Universidad Nacional de Rosario

Facultad de Ciencias Exactas, Ingeniería y Agrimensura

Resumen

Métodos Numéricos

Autor:
Arroyo, Joaquín

9 de diciembre de 2024

Índice

2. Errores Numéricos

2.1. Sistemas de Numeración Posicionales

Un **sistema de numeración posicional** es un método para representar números en el cual el valor de un dígito depende tanto de su valor intrínseco como de su posición dentro del número. En estos sistemas, un número se expresa como una suma ponderada de potencias de una base b.

En un sistema de base b, un número n con k dígitos se puede expresar como:

$$n = a_{k-1} \cdot b^{k-1} + a_{k-2} \cdot b^{k-2} + \cdots + a_1 \cdot b^1 + a_0 \cdot b^0$$

donde a_i representa el dígito en la posición i y b es la base del sistema. Algunos ejemplos comunes incluyen:

- **Sistema Decimal** (base 10): Es el sistema más utilizado en la vida cotidiana. Los dígitos pueden tomar valores entre 0 y 9.
- **Sistema Binario** (base 2): Utilizado en computación. Los dígitos pueden ser 0 o 1.
- **Sistema Hexadecimal** (base 16): También usado en computación, especialmente en programación y diseño de hardware. Los dígitos incluyen 0-9 y A-F.

La principal característica de los sistemas de numeración posicionales es que permiten representar cualquier número entero o fraccionario de manera compacta y sistemática.

2.2. Representación Computacional de Números en Punto Flotante

La **representación en punto flotante** es un método utilizado en computación para representar números reales que pueden tener una amplia gama de valores, desde números extremadamente pequeños hasta números muy grandes. Este sistema es especialmente útil para manejar cálculos científicos y gráficos por computadora.

Un número en punto flotante se representa en la forma:

$$x = m \times b^e$$

donde:

- m es la **mantisa** que contiene los dígitos significativos del número.
- b es la **base** del sistema de numeración (generalmente $b = 2$ en sistemas binarios).
- e es el **exponente** que determina la magnitud del número.

Por ejemplo, en un sistema binario, el número 13,25 se puede representar como:

$$13{,}25 = 1{,}325 \times 10^1 = 1101{,}01_2 = 1{,}10101_2 \times 2^3$$

2.3. Norma IEEE para Números en Punto Flotante

La **Norma IEEE 754** es un estándar ampliamente adoptado para la representación y el manejo de números en punto flotante en computadoras y sistemas digitales. Esta norma define formatos para la representación de números reales, así como reglas para operaciones aritméticas y manejo de excepciones.

Un número en punto flotante en este estándar se representa de la forma:

$$(-1)^s \times (1.m) \times 2^{(e-\text{bias})}$$

donde:

- s es el **bit de signo** (0 para positivo, 1 para negativo).
- m es la **mantisa**, que incluye los dígitos significativos.
- e es el **exponente**, ajustado por un **sesgo** (*bias*) para permitir exponentes negativos.

El sesgo depende del número de bits asignados al exponente.

La norma también define cómo deben realizarse las operaciones aritméticas y cómo deben manejarse las excepciones, como los desbordamientos, subdesbordamientos, y operaciones con números no definidos (*NaN*).

- **Desbordamiento** ocurre cuando el resultado de una operación excede el rango máximo representable.
- **Subdesbordamiento** ocurre cuando el resultado es menor que el valor mínimo representable.
- **NaN** (*Not a Number*) se utiliza para representar resultados indefinidos o inválidos, como $\sqrt{-1}$ o $0/0$.

El estándar también incluye modos de redondeo para tratar con los errores de precisión que surgen al realizar operaciones con números en punto flotante.

2.4. Truncamiento y Redondeo

En la representación y manipulación de números en punto flotante, es común que los números deban ser ajustados debido a las limitaciones en la precisión. Estos ajustes se realizan mediante **truncamiento** o **redondeo**.

El **truncamiento** es el proceso de cortar un número después de cierto número de dígitos significativos, eliminando los dígitos restantes sin considerar su valor. Por ejemplo, truncar 3,14159 a tres dígitos significativos da como resultado 3,14. Este método puede introducir un error sistemático hacia cero, lo que puede afectar la precisión de los cálculos.

El **redondeo** es un método más sofisticado que ajusta el valor de un número a la cantidad más cercana de dígitos significativos. Existen diferentes técnicas de redondeo, entre ellas:

- **Redondeo al número más cercano**: El dígito se ajusta al número más cercano posible.
- **Redondeo hacia arriba**: Se ajusta el número al siguiente valor mayor.
- **Redondeo hacia abajo**: Se ajusta el número al siguiente valor menor.
- **Redondeo hacia cero**: Se ajusta el número hacia cero, similar al truncamiento.
- **Redondeo de paridad** (o redondeo al número más cercano con ajuste de paridad): Cuando un número se encuentra exactamente en la mitad (por ejemplo, 2,5), se redondea al número par más cercano.

El **épsilon de la máquina**, denotado como $\epsilon_{\text{máquina}}$, es el número positivo más pequeño tal que $1 + \epsilon_{\text{máquina}}$ sea distinguible de 1 en la representación de punto flotante. Este valor define el límite de precisión relativa de un sistema de numeración en punto flotante.

Matemáticamente, $\epsilon_{\text{máquina}}$ se define como:

$$\epsilon_{\text{máquina}} = \text{mín}\{\epsilon > 0 : 1 + \epsilon > 1 \text{ en punto flotante}\}$$

El valor de $\epsilon_{\text{máquina}}$ depende del formato de precisión usado (simple, doble, etc.), y proporciona una medida de la precisión que se puede esperar en cálculos aritméticos.

La **unidad de redondeo** es la mitad de $\epsilon_{\text{máquina}}$, y representa la máxima cantidad de error que puede ser introducido debido al redondeo en una operación aritmética de punto flotante.

Se expresa como:

$$\text{Unidad de redondeo} = \frac{\epsilon_{\text{máquina}}}{2}$$

Este concepto es crucial para entender los límites de precisión en las operaciones de punto flotante y evaluar la estabilidad numérica de algoritmos.

Otra medida de precisión en los sistemas de punto flotante se relaciona con el número de bits en el significante (mantisa). Se define **el mayor entero positivo** M tal que todos los enteros x que satisfacen $0 \leq x \leq M$ se pueden representar de forma exacta en el formato de punto flotante. Formalmente, esto implica encontrar $M \in \mathbb{Z}^+$ tal que:

- Para todo $0 < x \leq M$ con $x \in \mathbb{Z}^+$, la representación en punto flotante de x es exacta, es decir, $\text{fl}(x) = x$.
- Para $M + 1$, la representación en punto flotante no es exacta, es decir, $\text{fl}(M + 1) \neq M + 1$.

Esto significa que M es el mayor entero tal que todos los enteros menores o iguales a M se representan sin error en el formato de punto flotante, mientras que $M + 1$ ya no puede ser representado con precisión exacta debido a las limitaciones de la mantisa.

2.5. Error Absoluto y Relativo

Al resolver un problema, buscamos obtener la solución exacta o verdadera, denotada por x_v. Sin embargo, al aplicar métodos numéricos, se obtiene generalmente una solución aproximada, x_a. El **error** en x_a se define como:

$$\text{Error} = x_v - x_a$$

El **error absoluto** y el **error relativo** en x_a se definen de la siguiente manera:

- **Error absoluto**: La magnitud del error se calcula como la diferencia absoluta entre la solución verdadera y la solución aproximada:

$$\text{Error absoluto} = |\text{Error}| = |x_v - x_a|$$

- **Error relativo**: El error absoluto se normaliza dividiéndolo por el valor verdadero, proporcionando una medida del error en relación con el tamaño de la solución verdadera:

$$\text{Error relativo} = \frac{\text{Error absoluto}}{|x_v|} = \frac{|x_v - x_a|}{|x_v|}$$

Estos conceptos son fundamentales para evaluar la precisión de métodos numéricos y la calidad de las soluciones aproximadas en comparación con la solución exacta.

2.6. Error de Truncamiento y Redondeo

Si $x \neq \text{fl}(x)$ y se utiliza truncamiento, entonces $\text{fl}(x) < x$ y el error $x - \text{fl}(x)$ es siempre positivo. Esto tiene implicaciones en el cálculo numérico, ya que no hay posibilidad de cancelación de errores y la propagación de errores es mayor. El truncamiento introduce un sesgo sistemático, lo que puede llevar a errores acumulativos significativos en cálculos sucesivos.

Con el redondeo, el error $x - \text{fl}(x)$ puede ser negativo para la mitad de los valores de x y positivo para la otra mitad. Además, el peor error posible por redondeo es la mitad que en el caso de truncamiento. A menudo, el error relativo se representa como:

$$\frac{x - \text{fl}(x)}{x} = -\epsilon, \text{ si } x \neq 0$$

donde $\text{fl}(x) = (1 + \epsilon)x$. Aquí, $\text{fl}(x)$ puede verse como un valor perturbado de x.

La siguiente proposición proporciona cotas sobre el error relativo ϵ para truncamiento y redondeo.

Proposición 1. Sea $x \in \mathbb{R}$ con $x \neq 0$. Las siguientes cotas sobre el error relativo ϵ son válidas, empleando las fórmulas de truncamiento y redondeo

- Para $\text{fl}(x)$ truncado:
$$-\beta^{-n+1} \leq \epsilon \leq 0$$
- Para $\text{fl}(x)$ redondeado:
$$-\frac{1}{2}\beta^{-n+1} \leq \epsilon \leq \frac{1}{2}\beta^{-n+1}$$

Donde β es la base del sistema y n es el número de dígitos en la mantisa.

2.7. Cifras Significativas

Las **cifras significativas** de un número son los dígitos que aportan información real sobre su valor, determinadas por la incertidumbre asociada. Al expresar un número, es crucial evitar cifras no significativas que no contribuyan a la precisión del valor.

Para identificar el número de cifras significativas en un número decimal, se siguen estas reglas:

- Cualquier dígito distinto de cero es significativo. Por ejemplo, 438 tiene tres cifras significativas.
- Los ceros entre dígitos distintos de cero son significativos. Por ejemplo, 402 tiene tres cifras significativas.
- Los ceros a la izquierda del primer dígito distinto de cero no son significativos. Por ejemplo, 0.0023 tiene dos cifras significativas.
- Los ceros a la derecha de un dígito distinto de cero y después de la coma son significativos. Por ejemplo, 10.00 tiene cuatro cifras significativas.
- En números enteros, los ceros a la derecha de un dígito distinto de cero pueden ser significativos, pero la cantidad exacta depende del contexto o de la notación científica utilizada.

Para un valor aproximado x_a en comparación con un valor verdadero x_v, se dice que x_a tiene m cifras significativas si el error $|x_v - x_a|$ es menor o igual a cinco unidades en el dígito (m + 1) de x_v, contando desde el primer dígito distinto de cero en x_v.

Redondear un número decimal x a m cifras significativas (o a m dígitos) es equivalente a redondear el número utilizando una mantisa de m dígitos en notación de punto flotante. Para ello, primero se escribe el número en la forma $x = \hat{x} \times 10^E$, con $0{,}1 \leq \hat{x} < 1$, y E un número entero. Luego se procede a redondear $\hat{x}$ a m dígitos después de la coma. El número redondeado es:

$$r_n(x) = \bar{x} \times 10^E$$

donde $\bar{x} = 0.a_1a_2 \cdots a_m$, con $a_1 \neq 0$ y todos los dígitos están después de la coma. Así, $r_n(x)$ tiene m cifras significativas. Además, el valor aproximado obtenido $x_a = r_n(x)$ tiene m cifras significativas con respecto al valor original $x_v = x$, cumpliendo la definición de cifras significativas anteriormente descrita.

2.8. Propagación de Errores

Al realizar cálculos con números sujetos a error, es importante considerar cómo estos errores se propagan a través de las operaciones aritméticas.

Sea ω una operación aritmética (+, -, $\times$, /) y $\hat{\omega}$ su versión computacional, que incluye redondeo o truncamiento. Supongamos que x_a y y_a son números con errores, siendo sus valores verdaderos $x_v = x_a + \epsilon$ y $y_v = y_a + \eta$, respectivamente. El error en el resultado de la operación $x_a \hat{\omega} y_a$ se define como:

$$x_v \omega y_v - x_a \hat{\omega} y_a = [x_v \omega y_v - x_a \omega y_a] + [x_a \omega y_a - x_a \hat{\omega} y_a]$$

Aquí, la primera cantidad entre corchetes es el **error propagado**, mientras que la segunda cantidad es el **error de redondeo o truncamiento**. Suponiendo que se emplea redondeo, tenemos que:

$$x_a \hat{\omega} y_a = \text{fl}(x_a \omega y_a)$$

Lo que significa que $x_a \omega y_a$ se calcula con exactitud y luego se redondea. Aplicando la cota para el error de redondeo:

$$|x_a \omega y_a - x_a \hat{\omega} y_a| \leq \frac{\beta^{-n+1}}{2} |x_a \omega y_a|$$

Donde β es la base del sistema y n es el número de dígitos en la mantisa.

3. Resolución de Sistemas de Ecuaciones No Lineales

Un **algoritmo** para resolver un problema matemático es un proceso iterativo que genera una sucesión de números o puntos siguiendo instrucciones precisas y un criterio de parada. Dado un vector $x_k \in \mathbb{R}^n$, el algoritmo produce un nuevo punto $x_{k+1} \in \mathbb{R}^n$. Este proceso se describe mediante un mapa algorítmico A, donde el punto inicial $x_0 \in \mathbb{R}^n$ genera la sucesión $x_1, x_2, x_3, \ldots$, con $x_{k+1} \in A(x_k)$ para $k = 0, 1, 2, \ldots$.

La transformación de x_k a x_{k+1} constituye una iteración del algoritmo. En general, A puede ser un mapa punto a conjunto, asignando a cada punto en el dominio X un subconjunto de X. Si A es un mapa punto a punto, se escribe $x_{k+1} = A(x_k)$.

3.1. Criterios de Parada

Para determinar cuándo detener un algoritmo, se pueden usar los siguientes criterios de parada, con una tolerancia $\epsilon > 0$:

- **Distancia luego de una iteración:** $\|x_k - x_{k-1}\| < \epsilon$
- **Distancia luego de N iteraciones:** $\|x_k - x_{k-N}\| < \epsilon$
- **Distancia relativa en una iteración:** $\frac{\|x_k - x_{k-1}\|}{\|x_{k-1}\|} < \epsilon$
- **Diferencia en el valor de una función luego de N iteraciones:** $|f(x_k) - f(x_{k-N})| < \epsilon$
- **Proximidad a cero de una función:** $|f(x_k)| < \epsilon$

Es importante tener en cuenta que estos criterios pueden presentar problemas, por lo que además de los criterios de parada, es aconsejable fijar un número máximo de iteraciones para evitar que el algoritmo no converja.

3.2. Orden de Convergencia

Definición. El **orden de convergencia** de una sucesión $\{x_k\}$ que converge a $\bar{x}$ es el mayor número $p \geq 1$ tal que

$$\lim_{k\to\infty} \frac{\|x_{k+1} - \bar{x}\|}{\|x_k - \bar{x}\|^p} = \beta < \infty$$

donde β es una constante. La **velocidad de convergencia** es mayor cuanto mayor sea p y menor sea β.

Se distinguen los siguientes casos particulares:

- **Convergencia lineal:** Si $p = 1$ y $\beta \in (0, 1)$, la tasa de convergencia es β.
- **Convergencia superlineal:** Si $p > 1$ o si $p = 1$ y $\beta = 0$.
- **Convergencia cuadrática (o de segundo orden):** Si $p = 2$ y $0 < \beta < \infty$.

3.3. Solución de Ecuaciones No Lineales de Una Variable

Encontrar una raíz de la ecuación $f(x) = 0$ es un problema común en matemáticas aplicadas. Dado que muchas ecuaciones no lineales no tienen solución analítica, se utilizan métodos numéricos iterativos para resolverlas.

Definición. Para una función no lineal $f : \mathbb{R} \to \mathbb{R}$, se llama raíz o cero a cualquier número $\alpha \in \mathbb{R}$ tal que $f(\alpha) = 0$.

Teorema 1 (Teorema de Bolzano). Si f es continua en el intervalo $[a, b] \subset \mathbb{R}$ y $f(a) \cdot f(b) < 0$, entonces existe al menos un $c \in (a, b)$ tal que $f(c) = 0$. Este teorema garantiza la existencia de al menos una raíz en el intervalo dado, aunque no especifica el número de raíces.

3.3.1. Método de la Bisección

Basado en el Teorema de Bolzano, este método busca una raíz en un intervalo $[a, b]$ donde $f(a)f(b) < 0$. Dado un error de tolerancia $\varepsilon > 0$, el procedimiento es:

- Definir $c = \frac{a+b}{2}$.
- Si $b - c \leq \varepsilon$, aceptar c como la raíz y detenerse.
- Si $b - c > \varepsilon$, comparar el signo de $f(c)$ con el de $f(a)$ y $f(b)$. Si $f(b)f(c) \leq 0$, reemplazar a con c. De lo contrario, reemplazar b con c. Regresar al paso 1.

En general, partiendo de $a_1 = a$ y $b_1 = b$, en cada iteración se evalúan las siguientes condiciones para actualizar los intervalos:

- Si $f(a_k) \cdot f(c_k) < 0$, entonces $a_{k+1} = a_k$, $b_{k+1} = c_k$, y $c_{k+1} = \frac{b_{k+1}+a_{k+1}}{2}$.
- Si $f(b_k) \cdot f(c_k) < 0$, entonces $a_{k+1} = c_k$, $b_{k+1} = b_k$, y $c_{k+1} = \frac{b_{k+1}+a_{k+1}}{2}$.
- Si $f(c_k) = 0$, entonces $\alpha = c_k$.

Acotación del Error

El ancho del intervalo se reduce en cada iteración:

$$b_{k+1} - a_{k+1} = \frac{1}{2}(b_k - a_k).$$

Por inducción, se obtiene:

$$b_k - a_k = \left(\frac{1}{2}\right)^{k-1} (b_1 - a_1).$$

Además, se tiene:

$$|\alpha - c_k| \leq b_k - c_k = c_k - a_k = \frac{1}{2}(b_k - a_k).$$

Entonces:

$$|\alpha - c_k| \leq \left(\frac{1}{2}\right)^{k} (b_1 - a_1).$$

Esta fórmula permite acotar el error de aproximación de la raíz. Dado que $\left(\frac{1}{2}\right)^k$ tiende a cero cuando k tiende a infinito, la fórmula muestra que c_k converge a la raíz α cuando $k \to \infty$. Para obtener una precisión de error no superior a $\varepsilon > 0$:

$$|\alpha - c_k| \leq \varepsilon$$

esto se cumple si:

$$\frac{1}{2^k}(b - a) \leq \varepsilon.$$

Es decir, para:

$$k \geq \frac{\ln\left(\frac{b-a}{\varepsilon}\right)}{\ln 2}.$$

Ventajas:

- **Convergencia Asegurada:** El método de la bisección siempre converge a una raíz si la función es continua en el intervalo $[a, b]$ y $f(a) \cdot f(b) < 0$.

- **Acotación del Error Garantizada:** El error se reduce en cada iteración, proporcionando una cota precisa del error en cada paso.
- **Velocidad de Convergencia Garantizada:** La cota del error se reduce a la mitad en cada iteración, asegurando una reducción constante del intervalo de búsqueda.

Desventaja:

- **Convergencia Relativamente Lenta:** Aunque el método es robusto y garantiza la convergencia, su velocidad de convergencia es más lenta en comparación con otros métodos iterativos, como el método de Newton-Raphson.

3.3.2. Método de Newton (o Método de Newton-Raphson)

Sea α una raíz de la ecuación $f(x) = 0$. Supongamos que $f \in C^2$ en $[a, b]$, y que $x_0 \in [a, b]$ es una estimación de α tal que $f'(x_0) \neq 0$ y x_0 está cerca de α. Consideramos el polinomio de Taylor para $f(x)$ expandido alrededor de x_0:

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x - x_0)^2}{2}f''(c_x),$$

donde c_x está entre x y x_0.

Para obtener una nueva estimación de α, denotada x_1, resolvemos:

$$p_1(x_1) = 0 = f(x_0) + (x_1 - x_0)f'(x_0),$$

lo que lleva a:

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

Gráficamente, x_1 corresponde a la intersección con el eje x de la recta tangente a la función $f(x)$ en el punto $(x_0, f(x_0))$.

La fórmula general de este método es:

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

Análisis del Error

Usando el desarrollo de Taylor con resto, tenemos:

$$0 = f(\alpha) = f(x_n) + (\alpha - x_n)f'(x_n) + \frac{1}{2}(\alpha - x_n)^2 f''(c_n),$$

donde c_n está entre α y x_n. Dividiendo por $f'(x_n)$ obtenemos:

$$0 = \frac{f(x_n)}{f'(x_n)} + \alpha - x_n + \frac{(\alpha - x_n)^2 f''(c_n)}{2f'(x_n)},$$

lo que implica:

$$\alpha - x_{n+1} = \frac{-f''(c_n)}{2f'(x_n)}(\alpha - x_n)^2.$$

Si el método de Newton converge, se tiene:

$$\lim_{n\to\infty} \frac{|\alpha - x_{n+1}|}{(\alpha - x_n)^2} = \left|\frac{-f''(\alpha)}{2f'(\alpha)}\right|,$$

lo que muestra que la convergencia del método de Newton es cuadrática, siempre que $f'(\alpha) \neq 0$.

Ventajas del método de Newton:

- Converge rápidamente en la mayoría de los casos.
- Formulación sencilla.
- Comportamiento fácil de entender.

Desventajas del método de Newton:

- Puede no converger si el valor inicial x_0 no está suficientemente cerca de la raíz α.
- Puede ocurrir que $f'(\alpha) = 0$.
- Se requiere conocer tanto $f(x)$ como $f'(x)$.

3.3.3. Método de la Secante

El método de la secante aproxima la función $f(x)$ mediante la línea secante que pasa por los puntos $(x_0, f(x_0))$ y $(x_1, f(x_1))$. La nueva aproximación x_2 se obtiene como la raíz de esta línea secante:

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

La fórmula general de la iteración es:

$$x_{n+1} = x_n - \frac{f(x_n)\cdot(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})}, \quad n \geq 1.$$

El método de la secante es una aproximación del método de Newton, utilizando una diferencia finita para aproximar la derivada:

$$f'(x) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}.$$

3.3.4. Análisis del Error

El error se puede expresar como:

$$\alpha - x_{n+1} = (\alpha - x_n)(\alpha - x_{n-1})\left[\frac{-f''(\xi_n)}{2f'(\rho_n)}\right],$$

donde ξ_n está entre x_{n-1}, x_n y α, y ρ_n entre x_{n-1} y x_n. Si x_n converge a α, el error tiene una tasa de convergencia superlineal:

$$\lim_{n\to\infty} \frac{|\alpha - x_{n+1}|}{|\alpha - x_n|^p} = \left|\frac{f''(\alpha)}{2f'(\alpha)}\right|^{p-1} = \beta,$$

con $p \approx 1{,}62$, lo que indica que la convergencia es superlineal, pero el método de Newton converge más rápido.

3.3.5. Ventajas y Desventajas

Ventajas:

- Converge más rápido que la convergencia lineal.
- No requiere conocer $f'(x)$.
- Solo necesita una evaluación de $f(x)$ por iteración, comparado con dos evaluaciones en el método de Newton.

Desventajas:

- Puede no converger.
- Puede tener problemas si $f'(\alpha) = 0$.
- Menos generalizable a sistemas de ecuaciones no lineales en comparación con el método de Newton.

3.3.6. Método de la Falsa Posición (o Regula Falsi)

El método de la falsa posición combina los métodos de la secante y de la bisección. Se eligen dos aproximaciones iniciales a y b tales que $f(a)f(b) < 0$.

- Se inicia con $a_1 = a$ y $b_1 = b$.
- Se obtiene una nueva aproximación c_1 aplicando el método de la secante a los puntos $(a_1, f(a_1))$ y $(b_1, f(b_1))$.
- Dependiendo de los valores de $f(c_1)$:
  - Si $f(a_1)f(c_1) < 0$, entonces $a_2 = a_1$ y $b_2 = c_1$.
  - Si $f(b_1)f(c_1) < 0$, entonces $a_2 = c_1$ y $b_2 = b_1$.
  - Si $f(c_1) = 0$, entonces $= c_1$.
- Se obtiene una nueva aproximación c_2 aplicando el método de la secante a a_2 y b_2.

3.3.7. Ventajas y Desventajas

Ventajas:

- Convergencia garantizada.

Desventajas:

- Es más lento que el método de la secante.

3.3.8. Métodos Iterativos de Punto Fijo

La fórmula general para los métodos iterativos de punto fijo es:

$$x_{n+1} = g(x_n),$$

donde $g(x)$ es una función continua apropiada.

Definición: Dada una función $g : \mathbb{R} \to \mathbb{R}$ continua, decimos que α es un punto fijo de g si $g(\alpha) = \alpha$. Gráficamente, los puntos fijos son los puntos donde la función $g(x)$ intersecta la recta $y = x$.

Ejemplo: El método de Newton es un caso específico de método iterativo de punto fijo, con la fórmula general:

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

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

Existencia de Soluciones de $x = g(x)$

Lema 1 (Existencia de Puntos Fijos): Sea $g(x)$ una función continua en $[a, b]$, y suponga que g satisface:

$$a \leq x \leq b \implies a \leq g(x) \leq b.$$

Entonces, la ecuación $x = g(x)$ tiene al menos una solución α en el intervalo $[a, b]$.

Demostración. Considerar la función $f(x) = x - g(x)$, $a \leq x \leq b$. Evaluando $f(x)$ en los puntos extremos, tenemos

$$f(a) \leq 0, \quad f(b) \geq 0$$

La función es continua en $[a, b]$ y $f(a)f(b) \leq 0$, luego, por el Teorema de Bolzano, $\exists\, \alpha \in [a, b]$ tal que $f(\alpha) = 0$.

Teorema 2 (Condición Suficiente de Convergencia): Sea $g : \mathbb{R} \to \mathbb{R}$, con $g \in C^1$ en $[a, b]$. Suponga que g satisface:

$$a \leq x \leq b \implies a \leq g(x) \leq b,$$

y que:

$$\lambda := \sup_{x \in [a,b]} |g'(x)| < 1.$$

Entonces:

i) Existe una solución única α de la ecuación $x = g(x)$ en $[a, b]$.

ii) Para cualquier valor inicial $x_0 \in [a, b]$, la iteración $x_{n+1} = g(x_n)$ converge a α.

iii) La distancia $|\alpha - x_n|$ se acota por:

$$|\alpha - x_n| \leq \frac{\lambda^n}{1 - \lambda} |x_0 - x_1|,$$

para $n \geq 0$.

iv) La relación de convergencia es:

$$\lim_{n \to \infty} \frac{|\alpha - x_{n+1}|}{|\alpha - x_n|} = |g'(\alpha)|.$$

Por lo tanto, para x_n cercano a α, se tiene:

$$\alpha - x_{n+1} \approx g'(\alpha)(\alpha - x_n).$$

Corolario 1: Suponga que $x = g(x)$ tiene una solución α y que $g(x)$ y $g'(x)$ son continuas en un intervalo alrededor de α. Luego:

a) Si $|g'(\alpha)| < 1$, la iteración $x_{n+1} = g(x_n)$ converge a α para x_0 suficientemente cercano a α.

b) Si $|g'(\alpha)| > 1$, la iteración $x_{n+1} = g(x_n)$ no converge a α.

c) Si $|g'(\alpha)| = 1$, no se pueden sacar conclusiones definitivas.

4. Resolución de Ecuaciones Lineales - Métodos Directos

Los métodos directos para resolver sistemas de ecuaciones lineales son métodos con un número finito de pasos, y obtienen la solución exacta provisto que todas las operaciones aritméticas sean exactas. El método directo más conocido y utilizado es la eliminación Gaussiana.

4.0.1. Eliminación de Gauss

El método de eliminación de Gauss transforma la matriz ampliada de un sistema lineal en una matriz triangular superior, lo que facilita su resolución por sustitución regresiva. El proceso consta de dos etapas:

1. **Eliminación progresiva**: Se transforma el sistema en uno triangular superior mediante operaciones elementales sobre las filas.

2. **Sustitución regresiva**: Una vez obtenido el sistema triangular, se resuelve comenzando por la última ecuación.

La matriz ampliada inicial $[A|b]$ se reduce en $n-1$ pasos hasta obtener una forma escalonada, que se resuelve usando sustitución regresiva.

En el proceso de eliminación, si el pivote $a_{kk}^{(k)} = 0$, se emplea **pivoteo parcial**. Esto consiste en intercambiar la ecuación $E_k^{(k)}$ con alguna $E_i^{(k)}$ para $i = k+1, \ldots, n$, donde $a_{ik}^{(k)} \neq 0$. Dado que la matriz A es no singular, al menos uno de estos elementos será diferente de cero. Tras el intercambio, se continúa con la eliminación.

El **número de operaciones** del método de eliminación de Gauss se divide en tres partes:

1. **Cálculo de $A^{(n)}$:** - El paso 1 requiere $n-1$ divisiones, $(n-1)^2$ multiplicaciones y $(n-1)^2$ sumas. - En general, el total de operaciones para calcular $A^{(n)}$ es aproximadamente $\frac{n^3}{3}$ para valores grandes de n.

2. **Modificación de $b^{(1)}$ a $b^{(n)}$:** - Requiere $\frac{n(n-1)}{2}$ sumas/restas y $\frac{n(n-1)}{2}$ multiplicaciones/divisiones.

3. **Solución de $A^{(n)}x = b^{(n)}$:** - Requiere $\frac{n(n-1)}{2}$ sumas/restas y $\frac{n(n+1)}{2}$ multiplicaciones/divisiones.

Por lo tanto, para valores grandes de n, el principal costo computacional está en la generación de $A^{(n)}$, con un orden de operaciones de aproximadamente $\frac{2n^3}{3}$.

Casos Especiales

Definición 1 Decimos que una matriz $A \in \mathbb{R}^{n\times n}$ es estrictamente diagonal dominante si

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \quad \forall i = 1, \ldots, n$$

Teorema 1 Toda matriz $A \in \mathbb{R}^{n\times n}$ estrictamente diagonal dominante es no singular. Para estas matrices, el sistema $Ax = b$ se puede resolver por eliminación de Gauss sin necesidad de pivoteo.

Teorema 2 Para toda matriz A simétrica y definida positiva, el sistema $Ax = b$ se puede resolver por eliminación de Gauss sin necesidad de pivoteo, siendo todos los elementos pivotes positivos.

Método de Gauss-Jordan

El método de Gauss-Jordan es una variante de la eliminación de Gauss, donde se eliminan las incógnitas tanto por encima como por debajo de la diagonal. La matriz ampliada $[A|b]$ se convierte en $[I|x]$ tras n pasos, obteniéndose así la solución x.

El método de Gauss-Jordan requiere un orden de $\frac{n^3}{2}$ multiplicaciones/divisiones y $\frac{n^3}{2}$ sumas/restas, por lo que tiene un mayor costo computacional que la eliminación de Gauss.

4.0.2. Factorización LU

Factorización LU a partir de la Eliminación Gaussiana

Mediante la eliminación de Gauss sin pivoteo, obtenemos el sistema triangular superior $Ux = g$, con $U = A^{(n)}$ y $g = b^{(n)}$:

$$U = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ 0 & u_{22} & \dots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & u_{nn} \end{bmatrix}, \quad u_{ij} = a_{ij}^{(i)}$$

Definimos la matriz triangular inferior L con los multiplicadores m_{ik} de la eliminación gaussiana:

$$L = \begin{bmatrix} 1 & 0 & \dots & 0 \\ m_{21} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & \dots & 1 \end{bmatrix}$$

Teorema 3 (Factorización LU): Sea $A \in \mathbb{R}^{n \times n}$ una matriz no singular, y sean L y U las matrices triangular inferior y superior obtenidas por eliminación de Gauss. Si U se genera sin pivoteo, entonces:

$$A = LU$$

Resolver $Ax = b$ equivale a resolver $LUx = b$, lo cual implica resolver dos sistemas triangulares:

- $Lg = b$ (Sistema triangular inferior, se resuelve por sustitución progresiva).
- $Ux = g$ (Sistema triangular superior, se resuelve por sustitución regresiva).

La sustitución progresiva y regresiva requieren alrededor de n^2 multiplicaciones/divisiones y n^2 sumas/restas. El costo para obtener las matrices L y U es el mismo que el de generar $A^{(n)} = U$ en la eliminación de Gauss. Una vez obtenida la factorización, los sistemas involucrando A se pueden resolver fácilmente para cualquier número de vectores b.

En la implementación de la factorización LU, los elementos $a_{ij}^{(k+1)}$ (con $j \geq i$) se almacenan reemplazando a los $a_{ij}^{(k)}$. Los multiplicadores m_{ij} se almacenan en los lugares de los elementos a_{ij} para $i > j$.

Unicidad de la factorización LU

Teorema 4 (Unicidad de la factorización LU): Si $A \in \mathbb{R}^{n \times n}$ es tal que la eliminación de Gauss puede realizarse sin pivoteo, entonces A puede factorizarse como $A = LU$, donde $U = A^{(n)}$ es el resultado final de la eliminación de Gauss aplicada a A, y L es una matriz triangular inferior con $l_{ii} = 1$, para $i = 1, \dots, n$, y $l_{ij} = m_{ij}$ para $i > j$. Dicha factorización es única.

Demostración. Demostraremos la unicidad de la factorización LU. Notemos que los factores L y U son no singulares ya que son matrices triangulares con elementos diagonales distintos de cero. Supongamos que $A = L_1U_1 = L_2U_2$ son dos factorizaciones LU de A. Entonces:

$$L_2^{-1}L_1U_1 = U_2$$

De aquí, podemos escribir:

$$L_2^{-1}L_1 = U_2U_1^{-1} \tag{1}$$

Sabemos que:

- La inversa de una matriz triangular inferior (superior) es una matriz triangular inferior (superior).
- El producto de dos matrices triangulares inferiores (superiores) es una matriz triangular inferior (superior).

Por lo tanto, $L_2^{-1}L_1$ es triangular inferior, mientras que $U_2U_1^{-1}$ es triangular superior. Esto implica que la ecuación (1) se puede expresar como:

$$L_2^{-1}L_1 = U_2U_1^{-1} = D,$$

siendo D una matriz diagonal. Como $[L_2]_{ii} = [L_2^{-1}]_{ii} = [L_1]_{ii} = 1$, tenemos que $D = I$, y por ende $L_1 = L_2$ y $U_1 = U_2$.

Factorización LU con Matriz de Permutación

Definición 2 (Matriz de permutación): Una matriz de permutación $P \in \mathbb{R}^{n\times n}$ es una matriz en la que hay exactamente una entrada cuyo valor es 1 en cada fila y en cada columna, siendo todas las demás entradas iguales a 0.

Teorema 5: Para toda matriz $A \in \mathbb{R}^{n\times n}$ no singular, existe una matriz de permutación P tal que PA posee una factorización LU, es decir, $PA = LU$.

Las matrices P, L y U se pueden generar al programar la eliminación de Gauss con pivoteo parcial, teniendo en cuenta los intercambios de filas requeridos. Una vez obtenida la factorización $PA = LU$, el sistema $Ax = b$ se resuelve permutando primero los elementos en b para construir $\tilde{b} = Pb$. Luego se resuelven los sistemas triangulares $Ly = \tilde{b}$ y $Ux = y$ por sustitución progresiva y regresiva, respectivamente.

Doolittle

El método de Doolittle es un procedimiento para factorizar una matriz A en el producto de dos matrices:

$$A = LU,$$

donde:

- L es una matriz triangular inferior con la diagonal principal compuesta únicamente por 1s.
- U es una matriz triangular superior.

El método se basa en descomponer cada elemento de A en términos de L y U mediante la relación:

$$a_{ij} = \sum_{k=1}^{i-1} l_{ik}u_{kj} + l_{ii}u_{ij},$$

aplicando las siguientes reglas:

- Los elementos de U se calculan como:

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}, \quad \text{para } i \leq j.$$

- Los elementos de L se calculan como:

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}}{u_{ii}}, \quad \text{para } i > j.$$

Este método es útil para resolver sistemas de ecuaciones lineales $Ax = b$, ya que permite descomponer el problema en dos sistemas más simples:

$$Ly = b \quad \text{y} \quad Ux = y.$$

Este método tiene una secuencia de operaciones óptima para ejecutarse en un ordenador con memoria caché, y es sencillo para programar.

Descomposición de Crout

Es una descomposición LU alternativa que usa una matriz U con números 1 en la diagonal. Se resuelve de manera similar al método de Doolittle. Ejemplificamos para un sistema de 3×3:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{pmatrix}$$

$$l_{11} = a_{11}, \quad l_{21} = a_{21}, \quad l_{31} = a_{31}$$
$$u_{12} = \frac{a_{12}}{l_{11}}, \quad u_{13} = \frac{a_{13}}{l_{11}}$$
$$l_{22} = a_{22} - l_{21}u_{12}, \quad l_{32} = a_{32} - l_{31}u_{12}$$
$$u_{23} = \frac{a_{23} - l_{21}u_{13}}{l_{22}}, \quad l_{33} = a_{33} - l_{31}u_{13} - l_{32}u_{23}$$

4.1. Cholesky

Teorema 6 La matriz $A \in \mathbb{R}^{n\times n}$ es definida positiva si y solo si existe una única matriz triangular superior R con elementos diagonales positivos tal que $A = R^T R$. Esta es la factorización de Cholesky de A.

Resolver el sistema $Ax = b$ es equivalente a resolver $R^T Rx = b$, lo cual es equivalente a resolver dos sistemas triangulares:

$R^T g = b$ (Sistema triangular inferior). Se resuelve por sustitución progresiva.

$Rx = g$ (Sistema triangular superior). Se resuelve por sustitución regresiva.

4.2. Factorización QR

Sea $A \in \mathbb{R}^{m\times n}$, $A = [a_1|a_2|\dots|a_n]$ una matriz con columnas $a_1, a_2, \dots, a_n$ linealmente independientes (esto implica $m \geq n$). Aplicando Gram-Schmidt a las columnas de A, resulta una base ortogonal normalizada $\{q_1, q_2, \dots, q_n\}$ del espacio columna de A, donde

$$A = QR$$

y

- $A \in \mathbb{R}^{m\times n}$: matriz con columnas linealmente independientes,
- $Q \in \mathbb{R}^{m\times n}$: base ortogonal normalizada del espacio columna de A,
- $R \in \mathbb{R}^{n\times n}$: matriz triangular superior con elementos diagonales positivos.

Teorema 7 (Factorización QR) Toda matriz $A \in \mathbb{R}^{m\times n}$ con columnas linealmente independientes puede factorizarse de manera única como $A = QR$ con Q y R definidas anteriormente. Si $A \in \mathbb{R}^{n\times n}$ es no singular, tenemos $Q^T = Q^{-1}$, ya que Q tiene columnas ortogonales normalizadas. El sistema $Ax = QRx = b$ es equivalente al sistema

$$Rx = Q^T b$$

el cual es un sistema triangular superior que se resuelve por sustitución regresiva.

5. Resolución de Ecuaciones Lineales - Métodos Iterativos

Los métodos iterativos generan una sucesión $\{x^{(k)}\}$ que converge a la solución del sistema lineal $Ax = b$. Estos métodos son eficientes para resolver sistemas lineales de grandes dimensiones, en especial, sistemas lineales dispersos como los que se presentan en los análisis de circuitos y en la solución numérica de sistemas de ecuaciones diferenciales parciales.

Para n grande, la eliminación de Gauss requiere aproximadamente $\frac{2}{3}n^3$ operaciones aritméticas, mientras que los métodos iterativos requieren del orden de n^2 operaciones para obtener una solución suficientemente precisa.

Comenzaremos describiendo los métodos iterativos de Jacobi y de Gauss-Seidel, métodos clásicos que datan de fines del siglo XVIII.

5.1. Método de Jacobi

El método de Jacobi es un método iterativo utilizado para resolver sistemas de ecuaciones lineales de la forma $Ax = b$, donde A es una matriz cuadrada y b es un vector. La idea principal del método es descomponer la matriz A en su parte diagonal y el resto, permitiendo calcular una aproximación de la solución de manera sucesiva.

Pasos del Método de Jacobi:

- **Inicialización:** Se elige un vector inicial $x^{(0)}$ como aproximación de la solución.
- **Iteración:** Para cada componente i de la solución, se actualiza $x^{(k+1)}$ usando la fórmula:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)} \right)$$

  donde a_{ii} es el elemento diagonal de la matriz A en la fila i, y la suma incluye todos los elementos de la fila i excepto a_{ii}.
- **Convergencia:** El proceso se repite hasta que la diferencia entre las iteraciones sucesivas sea menor que un umbral predefinido, lo que indica que se ha alcanzado una solución suficientemente precisa.

Propiedades:

La convergencia del método de Jacobi no está garantizada para todas las matrices. Sin embargo, si la matriz A es diagonal dominante o simétrica y definida positiva, el método convergerá.

Ventajas y Desventajas:

- **Ventajas:** Es fácil de implementar y adecuado para sistemas grandes y dispersos.
- **Desventajas:** Puede ser más lento en comparación con otros métodos, especialmente para sistemas muy grandes.

5.2. Método de Gauss-Seidel

El método de Gauss-Seidel es un método iterativo utilizado para resolver sistemas de ecuaciones lineales de la forma $Ax = b$. Este método es una mejora del método de Jacobi, ya que utiliza las soluciones más recientes tan pronto como están disponibles, lo que puede acelerar la convergencia.

Pasos del Método de Gauss-Seidel

- **Inicialización:** Se elige un vector inicial $x^{(0)}$ como aproximación de la solución.

- **Iteración:** Para cada componente i de la solución, se actualiza $x^{(k+1)}$ usando la fórmula:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j<i}}^{n} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right)$$

  donde a_{ii} es el elemento diagonal de la matriz A en la fila i. En esta fórmula, los valores actualizados $x_j^{(k+1)}$ se utilizan inmediatamente para las posiciones $j < i$.

- **Convergencia:** El proceso se repite hasta que la diferencia entre las iteraciones sucesivas sea menor que un umbral predefinido, lo que indica que se ha alcanzado una solución suficientemente precisa.

Propiedades:

La convergencia del método de Gauss-Seidel está garantizada si la matriz A es diagonal dominante o simétrica y definida positiva.

Ventajas y Desventajas:

- **Ventajas:** Generalmente converge más rápido que el método de Jacobi y es adecuado para sistemas grandes y dispersos.
- **Desventajas:** Puede no converger si la matriz A no cumple con las condiciones de convergencia, y puede ser más sensible a la elección de la matriz inicial.

5.3. Esquema General de los Métodos Iterativos

Sea $A \in \mathbb{R}^{n \times n}$ y el sistema a resolver $Ax = b$. Sea $N \in \mathbb{R}^{n \times n}$ no singular. Luego,

$$Nx = Nx - Ax + b$$

El proceso iterativo es de la forma:

$$Nx^{(k+1)} = (N - A)x^{(k)} + b, \quad k = 1, 2, 3, \ldots$$

Generalmente, N se elige tal que el sistema $Nz = f$ sea fácil de resolver. Para una matriz general $A \in \mathbb{R}^{n \times n}$, el método de Jacobi se define con:

$$N = D = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix}$$

Y el método de Gauss-Seidel se define con:

$$N = L + D = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

Para aplicar el método iterativo, la matriz N debe ser no singular. Si A es no singular, se puede lograr que N sea no singular intercambiando filas y/o columnas de A si es necesario.

5.4. Condiciones de Convergencia

Los métodos iterativos se pueden escribir en forma vectorial como:

$$Nx^{(k+1)} = (N - A)x^{(k)} + b$$

La solución del sistema cumple:

$$x = (I - N^{-1}A)x + N^{-1}b$$

Introduciendo el error $e^{(k)} = x - x^{(k)}$, y restando la ecuación anterior de la ecuación de solución, obtenemos:

$$e^{(k+1)} = (I - N^{-1}A)e^{(k)}$$

Teorema 1: Si $\|I - N^{-1}A\| < 1$, entonces la sucesión $\{x^{(k)}\}$, definida por el proceso iterativo, converge a la solución del sistema $Ax = b$ para cualquier estimación inicial $x^{(0)} \in \mathbb{R}^n$.

Demostración: Tomando la norma del error, tenemos:

$$\|e^{(k+1)}\| = \|(I - N^{-1}A)e^{(k)}\| \leq \|I - N^{-1}A\|\|e^{(k)}\|$$

$$\leq \|I - N^{-1}A\|\|(I - N^{-1}A)e^{(k-1)}\| \leq \cdots \leq \|I - N^{-1}A\|^{k+1}\|e^{(0)}\|$$

Siendo $\|I - N^{-1}A\| < 1$, se cumple que $\|I - N^{-1}A\|^{k+1} \to 0$ cuando $k \to \infty$, por lo que:

$$\lim_{k\to\infty} \|e^{(k+1)}\| = 0$$

Es decir, $x^{(k)} \to x$ cuando $k \to \infty$. □

La condición $\|I - N^{-1}A\| < 1$ es una condición suficiente de convergencia válida para cualquier norma matricial inducida.

Teorema 2 (Estabilidad asintótica de un proceso iterativo lineal): Sea $B \in \mathbb{R}^{n\times n}$. El proceso iterativo $x^{(k+1)} = Bx^{(k)}$ converge a $x = 0$ para todo vector inicial $x^{(0)} \in \mathbb{R}^n$ si y solo si $\rho(B) < 1$.

Corolario 1: La fórmula de iteración

$$Nx^{(k+1)} = (N - A)x^{(k)} + b$$

dará lugar a una sucesión que converge a la solución de $Ax = b$ para cualquier vector inicial $x^{(0)} \in \mathbb{R}^n$ si y solo si $\rho(I - N^{-1}A) < 1$.

Demostración: La demostración se obtiene aplicando el Teorema 2 al proceso iterativo dado por la ecuación

$$e^{(k+1)} = (I - N^{-1}A)e^{(k)}$$

□

La condición de que el radio espectral de la matriz del método iterativo, $\rho(I - N^{-1}A)$, sea menor que 1, es una condición necesaria y suficiente de convergencia.

Consideraremos ahora el caso especial de matrices diagonalmente dominantes.

Definición 1: Una matriz $A \in \mathbb{R}^{n\times n}$ es diagonalmente dominante si:

$$|a_{ii}| > \sum_{j\neq i} |a_{ij}|, \quad i = 1, \ldots, n$$

Teorema 3: Si la matriz $A \in \mathbb{R}^{n\times n}$ es diagonalmente dominante, entonces la sucesión $\{x^{(k)}\}$ generada por el método de Jacobi converge a la solución del sistema $Ax = b$ para cualquier $x^{(0)}$ inicial.

Teorema 4 Si la matriz $A \in \mathbb{R}^{n\times n}$ es diagonalmente dominante, entonces la sucesión $\{x^{(k)}\}$ generada por el método de Gauss-Seidel converge a la solución del sistema $Ax = b$ para cualquier $x^{(0)}$ inicial.

5.5. Métodos de Relajación

Utilizando la descomposición de A como $A = L + D + U$, como se usó en el Teorema 4, la ecuación en forma matricial de este método es la siguiente:

$$(D + \omega L)x^{(k+1)} = ((1-\omega)D - \omega U)\, x^{(k)} + \omega b$$

Si $(D + \omega L)^{-1}$ existe, entonces podemos expresar el método SOR de la forma:

$$x^{(k+1)} = T_\omega x^{(k)} + c_\omega$$

donde:

$$T_\omega = (D + \omega L)^{-1}((1-\omega)D - \omega U)$$

$$c_\omega = \omega(D + \omega L)^{-1} b$$

- Si $\omega = 1$, tenemos el método de Gauss-Seidel.
- Si $0 < \omega < 1$, se trata de un método de subrelajación. Estos métodos se pueden usar para obtener la convergencia de algunos sistemas que no son convergentes con el método de Gauss-Seidel.
- Si $\omega > 1$, se trata de un método de sobrerelajación.

Estos métodos se designan con la abreviatura **SOR** y se emplean para acelerar la convergencia en sistemas para los que el método de Gauss-Seidel converge.

El error del método **SOR** está determinado por:

$$e^{(k+1)} = T_\omega e^{(k)}$$

Por el Teorema 2, el método **SOR** converge a la solución de $Ax = b$ para todo vector inicial $x^{(0)}$ si y solo si $\rho(T_\omega) < 1$.

Para algunas matrices sencillas se puede determinar el valor de ω que minimiza $\rho(T_\omega)$, es decir, se puede elegir ω de manera óptima. En el siguiente teorema consideraremos el caso particular de las matrices definidas positivas y tridiagonales.

Teorema 5 Si A es definida positiva y tridiagonal, entonces la elección óptima de ω para el método SOR es:

$$\omega = \frac{2}{1 + \sqrt{1 - [\rho(T_J)]^2}}$$

donde $T_J = (I - D^{-1}A)$ es la matriz del método de Jacobi.

6. Aproximación de Autovalores

Teorema 1 (Polinomio Característico y Ecuación Característica) Sea $A \in \mathbb{R}^{n\times n}$. El polinomio característico de A es

$$p(\lambda) = \det(A - \lambda I)$$

El grado de $p(\lambda)$ es n. La ecuación característica de A es $p(\lambda) = 0$. Los autovalores de A son las soluciones de la ecuación característica o, de forma equivalente, las raíces del polinomio característico.

A tiene n autovalores, pero algunos pueden ser complejos (incluso si las entradas de A son reales), y algunos autovalores pueden repetirse. Si $A \in \mathbb{R}^{n\times n}$, entonces sus autovalores complejos deben ocurrir en pares conjugados. Este no es el caso si $A \in \mathbb{C}^{n\times n}$.

A^T posee el mismo polinomio característico que A, y por lo tanto, los mismos autovalores. Es decir, $\sigma(A) = \sigma(A^T)$ y $\det(A) = \det(A^T)$.

Definición 2 Una matriz cuadrada A se dice que es diagonalizable si es semejante a una matriz diagonal. Es decir, si mediante un cambio de base puede reducirse a una forma diagonal. En este caso, la matriz podrá descomponerse de la forma

$$A = PDP^{-1}$$

donde D es una matriz diagonal formada por los autovalores de A y P es una matriz invertible cuyos vectores columna son los autovectores de A.

Se dice que A es diagonalizable ortogonalmente si la matriz P es ortogonal, pudiendo descomponerse como

$$A = PDP^T$$

Toda matriz simétrica con coeficientes reales es diagonalizable ortogonalmente.

Teorema 2 (Teorema de la diagonalización) Una matriz $A \in \mathbb{R}^{n\times n}$ es diagonalizable si y solo si A tiene n autovectores linealmente independientes, o de forma equivalente, los autovectores de A conforman una base de $\mathbb{R}^n$.

Si la matriz $A \in \mathbb{R}^{n\times n}$ es diagonalizable, entonces cualquier vector $x \in \mathbb{R}^n$ puede expresarse como una combinación lineal única de los autovectores de A:

$$x = \sum_{i=1}^{n} c_i v_i$$

Luego, la transformación Ax puede expresarse como:

$$Ax = \sum_{i=1}^{n} \lambda_i c_i v_i$$

6.1. Teorema de Gershgorin

El teorema de Gershgorin permite determinar en qué rango se encuentran los autovalores de una matriz, y por lo tanto, permite acotar el radio espectral de la matriz.

Definición 3 (Círculos de Gershgorin) Sea $A \in \mathbb{C}^{n\times n}$. Se definen los círculos de Gershgorin como:

$$C_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq r_i\}, \quad \text{con} \quad r_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n$$

donde r_i es el radio del círculo C_i.

Teorema 3 (Teorema de Gershgorin) Sea $A \in \mathbb{C}^{n\times n}$ y sea λ un autovalor de A. Luego,

$$\lambda \in C_i \quad \text{para algún} \quad i = 1, \ldots, n,$$

donde C_i es un círculo de Gershgorin.

Demostración. Sea λ un autovalor de A y v un autovector asociado. Sea k la componente de v para la cual se tiene $|v_k| = \|v\|_\infty$. Luego, de la igualdad $Av = \lambda v$, se tiene para la k-ésima componente:

$$\sum_{j=1}^{n} a_{kj} v_j = \lambda v_k$$

$$(\lambda - a_{kk}) v_k = \sum_{\substack{j=1 \\ j\neq k}}^{n} a_{kj} v_j$$

$$|\lambda - a_{kk}||v_k| \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}||v_j| \leq r_k \|v\|_\infty$$

de donde

$$|\lambda - a_{kk}| \leq r_k.$$

Si bien el Teorema de Gershgorin es aplicable a matrices de coeficientes complejos, en el curso trabajaremos solamente con matrices de coeficientes reales. Estas matrices, en caso de no ser simétricas, pueden presentar autovalores complejos en pares conjugados.

Corolario 1 Sea $A \in \mathbb{R}^{n\times n}$ y sea λ un autovalor de A. Luego $\lambda \in C_i$ para algún $i = 1, \ldots, n$, donde C_i es un círculo de Gershgorin de A^T.

Demostración. Por el Teorema 1 sabemos que A y A^T poseen los mismos autovalores. Luego, por el Teorema 3, los autovalores de A deben encontrarse en los círculos de Gershgorin de las filas de A^T, que son las columnas de A.

Definición 4 Se dice que un subconjunto G de círculos de Gershgorin es un *grupo disjunto de círculos* si ningún círculo en G intersecta con un círculo que no pertenece a G.

Teorema 4 Si un grupo disjunto de círculos de Gershgorin G contiene k círculos, luego G contiene exactamente k autovalores (contando multiplicidades).

Demostración. Sea $A \in \mathbb{C}^{n\times n}$, y sea $\bar{A}(p)$ igual a la matriz A con sus elementos no diagonales multiplicados por la variable p, con p definida entre 0 y 1. La matriz $\bar{A}(0)$ tiene círculos de Gershgorin de radio 0 centrados en la ubicación de los elementos diagonales de A, y sus autovalores son los elementos diagonales de A. A medida que se incrementa p, los radios aumentan en base a p, y los autovalores también se moverán. El polinomio característico de $\bar{A}(p)$ será una función continua de la variable p, y sus raíces también serán continuas. Es decir, que los autovalores seguirán una trayectoria continua a medida que se incrementa el valor de p. De esta continuidad se desprende que a medida que p se incrementa de 0 a 1 no es posible que un autovalor se desplace de un grupo disjunto de círculos a otro grupo disjunto de círculos sin hallarse en el exterior de cualquier círculo, lo cual violaría el Teorema 3.

6.2. Método de la Potencia

El método de la potencia permite estimar el radio espectral de una matriz.

Teorema 5 (Método de la Potencia) Sea la matriz $A \in \mathbb{R}^{n\times n}$ y sean $\lambda_1, \ldots, \lambda_n$ sus autovalores, repetidos según su multiplicidad y ordenados según su módulo:

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n| \geq 0$$

Suponemos que existe un único autovalor de módulo máximo.

Sea $\{v_1, v_2, \ldots, v_n\}$ la base de autovectores correspondiente. Sea $z^{(0)}$ una estimación inicial del autovector v_1 (elegida como sea, incluso de forma aleatoria).

Definimos

$$w^{(n+1)} = Az^{(n)}$$

$$z^{(n+1)} = \frac{w^{(n+1)}}{\|w^{(n+1)}\|_\infty}, \quad n \geq 0$$

Entonces resulta que

$$z^{(n)} \to \frac{v_1}{\|v_1\|_\infty} \quad \text{cuando} \quad n \to \infty$$

y eligiendo una componente k no nula de $w^{(n-1)}$

$$\lambda^{(n)} = \frac{w_k^{(n)}}{z_k^{(n-1)}} \to \lambda_1 \quad \text{cuando} \quad n \to \infty.$$

7. Interpolación Polinómica

Problema de Interpolación Polinómica

Sea $f(x)$ una cierta función de la que posiblemente no se conoce una forma explícita, o bien es muy complicada para evaluarla, derivarla, integrarla, hallarle ceros, etc. Podemos aproximar $f(x)$ por funciones simples, y hacer los cálculos con estas aproximaciones.

Dados $n+1$ números distintos $a \leq x_1 < x_2 < \ldots < x_{n+1} \leq b$ de un intervalo $[a, b]$, llamados nodos de la interpolación, y $n+1$ números reales $y_1, y_2, \ldots, y_{n+1}$, con $y_i = f(x_i)$, para $i = 1, 2, \ldots, n+1$, llamados valores de la interpolación, el problema de interpolación trata de encontrar una función p, en una cierta clase prefijada de funciones F, tal que $p(x_i) = y_i$ para $i = 1, 2, \ldots, n+1$.

El caso particular más conocido es el problema de interpolación polinómica, en el que F es el conjunto de polinomios de grado menor o igual a n. Hemos supuesto que los números $x_1, x_2, \ldots, x_{n+1}$ están ordenados de menor a mayor, pero esto no es necesario. Lo importante es que sean números distintos.

Sea $x_{\min} = \text{mín}(x_1, x_2, \ldots, x_{n+1})$ y $x_{\max} = \text{máx}(x_1, x_2, \ldots, x_{n+1})$. Luego, si evaluamos $p(x)$ en $x \in [x_{\min}, x_{\max}]$, decimos que estamos interpolando, mientras que si evaluamos $p(x)$ en $x \notin [x_{\min}, x_{\max}]$, decimos que estamos extrapolando.

7.1. Interpolación Polinómica

Dados $n+1$ pares ordenados (x_i, y_i):

$$\{(x_i, y_i) : y_i = f(x_i), i = 0, 1, \ldots, n\},$$

también llamados puntos de la función f, donde $x_0, x_1, \ldots, x_n$ son números reales distintos, se trata de encontrar un polinomio $p(x)$ que interpole los datos, es decir, tal que:

$$p(x_i) = y_i, \quad i = 0, 1, \ldots, n \tag{1}$$

Surgen las siguientes preguntas: ¿Existe dicho polinomio $p(x)$, y si existe, de qué grado es? ¿Es único? ¿Cómo lo encontramos?

Consideremos el polinomio de grado m:

$$p(x) = a_0 + a_1 x + \cdots + a_m x^m$$

Vemos que hay $m+1$ parámetros independientes $a_0, a_1, \ldots, a_m$. Puesto que (1) impone $n+1$ condiciones sobre $p(x)$, es razonable considerar el caso en que $m = n$. Es decir, queremos encontrar $a_0, a_1, \ldots, a_n$ tales que:

$$\begin{aligned} a_0 + a_1 x_0 + a_2 x_0^2 + \cdots + a_n x_0^n &= y_0 \\ a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_n x_1^n &= y_1 \\ &\vdots \\ a_0 + a_1 x_n + a_2 x_n^2 + \cdots + a_n x_n^n &= y_n \end{aligned}$$

Luego, tenemos un sistema lineal de $n+1$ ecuaciones y $n+1$ incógnitas, que podemos escribir en forma matricial y vectorial como:

$$Xa = y$$

donde

$$X = \begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{pmatrix}, \quad a = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix}, \quad y = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

La matriz X es la matriz de Vandermonde. Puede demostrarse que para la matriz de Vandermonde se tiene:

$$\det(X) = \prod_{0 \leq j < i \leq n} (x_i - x_j) \tag{2}$$

Teorema 1 (Existencia y unicidad del polinomio interpolante) Dados $n+1$ puntos distintos $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$, con $x_0, x_1, \ldots, x_n$, números distintos, existe un polinomio $p(x)$ de grado menor o igual a n que interpola dichos puntos. Dicho polinomio es único en el conjunto de polinomios de grado menor o igual a n.

Demostración. Considerando la expresión del determinante de la matriz de Vandermonde, dado en la ecuación (2), se tiene que $\det(X) \neq 0$, porque si $i \neq j$, entonces $x_i \neq x_j$. Luego X es no singular y el sistema $Xa = y$ tiene solución única. El grado del polinomio interpolante puede ser menor o igual a n para los $n+1$ puntos dados, ya que algunos de los coeficientes a_i, $i = 0, 1, \ldots, n$, pueden ser iguales a cero.

7.2. Interpolación de Lagrange

Caso lineal

Encontrar un polinomio de primer grado que pase por los puntos distintos (x_0, y_0) y (x_1, y_1), donde $y_0 = f(x_0)$ y $y_1 = f(x_1)$. Definimos las funciones

$$L_0(x) = \frac{x - x_1}{x_0 - x_1} \quad \text{y} \quad L_1(x) = \frac{x - x_0}{x_1 - x_0}$$

Luego se define

$$p(x) = L_0(x)f(x_0) + L_1(x)f(x_1)$$

Como

$$L_0(x_0) = 1, \quad L_0(x_1) = 0, \quad L_1(x_0) = 0, \quad L_1(x_1) = 1,$$

tenemos

$$p(x_0) = y_0, \quad p(x_1) = y_1$$

Luego, p es la única función lineal que pasa por (x_0, y_0) y (x_1, y_1).

Caso general

Consideremos un polinomio de grado máximo n que pase por los $n+1$ puntos $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$.

Para $k = 0, 1, \ldots, n$, definimos

$$L_k(x) = \frac{(x - x_0)(x - x_1)\ldots(x - x_{k-1})(x - x_{k+1})\ldots(x - x_n)}{(x_k - x_0)(x_k - x_1)\ldots(x_k - x_{k-1})(x_k - x_{k+1})\ldots(x_k - x_n)} = \prod_{i=0, i\neq k}^{n} \frac{x - x_i}{x_k - x_i}$$

Las funciones $L_k(x)$ satisfacen las siguientes propiedades:

$$L_k(x_i) = 0, \quad \text{para} \quad i \neq k$$

$$L_k(x_k) = 1$$

El polinomio interpolador de Lagrange está dado por:

$$p(x) = L_0(x)y_0 + L_1(x)y_1 + \cdots + L_n(x)y_n = \sum_{k=0}^{n} L_k(x)y_k \tag{4}$$

Notar que $p(x)$ interpola los datos:

$$p(x_i) = y_i, \quad \text{para} \quad i = 0, 1, \ldots, n$$

Además, el grado de $p(x)$ es menor o igual a n ya que el grado de $L_k(x)$ para $k = 0, \ldots, n$ es igual a n. Luego, $p(x)$ es único en el conjunto de polinomios de grado menor o igual a n, de acuerdo al Teorema 1.

Desventajas de la interpolación de Lagrange

- **Requiere gran cantidad de cálculos.** Para cada valor de x es necesario reevaluar todas las funciones $L_k(x)$.
- **Si se agrega un punto, el polinomio $p_n(x)$ es de poca utilidad** para obtener el polinomio de grado superior.

7.3. Método de las Diferencias Divididas de Newton

Dados $n+1$ puntos $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$, se busca expresar el polinomio interpolador en la forma:

$$p(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots + a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1})$$

Dicho polinomio se puede obtener mediante un esquema recursivo:

$$\begin{aligned} p_1(x) &= a_0 + a_1(x - x_0) \\ p_2(x) &= p_1(x) + a_2(x - x_0)(x - x_1) \\ &\vdots \\ p_n(x) &= p_{n-1}(x) + a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1}) \end{aligned}$$

Para determinar el polinomio (5) se necesita conocer cómo calcular los coeficientes a_i, con $i = 0, 1, \ldots, n-1$. Imponiendo las condiciones de interpolación, $p_n(x_i) = y_i$, $i = 0, \ldots, n$, obtenemos:

$$p_n(x_0) = a_0 = y_0$$

$$p_n(x_1) = y_0 + a_1(x_1 - x_0) = y_1 \quad \text{de donde} \quad a_1 = \frac{y_1 - y_0}{x_1 - x_0}$$

$$p_n(x_2) = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1) = y_2$$

$$\text{de donde} \quad a_2 = \frac{\frac{y_2 - y_0}{x_2 - x_0} - \frac{y_1 - y_0}{x_1 - x_0}}{x_2 - x_1}$$

Vemos como a medida que i aumenta, el cálculo de los coeficientes a_i siguiendo esta estrategia comienza rápidamente a dificultarse. Para calcular los coeficientes a_i introduciremos el concepto de **diferencias divididas**.

Diferencia dividida de primer orden.

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

Diferencia dividida de segundo orden.

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$$

Diferencia dividida de orden k.

$$f[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{f[x_{i+1}, \ldots, x_{i+k}] - f[x_i, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}$$

Proposición 1. Sea $(i_0, i_1, \ldots, i_n)$ una permutación (o reubicación) de los enteros $(0, 1, \ldots, n)$. Luego,

$$f[x_{i_0}, x_{i_1}, \ldots, x_{i_n}] = f[x_0, x_1, \ldots, x_n]$$

Teorema 2. Suponga que $f(x)$ está definida en $[a, b]$ y que $x_0, x_1, \ldots, x_n$ son valores distintos en $[a, b]$. El polinomio de grado $\leq k$ que interpola $f(x)$ en $\{x_i, x_{i+1}, \ldots, x_{i+k}\} \subset \{x_0, x_1, \ldots, x_n\}$ está dado por

$$p_{i,k}(x) = f(x_i) + (x - x_i)f[x_i, x_{i+1}] + \cdots + (x - x_i)(x - x_{i+1})\cdots(x - x_{i+k-1})f[x_i, \ldots, x_{i+k}]$$

Teorema 3 (Error de la Interpolación Polinómica). Sean $x_0, x_1, \ldots, x_n$, $n+1$ números distintos en $[a, b]$, y sea $f(x) \in C^{n+1}$ en $[a, b]$. Luego, para todo $x \in [a, b]$ existe $\xi(x) \in (a, b)$ tal que

$$f(x) - p(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_n)}{(n+1)!} f^{(n+1)}(\xi(x))$$

donde $p(x)$ es el polinomio interpolante de grado menor o igual a n.

Acotación del error: Caso general. El procedimiento empleado en el Ejemplo 3 para acotar el error de interpolación se puede generalizar. Para $x_0, x_1, \ldots, x_n$ distintos en $[a, b]$ y $x \in [a, b]$, el error de interpolación está dado por

$$f(x) - p_n(x) = \frac{\Phi_n(x)}{(n+1)!} f^{(n+1)}(\xi(x))$$

donde

$$\Phi_n(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$$

Queremos hallar una cota del error $|f(x) - p_n(x)|$ en $[a, b]$:

$$|f(x) - p_n(x)| \leq \frac{1}{(n+1)!} \max_{x_0 \leq x \leq x_1} |\Phi_n(x)| \max_{x \in [a,b]} |f^{(n+1)}(x)|$$

Estas cotas no son fáciles de evaluar, salvo en casos particulares.

Fenómeno de Runge. El fenómeno de Runge ocurre cuando se usan polinomios de alto grado para interpolar funciones en intervalos amplios, especialmente en puntos equidistantes. En lugar de acercarse a la función original, el polinomio oscila fuertemente en los extremos del intervalo, causando grandes errores. Este problema se mitiga utilizando métodos como la interpolación en nodos de Chebyshev o evitando polinomios de alto grado.

8. Ajuste de Curvas

8.1. Polinomios de Chebyshev

Definición: Para $n \geq 0$ definimos la función

$$T_n(x) = \cos(n \times \cos^{-1}(x)), \quad -1 \leq x \leq 1 \tag{1}$$

De esta definición no es obvio que $T_n(x)$ es un polinomio en x de grado n, pero verificaremos este hecho a continuación. Introduciendo

$$\theta = \cos^{-1}(x) \quad \text{o} \quad x = \cos(\theta), \quad 0 \leq \theta \leq \pi$$

tenemos

$$T_n(x) = \cos(n\theta)$$

Evaluemos $T_n(x)$ para los primeros tres valores de n:
- Para $n = 0$:

$$T_0(x) = \cos(0 \times \theta) = 1$$

- Para $n = 1$:

$$T_1(x) = \cos(\theta) = x$$

- Para $n = 2$:

$$T_2(x) = \cos(2\theta) = 2\cos^2(\theta) - 1 = 2x^2 - 1$$

Hasta aquí vemos que efectivamente obtenemos un polinomio de grado n. Para analizar el caso general, consideraremos la siguiente fórmula trigonométrica:

$$\cos(\alpha \pm \beta) = \cos(\alpha)\cos(\beta) \mp \sin(\alpha)\sin(\beta) \tag{2}$$

Aplicando la fórmula (2) obtenemos para $n \geq 1$:

$$T_{n+1}(x) = \cos((n+1)\theta) = \cos(n\theta + \theta) = \cos(n\theta)\cos(\theta) - \sin(n\theta)\sin(\theta)$$
$$T_{n-1}(x) = \cos((n-1)\theta) = \cos(n\theta - \theta) = \cos(n\theta)\cos(\theta) + \sin(n\theta)\sin(\theta)$$

Sumando ambas expresiones obtenemos

$$T_{n+1}(x) + T_{n-1}(x) = 2\cos(n\theta)\cos(\theta) = 2xT_n(x)$$

Luego, la relación de recurrencia para los polinomios de Chebyshev es

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n \geq 1 \tag{3}$$

La ecuación (3) nos da la relación de recurrencia para los polinomios de Chebyshev. Siendo $T_0(x) = 1$ y $T_1(x) = x$, la relación de recurrencia demuestra que $T_n(x)$, dado por la ecuación (1), es en efecto un polinomio de grado n.

Propiedades: Veamos cuáles son las principales propiedades de los polinomios de Chebyshev. Para todo $n \geq 0$ se cumple:

- $|T_n(x)| \leq 1, \quad -1 \leq x \leq 1$
- $T_n(x) = 2^{n-1}x^n +$ términos de grado menor que n

- Todas las raíces de los polinomios de Chebyshev se encuentran entre $x = -1$ y $x = 1$, y el valor absoluto de los polinomios de Chebyshev en el intervalo $[-1, 1]$ es menor o igual a 1.

Teorema 1: Sea $n \in \mathbb{Z}$, $n \geq 1$. De entre todos los posibles polinomios mónicos de grado n, $\tilde{T}_n(x)$ es el polinomio mónico con menor máximo en $[-1, 1]$, y su valor máximo en $[-1, 1]$ es $\frac{1}{2^{n-1}}$.

8.2. Aproximación Mediante Polinomios de Chebyshev

Supongamos que se desea utilizar un polinomio de interpolación para aproximar una función $f(x)$ en un intervalo $[a, b]$. El fenómeno de Runge es un claro indicio de que no es conveniente elegir los nodos de interpolación uniformemente espaciados.

Lo ideal sería elegir los nodos de forma de minimizar el máximo error de interpolación en el intervalo. Definiendo

$$E(p) = \max_{x \in [a,b]} |f(x) - p(x)|$$

queremos hallar la mejor aproximación posible para un grado dado n. Definimos el error minimax como:

$$\rho_n(f) = \min_{\text{grado}(p) \leq n} E(p) = \min_{\text{grado}(p) \leq n} \left(\max_{x \in [a,b]} |f(x) - p(x)| \right)$$

Es decir, $\rho_n(f)$ es la menor cota superior sobre el error que se puede obtener al aproximar $f(x)$ por un polinomio de grado $\leq n$ en el intervalo $[a, b]$. El polinomio correspondiente, denotado $m_n(x)$, es el polinomio minimax. Es decir, $E(m_n) = \rho_n(f)$.

Sin embargo, el polinomio minimax es muy difícil de obtener, ya que para ello se necesita resolver el problema de optimización no lineal minimax dado por la ecuación anterior. En su lugar, se puede obtener una aproximación del polinomio minimax eligiendo los nodos de interpolación en base a las raíces del polinomio de Chebyshev.

8.3. Aproximación de Mínimos Cuadrados

El problema de mínimos cuadrados consiste en aproximar una función $g(x)$ desconocida a partir de un conjunto de datos $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, donde $y_i = g(x_i) + v_i$, siendo v_i el error de medición.

Queremos aproximar $g(x)$ mediante una función $f(x)$ de la forma:

$$f(x) = a_1\Phi_1(x) + a_2\Phi_2(x) + \cdots + a_p\Phi_p(x)$$

donde $a_1, a_2, \ldots, a_p$ son coeficientes ajustables y $\Phi_1(x), \Phi_2(x), \ldots, \Phi_p(x)$ son funciones dadas.

El objetivo es minimizar la suma de los errores al cuadrado:

$$G(a_1, a_2, \ldots, a_p) = \sum_{i=1}^{m} \left(a_1\Phi_1(x_i) + \cdots + a_p\Phi_p(x_i) - y_i\right)^2$$

El problema de optimización se puede resolver encontrando el mínimo de $G(x)$, que es una función convexa. El mínimo de $G(x)$ se alcanza cuando su gradiente es cero:

$$\frac{\partial G}{\partial a_i} = 0, \quad i = 1, 2, \ldots, p$$

Esto lleva al sistema lineal de p ecuaciones:

$$\sum_{j=1}^{m} \Phi_k(x_j)\Phi_i(x_j)a_k = \sum_{j=1}^{m} y_j\Phi_i(x_j), \quad i = 1, 2, \ldots, p$$

La solución de este sistema proporciona los coeficientes $a_1, a_2, \ldots, a_p$ que minimizan el error de la aproximación.

8.3.1. Aproximación Lineal de Mínimos Cuadrados

En la aproximación lineal de mínimos cuadrados, se aproximan los datos mediante un polinomio de primer grado:

$$f(x) = a_1 + a_2 x$$

es decir, tomando en la ecuación (7) $\Phi_1(x) = 1$ y $\Phi_2(x) = x$. En este caso, el sistema de ecuaciones se reduce a:

$$m a_1 + \left(\sum_{j=1}^{m} x_j \right) a_2 = \sum_{j=1}^{m} y_j$$

$$\left(\sum_{j=1}^{m} x_j \right) a_1 + \left(\sum_{j=1}^{m} x_j^2 \right) a_2 = \sum_{j=1}^{m} x_j y_j$$

Este sistema de ecuaciones se resuelve para obtener los coeficientes a_1 y a_2 que minimizan el error de la aproximación lineal.

8.3.2. Aproximación de Mínimos Cuadrados en Forma Vectorial

El error de aproximación dado por la ecuación se puede escribir como:

$$\varepsilon = Ax - b$$

con

$$A = \begin{bmatrix} \Phi_1(x_1) & \dots & \Phi_p(x_1) \\ \vdots & \ddots & \vdots \\ \Phi_1(x_m) & \dots & \Phi_p(x_m) \end{bmatrix} \in \mathbb{R}^{m \times p}, \quad b = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m, \quad x = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \in \mathbb{R}^p$$

El problema de mínimos cuadrados consiste en hallar un vector x que minimice:

$$G(x) = \sum_{i=1}^{m} \varepsilon_i^2 = \varepsilon^T \varepsilon = (Ax - b)^T (Ax - b)$$

El gradiente de un término lineal $f(x) = b^T x$ es $\nabla f(x) = b$, y el gradiente de un término cuadrático $g(x) = x^T Q x$ es $\nabla g(x) = 2Qx$.

Teorema 2: El conjunto solución del problema de mínimos cuadrados es el conjunto de soluciones del sistema:

$$A^T A x = A^T b$$

Si rango$(A) = n$, existe una solución única dada por:

$$x = (A^T A)^{-1} A^T b$$

8.3.3. Aproximación Polinomial de Mínimos Cuadrados

Polinomio de primer grado:

$$f(x) = a_1 + a_2 x, \quad \Phi_1(x) = 1, \quad \Phi_2(x) = x$$

Polinomio de segundo grado:

$$f(x) = a_1 + a_2 x + a_3 x^2, \quad \Phi_1(x) = 1, \quad \Phi_2(x) = x, \quad \Phi_3(x) = x^2$$

8.3.4. Casos Especiales

Función exponencial natural: Cuando los datos tienen una relación exponencial $f(x) = a_2 e^{a_1 x}$, el error es:

$$\varepsilon_i = \ln(f(x_i)) - \ln(y_i) = \ln(a_2) + a_1 x_i - \ln(y_i)$$

En forma vectorial:

$$\varepsilon = Ax - b$$

con

$$A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \in \mathbb{R}^{m\times 2}, \quad b = \begin{bmatrix} \ln(y_1) \\ \ln(y_2) \\ \vdots \\ \ln(y_m) \end{bmatrix} \in \mathbb{R}^m, \quad x = \begin{bmatrix} \ln(a_2) \\ a_1 \end{bmatrix} \in \mathbb{R}^2$$

Función potencial: Si la función es de la forma $f(x) = a_2 x^{a_1}$, tomando el logaritmo se obtiene:

$$\ln(f(x)) = \ln(a_2) + a_1 \ln(x)$$

El error es:

$$\varepsilon_i = \ln(f(x_i)) - \ln(y_i) = \ln(a_2) + a_1 \ln(x_i) - \ln(y_i)$$

En forma vectorial:

$$\varepsilon = Ax - b$$

con

$$A = \begin{bmatrix} 1 & \ln(x_1) \\ 1 & \ln(x_2) \\ \vdots & \vdots \\ 1 & \ln(x_m) \end{bmatrix} \in \mathbb{R}^{m\times 2}, \quad b = \begin{bmatrix} \ln(y_1) \\ \ln(y_2) \\ \vdots \\ \ln(y_m) \end{bmatrix} \in \mathbb{R}^m, \quad x = \begin{bmatrix} \ln(a_2) \\ a_1 \end{bmatrix} \in \mathbb{R}^2$$

8.3.5. Aplicación de la Factorización QR al Problema de Mínimos Cuadrados

El sistema de mínimos cuadrados $A^T Ax = A^T b$ se puede resolver usando la factorización QR. Si $A = QR$, entonces:

$$A^T A = (QR)^T QR = R^T Q^T QR = R^T R$$

Por lo tanto, se tiene:

$$R^T Rx = R^T Q^T b$$

Resolviendo el sistema triangular superior:

$$Rx = Q^T b$$

Finalmente, la solución es:

$$x = R^{-1} Q^T b = (A^T A)^{-1} A^T b$$

que es la solución de mínimos cuadrados cuando el sistema es inconsistente.

9. Integración Numérica

Dada una función $f : [a, b] \to \mathbb{R}$, se quiere calcular la integral definida:

$$I(f) = \int_a^b f(x)\,dx$$

Por el Teorema Fundamental del Cálculo:

$$I(f) = F(b) - F(a)$$

donde $F(x)$ es cualquier antiderivada de $f(x)$. Sin embargo, muchos integrandos no tienen una antiderivada explícita o no es fácil de obtener, por lo que se requieren otros métodos, como la *cuadratura* o *integración numérica*, que aproxima la integral por una suma:

$$\sum_{i=0}^{n} a_i f(x_i)$$

9.1. Integración Numérica Basada en Polinomios Interpolantes

Sea $\{x_0, x_1, \ldots, x_n\}$ un conjunto de $n+1$ nodos distintos en $[a, b]$. Tenemos:

$$f(x) = p_n(x) + \frac{1}{(n+1)!}\prod_{i=0}^{n}(x - x_i) f^{(n+1)}(\xi(x))$$

donde $p_n(x)$ es el polinomio interpolante de Lagrange:

$$p_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x)$$

Integrando en el intervalo $[a, b]$, obtenemos:

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} a_i f(x_i) + \frac{1}{(n+1)!}\int_a^b \prod_{i=0}^{n}(x - x_i) f^{(n+1)}(\xi(x))\,dx$$

donde:

$$a_i = \int_a^b L_i(x)\,dx, \quad i = 0, 1, \ldots, n$$

9.2. Regla del Trapecio

Aproximamos $f(x)$ mediante un polinomio lineal. Sean $x_0 = a$, $x_1 = b$, y $h = b - a$. El polinomio de primer grado que interpola estos nodos es:

$$p_1(x) = \frac{(x - x_1)}{(x_0 - x_1)} f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)} f(x_1)$$

Integrando, obtenemos:

$$\int_a^b f(x)\,dx = \int_{x_0}^{x_1} \left[\frac{(x - x_1)}{(x_0 - x_1)} f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)} f(x_1)\right] dx$$

Aplicando el teorema del valor medio ponderado de las integrales:

$$\int_{x_0}^{x_1} f''(\xi(x))(x - x_0)(x - x_1)\,dx = f''(c)\int_{x_0}^{x_1} (x - x_0)(x - x_1)\,dx$$

y evaluando la integral:

$$\int_{x_0}^{x_1} (x - x_0)(x - x_1)\,dx = \frac{(x_1 - x_0)^3}{6}$$

Finalmente, obtenemos la regla del trapecio:

$$\int_{x_0}^{x_1} f(x)\,dx = \frac{h}{2}\left[f(x_0) + f(x_1)\right] - \frac{h^3}{12} f''(c)$$

9.3. Método Compuesto del Trapecio

En el método compuesto del trapecio se utilizan varios subintervalos de igual longitud. Sea n el número de subintervalos, con:

$$h = \frac{b-a}{n}, \quad x_j = a + jh, \quad j = 0, 1, \ldots, n$$

La integral se aproxima por la suma de integrales en cada subintervalo:

$$I(f) = \int_a^b f(x)\,dx = \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} f(x)\,dx$$

Aproximando cada integral por un trapecio:

$$T_n(f) = \frac{h}{2}\left(f(x_0) + f(x_1)\right) + \frac{h}{2}\left(f(x_1) + f(x_2)\right) + \cdots + \frac{h}{2}\left(f(x_{n-1}) + f(x_n)\right)$$

Finalmente, la fórmula de integración numérica trapezoidal es:

$$T_n(f) = h\left[\frac{1}{2}f(x_0) + f(x_1) + f(x_2) + \cdots + f(x_{n-1}) + \frac{1}{2}f(x_n)\right]$$

9.4. Error de la Integración Numérica Trapezoidal

Teorema 1: Sea $f : \mathbb{R} \to \mathbb{R}$, $f \in C^2$ en $[a, b]$. Luego, el error de la integración trapezoidal $E_T^n(f)$ está dado por:

$$E_T^n(f) := \int_a^b f(x)\,dx - T_n(f) = -\frac{h^2(b-a)}{12} f''(c_n)$$

para algún $c_n \in [a, b]$.

9.5. Regla de Simpson

Aproximamos $f(x)$ mediante el polinomio de interpolación de grado 2, con los nodos $x_0 = a$, $x_1 = a + h$, $x_2 = b$, con $h = \frac{b-a}{2}$.

El polinomio interpolante de Lagrange está dado por:

$$p_2(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} f(x_1) + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} f(x_2)$$

Luego,

$$\int_a^b f(x)\,dx = \int_{x_2}^{x_0} p_2(x)\,dx + \int_{x_2}^{x_0} \frac{(x-x_0)(x-x_1)(x-x_2)}{6} f^{(3)}(\xi(x))\,dx$$

La integración de un polinomio de segundo grado no presenta dificultades. Se obtiene:

$$\int_{x_2}^{x_0} p_2(x)\,dx = \frac{h}{3}\left(f(x_0) + 4f(x_1) + f(x_2)\right)$$

Por otra parte, se puede demostrar que el error de integración está dado por (la demostración no es directa):

$$\int_{x_2}^{x_0} \frac{(x-x_0)(x-x_1)(x-x_2)}{6} f^{(3)}(\xi(x))\,dx = -\frac{h^5}{90} f^{(4)}(\xi)$$

para algún $\xi \in [x_0, x_2]$.

Con lo cual obtenemos la siguiente Regla de Simpson:

$$\int_a^b f(x)\,dx = \frac{h}{3}\left(f(x_0) + 4f(x_1) + f(x_2)\right) - \frac{h^5}{90} f^{(4)}(\xi) \quad \text{para algún } \xi \in [x_0, x_2].$$

9.6. Método Compuesto de Simpson

En el método compuesto de Simpson se divide el intervalo $[a, b]$ en n subintervalos de igual longitud, siendo n un número par.

$$h = \frac{b-a}{n}, \quad x_j = a + jh, \quad j = 0, 1, \ldots, n$$

La integral definida se aproxima aplicando la regla de Simpson a cada par de subintervalos adyacentes:

$$I(f) = \int_a^b f(x)\,dx = \int_{x_2}^{x_0} f(x)\,dx + \int_{x_4}^{x_2} f(x)\,dx + \cdots + \int_{x_n}^{x_{n-2}} f(x)\,dx$$

Aproximando cada integral mediante la regla de Simpson, obtenemos:

$$\begin{aligned} S_n(f) = &\frac{h}{3}\left(f(x_0) + 4f(x_1) + f(x_2)\right) \\ &+ \frac{h}{3}\left(f(x_2) + 4f(x_3) + f(x_4)\right) \\ &+ \ldots \\ &+ \frac{h}{3}\left(f(x_{n-4}) + 4f(x_{n-3}) + f(x_{n-2})\right) \\ &+ \frac{h}{3}\left(f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)\right) \end{aligned}$$

De donde obtenemos la siguiente fórmula de iteración numérica de Simpson:

$$S_n(f) = \frac{h}{3}\left[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \cdots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)\right]$$

9.7. Error de la Integración Numérica de Simpson

Teorema 2. Sea $f : \mathbb{R} \to \mathbb{R}$, $f \in C^4$ en $[a, b]$, y sea $n \in \mathbb{Z}^+$ un número par. Luego, el error de la integración numérica de Simpson está dado por:

$$E_S^n(f) := \int_a^b f(x)\,dx - S_n(f) = -\frac{h^4(b-a)}{180} f^{(4)}(c_n)$$

para algún $c_n \in [a, b]$.

9.8. Integración Numérica en Dominio Bidimensional

Se desea calcular la integral de una función $f(x, y)$ en un dominio bidimensional $Q = \{(x, y) \in \mathbb{R}^2 : a \le x \le b, c(x) \le y \le d(x)\}$.

$$I = \int_a^b \int_{c(x)}^{d(x)} f(x, y)\,dy\,dx$$

Definimos:

$$G(x) = \int_{c(x)}^{d(x)} f(x, y)\,dy$$

Entonces, la integral total se expresa como:

$$I = \int_a^b G(x)\,dx$$

Lo que puede aproximarse mediante la siguiente fórmula:

$$I \approx \sum_{i=1}^{n} w_i G(x_i)$$

donde w_i son los factores de ponderación del método específico utilizado y x_i son los nodos. Por otra parte, $G(x_i)$ se puede aproximar como:

$$G(x_i) = \int_{c(x_i)}^{d(x_i)} f(x_i, y_i)\, dy \approx \sum_{j=1}^{n_i} a_{ij} f(x_i, y_i)$$

para $i = 1, 2, \ldots, n$.