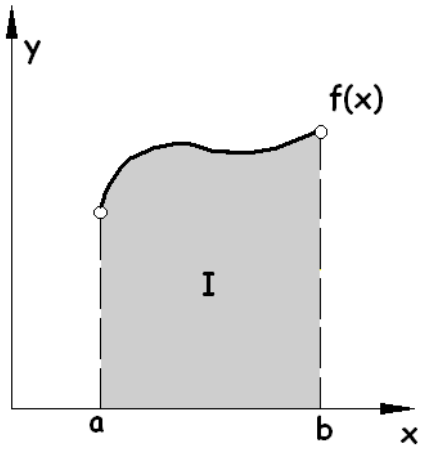

Figura 5.1: Determinación del área bajo una función

$$I_{inf}(h) = hf(x_0) + hf(x_1) + ... + hf(x_{n-1}) = h\sum_{i=0}^{n-1} f(x_i) \qquad (5.2)$$

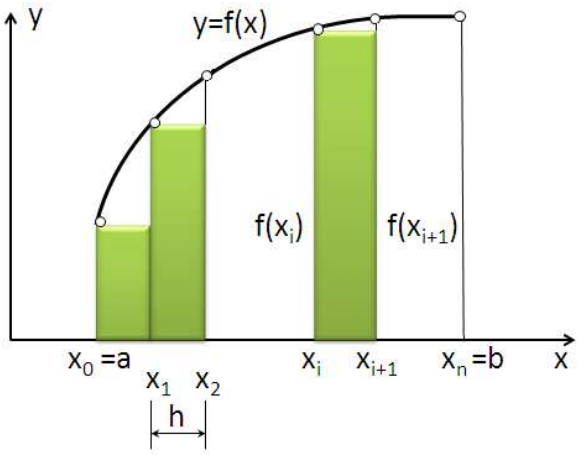

Figura 5.2: Aproximación rectangular inferior

De acuerdo a lo anterior la *aproximación rectangular superior* es la suma de las áreas de todos los rectángulos superiores

$$I_{sup}(h) = hf(x_1) + hf(x_2) + ... + hf(x_n) = h\sum_{i=0}^{n-1} f(x_{i+1}) \qquad (5.3)$$

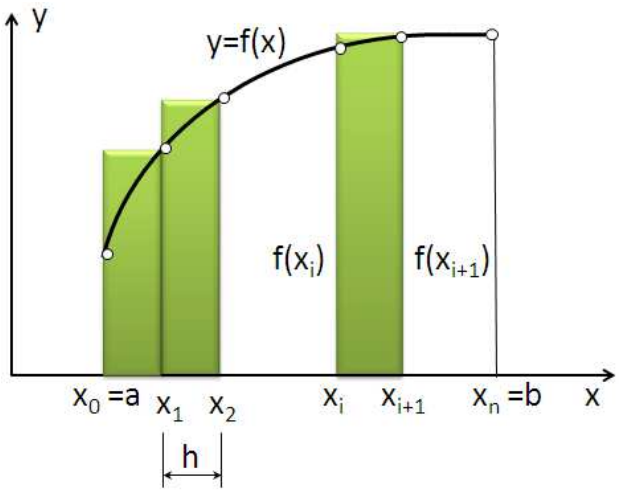

Figura 5.3: Aproximación rectangular superior

2. Reglas trapezoidales

Como ya se mencionó, se pueden utilizar aproximaciones rectangulares para el cálculo de la integral I, pero es mejor utilizar como aproximación del área de I el trapezoide formado por $PQRS$ (figura 5.4).

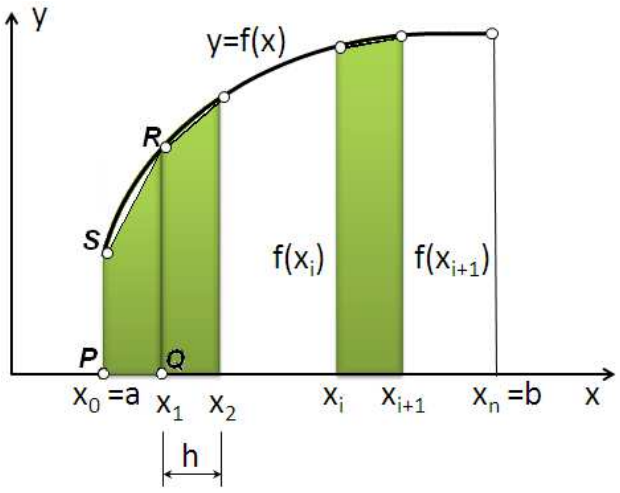

Figura 5.4: Aproximación mediante trapezoides

2.1. Regla trapezoidal simple

Es posible emplear un solo trapezoide (ver figura 5.5), que abarque el intervalo $[a; b]$ que puede calcularse como:

$$A_i = (x_{i+1} - x_i)\frac{f(x_i) + f(x_{i+1})}{2} \tag{5.4}$$

Como puede observarse, el error es significativo con respecto al área real, dependiendo de la forma de la curva el error puede ser por exceso (área mayor a la real) o por defecto (área menor a la real), como en este caso.
Si se divide el intervalo en más sub-áreas el error en el cálculo de la integral disminuye.
La forma mas sencilla de disminuir el error en el calculo del área es dividir el intervalo definido en n sub-intervalos de menor tamaño y aproximar el área como la suma de las áreas de cada uno de los trapezoides que se forman.
Ampliando la regla del trapezoide se subdivide el intervalo $[a; b]$ en n sub-intervalos de la misma longitud ($h = (b - a)/n$).

Regla trapezoidal de segmentos múltiples

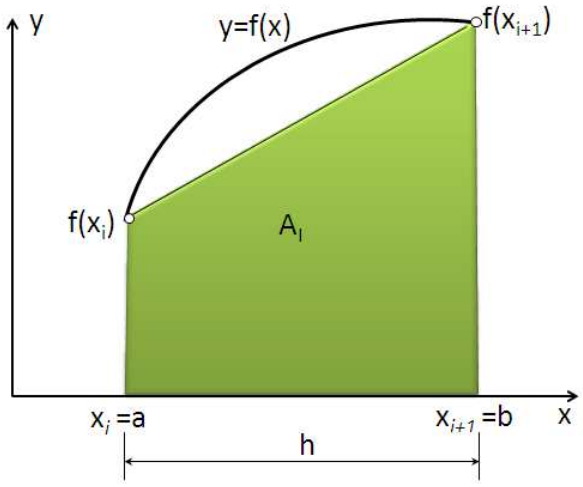

Figura 5.5: Aproximación mediante un trapezoide

El método trapezoidal de segmentos múltiples consiste en aproximar la integral I mediante la suma $I_T(h)$ de las áreas de todos los trapezoides mostrados en la figura 5.6

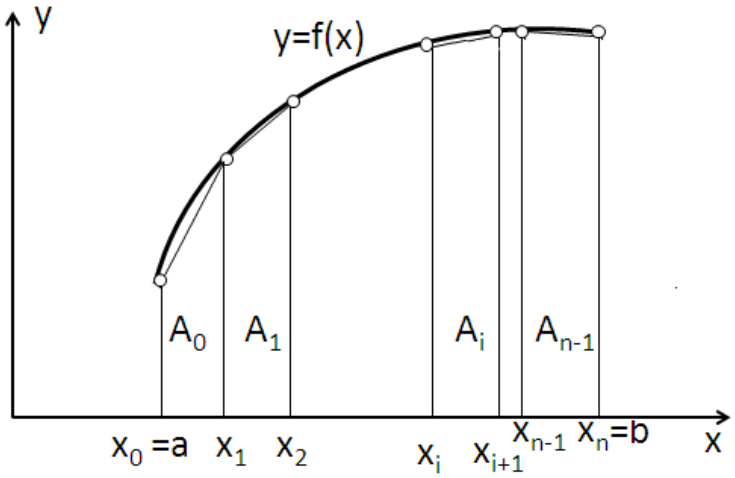

Figura 5.6: Método de trapecios múltiples

$$I_T(h) = A_0 + A_1 + A_2 + ...A_{n-1} = h\frac{f(x_0)}{2} + f(x_1) + f(x_2) + ... + f(x_{n-1}) + \frac{f(x_n)}{2} \quad (5.5)$$

Sustituyendo el valor de h y representando por medio de una sumatoria, finalmente tenemos:

$$I_T(h) = \int_a^b f(x)dx = (b-a)\left[\frac{f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n)}{2n}\right] \tag{5.6}$$

En la figura 5.6 puede verse que este método permite tratar funciones no monótonas de manera directa, sin necesidad de distinguir entre los tramos crecientes y los tramos decrecientes.

Algoritmo 4.2.1:
Regla trapezoidal de segmentos múltiples (pseudocódigo)

Entradas:
f: función a integrar
a límite inferior
b: límite superior
n: número de intervalos
tol: margen de error
$iterMax$: número de iteraciones

Salidas:
i: valor de la integral

Paso 1: Leer $a, b, n, tol, iterMax$
Paso 2: Calcular $h = b - a$.
Paso 3: Calcular $i_0 = \frac{h}{2} * [f(a) + f(b)]$
Paso 4: Calcular $error = tol + 1$
Paso 5: Asignar $contador = 1$
Paso 6: while $error > tol \cap contador < iterMax$ do
$h = \frac{b-a}{n}$
$sum = 0$
for $i = 1$ to $n - 1$ do
$sum = sum + f(a + (i * h))$
$i_1 = \frac{h}{2} * [f(a) + 2 * sum + f(b)]$
$error = |i_1 - i_0|$ ó $|\frac{i_1 - i_0}{i_1}|$
$i_0 = i_1$
$n = n * 2$
$contador = contador + 1$
Paso 7: Si $error \leq tol$ entonces
SALIDA ("La integral de f en el intervalo [a,b] es i con un error igual a $error$")
sino
SALIDA("Demasiadas iteraciones")
PARAR

Ejemplo (- 26)

Usar la regla del trapezoide para aproximar el valor de la integral de la función $f(x) = e^x/x$, cuyo intervalo esta dado entre $[2; 4]$.

$$A = \int_2^4 \frac{e^x}{x} dx$$

Para un solo trapezoide (ecuación 5.4):

$$A = \int_2^4 \frac{e^x}{x} dx \approx (4-2)\left[\frac{f(2)+f(4)}{2}\right] = \frac{e^2}{2} + \frac{e^4}{4} = 17.3441$$

Considerando 5 intervalos (ecuación 5.6), tenemos:

$$\begin{aligned} A &= \int_2^4 \frac{e^x}{x} dx \approx (4-2)\left[\frac{f(2)+2[f(2.5)+f(3)+f(3.5)]+f(4)}{2*5}\right] \\ &= \frac{e^2}{5} + \frac{2e^{2.5}}{5} + \frac{2e^3}{5} + \frac{2e^{3.5}}{5} + \frac{e^4}{5} \\ &= 35.53800383 \end{aligned}$$

2.2. Reglas de Simpson

Otra forma más adecuada de obtener una aproximación más cercana a la real de una integral, es usar polinomios de grado superior para unir los puntos y aproximar la función real.
El método de Simpson no busca incurrir en un mayor número de subdivisiones, sino de ajustar una curva de orden superior en lugar de una línea recta como en la Regla Trapezoidal.

Regla de Simpson 1/3 simple

Dada una función $f(x)$, si entre $f(a)$ y $f(b)$ existe un tercer punto, entonces será posible ajustar por ellos una parábola, en la misma forma, si existen dos puntos entre $f(a)$ y $f(b)$, entonces por esos cuatro puntos se podrá ajustar una curva de grado tres, y así sucesivamente.

En la figura 5.7, se muestra la función que es una parábola que aproxima a la función real. En este caso se calcula el área o la integral bajo la parábola que une los tres puntos. Note que hay tres puntos y dos segmentos, por lo que esta integral se resuelve con regla de Simpson 1/3.

Por lo tanto las fórmulas que resultan de tomar integrales bajo estos polinomios se conocen como regla de Simpson.

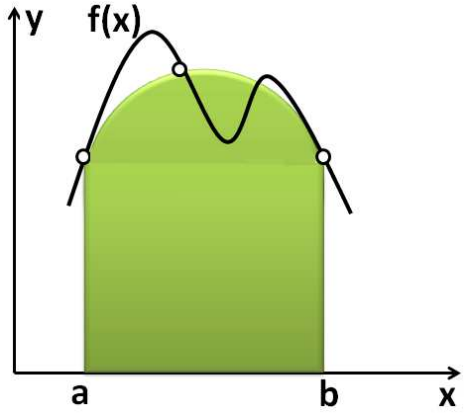

Figura 5.7: Método Simpson 1/3 simple

Esta regla resulta cuando se utiliza una interpolación polinomial de segundo orden:

$$h = \int_a^b f(x)dx = \int_a^b f_2(x)dx \tag{5.7}$$

La función f_2, es la interpolación polinomial de segundo orden. Esto se logra con el polinomio de Lagrange de segundo grado.
Para b hacemos la siguiente sustitución:

$$h = \frac{b-a}{2} \Rightarrow b = 2h + a \tag{5.8}$$

Se tiene que:

$$I(f) \approx \frac{h}{3}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] \tag{5.9}$$

La ecuación anterior se conoce como la regla de Simpson 1/3, y la especificación 1/3 se origina del hecho que h está dividida en tres intervalos.

Regla de Simpson 1/3 de segmentos múltiples

La aplicación múltiple utiliza la misma idea que la regla de Simpson con la diferencia que se divide el intervalo de integración en muchos segmentos o subintervalos, como se observa en la figura 5.8. Es decir en lugar de 2 segmentos se hace para n segmentos donde n es de la forma 2k. Por lo tanto tomamos $h = (b-a)/n$.

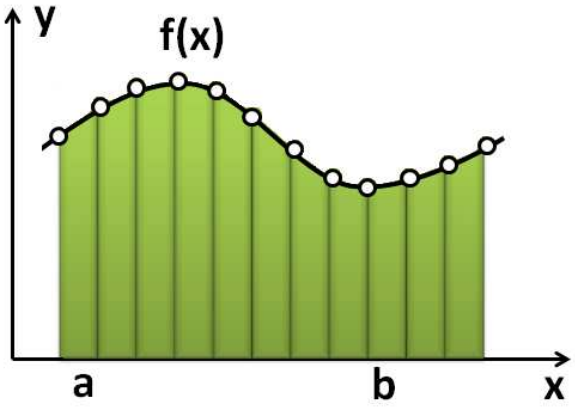

Figura 5.8: Método Simpson 1/3 de segmentos múltiples

Por lo tanto, aplicando la regla de Simpson a cada subintervalo se obtiene.

$$I_T(f) = \int_{x_0}^{x_2} f(x)dx + \int_{x_2}^{x_4} f(x)dx + ... + \int_{x_{n-1}}^{x_n} f(x)dx \tag{5.10}$$

La ecuación anterior es la regla de Simpson 1/3 generalizada a un número par de segmentos e impar de puntos.
Para multiples segmentos se tiene:

$$I_T(f) \approx \frac{f(x_0) + 4\sum_{i=1,3,5...}^{n-1} f(fx_i) + 2\sum_{j=2,4,6...}^{n-2} f[x_j] + f(x_n)}{3n} \tag{5.11}$$

La ecuación anterior es la regla de Simpson 1/3 generalizada a un número par de segmentos e impar de puntos.

Regla de Simpson 3/8 simple

A continuación se describe la regla de integración de Simpson 3/8 para la *integración cerrada*, es decir, para cuando los valores de la función en los extremos de los límites de integración son conocidos.
Además de aplicar la regla trapezoidal con segmentación más fina, otra forma de obtener una estimación más exacta de una integral es con el uso de polinomios de orden superior para conectar los puntos (en lugar de utilizar líneas para conectarlos).
Las reglas de Simpson son las fórmulas que resultan al tomar las integrales bajo los polinomios que conectan a los puntos.
La derivación de la Regla de los 3/8 de Simpson es similar a la regla de 1/3, excepto que se determina el área bajo una parábola de tercer grado que conecta 4 puntos sobre una curva dada. La forma general de la parábola de tercer grado es:

$$Y = aX^3 + bX^2 + cX + d \tag{5.12}$$

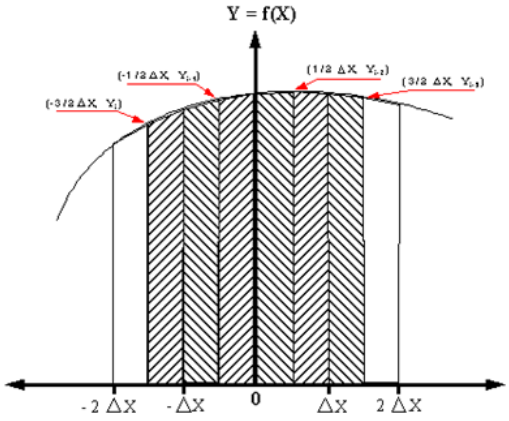

Figura 5.9: Descripción de la gráfica de la regla de Simpson 3/8

En la derivación, las constantes se determinan requiriendo que la parábola pase a través de los cuatro puntos indicados sobre la curva mostrada en la figura 5.9. El intervalo de integración es de $-3\Delta X/2$ a $3\Delta X/2$, lo que produce:

$$A_T = \int_{\frac{-3\Delta}{2}}^{\frac{3\Delta}{2}} (aX^3 + bX^2 + cX + d)dx = \frac{3\Delta X}{8}(Y_i + 3Y_{i+1} + 3Y_{i+2} + Y_{i+3}) \tag{5.13}$$

que es la regla de los 3/8 de Simpson.

Algoritmo 4.2.2:
Regla de Simpson 1/3 (pseudocódigo)

Entradas:
f: función a integrar
a límite inferior
b: límite superior
n: número de intervalos
tol: margen de error
$iterMax$: número de iteraciones

Salidas:
i: valor de la integral

Paso 1: Leer $a, b, n, tol, iterMax$
Paso 2: if $n\%2 \neq 0$ entonces
SALIDA("Número de particiones impar")
sino $h = \frac{b-a}{2}$
$m = \frac{a+b}{2}$
$error = tol + 1$
$contador = 1$
$i_0 = \frac{h}{3} * [f(a) + 4 * f(m) + f(b)]$
while $error > tol \cap contador < iterMax$ do
$h = \frac{b-a}{n}$
$sump = 0$
$sumi = 0$
for $i = 1$ to $n - 1$ do
si $i\%2 = 0$ entonces
$sump = sump + f(a + (i * H))$
sino $sumi = sumi + f(a + (i * h))$
$i_1 = \frac{h}{3}[f(a) + 4 * sumi + 2 * sump + f(b)$
$i_0 = i_1$
$n = n * 2$
$contador = contador + 1$
si $error \leq tol$ entonces
SALIDA ("La integral de f en el intervalo [a,b] es i con un error igual a $error$")
sino
SALIDA("Demasiadas iteraciones")
PARAR

Ejemplo (- 27)

Use la regla de Simpson 1/3 y 3/8 para integrar la siguiente función:

$$f(x) = 0.2 + 25x - 200x^2 + 675x^3 - 900x^4 + 400x^5$$

Desde $a = 0$ hasta $b = 0.8$. La integral exacta es 1.640533.

Por **Simpson 1/3**

$x_0 = 0$
$x_2 = 0.8$
$x_1 = (0 + 0.8)/2 = 0.4$

$f(x_0) = f(0) = 0.2$
$f(x_1) = f(0.4) = 2.456$
$f(x_2) = f(0.8) = 0.232$

Sustituimos los valores en la ecuación:
$I \approx (b - a)\frac{f(x_0)+4f(x_1)+f(x_2)}{6}$
$I \approx (0.8)\frac{0.2+4(2.456)+0.232}{6}$
$I \approx 1.367467$

Por **Simpson 3/8**

Cada separación va a tener:
$x = \frac{(0+0.8)}{3} = 0.2667$
$x_0 = 0$
$x_1 = (0 + 0.2667) = 0.2667$
$x_2 = (0.2667 + 0.2667) = 0.5333$
$x_3 = 0.8$

$f(x_0) = f(0) = 0.2$
$f(x_1) = f(0.2667) = 1.432724$
$f(x_2) = f(0.5333) = 3.487177$
$f(x_3) = f(0.8) = 0.232$

Sustituimos los valores en la ecuación:
$I \approx (b - a)\frac{f(x_0)+3f(x_1)+3f(x_2)+f(x_3)}{8}$
$I \approx (0.8)\frac{0.2+3(1.432724)+3(3.487177)+0.232}{8}$
$I \approx 1.519170$

3. Errores en fórmulas de cuadratura

CUando se emplea un solo segmento de línea recta (figura 5.4) para estimar la integral bajo una curva $f(x)$, se incurre en un error que puede ser sustancial.

3.1. Error de truncamiento en la Regla Trapezoidal

Una estimación del *error por truncamiento*, de una sola aplicación de la regla trapezoidal está dada por,

$$E_T(h) = -\frac{1}{12} f''(\xi)(b-a) \tag{5.14}$$

donde ξ es un punto cualquiera dentro del intervalo de integración; para efectos prácticos podría ser el punto medio del segmento $[a;b]$. La ecuación anterior indica que si la función que se está integrando es lineal, el método trapecial proporcionará valores exactos, ya que, la segunda derivada de una recta es cero; de otra manera ocurrirá un error, para funciones curvas.
Se puede demostrar que el error que se comete al aproximar la integral exacta I por I_T (ecuación 5.6) es $O(h^2)$. Esto indica que el método compuesto del trapecio es cuadrático: a medida que h tiende a cero, el error tiende cuadráticamente a cero. Para ver cuál es la implicación practica de este resultado teórico, se utiliza el método compuesto del trapecio para calcular numéricamente la integral 5.4.

3.2. Error de redondeo en la Regla Trapezoidal

Otro error que se introduce al obtener el área aproximada de cada subarea es el *error por redondeo*. Este se produce cuando las operaciones aritméticas requeridas se efectúan con valores numéricos que tienen un número limitado de dígitos significativos. Se puede demostrar que una aproximación al límite del error por redondeo es:

$$e_R \leq= -\frac{ye(b-a)^2}{2}\left(\frac{1}{\Delta X}\right) \tag{5.15}$$

Se tiene entonces que el límite en el error por redondeo aumenta proporcionalmente a $(1/\Delta X)$, lo cual pronto domina al error por truncamiento que es proporcional a ΔX^2. En realidad, el error por redondeo en sí no crece proporcionalmente con ΔX^{-1} sino con ΔX^{-p} en donde $0 < p < 1$, pero sin embargo aún supera al error por truncamiento si ΔX decrece lo suficiente.

3.3. Error de truncamiento en la Regla de Simpson

En la regla de Simpson el error que resulta de aproximar el área verdadera de subintervalos bajo la curva $f(X)$ comprendida entre $X_i - 1$ y $X_i + 1$ mediante el área bajo una parábola de segundo grado, se demuestra que es:

$$e_T = -\frac{-1}{90}(f^{VI}(\xi)(X^5) \tag{5.16}$$

Este error por truncamiento es la cantidad que se debe agregar al área aproximada de dos subintervalos, que se obtiene mediante la regla de un tercio de Simpson, para obtener el área real bajo la curva en ese intervalo. El término mostrado del error por truncamiento generalmente no se puede valuar en forma directa. Sin embargo, se puede obtener una buena estimación de su valor para cada intervalo suponiendo que es suficientemente constante en el intervalo (se supone que las derivadas de orden superior son despreciables) y valuando para . La estimación del error por truncamiento para toda la integración se obtiene sumando las estimaciones correspondientes a cada dos fajas. Si la estimación del error total por truncamiento es mayor de lo que se puede tolerar, se deben utilizar intervalos de dos fajas menores. Considerando el error por redondeo que también aparece, existe un ancho óptimo de la faja para obtener un error total mínimo en la integración.

Bibliografía

[Chapra & Canale, 2008] Steven C. Chapra, Raymond P. Canale *"Métodos numéricos para ingenieros, 5ª Edición*"; Mc Graw Hill; (2008).

[Diccionario Larousse, 2003] Diccionario Larousse; Edición Premium; EDICIONES LAROUSSE MÉXICO y SPS EDITORIAL BARCELONA; (2003).

[Golub & Ortega, 1992] Gene H. Golub and James M. Ortega. *"Scientific Computing and Differential Equations - An Introduction to Numerical Methods*". Academic Press, 1992.

[Gustafsson, 2018] Gustafsson B., *"Scientific Computing A Historical Perspective*", Texts in Computational Science and Enginnering, Springer, (2018).

[Rodríguez, 2006] Rodríguez A., *"Arquímedes. El genio de Siracusa*"; Facultad de Ciencias; Universidad Autónoma de Madrid; (2006).

[Kubicek, et al. 2008] Milan Kubicek, Drahoslava Janovska, Miroslava Dubcov *"NUMERICAL METHODS AND ALGORITHMS*"; http://old.vscht.cz/mat/NM-Ang/NM-Ang.pdf (2008).

[US Department of Labor] United States Department of Labor; http://www.dol.gov/