Métodos Numéricos

Ignacio Rimini

Agosto 2024

Índice

1. Sucesiones y series numéricas.

1.1. Sucesiones numéricas.

1.1.1. Definición. Sucesión numérica.

Una sucesión es una función de $\mathbb{N}$ en $\mathbb{R}$, $f : \mathbb{N} \to \mathbb{R}$ que genera una lista ordenada e ilimitada de números:

$$f(1), f(2), f(3), ...f(n), ...$$

Donde $f(n)$ es el término n-ésimo de la sucesión. Es frecuente utilizar notaciones del tipo $\{f(n)\}$, $\{f_n\}$ o $\{f_n\}_{n=1}^{\infty}$ para denotar una sucesión.

Ejemplo de sucesiones.

- $f_n = a : \ a, a, a, a, a....$
- $f_n = n : \ 1, 2, 3, ...$
- $f_n = \frac{1}{n} : \ 1, \frac{1}{2}, \frac{1}{3}, ...$
- $f_n = (-1)^n : \ -1, 1, -1, 1, ...$
- **Sucesión aritmética:** es una sucesión donde cada término se obtiene sumando un número fijo (llamado diferencia común) al término anterior. Ej: $2, 5, 8, 11, 14, ...$ donde la diferencia común es 3.
- **Sucesión geométrica:** es una sucesión donde cada término se obtiene multiplicando el término anterior por un número fijo (llamado razón común). Ej: $3, 9, 27, 81, ...$ donde la razón común es 3.

1.1.2. Definición. Convergencia de sucesiones.

Definición informal. La **convergencia de sucesiones** se refiere al comportamiento de una sucesión a medida que sus términos avanzan hacia el infinito. Decimos que una sucesión $\{f_n\}$ de números reales **converge** a un número real L si, a medida que n (el índice de la sucesión) se hace grande, los términos de la sucesión se acercan cada vez más a L.

Definición formal. Una sucesión $\{f_n\}$ es **convergente** si existe un número real L tal que para cada $\epsilon > 0$ se puede encontrar un número natural N tal que $\forall n \geq N$ se verifica $|f_n - L| < \epsilon$.

Se dice entonces que L es el **límite** de la sucesión $\{f_n\}$, y se escribe $L = \lim_{n\to\infty} f_n$. También decimos que la sucesión $\{f_n\}$ **converge** a L. Una sucesión que no converge se dice que es **divergente.**

Ejemplos.

- La sucesión $\{\frac{1}{n}\}$ converge a 0.
- La sucesión $\{n!\}$ es divergente.
- La sucesión $\{(-1)^n\}$ es divergente.

1.1.3. Teorema. Unicidad del límite en sucesiones convergentes.

Una sucesión convergente tiene uno y sólo un límite.

Demostración.

Sean $a, b \in \mathbb{R}$ tales que $a = \lim_{n\to\infty} f_n$ y $b = \lim_{n\to\infty} f_n$. Supongamos por absurdo que $a \neq b$.

Sin pérdida de generalidad, supongamos que $a < b$. Luego, debe existir $z \in \mathbb{R}$ tal que $a < z < b$.

Puesto que $z < b$ y b es el límite de f_n, debe existir un N' tal que para todo $n > N'$ sea $f_n > z$. Igualmente, puesto que $a < z$ y a es el límite de f_n, debe existir un N'' tal que para todo $n > N''$ sea $f_n < z$.

Tomando $N = max\{N', N''\}$, llegamos a una contradición, ya que se tendría que cumplir $z < f_n < z$ para todo $n > N$.

Esta contradicción surge de suponer que $a \neq b$. Por lo tanto, resulta $a = b$ y luego el límite de una sucesión convergente es único.

1.1.4. Definición. Sucesión acotada.

Una sucesión $\{f_n\}$ se dice que está **acotada superiormente** si existe algún número $c \in \mathbb{R}$ tal que $\forall n \in \mathbb{N}$, se cumple que $f_n \leq c$.

Se dice que una sucesión está **acotada inferiormente** si existe algún numero $k \in \mathbb{R}$ tal que $\forall n \in \mathbb{N}$, se cumple que $k \leq f_n$.

Decimos que una sucesión está **acotada** si lo está superior e inferiormente. Esto equivale a que exista algún núumero $M > 0$ tal que $\forall n \in \mathbb{N}$, se cumple que $|f_n| \leq M$.

Ejemplo.

- La sucesión $\{(-1)^n\}$ esta acotada inferiormente por -1 y superiormente por 1.
- La sucesión $\{\frac{1}{n}\}$ es acotada porque todos los términos están entre 0 y 1.

1.1.5. Definición. Crecimiento y decrecimiento de una sucesión.

- Una sucesión $\{f_n\}$ es **monótona creciente** si $f_n \leq f_{n+1}$ $\forall n \in \mathbb{N}$.
- Una sucesión $\{f_n\}$ es **monótona decreciente** si $f_{n+1} \leq f_n$ $\forall n \in \mathbb{N}$.
- Una sucesión $\{f_n\}$ es **monótona** cuando es creciente o decreciente únicamente.
- Una sucesión $\{f_n\}$ es **estrictamente creciente** si $f_n < f_{n+1}$ $\forall n \in \mathbb{N}$.
- Una sucesión $\{f_n\}$ es **estrictamente creciente** si $f_{n+1} < f_n$ $\forall n \in \mathbb{N}$.

1.1.6. Teorema. Sucesión monótona y acotada siempre converge.

1. Toda sucesión monótona creciente y acotada superiormente es convergente.
2. Toda sucesión monótona decreciente y acotada inferiormente es convergente.

Observación. Esto equivale a decir que una sucesión monótona converge si y sólo si es acotada. Sin embargo, no significa que el límite de convergencia es la cota.

1.1.7. Teorema. Operaciones con límites de sucesiones convergentes.

Sean $\{a_n\}$ y $\{b_n\}$ sucesiones convergentes con límites

$$a = \lim_{n\to\infty} a_n,\ b = \lim_{n\to\infty} b_n$$

Entonces las siguientes reglas se cumplen:

1. **Regla de la suma:** $\lim_{n\to\infty}(a_n + b_n) = a + b$
2. **Regla de la diferencia:** $\lim_{n\to\infty}(a_n - b_n) = a - b$
3. **Regla del producto:** $\lim_{n\to\infty}(a_n b_n) = ab$
4. **Regla de la multiplicación por escalar:** $\lim_{n\to\infty}(cb_n) = cb$, donde $c \in \mathbb{R}$
5. **Regla del cociente:** $\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{a}{n}$, si $b \neq 0$.

1.1.8. Teorema del sandwich para sucesiones.

Sean $\{a_n\}$, $\{b_n\}$ y $\{c_n\}$ sucesiones de números reales. Si $a_n \leq b_n \leq c_n$ para todo n mayor que algún índice N, y además

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} c_n = L$$

Entonces también vale $\lim_{n\to\infty} b_n = L$.

1.1.9. Teorema. Regla de L'Hopital para sucesiones.

Supongamos que $f(x)$ es una función definida para todo $x \geq N$, $x \in \mathbb{R}$ y que $\{f_n\}$ es una sucesión de números reales tal que $f_n = f(n)$ para $n \geq N$. Entonces:

$$\lim_{x\to\infty} f(X) = L \ \Rightarrow\ \lim_{n\to\infty} f_n = L$$

Observación. Es decir, si existe una función de variable real que es igual a los puntos de la sucesión para los mismos $n \in \mathbb{N}$, entonces podemos trabajar con el límite de la función real y ver que si esta función tiene límite, también lo tiene la sucesión.

1.1.10. Definición. Subsucesión.

Una sucesión $\{b_n\}$ es una **subsucesión** de una sucesión $\{a_n\}$ si existe una aplicación $\sigma : \mathbb{N} \to \mathbb{N}$ estrictamente creciente tal que $b_n = a_{\sigma(n)}$ para cada $n \in \mathbb{N}$.

Ejemplos.

- La subsucesión de los números pares se obtiene tomando $\sigma(n) = 2n$:

$$\{a_{2n}\} = a_2, a_4, a_6, a_8, ...a_{2n}, ...$$

- La subsucesión de los números impares se obtiene tomando $\sigma(n) = 2n - 1$:

$$\{a_{2n}\} = a_1, a_3, a_5, a_7, ..., a_{2n-1}, ...$$

1.1.11. Proposición. Sucesión convergente si todas sus subsucesiones son convergentes al mismo límite.

Una sucesión $\{a_n\}$ es **convergente** si y sólo si todas sus subsucesiones son convergentes y convergen a un mismo límite.

Demostración.

- ($\Rightarrow$) Supongamos primero que $\{a_n\}$ es convergente y sea $\{b_n\}$ uns subsucesión de $\{a_n\}$.

  Luego, $b_n = a_{\sigma(n)}$ para alguna función $\sigma_{\mathbb{N}} \to \mathbb{N}$ estrictamente creciente. Vamos a probar que $\{b_n\}$ es convergente.

  Como $\{a_n\}$ es convergente, sea $\epsilon > 0$, entonces existe $N \in \mathbb{N}$ tal que para todo $n > N$ se cumple que:

  $$|a_n - L| < \epsilon$$

  Luego, como σ es una función creciente, existe $M \in \mathbb{N}$ tal que $\sigma(k) > N$ para todo $k > M$. Entonces para todo $k > M$ se cumple:

  $$|a_{\sigma(k)} - L| < \epsilon$$

  Llamando $\sigma(M) = N'$, tenemos que para todo $n > N'$:

  $$|L - b_n| < \epsilon$$

  Por lo tanto, la subsucesión $b_n = a_{\sigma(n)}$ es convergente. Más aún, probamos que todas las subsucesiones convergen a L.

- ($\Leftarrow$) Supongamos ahora que todas las subsucesiones de $\{a_n\}$ son convergentes y convergen a un límite L. Tomando $\sigma : \mathbb{N} \to \mathbb{N} : \sigma(n) = n$, tenemos que la subsucesión $b_n = a_{\sigma(n)} = a_n$ es la sucesión original.

  Luego por hipótesis, la sucesión $\{a_n\}$ es convergente y converge a L.

1.2. Series numéricas.

1.2.1. Definición. Serie numérica.

Dada una sucesión de números reales $\{a_n\}$, se puede formar otra sucesión $\{s_n\}$ para la cual:

$$s_n = a_1 + a_2 + ... + a_n = \sum_{k=1}^{n} a_k$$

La sucesión de las sumas parciales $\{s_n\}$ se llama **serie infinita** o simplemente **serie.** Para representar una serie se utiliza la notación $\sum_{n=1}^{\infty} a_n$.

1.2.2. Definición. Serie positiva y serie alternada.

- Una serie $\sum_{n=1}^{\infty} a_n$ es una **serie de términos positivos** cuando $a_n > 0$ para cada n.
- Una serie $\sum_{n=1}^{\infty} a_n$ es una **serie alternada** cuando $a_n = (-1)^n c_n$, para alguna sucesión $\{c_n\}$ tal que $c_n > 0$ para cada n.

1.2.3. Definición. Convergencia y divergencia de una serie.

Se dice que la serie $\sum_{n=1}^{\infty} = a_n$ **converge** cuando la sucesión de sumas parciales $\{s_n\}$ tiene límite finito. Esto es:

$$\lim_{n\to\infty} \sum_{k=1}^{n} a_k = \lim_{n\to\infty} s_n = s$$

El límite s es la **suma** de la serie. Una serie que no tiene límite finito se dice que es **divergente.**

1.2.4. Teorema. Condición necesaria para la convergencia de una serie.

- Si la serie $\sum_{n=1}^{\infty} a_n$ converge, entonces $\lim_{n\to\infty} a_n = 0$.
- Si $\lim_{n\to\infty} a_n \neq 0$, entonces la serie $\sum_{n=1}^{\infty} a_n$ diverge.

Es decir, si una serie infinita de términos a_n converge, entonces el límite de los términos individuales de la sucesión debe ser cero.

Demostración.

Supongamos que la serie $\sum_{n=1}^{\infty} a_n$ converge, es decir, $\lim_{n\to\infty} s_n = s$. Veamos que $\lim_{n\to\infty} a_n = 0$.

$$a_n = s_n - s_{n-1} \Rightarrow \lim_{n\to\infty} a_n = \lim_{n\to\infty} s_n - \lim_{n\to\infty} s_{n-1} = s - s = 0 \tag{1}$$

Es decir, partimos de que la serie converge y llegamos a que $\lim_{n\to\infty} a_n = 0$. Con lo cual, el teorema queda demostrado.

1.2.5. Teorema. Propiedad de linealidad de una serie.

Sean $\sum_{n=1}^{\infty} a_n$, $\sum_{n=1}^{\infty} b_n$ series convergentes con sumas s' y s'', respectivamente.

Si α, β son constantes, entonces la serie $\sum_{n=1}^{\infty}(\alpha a_n + \beta b_n)$ es convergente con suma $s = \alpha s' + \beta s''$.

1.2.6. Corolario. Convergencia de suma de series donde una diverge.

Si $\sum_{n=1}^{\infty} a_n$ converge y $\sum_{n=1}^{\infty} b_n$ diverge, entonces $\sum_{n=1}^{\infty}(a_n + b_n)$ diverge.

Demostración.

Supongamos por absurdo que $\sum_{n=1}^{\infty}(a_n + b_n)$ es convergente. Luego, dado que también $\sum_{n=1}^{\infty} a_n$ es convergente, por el teorema anterior podríamos escribir:

$$\sum_{n=1}^{\infty}(a_n + b_n) + \sum_{n=1}^{\infty}(-a_n) = \sum_{n=1}^{\infty} b_n$$

Entonces por el teorema anterior $\sum_{n=1}^{\infty} b_n$ sería convergente, lo cual es una contradicción. Absurdo de suponer que $\sum_{n=1}^{\infty}(a_n + b_n)$ es convergente, por lo tanto, diverge.

1.2.7. Definición. Propiedad telescópica de una serie.

La **propiedad telescópica de una serie** es una característica que permite simplificar la suma total al hacer que muchos términos se cancelen entre sí. Una serie se dice **telescópica** cuando, al escribir los términos de manera explícita y sumarlos, la mayoría de los términos se cancela, dejando sólo unos pocos términos que determinan la suma total.

La serie $\sum_{n=1}^{\infty} a_n$ es **telescópica** cuando la podemos representar de la forma $\sum_{n=1}^{\infty}(b_n - b_{n+1})$, es decir, cuando para una sucesión $\{b_n\}$ se cumple $a_n = b_n - b_{n+1}$. En tal caso tenemos:

$$\sum_{k=1}^{n}(b_k - b_{k+1}) = b_1 - b_{n+1}$$

1.2.8. Teorema. Suma de series telescópicas.

Sean $\{a_n\}$ y $\{b_n\}$ dos sucesiones tales que $a_n = b_n - b_{n+1}$, $n = 1, 2, 3, ...$ Entonces la serie $\sum_{n=1}^{\infty} a_n$ es telescópica y converge si y sólo si la sucesión $\{b_n\}$ converge, en cuyo caso vale:

$$\sum_{n=1}^{\infty} a_n = b_1 - L, \text{ donde } L = \lim_{n\to\infty} b_n$$

1.3. Resultados importantes.

1.3.1. Desigualdad de sumas parciales en serie armónica.

La **serie armónica** es $\sum_{n=1}^{\infty} \frac{1}{n}$. Para esta serie tenemos:

$$a_n = \frac{1}{n} \text{ y } s_n = \sum_{k=1}^{n} \frac{1}{k}$$

A continuación obtendremos una desigualdad que se cumple para la serie armónica, y que será utilizada más adelante:

$$\begin{aligned} s_{2n} - s_n = \frac{1}{n+1} + ... + \frac{1}{2n} \geq \frac{1}{2n} + ... + \frac{1}{2n} = \frac{n}{2n} = \frac{1}{2} \\ s_{2n} - s_n \geq \frac{1}{2} \end{aligned} \tag{2}$$

1.3.2. Suma parcial de la serie alternada.

Para la serie $\sum_{n=1}^{\infty}(-1)^n$, tenemos $a_n = (-1)^n$, por lo cual es una serie alternada. Además:

$$s_n = \sum_{k=1}^{n}(-1)^k = -1 + 1 - 1 + ... + (-1)^n = \begin{cases} -1 & si \quad n \text{ } es \text{ } impar \\ 0 & si \quad n \text{ } es \text{ } par \end{cases}$$

1.3.3. Suma parcial de la serie geométrica.

La **serie geométrica** de razón r es:

$$\sum_{n=0}^{\infty} r^n, \text{ con } a_n = r^n$$

Si $r > 0$ se trata de una serie de términos positivos.

Si $r < 0$ se trata de una serie alternada.

Para la serie geométrica se puede obtener una expresión analítica de las sumas parciales:

$$\begin{aligned} (1-r)s_n &= s_n - rs_n \\ &= (1 + r + ... + r^n) - (r + r^2 + ... + r^{n+1}) \\ &= 1 - r^{n+1} \end{aligned} \tag{3}$$

Luego, si $r \neq 1$ tenemos que $s_n = \frac{1-r^{n+1}}{1-r}$

1.3.4. Divergencia de serie armónica.

En un ejemplo anterior hemos visto que para la serie armónica $\sum_{n=1}^{\infty} \frac{1}{n}$ se cumple $\frac{1}{2} \leq s_{2n} - s_n$.

Supongamos que la serie armónica converge, es decir, que existe s tal que $\lim_{n\to\infty} s_n = s$. Luego:

$$\frac{1}{2} \leq \lim_{n\to\infty} s_{2n} - \lim_{n\to\infty} s_n = s - s = 0$$

Llegamos a una contradicción que surge de suponer que la serie armónica converge. Por lo tanto, la serie armónica es **divergente.**

1.3.5. Convergencia de serie geométrica.

Consideremos la serie geométrica (de razón r):

$$\sum_{n=0}^{\infty} r^n = \sum_{n=1}^{\infty} r^{n-1}$$

- Si $|r| \geq 1$ entonces $\lim_{n\to\infty} r^n \neq 0$ y por criterio necesario de convergencia resulta que la serie diverge.
- En la sección de suma parcial de la serie geométrica hemos probado que si $r \neq 1$, entonces $s_n = \frac{1-r^{n+1}}{1-r}$.

  Luego, si $|r| < 1$ tenemos:

  $$s = \lim_{n\to\infty} s_n = \lim_{n\to\infty} \frac{1-r^{n+1}}{1-r} = \frac{1}{1-r}$$

  De donde la última igualdad surge de que $\lim_{n\to\infty} r^{n+1} = 0$ pues $|r| < 1$. Luego, notar que es posible obtener una expresión analítica para la suma de una serie geométrica de razón $|r| < 1$.

1.3.6. Reindexar términos de una serie.

Reindexar los términos en una serie significa cambiar el índice de los términos sin alterar su valor ni el orden de la secuencia. Esto se hace generalmente para simplificar la expresión, cambiar el punto de inicio, o alinearla con otra serie.

- Supongamos que tenemos una sucesión $\{a_n\}$ que comienza con $n = 1$: $a_1, a_2, a_3, ...$ Podemos decidir reindexar la sucesión para que comience con $n = 0$, definiendo una nueva sucesión $\{b_n\}$ donde $b_n = a_{n+1}$. Entonces la sucesión reindexada sería:

  $$b_0 = a_1, b_1 = a_2, b_2 = a_3, ...$$

- Supongamos que tenemos la serie infinita $\sum_{n=1}^{\infty} a_n$.

  Si queremos reindexar esta serie para que empiece en $n = 0$, podemos cambiar el índice de la siguiente manera. Definimos un nuevo índice $m = n - 1$. Entonces, cuando $n = 1$, $m = 0$ y cuando $n = \infty$, m sigue siendo infinito;

  $$\sum_{n=1}^{\infty} a_n = \sum_{m=0}^{\infty} a_{m+1}$$

1.3.7. Adición o supresión de términos.

En una serie, siempre podemos agregar o suprimir un número finito de términos sin alterar su convergencia o divergencia. En el caso de la convergencia, esto suele modificar la suma.

Ejemplo. Si $\sum_{n=1}^{\infty}$ converge, entonces $\sum_{n=k}^{\infty} a_n$ converge para cualquier $k > 1$ y se cumple:

$$\sum_{n=1}^{\infty} a_n = a_1 + a_2 + ... + a_{k-1} + \sum_{n=k}^{\infty} a_n$$

1.4. Criterios de convergencia para series de términos no negativos.

1.4.1. Teorema. Criterio de acotación.

Si $a_n \geq 0$ para cada $n \geq 1$, la serie $\sum_{n=1}^{\infty} a_n$ converge **si y sólo si** la sucesión de sus sumas parciales $\{s_n\}$ está acotada superiormente.

Es decir, si los términos de la sucesión son no negativos, las sumas parciales también lo son, y entonces la serie converge si y sólo si está acotada.

Demostración.

Para cada $n \in \mathbb{N}$ se cumple:

$$s_{n+1} - sn = a_{n+1} \geq 0$$

Es decir, $s_{n+1} \geq s_n$, con lo cual, la sucesión $\{s_n\}$ es monótona creciente. Luego, por el teorema: **Sucesión monótona y acotada siempre converge**, la sucesión $\{s_n\}$ converge si y sólo si está acotada superiormente.

1.4.2. Criterio de comparación.

Si $a_n \geq 0$, $b_n \geq 0$, y existe una constante $c > 0$ tal que $a_n \leq cb_n$, si $n \geq N$, entonces:

- Si $\sum_{n=1}^{\infty} b_n$ converge $\Rightarrow \sum_{n=1}^{\infty} a_n$ converge.
- Si $\sum_{n=1}^{\infty} a_n$ diverge $\Rightarrow \sum_{n=1}^{\infty} b_n$ diverge.

Demostración.

Sean las sumas parciales $s_n = a_1 + ... + a_n$, $t_n = b_1 + ... + b_n$. Entonces $a_n \leq cb_n$ implica $s_n \leq ct_n$.

Si $\sum_{n=1}^{\infty} b_n$ converge, por teorema **Criterio de acotación**, sus sumas parciales están acotadas. Si M es una cota de las sumas parciales t_n, se tiene entonces $s_n \leq cM$, y por lo tanto, las sumas parciales s_n de la sucesión a_n también estan acotadas.

Por lo tanto, por teorema **Criterio de acotación**, la suma $\sum_{n=1}^{\infty} a_n$ también es convergente.

Ejemplo.

Consideremos la serie $\sum_{n=1}^{\infty} \frac{2+cos(n^3)}{2^n+n}$. Podemos escribir la siguiente desigualdad:

$$0 \leq \frac{2+cos(n^3)}{2^n+n} \leq \frac{3}{2^n} = 3(\frac{1}{2})^n$$

Luego, sean $a_n = \frac{2+cos(n^3)}{2^n+n}$ y $b_n = 3(\frac{1}{2})^n$. Entonces se cumple $0 \leq a_n \leq b_n$.

Como $\sum_{n=1}^{\infty} b_n = 3\sum_{n=1}^{\infty}(\frac{1}{2})^n$ es una serie geométrica de razón menor que 1, entonces converge. Luego, por el **criterio de comparación**, esto implica que la serie $\sum_{n=1}^{\infty} a_n$ también converge.

1.4.3. Teorema. Criterio del límite.

Sean $\{a_n\}$, $\{b_n\}$ dos sucesiones tales que $a_n \geq 0$, $b_n > 0$ y sea:

$$\lambda = \lim_{n\to\infty} \frac{a_n}{b_n}$$

Si λ es finito y $\lambda \neq 0$, entonces $\sum_{n=1}^{\infty} b_n$ converge si y sólo si $\sum_{n=1}^{\infty} a_n$ converge.

Es decir las dos series tiene el mismo carácter: ambas convergen o ambas divergen. (Notar que $0 < \lambda < \infty$).

Demostración.

Sea $c \in (\lambda, \infty)$. Entonces existe algún $N \in \mathbb{N}$ tal que $\frac{a_n}{b_n} \leq c$ para todo $n \geq N$, es decir, $0 \leq a_n \leq cb_n$ para todo $n \geq N$.

Por lo tanto, por el **criterio de comparación**, si la serie $\sum_{n=1}^{\infty} b_n$ converge, entonces la serie $\sum_{n=1}^{\infty} a_n$ también converge.

Por otra parte, sea $d \in (0, \lambda)$, entonces existe algún $N' \in \mathbb{N}$ tal que $\frac{a_n}{b_n} \geq d$ para todo $n \geq N'$, es decir, $a_n \geq db_n \geq 0$ para todo $n \geq N'$.

Por lo tanto, por el **criterio de comparación**, si la serie $\sum_{n=1}^{\infty} b_n$ diverge, entonces la serie $\sum_{n=1}^{\infty} a_n$ también diverge.

1.4.4. Teorema. Criterio de la raíz.

Sea $\{a_n\}$ una sucesión tal que $a_n \geq 0$ y sea:

$$\alpha = \lim_{n\to\infty} \sqrt[n]{a_n}$$

Entonces:

1. Si $\alpha < 1$, la serie $\sum_{n=1}^{\infty} a_n$ converge.
2. Si $\alpha > 1$, la serie $\sum_{n=1}^{\infty} a_n$ diverge.
3. Si $\alpha = 1$, el criterio no decide.

Demostración.

1. Si $\alpha < 1$, elegimos L tal que $\alpha < L < 1$. Entonces $0 \leq \sqrt[n]{a_n} \leq L$ para todo $n \geq N$.

   Por lo tanto, $a_n \leq L^n$ para todo $n \geq N$. Sabemos que la serie $\sum_{n=1}^{\infty} L^n$ es convergente, por ser una serie geométrica con $L < 1$.

   Luego, por el **criterio de la comparación**, la serie $\sum_{n=1}^{\infty} a_n$ también converge.

2. Como $\alpha > 1$, entonces $a_n > 1$ para una infinidad de valores de n y por lo tanto a_n no puede tender a 0. Por lo cual, no se cumple la condición necesaria de convergencia y por lo tanto la serie $\sum_{n=1}^{\infty} a_n$ diverge.

3. Consideremos dos ejemplos en los que $a_n = \frac{1}{n}$ y $a_n = \frac{1}{n^2}$. En ambos casos $\lim_{n\to\infty} \sqrt[n]{a_n} = \alpha = 1$.

   Sin embargo, $\sum_{n=1}^{\infty} \frac{1}{n}$ diverge, mientras que $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converge. Por lo tanto, si $\alpha = 1$ el criterio no dice nada.

1.4.5. Teorema. Criterio del cociente.

Sea $\{a_n\}$ una sucesión tal que $a_n > 0$ y sea $L = \lim_{n\to\infty} \frac{a_{n+1}}{a_n}$. Entonces:

1. Si $L < 1$, la serie $\sum_{n=1}^{\infty} a_n$ converge.
2. Si $L > 1$, la serie $\sum_{n=1}^{\infty} a_n$ diverge.
3. Si $L = 1$, el criterio no decide.

Demostración.

1. Si $L < 1$, elegimos z tal que $L < z < 1$. Entonces ha de existir un N tal que $\frac{a_{n+1}}{a_n} < z$ para todo $n \geq N$. Luego:

$$\frac{a_{n+1}}{a_n} < z = \frac{z^{n+1}}{z^n} \Rightarrow \frac{a_{n+1}}{z^{n+1}} < \frac{a_n}{z_n}, \forall n \geq N$$

   Es decir, la sucesión $\{\frac{a_n}{z_n}\}$ es decreciente para $n \geq N$. En particular, $\frac{a_n}{z^n} \leq \frac{a_N}{z^N}$ para $n \geq N$, o de otro modo $a_n \leq cz^n$, donde $c = \frac{a_N}{z^N}$.

   Como la serie $\sum_{n=1}^{\infty} z^n$ converge por ser la serie geométrica de razón $z < 1$, entonces por el **criterio de comaración**, $\sum_{n=1}^{\infty} a_n$ converge.

2. Si $L > 1$ eso implica que $a_{n+1} > a_n$ para todo $n \geq N$, y por lo tanto a_n no tiende a 0.

   Luego, por no cumplir la **condición necesaria**, $\sum_{n=1}^{\infty} a_n$ diverge.

3. Se pueden tomar los mismos ejemplos que en el **criterio de la raíz.**

1.4.6. Teorema. Criterio de la integral.

Sea f una función positiva y estrictamente decreciente definida en $[1, +\infty)$ tal que $f(n) = a_n$ para todo $n \in \mathbb{N}$.

La serie $\sum_{n=1}^{\infty} a_n$ converge si y sólo si la integral $\int_1^{+\infty} f(x)dx$ converge.

Ejemplo.

Consideremos la serie $\sum_{n=1}^{\infty} \frac{1}{n^2}$.

Sea $f(x) = \frac{1}{x^2}$, entonces f es positiva para todo x. Además, como $f'(x) = -2x^{-3} < 0$ si $x > 0$, entonces por análisis de monotonía por derivada, f es estrictamente decreciente en $[1, \infty)$. Luego veamos la integral:

$$\int_1^{+\infty} \frac{1}{x^2}dx = \lim_{b\to\infty} \int_1^b \frac{1}{x^2}dx = \lim_{b\to\infty}(\frac{-1}{b} - \frac{-1}{1}) = \lim_{b\to\infty} 1 - \frac{1}{b} = 1$$

Entonces $\int_1^{+\infty} \frac{1}{x^2}dx$ converge y por lo tanto la serie $\sum_{n=1}^{\infty} \frac{1}{n^2}$ también converge.

1.5. Criterios de convergencia para series alternadas.

Las series alternadas son de la forma:

$$\sum_{n=1}^{\infty}(-1)^{n-1}a_n = a_1 - a_2 + a_3 - a_4...$$

donde cada $a_n > 0$.

1.5.1. Teorema. Criterio de Leibniz.

Si $\{a_n\}$ es una sucesión monótona decreciente con límite 0, la serie alternada $\sum_{n=1}^{\infty}(-1)^{n-1}a_n$ converge.

Demostración.

2. Errores numéricos.

2.1. Sistemas de numeración posicionales.

Los sistemas de numeración son **posicionales** cuando el valor de cada dígito del número depende de la posición en la que se encuentra. Ejemplos de sistemas posicionales: *binario, decimal, octal y hexadecimal.* Un ejemplo de sistema de numeración no posicional es el *sistema romano.*

El número de símbolos permitidos en un sistema de numeración posicional se conoce como **base** del sistema de numeración. Si un sistema de numeración posicional tiene base β significa que disponemos de β símbolos diferentes para escribir los números.

Tabla 1: Sistemas de numeración posicionales

Sistema	Base	Cifras que utiliza
Binario	2	0, 1
Ternario	3	0, 1, 2
Cuaternario	4	0, 1, 2, 3
Quinario	5	0, 1, 2, 3, 4
Senario	6	0, 1, 2, 3, 4, 5
Septario o Hectal	7	0, 1, 2, 3, 4, 5, 6
Octal	8	0, 1, 2, 3, 4, 5, 6, 7
Nonario	9	0, 1, 2, 3, 4, 5, 6, 7, 8
Decimal	10	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Undecimal	11	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A
...	...	...
Hexadecimal	16	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F

2.1.1. Escritura de un número en sistema de numeración posicional.

En general, un número con parte entera finita se representa con la base β como:

$$(-1)^{\sigma}(a_n a_{n-1}...a_1 a_0.a_{-1}a_{-2})_{\beta}$$

Donde los coeficientes a_i son los valores (posición) de los dígitos en el sistema con base β, es decir, enteros positivos tales que $0 \leq a_i \leq \beta - 1$, y σ es una variable binaria que representa el signo del número ($\sigma = 0$ si el número es positivo y $\sigma = 1$ si es negativo).

La fórmula general para construir un número real x en un sistema de numeración posicional de base β es la siguiente:

$$x = (-1)^{\sigma}(a_n\beta^n + a_{n-1}\beta^{n-1} + ... + a_1\beta^1 + a_0\beta^0 + a_{-1}\beta^{-1} + a_{-2}\beta^{-2} + ...)$$

Ejemplo. El número 213,58 en sistema decimal puede construirse como:

$$213,58 = (-1)^0 \times (2 \times 10^2 + 1 \times 10^1 + 3 \times 10^0 + 5 \times 10^{-1} + 8 \times 10^{-2})$$

2.2. Sistema binario.

En el **sistema binario**, de base $\beta = 2$, los números se representan utilizando solamente dos cifras: cero (0) y uno (1). Un dígito binario, o **bit**, puede representar uno de estos dos valores.

Un número binario se representa utilizando el subíndice $\beta = 2$, como por ejemplo: $(10101,1101)_{\beta}$.

2.2.1. Conversión de binario a decimal.

Para la conversión de binario a decimal empleamos la fórmula general de números en sistemas posicionales con la base $\beta = 2$.

Por ejemplo, $(10101, 1101)_2$ es la representación binario del número $21,8125$, puesto que:

$$(10101, 1101)_2 = 2^4 + 2^2 + 2^0 + 2^{-1} + 2^{-2} + 2^{-4} = 21,8125$$

Ejemplos.

- **Entero binario con m unos.**

  $$x = (111..,1)_2 = 2^{m-1} + 2^{m-2} + ... + 2^1 + 2^0 = \sum_{k=0}^{m-1} 2^k$$

  Lo cual representa una suma parcial de una serie geométrica de razón 2. Luego, como conocemos el valor de esta suma parcial tenemos que $x = \frac{1-2^m}{1-2} = 2^m - 1$.

- **Binario periódico (0,01010101...)**

  $$x = (0, 01010101...)_2 = 2^{-2} + 2^{-4} + 2^{-6}... = 2^{-2}(1 + 2^{-2} + 2^{-4} + ...) = \frac{1}{4}\sum_{n=0}^{\infty}(\frac{1}{4})^n$$

  Lo cual representa una serie geométrica de razón $r = \frac{1}{4} < 1$. Sabemos que dicha serie es convergente, y conocemos la expresión de su suma:

  $$x = \frac{1}{4}\sum_{n=0}^{\infty}(\frac{1}{4})^n = \frac{1}{4}\frac{1}{1-\frac{1}{4}} = \frac{1}{4-1} = \frac{1}{3} = 0,333...$$

- **Binario periódico (0,1100110011...)**

$$\begin{aligned} x = (0, 1100110011...)_2 &= 2^{-1} + 2^{-2} + 2^{-5} + 2^{-6}... \\ &= \sum_{n=1}^{\infty}(\frac{1}{2})^{4n-2} + \sum_{n=1}^{\infty}(\frac{1}{2})^{4n-3} \\ &= \sum_{n=1}^{\infty}((\frac{1}{2})^4)^n \cdot (\frac{1}{2})^{-2} + \sum_{n=1}^{\infty}((\frac{1}{2})^4)^n \cdot (\frac{1}{2})^{-3} \\ &= \sum_{n=1}^{\infty}(\frac{1}{16})^n \cdot 4 + \sum_{n=1}^{\infty}(\frac{1}{16})^n \cdot 8 \\ &= 4 \cdot \sum_{n=1}^{\infty}(\frac{1}{16})^n + 8 \cdot \sum_{n=1}^{\infty}(\frac{1}{16})^n \\ &= 12\frac{\frac{1}{16}}{1-\frac{1}{16}} = \frac{12}{15} = 0,8 \end{aligned} \tag{4}$$

2.2.2. Conversión de decimal a binario.

Un número entero x se convierte a binario dividiendo sucesivamente por dos hasta que el cociente sea 0, y registrando el valor de los restos, de acuerdo al siguiente procedimiento:

- Dividir x por 2, llamar al cociente x_1, y al resto a_0.
- Dividir x_1 por 2, llamar al cociente x_2, y al resto a_1.
- Dividir x_2 por 2, llamar al cociente x_3, y al resto a_2.

Un número fraccionario x se convierte a binario multiplicando sucesivamente por dos, y regristrando la parte entera del número resultante, de acuerdo al siguiente procedimiento:

- Multiplicar x por 2. La parte entera es a_{-1} y la parte fraccionaria es x_1.
- Multiplicar x_1 por 2. La parte entera es a_2 y la parte fraccionaria es x_2.
- Multiplicar x_2 por 2. La parte entera es a_{-3} y la parte fraccionaria es x_3.

Ejemplos.

- Para convertir el número $(11)_{10}$ a binario, tenemos:

$$11/2 = 5 + 1 \Rightarrow a_0 = 1$$
$$5/2 = 2 + 1 \Rightarrow a_1 = 1$$
$$2/2 = 1 + 0 \Rightarrow a_2 = 0$$
$$1/2 = 0 + 1 \Rightarrow a_3 = 1$$

  Luego, $(11)_{10} = (1011)_2$.

- Para convertir el número $x = (0, 2)_{10}$ a binario, tenemos:

$$0,2 \cdot 2 = 0,4 \Rightarrow a_{-1} = 0$$
$$0,4 \cdot 2 = 0,8 \Rightarrow a_{-2} = 0$$
$$0,8 \cdot 2 = 1,6 \Rightarrow a_{-3} = 1$$
$$0,6 \cdot 2 = 1,2 \Rightarrow a_{-4} = 1$$

  Luego, $(0, 2)_{10} = (0, 001100110011...)_2$. Notar que obtenemos una fracción binaria periódica.

2.3. Representación computacional de números en punto flotante.

2.3.1. Representación general.

Sea β la base del sistema de numeración empleado en la computadora. La mayoría de las computadoras emplean $\beta = 2$, o también $\beta = 8$ o 16. Un número x se representa en la computadora como un número en punto flotante, $fl(x)$ de la forma:

$$fl(x) = (-1)^{\sigma}(, a_1\ a_2\ ...\ a_n)_{\beta} \times \beta^{E-s}$$

- **Mantisa.** La mantisa m es la parte fraccional del número, definida por los n dígitos a_i, como:

$$m = (, a_1\ a_2\ ...\ a_n) = \frac{a_1}{b^1} + \frac{a_2}{\beta^2} + ... + \frac{a_n}{\beta^n}$$

- **Exponente.** El exponente E es un número entero positivo, definido por los t dígitos c_j como:

$$E = (c_1\ c_2\ ...\ c_t)_{\beta} = c_1\beta^1 + c_2\beta^2 + ... + c_t\beta^t$$

  El uso exclusivo de enteros positivos para el exponente no permitiría una representación adecuado de números con magnitud pequeña. Para garantizar que estos números también sean representables, se resta el **sesgo** s del exponente, el cual es una constante para una representación dada.

- **Signo.** El signo del número está definido por la variable binaria σ, introducida previamente, de forma que $\sigma = 0$ si el número es positivo y $\sigma = 1$ si el número es negativo.

Observación.

El cero no puede ser representado como punto flotante normalizado y se representa como caso particular.

Para una representación dada, existen límites en los exponentes que se pueden representar:

$$L \leq E - s \leq U$$

Con $L < 0$ y $U > 0$. Si el exponente de un número x viola la cota inferior, es decir, si $E - s < L$, ocurre un desbordamiento a cero o **underflow.** En este caso, se toma $fl(x) = 0$ y los cálculos continuan.

Si el exponente de x viola la cota superior, es decir, si $E - s > U$, luego x no se puede representar como $fl(x)$ y ocurre un desbordamiento u **overflow**. Esto representa un error fatal y el cálculo (programa) se interrumpe.

2.3.2. Norma IEEE para números en punto flotante.

(página 7)

2.3.3. Truncamiento y redondeo.

La mayoría de los números reales no se pueden representar en forma exacta en la representación en punto flotante introducida previamente. Por lo tanto deben aproximarse por un número cercano que sea representable.

Dado un número real arbitrario x, existen dos maneras principales de generar $fl(x)$ a partir de x: el truncamiento y el redondeo.

Cualquier número real x se puede escribir como:

$$x = (-1)^{\sigma}(, a_1\ a_2\ ...\ a_n\ a_{n+1}\ ...)_{\beta} \times \beta^{E-s},\ a_1 \neq 0$$

Truncamiento. Este método para conseguir representar un número x en $fl(x)$ consiste en cortar los números $a_{n+1}, a_{n+2}, ...$ Es decir, si la mantisa tiene n bits, para representar x utilizamos los primeros n bits y los restantes los cortamos.

$$fl(x) = (-1)^{\sigma}(, a_1\ a_2\ ...\ a_n)_{\beta} \times \beta^{E-s}$$

Redondeo. En el caso de un número redondeado, tenemos que si la mantisa tiene n dígitos, debemos redondear el dígito a_n dependiendo del valor de a_{n+1}.

Si a_{n+1} tiene valor 0, entonces el redondeo funciona igual que el truncamiento: se dejan los n primeros bits tal cual como están en el número original. En cambio si a_{n+1} tiene valor 1, debemos sumar el número fraccionario de todos 0 con un bit 1 en la posición a_n al número original.

$$fl(x) = \begin{cases} (-1)^{\sigma}(, a_1\ a_2\ ...\ a_{n-1}\ a_n)_{\beta} \times \beta^{E-s} & 0 \leq a_{n+1} \leq \frac{\beta}{2} \\ (-1)^{\sigma}[(, a_1\ a_2\ ...\ a_{n-1}\ a_n)_{\beta} + (, 0\ 0\ ...\ 0\ 1)_{\beta}] \times \beta^{E-s} & \frac{\beta}{2} \leq a_{n+1} < \beta \end{cases}$$

2.4. Medidas de precisión de la representación en punto flotante.

2.4.1. Épsilon de máquina.

El **épsilon de máquina** es el número positivo más pequeño que, sumado a 1 en la representación numérica de la máquina, produce un resultado distinto de 1. En otras palabras, es el número más pequeño ϵ_m tal que:

$$1 + \epsilon_m > 1$$

Empleando la norma IEEE se tiene que:

- $1 = (1, 0\ 0\ ...\ 0\ 0)_2 \times 2^0$
- $y = (1, 0\ 0\ ...\ 0\ 1)_2 \times 2^0 = 1 + 2^{-n} > 1$

Luego, $\epsilon_m = 2^{-n}$. En precisión simple $\epsilon_m = 2^{-23} \approx 1,19 \times 10^{-7}$.

2.4.2. Unidad de redondeo.

La **unidad de redondeo** de un computador es un número δ que satisface:

1. Es un número positivo en punto flotante.
2. Es el menor número tal que $fl(1 + \delta) > 1$.

Observación.

Esto significa que $1+\delta$ puede representarse con suficiente precisión para que la máquina lo distinga de 1. Si tomamos un número más pequeño $\delta' < \delta$, entonces ocurre lo siguiente: $fl(1 + \delta') = 1$.

Esto significa que, aunque estamos sumando un número pequeño δ' a 1, en la aritmética de punto flotante, el resultado sigue siendo exactamente igual a 1. La máquina no puede distinguir entre 1 y $1 + \delta'$ debido a la limitación de la precisión.

Es decir, δ es el límite de cuánto podemos sumar a 1 antes de que la máquina deje de reconocer el número como 1. Es por esto que δ mide el **ancho del cero**, esto se refiere a cuánto debemos sumar a 1 para que la máquina lo perciba como un número distinto de 1.

Valor de la unidad de redondeo.

En el caso de la norma IEEE 754, la posición $n+1$ en la mantisa del número 1 es donde se produce el primer cambio significativo que afectará el resultado. Si en esa posición se añade un 1, el número 1 más el número fraccionario de todos 0 con un 1 en la posición a_{n+1} se vuelve mayor que 1 al redondearlo:

$$1 + 2^{-n-1}$$

Este es el número más pequeño que, sumado a 1, es mayor que 1. Por lo tanto, $\delta = 2^{-n-1}$. En precisión simple, $\delta \approx 5,96 \times 10^{-8}$.

2.4.3. Mayor entero positivo representable en forma exacta.

Otra medida de precisión relacionada con el número de bits del significante consiste en hallar el mayor entero M tal que todo entero x que satisface $0 \leq x \leq M$ se puede representar en forma exacta en punto flotante.

Es decir, se trata de hallar $M \in \mathbb{Z}^+$ tal que:

1. $0 < x \leq M, x \in \mathbb{Z}^+$ implica $fl(x) = x$.

2. $fl(M+1) \neq M+1$

Valor del mayor entero positivo.

En simple precisión, la mantisa tiene 24 bits (esto incluye los 23 bits quue se almacenan explícitamente, más un 1 que no se almacena pero se asume implícitamente en la representación). Estos 24 bits determinan cuántos dígitos significativos podemos tener en la representación binaria.

Si usamos todos los 24 bits de la mantisa (el mayor número posible con estos bits), obtenemos el valor máximo entero que puede representarse exactamente.

El mayor número que podemos representar con 24 bits es el número en el que todos los bits de la mantisa son 1. En binario, esto se escribe como:

$$(1,1\ 1\ ...\ 1)_2 \times 2^{23}$$

La parte 2^{23} indica que estamos moviendo el punto decimal 23 posiciones a la izquierda.

Si convertimos este número a decimal vemos que $(1,1\ 1\ ...\ 1)_2 \times 2^{23} = (1\ 1\ ...\ 1)_2$ con 23 unos.

Luego, en decimal: $2^{23} + 2^{22} + 2^{21} + ... + 2^1 + 2^0 = \sum_{i=0}^{23} 2^i = 2^{24} - 1$.

Por lo tanto, el mayor número entero que se puede representar en precisión simple es:

$$M = 2^{24} - 1 = 2^{n+1} - 1 = 16{,}777{,}216$$

2.5. Errores numéricos.

2.5.1. Error absoluto y relativo.

Al resolver un problema, buscamos obtener la solución exacta o verdadera, que denotamos x_v. Sin embargo, aplicando métodos numéricos se obtiene por lo general una solución aproximada que llamamos x_a. Definimos el **error** en x_a como:

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

Y definimos el **error absoluto** y el **error relativo** en x_a como:

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

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

2.5.2. Error de truncamiento y redondeo.

Si $x \neq fl(x)$ y se utiliza truncamiento, luego $fl(x) < x$ y el error $x - fl(x)$ es siempre positivo. Esto trae consecuencias en el cálculo numérico, ya que no hay posibilidad de cancelación de errores y la propagación de errores es mayor.

Con el redondeo, el error $x - fl(x)$ es negativo para la mitad de los valores de x y positivo para la otra mitad de los valores posibles de x. Además, el peor error posible por redondeo es la mitad que en el caso de truncamiento.

A menudo se representa el error relativo como:

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

De donde $fl(x) = (1+\epsilon)x$. Esto puede verse como que $fl(x)$ es un valor perturbado de x.

2.5.3. Proposición. Cotas para error relativo de truncamiento y redondeo.

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 dadas en la sección 2.3.3, respectivamente.

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

Demostración.

1. Supondremos $\sigma = 0$ (el caso de $\sigma = 1$ no cambia el signo de ϵ). En el caso de truncamiento tenemos:

$$x - fl(x) = (,0\ 0\ ...\ 0\ a_{n+1}\ a_{n+2}\ ...)_\beta \times \beta^e \text{ con } e = E - s$$

Luego, declaramos $\gamma = \beta - 1$, es decir, γ es el mayor dígito posible para una base dada (en binario, $\gamma = 1$). Por lo tanto, el número $(,0\ 0\ ...\ a_{n+1}\ a_{n+2}\ ...) \leq (,0\ 0\ ...\ \gamma\ \gamma\ ...)$. Luego tenemos:

$$\begin{aligned} 0 \leq x - fl(x) \leq (,0\ 0\ ...\ 0\ \gamma\ \gamma\ ...)_\beta \times \beta^e &= \gamma \left[\sum_{i=1}^{\infty} \frac{1}{\beta^{n+i}}\right] \beta^e \\ &= \frac{\gamma}{\beta^n} \left[\sum_{i=1}^{\infty} (\frac{1}{\beta})^i\right] \beta^e \ (serie\ geometrica) \\ &= \frac{\gamma}{\beta^n}\ \frac{\frac{1}{\beta}}{1 - \frac{1}{\beta}} \beta^e \\ &= \frac{\gamma}{\beta^n}\ \frac{\frac{1}{\beta}}{\frac{\beta-1}{\beta}} \beta^e \\ &= \frac{\gamma}{\beta^n}\ \frac{\frac{\beta}{\beta}}{\beta - 1} \beta^e \\ &= \frac{\gamma}{\beta^n}\ \frac{1}{\gamma} \beta^e \\ &= \frac{\beta^e}{\beta^n} = \beta^{-n+e} \end{aligned} \tag{5}$$

Dividiendo por x la desigualdad anterior, tenemos:

$$0 \leq \frac{x - fl(x)}{x} \leq \frac{\beta^{-n+e}}{(,a_1\ a_2\ ...)_\beta \times \beta^e} \Rightarrow 0 \leq -\epsilon \leq \frac{\beta^{-n}}{(,1\ 0\ 0\ 0\ ...)_\beta} = \beta^{-n+1} \tag{6}$$

Y luego, $-\beta^{-n+1} \leq \epsilon \leq 0$ al dar vuelta las desigualdades por multiplicar por (-1).

2. Demostrar.

2.6. Cifras significativas.

En un trabajo científico se considera que las cifras significativas de un número son aquellas que tienen un significado real o aportan alguna información. Las cifras siggnificativas de un número vienen determinadas por su incertidumbre.

Por ejemplo, si expresamos un número con un error del orden de décimas, es evidente que todas las cifras del número que ocupan una posición menor que las décimas no aportan ninguna información.

No tiene sentido dar el número con una exactitud de dies milésimas si afirmamos que el error es de casi un metro. Cuando se expresa un número debe evitarse siempre la utilización de cifras no significativas.

2.6.1. Cifras significativas de un número.

Para conocer el número de cifras significativas de un número decimal, se siguen las siguientes reglas:

- **Dígitos no nulos:** todos los dígitos diferentes de cero son siempre cifras significativas.

  Ej: en el número 457 hay 3 cifras significativas.

- **Ceros entre dígitos significativos:** los ceros situados entre dígitos no nulos son cifras significativas.

  Ej: en el número 4067 hay 4 cifras significativas.

- **Ceros a la izquierda de los dígitos no nulos:** los ceros que están a la izquierda de un dígito no nulo **NO** cuentan como cifras significativas, porque solo sirven para ubicar el punto decimal.

  Ej: en el número 0.0056, solo 56 son cifras significativas.

- **Ceros a la derecha de un dígito no nulo: si hay un punto decimal,** los ceros a la derecha de un dígito no nulo son cifras significativas.

  Ej: en el número 45.00 hay 4 cifras significativas, porque los ceros indican precisión.

  Si no hay punto decimal, los ceros a la derecha de los dígitos no nulos pueden o no ser significativos, dependiendo del contexto.

  Ej: en el número 5000 puede haber 1,2,3 o 4 cifras significativas, dependiendo de cómo se exprese el número.

2.6.2. Cifras significativas de un número aproximado.

Sea x_v el valor verdadero de un número y x_a un valor aproximado. Decimos que x_a tiene m cifras significativas con respecto a x_v si el error absoluto $|x_v - x_a|$ tiene una magnitud menor o igual a cinco unidades en el dígito $(m+1)$ de x_v contando de izquierda a derecha desde el primer dígito distinto de cero en x_v.

Es decir, si $x_v = a_1\ a_2\ a_3 ... a_m\ a_{m+1} ...$ donde $a_1, a_2, ... a_m$ son los primeros m dígitos significativos de x_v, decimos que x_a tiene **m cifras significativas** con respecto a x_v si el error $|x_v - x_a|$ es menor o igual a 5 unidades en la posición $m+1$. Esto significa que el valor aproximado x_a es preciso hasta el dígito m sin un error que afecte significativamente al dígito $m+1$.

Ejemplos.

- $x_v = 0,02144 \quad x_a = 0,02138 \Rightarrow |x_v - x_a| = 0,00006.$

  Decimos que x_a tiene dos cifras significativas (y no tres) respecto a x_v pues tomando las cifras 44 y 38, hay una diferencia de 6.

- $x_v = 23,496 \quad x_a = 23,494 \Rightarrow |x_v - x_a| = 0,002.$

  Decimos que x_a tiene cuatro cifras significativas con respecto a x_v.

Generalización usando error relativo.

Para medir el número de cifras significativas de un valor aproximado se suele emplear la siguiente desigualdad. Si

$$\left|\frac{x_v - x_a}{x_v}\right| \leq 5 \times 10^{-m-1}$$

Entonces decimos que x_a tiene m cifras significativas con respecto a x_v.

Notar que esta desigualdad es una condición suficiente pero no necesaria para que x_a tenga m cifras significativas con respecto a x_v. Hay ejemplos que tienen un mayor número de cifras significativas que las indicadas por esta condición.

2.6.3. Redondeo a m cifras significativas.

Redondear un número decimal x a m cifras significativas es equivalente a redondear el número utilizando en notación de punto flotante una mantisa de m dígitos.

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 $\hat{x}$ con m dígitos después de la coma.

El número redondeado es $rn(x) = \bar{x} \times 10^E$ con $\bar{x} = 0, a_1\ a_2\ ...\ a_m$. Puesto que $a_1 \neq 0$ y todos los dígitos se encuentran después de la coma, $rn(x)$ tiene m cifras significativas.

Además, el valor aproximado que se obtiene $x_a = rn(x)$ tiene m cifras significativas con respecto al valor original $x_v = x$, puesto que al redondear un número se cumple la definición de la diferencia menor a 5.

Ejemplos.

- Redondeo a 5 cifras significativas.

  $x_v = 1,123456 \quad x_a = 1,1235 \Rightarrow |x_v - x_a| = 0{,}000044$

  Luego x_a tiene cinco cifras significativas con respecto a x_v.

- Redondeo con 2 cifras significativas.

  $x_v = 0,20004 \quad x_a = 0,20 \Rightarrow |x_v - x_a| = 0{,}00004$

  Luego x_a tiene dos cifras significativas (y no cuatro porque los ceros no cuentan) con respecto a x_v.

- Redondeo con 4 cifras significativas.

  $x_v = 0,20005 \quad x_a = 0,2001 \Rightarrow |x_v - x_a| = 0{,}00005$

  Luego x_a tiene cuatro cifras significativas con respecto a x_v.

2.7. Propagación de errores.

Sea ω una de las operaciones aritméticas (suma, resta, multiplicación, división) y sea $\hat{\omega}$ la versión computacional de la misma operación, la cual incluye redondeo o truncamiento. Sean x_a e y_a números usados en los cálculos, y suponga que ya presentan error, siendo sus valores verdaderos:

$$x_v = x_a + \epsilon \qquad y_v = y_a + \eta$$

Luego, $x_a \hat{\omega} y_a$ es el número calculado, y su error está dado por:

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

Donde la primera cantidad entre corchetes es llamada **error propagado**, mientras que la segunda cantidad es el error de redondeo o truncamiento. Supondremos en los sucesivos pasos que se emplea redondeo. Para esta segunda cantidad, usualmente tendremos $x_a \hat{\omega} y_a = fl(x_a \omega y_a)$.

Lo cual significa que $x_a \omega y_a$ se calcula con exactitud y luego se redondea. Aplicando la cota del error relativo por redondeo tenemos:

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

2.7.1. Error propagado en la multiplicación.

Para el error $x_a y_a$ tenemos:

$$\begin{aligned} x_v y_V - x_a y_a &= x_v y_v - (x_v - \epsilon)(y_v - \eta) \\ &= x_v y_v - x_v y_v + x_v \eta + y_v \epsilon - \epsilon \eta \\ &= x_v \eta + y_v \epsilon - \epsilon \eta \end{aligned} \tag{7}$$

Definiendo el error relativo, $Rel(x_a) = \frac{\epsilon}{x_v}$, tenemos:

$$\begin{aligned} Rel(x_a y_a) &= \frac{\epsilon}{x_v y_v} \\ &= \frac{x_v y_v - x_a y_a}{x_v y_v} \\ &= \frac{x_v \eta + y_v \epsilon - \epsilon \eta}{x_v y_v} \\ &= \frac{\eta}{y_v} + \frac{\epsilon}{x_v} - \frac{\epsilon}{x_v} \frac{\eta}{y_v} \\ &= Rel(y_a) + Rel(x_a) - Rel(x_a) Rel(y_a) \end{aligned} \tag{8}$$

Tomando $|Rel(x_a)|$, $|Rel(y_a)|$ mucho menores que 1, entonces:

$$Rel(x_a y_a) \approx Rel(x_a) + Rel(y_a)$$

2.7.2. Error propagado en la división.

3. Resolución de Ecuaciones No Lineales.

3.1. Algoritmos y Convergencia.

3.1.1. Definición. Algoritmo.

Un **algoritmo**, para resolver un problema matemático, es un proceso iterativo que genera una sucesión de números (o puntos) de acuerdo a un conjunto de instrucciones precisas, junto con un criterio de parada.

Si comenzamos con un punto inicial $x_0 \in \mathbb{R}^n$, cada iteración del algoritmo produce un nuevo punto $x_1, x_2, x_3, ...$ y así sucesivamente. Esta secuencia busca aproximarse a una solución del problema y a este proceso se lo llama **mapa algorítmico.**

El concepto de **mapa algorítmico** $\mathcal{A}$ describe el proceso que transforma un punto x_k en el siguiente punto x_{k+1}. Este mapa es simplemente una función que toma como entrada un punto x_k y genera un nuevo punto x_{k+1}. El proceso de aplicar esta transformación en cada iteración se expresa como:

$$x_{k+1} \in A(x_k)$$

La transformación de x_k a x_{k+1} constituye una **iteración** del algoritmo. Si el algoritmo está bien diseñado, la secuencia de puntos $x_0, x_1, x_2, ...$ debería **converger** a una solución x^*.

En el caso general, el mapa algorítmico $\mathcal{A}$ puede ser **punto a conjunto**. Esto significa que, dado un punto x_k, el mapa puede generar un conjunto de puntos como posibles opciones para x_{k+1}. Si el mapa es **punto a punto**, significa que $A(x_k)$ genera un único punto x_{k+1} para cada x_k. En este último caso escribimos $x_{k+1} = A(x_k)$.

3.1.2. Criterios de parada.

A continuación se describen alunos criterios de parada que pueden aplicarse en algún paso de un algoritmo. Se elige una tolerancia $\epsilon > 0$ y se realzian nuevas iteraciones del algoritmo hasta que se cumpla una de las siguientes condiciones:

- **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:** $| f(x_k) | < \epsilon$

Al usar cualquiera de estos criterios de parada pueden surgir problemas. Por ejemplo, existen sucesiones $\{x_k\}$ con la propiedad de que las diferencias $x_k - x_{k-1}$ convergen a cero, mientras que la sucesión diverge.

Este es el caso de la sucesión de sumas parciales de la serie armónica, dada por $x_k = \sum_{n=1}^{k} \frac{1}{n}$. Es conocido que la serie armónica diverge, aún cuando se tiene que $\lim_{k\to\infty}(x_k - x_{k-1}) = \lim_{k\to\infty} \frac{1}{k} = 0$.

También es posible que $f(x_k)$ sea cercano a cero, mientras que x_k difiere significativamente de la raíz α de la función f.

Al programar un algoritmo iterativo, además de incluir como criterio de parada una de las tolerancias del error vistas, se debe parar el programa cuando se supere un máximo número de iteraciones. Esto es así debido a que el algoritmo puede no converger.

3.1.3. Definición. Orden de convergencia.

El **orden de convergencia** de un algoritmo describe **cuán rápidamente** se caerca una secuencia de aproximaciones a la solucion exacta conforme el número de iteraciones aumenta. Es una medida del **ritmo de convergencia** de un método iterativo, y es especialmente útil cuando se estudian algoritmos numéricos.

Dado un algoritmo iterativo que genera una secuencia de aproximaciones $x_1, x_2, x_3, ...$ a la solución exacta x^*, se dice que la secuencia tiene un orden de convergencia p si existe una constante $0 < C < \infty$ tal que:

$$\lim_{k\to\infty} \frac{|x_{k+1}-x^*|}{|x_k-x^*|} = \mathrm{C}$$

- x_k es la k-ésima aproximación de la solución.
- x^* es la solución exacta.
- p es el orden de convergencia. Este valor describe cuántas veces más rápido se acerca la secuencia a x^* conforme las iteraciones avanzan.
- C es una constante positiva.

Tipos de convergencia.

La velocidad de convergencia es mayor cuanto mayor sea p y menor sea C. Podemos distinguir los siguientes casos particulares.

- **Convergencia lineal:** se da si $p = 1$ y $C \in (0, 1)$.

  Esto significa que la secuencia se aproxima a la solución con un ritmo constante. Cada nueva aproximación reduce el error en una cantidad proporcional al error anterior. En este caso, llamamos C como la tasa de convergencia.

  Matemáticamente se cumple: $|x_{k+1} - x^*| \approx C\ |x_k - x^*|$

- **Convergencia superlineal:** se da si $1 < p < 2$ o si $p = 1$ y $C = 0$.

  En este caso la secuencia se aproxima a la solución más rápidamente que en la convergencia lineal, pero sin alcanzar la cuadrática.

- **Convergencia cuadrática:** se da si $p = 2$ y $0 < C < \infty$.

  La secuencia se aproxima a la solución muy rápidamente. El error se reduce **exponencialmente,** es decir, el error en la $k + 1$-ésima iteración es aproximadamente el **cuadrado** del error en la iteración k.

  Matemáticamente se cumple: $|x_{k+1} - x^*| \approx C\ |x_k - x^*|^2$

Ejemplos.

- *Convergencia lineal:* $a_n = (\frac{1}{10})^n \to L = 0$

  $$\{a_n\} = 0{,}1,\ 0{,}01,\ 0{,}001\ ...$$
  $$\frac{|a_{n+1}-L|}{|a_n-L|} = 0, 1$$

  Como $p = 1$ y $C \in (0, 1)$ entonces la convergencia es lineal.

- *Convergencia superlineal:* $a_n = \frac{1}{n!} \to L = 0$

  $$\{a_n\} = 1, \tfrac{1}{2}, \tfrac{1}{6}, \tfrac{1}{24}, ...$$
  $$\frac{|a_{n+1}-L|}{|a_n-L|} = \frac{n!}{(n+1)!} = \frac{1}{n+1} \to 0 \text{ cuando } n \to \infty$$

  Como $p = 1$ y $C = 0$ entonces la convergencia es superlineal.

- *Convergencia cuadrática:* $a_n = \frac{1}{2^{2^n}} \to L = 0$

$$\frac{|a_{n+1}-L|}{|a_n-L|^2} = \frac{(2^{2^n})^2}{2^{2^{n+1}}} = \frac{2^{2n+1}}{2^{2n+1}} = 1$$

Como $p = 2$ y $C = 1$ entonces la convergencia es cuadrática.

3.1.4. Definición. Caracterización de convergencia superlineal.

La convergencia es superlineal si:

$$\lim_{k\to\infty} \frac{|x_{k+1}-x^*|}{|x_k-x^*|} = 0$$

3.2. Solución de Ecuaciones No Lineales de Una Variable.

3.2.1. Definición. Raíz de una función.

Sea $f : \mathbb{R} \to \mathbb{R}$ una función no lineal. Sea llama **raíz** o **cero** de la ecuación no lineal $f(x) = 0$ a todo número $\alpha \in \mathbb{R}$ tal que $f(\alpha) = 0$.

3.2.2. Teorema de Bolzano.

Sea f continua en $[a, b] \subset \mathbb{R}$ tal que $f(a)f(b) < 0$, entonces existe $c \in (a, b)$ tal que $f(c) = 0$.

Intuitivamente, el resultado afirma que, si una función es continua en un intervalo, entonces toma todos los valores intermedios comprendidos entre los extremos del intervalo. El teorema no especifica el número de raíces, solo afirma que como mínimo existe una.

3.3. Método de la Bisección para encontrar raíces.

3.3.1. Algoritmo.

Este método para encontrar la raíz de una función no lineal se basa en el Teorema de Bolzano. Suponemos que $f(x)$ es continua en $[a, b]$ y que $f(a)f(b) < 0$. Luego $f(x) = 0$ tiene al menos una raíz en dicho intervalo. Dada una tolerancia del error $\epsilon > 0$, el método de la bisección consiste en los siguientes pasos:

1. Definir $c = \frac{a+b}{2}$
2. Si $b - c \leq \epsilon$, aceptar c como la raíz y detenerse.
3. Si $b - c > \epsilon$, 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. En caso contrario reemplazar b con c. Regresar al paso 1.

En general, partiendo de $a_1 = a$ y $b_1 = b$, en cada iteración evaluamos:

- $f(a_k)f(c_k) < 0 \Rightarrow a_{k+1} = a_k, \quad b_{k+1} = c_k, \quad c_{k+1} = \frac{b_{k+1}+a_{k+1}}{2}$
- $f(b_k)f(c_k) < 0 \Rightarrow a_{k+1} = c_k, \quad b_{k+1} = b_k, \quad c_{k+1} = \frac{b_{k+1}+a_{k+1}}{2}$
- $f(c_k) = 0 \Rightarrow \alpha = c_k$

3.3.2. Acotación del error.

Primero observemos que:

$$b_2 - a_2 = \tfrac{1}{2}(b_1 - a_1)$$
$$b_3 - a_3 = \tfrac{1}{2}(b_2 - a_2) = \tfrac{1}{2}\tfrac{1}{2}(b_1 - a_1) = (\tfrac{1}{2})^2(b_1 - a_1)$$

Si procedemos por inducción obtenemos:

$$b_k - a_k = (\tfrac{1}{2})^{k-1}\ (b_1 - a_1)$$

Además,

$$|\alpha - c_k| \leq b_k - c_k = c_k - a_k = \tfrac{1}{2}\ (b_k - a_k)$$
$$\boxed{|\alpha - c_k| \leq (\tfrac{1}{2})^k\ (b_1 - a_1)}$$

Esta fórmula permite acotar el error de aproximación de la raíz. Puesto que $(\frac{1}{2})^k$ tiende a cero cuando k tiende a infinito, la fórmula demuestra que c_k converge a la raíz α cuando k tiende a infinito.

Supongamos que queremos obtener la raíz con un error no superior a $\epsilon > 0$:

$$|\alpha - c_k| \leq \epsilon$$

Esto se cumple si:

$$\begin{aligned} \frac{1}{2^k}(b-a) \leq \epsilon &\Rightarrow b - a \leq \epsilon \cdot 2^k \\ &\Rightarrow \frac{b-a}{\epsilon} \leq 2^k \\ &\Rightarrow \log_2(\frac{b-a}{\epsilon}) \leq k \end{aligned} \tag{9}$$

3.3.3. Ventajas y desventajas del método de la bisección.

Ventajas.

- Converge siempre.
- Acotación del error garantizada. El error disminuye en cada iteración.
- Velocidad de convergencia garantizada. La cota del error se reduce a la mitad en cada iteración.

Desventajas.

- La convergencia es relativamente lenta en comparación con otros métodos.

3.4. Método de Newton para encontrar raíces.

3.4.1. Explicación.

Método iterativo que se utiliza para encontrar aproximaciones a las raíces de una función $f(x)$. El método de Neton utiliza la **tangente** de la función en un punto x_n para aproximar la raíz de la función. A partir de una estimación inicial x_0, el método genera sucesivas aproximaciones $x_1, x_2, x_3, ...$ que convergen hacia la raíz.

Sea α una raíz de la ecuación $f(x) = 0$. Supongamos que $f \in \mathbb{C}^2$ (es decir, f es 2 veces continuamente derivable) en $[a, b]$. Sea $x_0 \in [a, b]$ una estimación de α tal que $f'(x_0) \neq 0$ y x_0 es *cercano* a α.

Consideramos el polinomio de Taylor para $f(x)$ expandido alrededor de x_0, donde c_x está entre x y x_0:

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

Obtenemos una nueva estimación de α igual a x_1 de la siguiente manera:

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

de donde

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

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

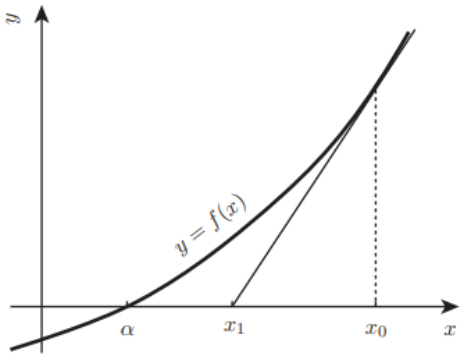


Figura 1: Método de Newton.

Repitiendo este procedimiento, obtenemos la fórmula general de la iteración de Newton:

$$\boxed{x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}} \; n = 0, 1, 2, ...$$

3.4.2. Algoritmo.

1. Calcular x_{n+1} a partir de $x_n, f(x_n), f'(x_n)$.
2. Si $f(x_{n+1}) < \epsilon$ (tolerancia) entonces x_{n+1} es la raíz aproximada.
3. Si $f(x_{n+1}) > \epsilon$ volver al paso 1.

3.4.3. Análisis del error.

Aplicando el desarrollo de Taylor con resto con respecto a x_n, 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. Diviendo todo por $f'(x_n)$ tenemos:

$$
\begin{aligned}
0 &= \frac{f(x_n)}{f'(x_n)} + \alpha - x_n + (\alpha - x_n)^2 \frac{f''(c_n)}{2f'(x_n)} \\
&= x_n - x_{n+1} + \alpha - x_n + (\alpha - x_n)^2 \frac{f''(c_n)}{2f'(x_n)}
\end{aligned} \tag{10}
$$

De donde el paso $\frac{f(x_n)}{f'(x_n)} = x_n - x_{n+1}$ surge del paso iterativo del método de Newton $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$

Luego, siguiendo con la igualdad anterior tenemos que el error es:

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

Suponiendo que el método de Newton converge tenemos que la convergencia es **cuadrátrica**:

$$\lim_{n\to\infty} \frac{|\alpha - x_{n+1}|}{|\alpha - x_n|^2} = \lim_{n\to\infty} \left|\frac{-f''(c_n)}{2f'(x_n)}\right| = \left|\frac{-f''(\alpha)}{2f'(\alpha)}\right| \text{ si } f'(\alpha) \neq 0$$

3.4.4. Ventajas y desventajas del método de Newton.

Ventajas.

- Converge rápidamente en la mayoría de los casos (convergencia cuadrátrica).
- Formulación sencilla y comportamiento fácil de entender.

Desventajas.

- Puede no converger. La convergencia del método de Newton no está garantizada a partir de cualquier valor inicial x_0. Sin embargo, la convergencia está garantizada para x_0 suficientemente cercano a la raíz α.

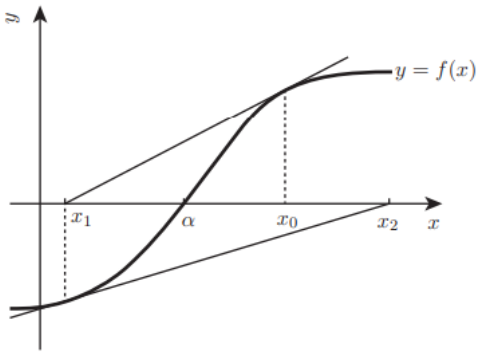


Figura 2: Ejemplo de divergencia del Método de Newton.

- Puede ocurrir que $f'(\alpha) = 0$.
- Se necesita conocer tanto $f(x)$ como $f'(x)$.

3.5. Método de la secante para encontrar raíces.

3.5.1. Explicación.

Aproximamos la función $f(x)$ mediante la línea secante que pasa por los dos puntos $(x_0, f(x_0))$ y $(x_1, f(x_1))$. El punto x_2 se obtiene como la raíz de la línea secante, y constituye una nueva aproximación de α. Igualando las pendientes se tiene:

$$\frac{f(x_1)-f(x_0)}{x_1-x_0} = \frac{f(x_2)-f(x_0)}{x_2-x_0} = \frac{0-f(x_1)}{x_2-x_1}$$

Despejando obtenemos:

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

Luego, la fórmula general del algoritmo para calcular una nueva aproximación de la raíz es:

$$\boxed{\mathrm{x}_{n+1} = x_n - f(x_n)\ \frac{x_n-x_{n-1}}{f(x_n)-f(x_{n-1})}}\ n \geq 1$$

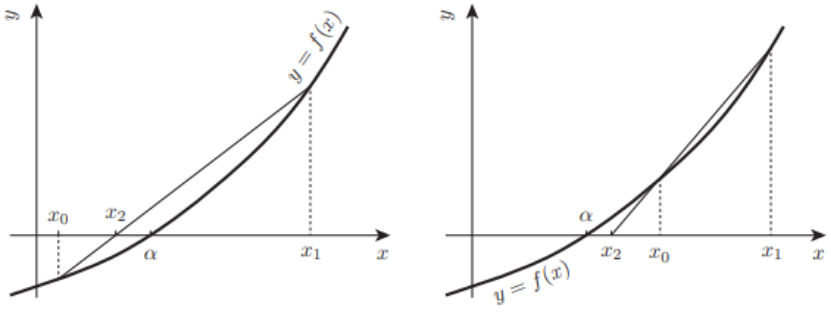


Figura 3: Método de la secante con distintos puntos iniciales.

Observación. El método de la secante es una aproximación del método de Newton, donde la derivada de f se aproxima mediante una diferencia finita: $f'(x) \approx \frac{f(x_n)-f(x_{n-1})}{x_n-x_{n-1}}$

Es un método de dos puntos, igual que el método de la bisección, pero casi siempre converge más rápido que el método de la bisección.

3.5.2. Algoritmo.

1. Elegir dos valores iniciales x_0 y x_1 que sean aproximaciones cercanas a la raíz de la función.
2. Calcular los valores de $f(x_0)$ y $f(x_1)$.
3. Aplicar la fórmula de la secante para conseguir el punto x_2:
$$x_{n+1} = x_n - f(x_n)\frac{x_n-x_{n-1}}{f(x_n)-f(x_{n-1})}$$
4. Si $|x_{n+1} - x_n|$ es menor que una tolerancia ϵ, entonces consideramos a x_{n+1} como la raíz. Alternativamente, podemos verificar si $|f(x_{n+1})| < \epsilon$ para asegurarnos de que el valor de la función en x_{n+1} es cercano a 0.
5. Repetir. Actualizar x_{n-1} con x_n y x_n con x_{n+1} y repetir el proceso hasta que se cumpla la condición de parada o se alcance el número máximo de iteraciones.

3.5.3. Análisis del error.

Se puede demostrar que:

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

Donde ξ_n se encuentra entre el mínimo y el máximo de x_{n-1}, x_n y α, mientras que ρ_n se encuentra entre x_{n-1} y x_n.

Suponiendo que $f \in \mathbb{C}^2$ (dos veces continuamente derivable) y $f'(\alpha) \neq 0$, se puede desmotrar que si x_n converge a α, luego:

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

con $p = (\sqrt{5}+1)/2 \approx 1,62$ (convergencia superlineal). Notar que el método de Newton converge más rápido que el método de la secante.

3.5.4. Ventajas y desventajas del método de la secante.

Ventajas.

- Converge más rápido que la convergencia lineal.
- No requiere conocer o estimar $f'(x)$.
- Requiere una sola evaluación de $f(x)$ por iteración comparado con el método de Newton que requiere dos evaluaciones para estimar $f'(x)$.

Desventajas.

- Puede no converger.
- Puede tener dificultades si $f'(\alpha) = 0$.
- El método de Newton se puede generalizar más fácilmente a sistemas de ecuaciones no lineales.

3.6. Método de la falsa posición para encontrar raíces.

3.6.1. Explicación.

Este método es una combinación de los métodos de la secante y de la bisección. Elegimos las aproximaciones iniciales a y b tales que $f(a)f(b) < 0$. Luego, $a_1 = a$ y $b_1 = b$.

Así, obtenemos c_1 aplicando el método de la secante.

- Si $f(a_1)f(c_1) < 0 \Rightarrow a_2 = a_1,\ b_2 = c_1$
- Si $f(b_1)f(c_1) < 0 \Rightarrow a_2 = c_1,\ b_2 = b_1$
- Si $f(c_1) = 0 \Rightarrow \alpha = c_1$

Luego obtenemos c_2 aplicando el metodo de la secante a a_2 y b_2.

3.6.2. Algoritmo.

1. Elegir dos valores iniciales a y b tales que $f(a)f(b) < 0$.
2. Calcular $f(a)$ y $f(b)$.
3. Usar la fórmula de la falsa posición para calcular el nuevo punto c_1:

$$c_1 = b - f(b)\frac{b-a}{f(b)-f(a)}$$

4. Verificar la condición de parada. Si $|f(c_1)|$ es menor que una tolerancia ϵ entonces c_1 es la raíz. Si no, seguimos al siguiente paso.
5. Actualizar el intervalo. Si $f(a)f(c_1) < 0$ entonces la raíz está en el intervalo $[a, c_1]$, por lo que actualizamos $b = c_1$.

   Si $f(b)f(c_1) < 0$ entonces la raíz está en el intervalo $[c_1, b]$, por lo que actualizamos $a = c_1$.
6. Repetimos desde el paso 2 hasta que se cumpla la condición de parada o se alcance el número máximo de iteraciones.

3.6.3. Ventajas y desventajas del método de la falsa posición.

Ventajas.

- Convergencia garantizada.

Desventajas.

- Es más leno que el método de la secante.

3.7. Métodos iterativos de punto fijo.

El **método de punto fijo** es un método iterativo que se utiliza para encontrar puntos fijos de una función $g(x)$. También puede aplicarse para resolver ecuaciones no lineales de la forma $f(x) = 0$ al reformular la ecuación en términos de un punto fijo $x = g(x)$.

3.7.1. Definición. Punto fijo.

Dada una funcion $g : \mathbb{R} \to \mathbb{R}$ 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 a la función $y = x$ (lineal).

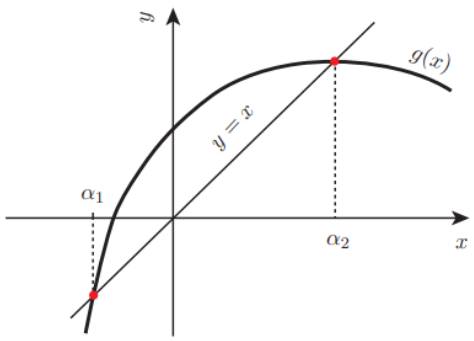


Figura 4: Puntos fijos de una función g.

3.7.2. Algoritmo básico de los métodos iterativos de punto fijo.

1. **Reformular la ecuación.** Si queremos resolver una ecuación $f(x) = 0$ primero la reformulamos como $x = g(x)$, de manera que la solución de $f(x) = 0$ es equivalente a encontrar un punto fijo de $g(x)$.
2. **Elección de un punto inicial.** Elegimos un valor inicial x_0.
3. **Iteración.** Se usa la relación $x_{n+1} = g(x_n)$ para generar una secuencia de aproximaciones al punto fijo, que será la raíz de la ecuación $f(x) = 0$.
4. **Condición de parada.** El proceso iterativo continúa hasta que $|x_{n+1} - x_n|$ sea menor que una tolerancia ϵ, es decir, hasta que la secuencia converja a un valor x^*

3.7.3. Ejemplo de método iterativo.

Se desea resolver la ecuación $f(x) = 3x^2 - e^x = 0$, hallando una raíz $\alpha > 0$. Esto es equivalente a resolver la ecuación $3x^2 = e^x$.

Esta ecuación tiene 3 soluciones que corresponden a los puntos de intersección de las funciones $3x^2$ y e^x.

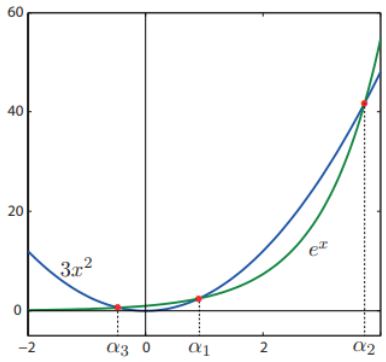


Figura 5: Soluciones de la ecuación $3x^2 - e^x = 0$.

Si queremos hallar una raíz positiva de $f(x)$ empleando un método de punto fijo, existen numerosas elecciones posibles de la función g. Consideramos las siguientes tres alternativas, que surgen de despejar x de la ecuación $3x^2 - e^x = 0$:

- $x = g_1(x) = log(3x^2)$
- $x = g_2(x) = \frac{e^{x/2}}{\sqrt{3}}$
- $x = g_3(x) = \frac{e^x}{3x}$

A partir de la gráfica de $g_1(x)$, vemos que las iteraciones de $x_{n+1} = g_1(x_n)$ convergen a α_2 para cualquier punto inicial $x_0 > \alpha_1$. Luego vemos que las iteraciones de $x_{n+1} = g_2(x_n)$ convergen a α_1 para cualquier punto inicial $x_0 < \alpha_2$. Por otra parte, las iteraciones de $x_{n+1} = g_3(x_n)$ convergen a α_1 para cualquier punto inicial $x_0 \in (\rho, \alpha_2)$, y la velocidad de convergencia cerca de α_1 es mayor que en los casos anteriores.

En resumen, este ejemplo nos permite realizar las siguientes observaciones:

- El método puede converger o diverger dependiendo del valor inicial x_0.

- El método puede converger a una raíz u otra dependiendo de la elección de $g(x)$.
- La convergencia puede ser más rápida o más lenta dependiendo de la elección de $g(x)$.
- Si $g(x) \leq f(x) = x$, entonces converge a la menor raíz (la que está más a la izquierda). Y si $g(x) \geq f(x) = x$ entonces converge a la raíz mayor (la de más a la derecha).
- Los escalones de convergencia siguen a la función identidad.

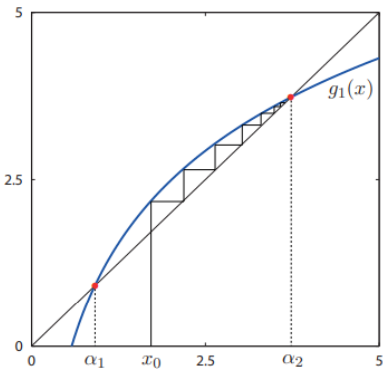


Figura 6: Iteración de punto fijo con $g_1(x)$.

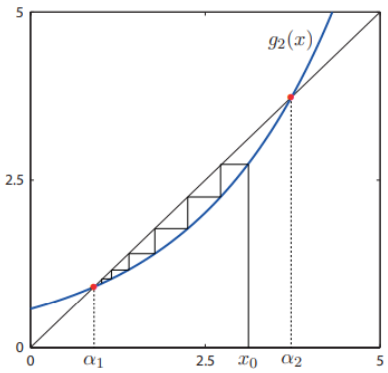


Figura 7: Iteración de punto fijo con $g_2(x)$.

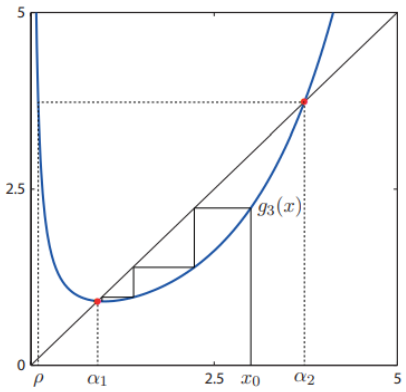


Figura 8: Iteración de punto fijo con $g_3(x)$.

3.7.4. Ejemplo. Método de Newton.

El **método de Newton** se considera un **método iterativo de punto fijo** porque reformula el problema de encontrar una raíz de la función $f(x)$ en términos de encontrar un punto fijo de una nueva función derivada de $f(x)$.

El método de Newton busca la raíz de una función $f(x) = 0$ mediante la fórmula iterativa $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$.

Podemos interpretar esta fórmula del método de Newton como una **iteración de punto fijo** si la expresamos de la siguiente manera:

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

Donde $g(x)$ es la **función iteración** derivada de $f(x)$ para el método de Newton:

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

El objetivo es encontrar un valor x^* tal que $g(x^*) = x^*$, lo cual equivale a resolver:

$$x^* - \frac{f(x^*)}{f'(x^*)} = x^*$$

3.7.5. Lema. Existencia de puntos fijos.

Sea $g(x)$ una función continua en $[a, b]$, y suponga que g satisface la siguiente propiedad:

$$a \leq x \leq b \Rightarrow a \leq g(x) \leq b$$

Luego, 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 extemos tenemos:

- $f(a) = a - g(a) \leq 0$ pues $a \leq g(x)$.
- $f(b) = b - g(b) \geq 0$ pues $g(x) \leq b$.

La función f es continua en $[a, b]$ por ser una resta de funciones continuas en $[a, b]$. Además, como $f(a)$ y $f(b)$ tienen distinto signo, entonces $f(a)f(b) \leq 0$ y por lo tanto se cumplen las hipótesis del Teorema de Bolzano.

Es decir, existe $\alpha \in [a, b]$ tal que $f(\alpha) = 0$. Por lo tanto, $f(\alpha) = \alpha - g(\alpha) = 0 \Rightarrow \alpha = g(\alpha)$. Y así demostramos la existencia de puntos fijos en g.

3.7.6. Teorema. Condición suficiente de convergencia.

Sea $g : \mathbb{R} \to \mathbb{R}$, $g \in \mathbb{C}^1$ (derivable) en $[a, b]$. Supongamos que g staisface:

- $a \leq x \leq b \Rightarrow a \leq g(x) \leq b$ (condición de existencia de punto fijo)
- $\lambda = \sup_{x \in [a,b]} |g'(x)| < 1$

Entonces:

1. Existe una solución única α de la ecuación $x = g(x)$ en $[a, b]$.
2. Para cualquier valor inicial $x_0 \in [a, b]$, la iteración $x_{n+1} = g(x_n)$ converge a α.
3. Cota de error: $|\alpha - x_n| \leq \frac{\lambda^n}{1-\lambda} |x_0 - x_1|$, $n \geq 0$
4. $\lim_{n\to\infty} \frac{\alpha - x_{n+1}}{\alpha - x_n} = g'(\alpha)$

   Por lo tanto, para x_n cercano a α:

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

Demostración.

Resultado previo.

Las hipótesis sobre g permiten aplicar el lema de existencia de punto fijo, lo que nos permite afirmar que existe al menos una solución de $x = g(x)$ en $[a, b]$.

Como $[a, b]$ es un intervalo cerrado y acotado, y g es una función continua en $[a, b]$ y derivable (condición de que $g \in \mathbb{C}^1$), entonces también es continua y derivable en $[w, z]$ para algunos $w, z \in [a, b]$. Luego, g cumple las hipótesis del Teorema de Valor Medio de Lagrange, y por lo tanto existe un $c \in (w, z)$ tal que:

$$g'(c) = \frac{g(z) - g(w)}{z - w} \Rightarrow g(z) - g(w) = g'(c)(z - w)$$

Luego, utilizando el hecho de que $\lambda = \sup_{x\in[a,b]} |g'(x)| < 1$, sigue que:

$$\begin{aligned} |g(z) - g(w)| &= |g'(c)||z - w| \\ &\leq \lambda|z - w|, \quad a \leq w, z \leq b \end{aligned} \tag{11}$$

Demostración.

1. Supongamos que existen 2 soluciones α y β en el intervalo $[a, b]$. Luego, $\alpha = g(\alpha)$ y $\beta = g(\beta)$. Restando ambas igualdades:

   $$\alpha - \beta = g(\alpha) - g(\beta)$$

   Tomando el valor absoluto y aplicando la desigualdad del resultado previo obtenemos:

   $$\begin{aligned} |\alpha - \beta| &= |g(\alpha) - g(\beta)| \\ &\leq \lambda|\alpha - \beta| \\ |\alpha - \beta| - \lambda|\alpha - \beta| &\leq 0 \\ (1 - \lambda)|\alpha - \beta| &\leq 0 \end{aligned} \tag{12}$$

   Como $\lambda < 1$, entonces $(1 - \lambda) > 0$ y como el valor absoluto siempre es positivo, para que se cumpla la desigualdad menor igual a cero, solo queda que $|\alpha - \beta| = 0$, es decir, que $\alpha = \beta$.

   Por lo tanto, concluimos que $x = g(x)$ tiene una solución única en $[a, b]$.

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

   Veamos que la propiedad que asumimos como hipótesis de g de que $a \leq x \leq b \Rightarrow a \leq g(x) \leq b$ implica que dado un valor inicial $x_0 \in [a, b]$, las iteraciones x_k pertenecen también a $[a, b]$, para $k \geq 1$.

   Por ejemplo, si $a \leq x_0 \leq b$, entonces $a \leq g(x_0) \leq b$. Como $g(x_0) = x_1$, esto demuestra que $x_1 \in [a, b]$. Para demostrar que las iteraciones converen, restamos ambos miembros de las igualdades $x_{n+1} = g(x_n)$ de $\alpha = g(\alpha)$:

   $$\begin{aligned} \alpha - x_{n+1} &= g(\alpha) - g(x_n) \\ &= g'(c_n)(\alpha - x_n) \end{aligned} \tag{13}$$

   Para algún c_n entre α y x_n. Este último paso se obtuvo aplicando el Teorema del Valor Medio de Lagrange como en el principio. Aplicando valor absoluto y la desigualdad del **resultado previo** obtenemos:

   $$|\alpha - x_{n+1}| = |g'(c_n)(\alpha - x_n)| \leq \lambda|\alpha - x_n|, \quad n \geq 0$$

   Por inducción tenemos:

   - $|\alpha - x_1| \leq \lambda|\alpha - x_0|$
   - $|\alpha - x_2| \leq \lambda|\alpha - x_1| \leq \lambda\lambda|\alpha - x_0|$
   - $|\alpha - x_3| \leq \lambda|\alpha - x_2| \leq \lambda\lambda\lambda|\alpha - x_0|$
   - $|\alpha - x_n| \leq \lambda^n|\alpha - x_0|, \quad n \geq 0$

Como queremos ver convergencia, debemos ver que sucede cuando $n \to \infty$. Veamos que como $\lambda < 1$, entonces cuando $n \to \infty$, $\lambda^n \to 0$ y por lo tanto $\alpha - x_n \to 0$ y entonces $x_n \to \alpha$.

3. Por propiedad del valor absoluto:

$$|\alpha - x_0| \leq |\alpha - x_1| + |x_1 - x_0|$$

Luego, utilizando esta última desigualdad y la desigualdad de que $|\alpha - x_{n+1}| \leq \lambda|\alpha - x_n|$ del item anterior, con $n = 0$, resulta:

$$\begin{aligned} |\alpha - x_0| \leq \lambda|\alpha - x_0| + |x_1 - x_0| &\Rightarrow |\alpha - x_0| - \lambda|\alpha - x_0| \leq |x_1 - x_0| \\ &\Rightarrow (1-\lambda)|\alpha - x_0| \leq |x_1 - x_0| \\ &\Rightarrow |\alpha - x_0| \leq \frac{1}{1-\lambda}|x_1 - x_0| \end{aligned} \tag{14}$$

Combinando esta última expresión con la ecuación $|\alpha - x_n| \leq \lambda^n|\alpha - x_0|$ del item anterior, obtenemos:

$$|\alpha - x_n| \leq \tfrac{\lambda^n}{1-\lambda}|x_0 - x_1|, \quad n \geq 0$$

4. En el item 2 hemos visto que $\alpha - x_{n+1} = g'(c_n)(\alpha - x_n)$ para algún c_n entre α y x_n. Luego:

$$\lim_{n\to\infty} \frac{\alpha - x_{n+1}}{\alpha - x_n} = \lim_{n\to\infty} g'(c_n) = g'(\alpha) \tag{15}$$

Esto surge pues c_n está entre α y x_n y los x_n convergen a α, entonces c_n también converge a α.

3.7.7. Corolario. Caracterización de la condición de convergencia.

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:

- Si $|g'(\alpha)| < 1$, la iteración $x_{n+1} = g(x_n)$ converge a α para x_0 suficientemente cercano a α.
- Si $|g'(\alpha)| > 1$, la iteración $x_{n+1} = g(x_n)$ no converge a α.
- Si $|g'(\alpha)| = 1$, no se pueden sacar conclusiones.

Demostración.

- Vimos que $\alpha - x_{n+1} = g'(c_n)(\alpha - x_n)$ para algún c_n entre α y x_n. Luego:

$$|\alpha - x_{n+1}| = |g'(c_n)||\alpha - x_n|$$

Siendo $g'(x)$ continua y $|g'(\alpha)| < 1$ (hipótesis), entonces existe $\epsilon > 0$ tal que:

$$|g'(x)| < 1, \quad \forall x \in [\alpha - \epsilon, \alpha + \epsilon]$$

Supongamos que $x_n \in [\alpha - \epsilon, \alpha + \epsilon]$, luego c_n también pertenece a $[\alpha - \epsilon, \alpha + \epsilon]$ y por lo tanto $|g'(c_n)| < 1$.

Por lo tanto, x_{n+1} está más próximo a α que x_n y $x_{n+1} \in [\alpha - \epsilon, \alpha + \epsilon]$. Siendo $x_{n+1} = g(x_n)$, se verifica la condición:

$$x_n \in [\alpha - \epsilon, \alpha + \epsilon] \Rightarrow g(x_n) \in [\alpha - \epsilon, \alpha + \epsilon]$$

Vemos que se cumplen las condiciones del teorema de Condición Necesaria de Convergencia usando $[a, b] = [\alpha - \epsilon, \alpha + \epsilon]$, lo cual demuestra que las iteraciones convergen a α para x_0 suficientemente cercano a α.

- Del mismo modo, siendo $g'(x)$ continua y $|g'(\alpha)| > 1$, entonces existe $\epsilon > 0$ tal que:

$$|g'(x)| > 1, \quad \forall x \in [\alpha - \epsilon, \alpha + \epsilon]$$

Supongamos que $x_n \in [\alpha - \epsilon, \alpha + \epsilon]$, luego c_n también pertenece a $[\alpha - \epsilon, \alpha + \epsilon]$ y $|g'(c_n)| > 1$.

Por lo tanto, de la fórmula $|\alpha - x_{n+1}| = |g'(c_n)||\alpha - x_n|$, como $|g'(c_n)| > 1$ entonces x_{n+1} está más alejado de α que x_n, y por lo tanto las iteraciones no convergen a α.

- Depende de cada caso en particular. En caso de converger, el método sería impráctico ya que la convergencia sería demasiado lenta.

Ejemplo.

Vimos que el método de Newton es un método iterativo de punto fijo con:

$$g(x) = x - \tfrac{f(x)}{f'(x)}$$

Diferenciando obtenemos:

$$g'(x) = 1 - \frac{f'(x)}{f'(x)} + \frac{f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2} \tag{16}$$

Evaluando $g'(\alpha) = 0$ ya que $f(\alpha) = 0$ (pues α es la raíz). Siendo $|g'(\alpha)| = 0 < 1$, queda demostrado que el método de Newton converge a la raíz α para x_0 suficientemente cercano a α.

3.8. Solución de Sistemas de Ecuaciones No Lineales.

4. Conceptos Preliminares del Álgebra Lineal.

4.1. Sistemas de Ecuaciones Lineales.

Consideramos sistemas de n ecuaciones y n incógnitas:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n &= b_2 \\ &... \\ a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n &= b_n \end{aligned} \tag{17}$$

En forma matricial:

$$Ax = b$$

$$\mathsf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \in \mathbb{R}^{n\times n}, \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \in \mathbb{R}^n, \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n,$$

4.1.1. Representación computacional de una matriz.

El almacenamiento de una matriz de $n \times n$ en doble precisión requiere $8n^2$ bytes. Para matrices muy grandes esto no es despreciable. Por ejemplo, si $n = 8000$ se requieren 512 MB.

Introduciremos dos conceptos de matrices:

- **Matriz plena (densa):** es una matriz en la que la mayoría de los elementos son **diferentes de cero.** La representación computacional de este tipo de matrices ocupa más espacio en la memoria, ya que se almacenan todos los elementos, incluso los ceros.
- **Matriz rala (dispersa):** es una matriz en la que la mayoría de los elementos son **ceros,** y solo una pequeña fracción de las entradas tiene valores diferentes de cero.

  La representación computacional de este tipo de matrices busca almacenar solamente los índices de los elementos no nulos junto con sus valores, para así, ahorrar memoria al evitar almacenar los ceros.

4.1.2. Definición. Matriz p-banda.

Decimos que $A \in \mathbb{R}^{n\times n}$ es una **matriz p-banda** si existe un $p \in \mathbb{Z}$ con $1 \leq p < n$, tal que:

$$|i - j| \geq p \Rightarrow a_{ij} = 0$$

Es decir, es una matriz en la que todos los elementos fuera del rango determinado p alrededor de la diagonal principal son ceros. Este número p indica el ancho de la banda, es decir, cuántas filas y columnas a ambos lados de la diagonal principal contienen elementos no nulos.

Ejemplos.

- Si $p = 1$, entonces la matriz es diagonal.

- Si $p = 2$, entonces la matriz es tridiagonal.

$$\begin{bmatrix} a_1 & c_1 & 0 & ... & ... & 0 \\ b_2 & a_2 & c_2 & 0 & ... & 0 \\ 0 & b_3 & a_3 & c_3 & ... & 0 \\ & & & ... & & \\ 0 & ... & ... & 0 & b_n & a_n \end{bmatrix} \tag{18}$$

4.1.3. Sistemas de Ecuaciones Lineales Equivalentes.

Sean $A, B \in \mathbb{R}^{n\times n}$. Decimos que dos sistemas de ecuaciones lineales $Ax = b$ y $Bx = d$ son equivalentes si poseen exactamente las mismas soluciones.

Si podemos reducir un sistema de ecuaciones $Ax = b$ a un sistema más simple $Bx = d$ que sea más sencilo de resolver, podemos resolver este último y obtendremos de este modo las soluciones del sistema original.

Se obtienen sistemas de ecuaciones lineales equivalentes realizando operaciones elementales por filas sobre una matriz.

4.1.4. Definición. Operaciones Elementales por Filas.

Sea $A \in \mathbb{R}^{n\times n}$. Llamaremos operaciones elementales por filas sobre A a cada una de las siguientes operaciones:

1. La ecuación E_i puede multiplicarse por una constante $\lambda \neq 0$ y la ecuación resultante se emplea en vez de E_i. Esta operación se denota por $(\lambda E_i) \to (E_i)$.
2. La ecuación E_j puede multiplicarse por cualquier constante λ y sumarse a la ecuación E_i, la ecuación resultante se emplea en vez de E_i. Esta operación se denota por $(E_i + \lambda E_j) \to (E_i)$.
3. El orden de las ecuaciones E_i y E_j puede intercambiarse. Esta operacion se denota por $(E_i) \leftrightarrow (E_j)$.

4.2. Determinantes.

4.2.1. Definición. Determinante de una Matriz.

El **determinante** es un número real asociado a una matriz cuadrada, que permite calcular varias propiedades como la singularidad, el polinomio característico, y los autovalores de la matriz.

- El determinante de una matriz de 1×1 es $det(A) = a$.
- El determinante de una matriz de 2×2 es:

$$det(A) = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

- El determinante de una matriz de 3×3 es:

$$det(A) = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

4.2.2. Definición. Menor adjunto de una matriz.

Sea $A = (a_{ij})_{n\times n}$ una matriz de $n \times n$. Dado un par de índices (i, j) representamos por A_{ij} a la matriz que resulta al eliminar la i-ésima fila y la j-ésima columna de A.

El **menor adjunto** δ_{ij} de A es el número real dado por:

$$\delta_{ij} = (-1)^{i+j}|A_{ij}|$$

Emplearemos el menor adjunto para calcular el determinante de una matriz cuadrada cualquiera. Veamos el siguiente teorema.

4.2.3. Teorema de Laplace.

Dada una matriz $A \in \mathbb{R}^{n\times n}$, su determinante se puede calcular mediante el desarrollo de Laplace por una fila (o columna) cualquiera, de la siguiente forma:

- Si elegimos la i-ésima fila, $|A| = a_{i1}\delta_{i1} + ... + a_{in}\delta_{in}$
- Si elegimos la j-ésima columna, $|A| = a_{1j}\delta_{1j} + ... + a_{nj}\delta_{nj}$

Observación. El desarrollo de Laplace no es práctico para matrices de órdenes superiores ya que su grado de complejidad es O($n!$). Resulta más práctico convertir la matriz a triangular superior mediante la Eliminación de Gauss y multiplicar la diagonal, ya que en ese caso la complejidad se reduce a O(n^3).

Propiedades de los determinantes.

- En el teorema de Laplace vemos que el determinante no cambia cuando se intercambian las filas por columnas, es decir, $|A| = |A^t|$.
- Si $A \in \mathbb{R}^{n\times n}$ es triangular, su determinante es el producto de los elementos de la diagonal principal:

$$A \in \mathbb{R}^{n\times n} \text{ triangular} \Rightarrow det(A) = a_{11}a_{22}...a_{nn}$$

- Teniendo en cuenta que la matriz identidad es triangular, se deduce que $det(I) = 1$.

4.3. Rango de una Matriz.

4.3.1. Definición. Rango de una matriz.

El **rango de una matriz** se define como (a) el máximo número de columnas linealmente independientes de la matriz, o (b) el máximo número de filas linealmente independientes de la matriz. Ambas definiciones son equivalentes.

Un conjunto de vectores es linealmente independiente si ninguno de ellos puede expresarse como una combinación lineal de los demás.

4.3.2. Cotas en el rango de un producto de matrices.

Si $A \in \mathbb{R}^{m\times n}$ y $B \in \mathbb{R}^{n\times p}$, luego:

$$\boxed{\text{rank(A) + rank(B) - n} \leq rank(AB) \leq min\{rank(A), rank(B)\}}$$

En particular, si $A, B \in \mathbb{R}^{n\times n}$ son no singulares, luego:

$$n \leq rank(AB) \leq n$$

Es decir, $rank(AB) = n$ y el producto AB es también no singular.

4.3.3. Rango en productos con matriz transpuesta.

Para $A \in \mathbb{R}^{m \times n}$ se cumple:

$$\boxed{\text{rank}(\text{A}^t A) = rank(A) = rank(AA^t)}$$

En particular, si $n \leq m$ y $rank(A) = n$, luego $A^t A$ es no singular.

4.4. Matriz Inversa.

4.4.1. Definición. Matriz Inversa.

Sea $A \in \mathbb{R}^{n \times n}$. Decimos que A^{-1} es la inversa de A si:

$$AA^{-1} = A^{-1}A = I$$

Además, si la inversa de A existe, esta es única.

Observación.

Para una matriz de 2×2, la matriz inversa está dada por:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}, \text{ si } ad - bc \neq 0$$

4.4.2. Teorema. Enunciados equivalentes de matriz invertible.

Sea $A \in \mathbb{R}^{n \times n}$. Los siguientes enunciados son equivalentes:

1. A tiene inversa.
2. $det(A) \neq 0$.
3. Las filas de A son linealmente independientes (forman una base de $\mathbb{R}^n$).
4. Las columnas de A son linealmente independientes (forman una base de $\mathbb{R}^n$).
5. El sistema $Ax = 0$ donde $x \in \mathbb{R}^n$ tiene como única solución $x = 0$.
6. Para cada $b \in \mathbb{R}^n$, el sistema $Ax = b$, posee una única solución $x \in \mathbb{R}^n$.
7. $rank(A) = n$
8. Todos los autovalores de A son distintos de cero.

Observación. Un sistema de ecuaciones lineales $Ax = b$ con $A \in \mathbb{R}^{n \times n}$ y $det(A) \neq 0$ se puede resolver invirtiendo la matriz A como sigue:

$$x = A^{-1}b$$

Sin embargo, mediante la Eliminación de Gauss se puede resolver el sistema de forma más eficiente sin calcular la inversa.

4.5. Matriz Simétrica.

4.5.1. Definición. Matriz Simétrica.

Una matriz $A \in \mathbb{R}^{n\times n}$ es **simétrica** si y sólo si:

$$a_{ij} = a_{ji},\ 1 \leq i,\ j \leq n$$

En otras palabras, A es simétrica si $A^t = A$.

4.5.2. Teorema. Matriz simétrica y autovalores.

Sea $A \in \mathbb{R}^{n\times n}$ una matriz simétrica real. Luego, existe un conjunto de n pares autovalor-autovector $\{\lambda_i, v_i\}$ que satisfacen las siguientes propiedades:

1. Los autovalores $\lambda_1, ...\lambda_n$ son las raíces del polinomio característico $f(\lambda)$ de A. Todos los autovalores λ_i son números reales y pueden repetirse de acuerdo a su multiplicidad.
2. Los autovectores $v_1, ...v_n$ son ortogonales entre sí, y pueden elegirse de longitud unitaria:
$$v_i^t v_j = 0, \quad v_i^t v_i = 1$$
3. Para cada vector $\bar{x} = [x_1, x_2, ...x_n]^t$ existe un único vector $\bar{c} = [c_1, c_2, ...c_n]^t$ tal que:
$$\bar{x} = c_1\bar{v}_1 + ... + c_n\bar{v}_n$$
Si los autovectores son de longitud unitaria, las constantes están dadas por:
$$c_i = \bar{x}^t\bar{v}_i$$
4. Sea la matriz de autovectores $P = [v_1, v_2, ...v_n]$. Luego:
$$P^tAP = D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}$$
Además,
$$PP^t = P^tP = I$$
$$A = PDP^t$$

4.6. Matriz Definida Positiva.

4.6.1. Definición. Matriz Definida Positiva.

Sea $A \in \mathbb{R}^{n\times n}$ una matriz simétrica:

- A es **definida positiva** si sus autovalores son positivos, es decir, $\lambda_i > 0$.
- A es **semidefinida positiva** si sus autovalores son no negativos, es decir, $\lambda_i \geq 0$.
- A es **definida negativa** si sus autovalores son negativos, es decir, $\lambda_i < 0$.
- A es **semidefinida negativa** si sus autovalores son no positivos, es decir, $\lambda_i \leq 0$.
- A es **indefinida** si no cumple ninguna de las definiciones anteriores.

4.6.2. Teorema. Caracterización de matriz semidefinida positiva.

Una matriz $A \in \mathbb{R}^{n \times n}$ simétrica es **semidefinida positiva** si y sólo si A se puede factorizar como $A = B^t B$, y A es **definida positiva** si y sólo si esa matriz B es no singular.

Demostración.

4.6.3. Teorema. Caracterización de matriz definida positiva.

Una matriz simétrica $A \in \mathbb{R}^{n \times n}$ es **definida positiva** si y sólo si $x^t Ax > 0$ para todo $x \in \mathbb{R}^n$ y $x \neq \bar{0}$.

Demostración.

4.6.4. Teorema. Enunciados equivalentes de matriz definida positiva.

Para matrices A reales simétricas, los siguientes enunciados son equivalentes y sirven como definición de matriz definida positiva.

- $x^t Ax > 0$, $\forall x \in \mathbb{R}^n$
- Todos los autovalores de A son estrictamente positivos.
- $A = B^t B$ para alguna matriz B no singular. B no es única, pero 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.
- Toda submatriz principal de A es definida positiva.

5. Resolución de Sistemas de Ecuaciones Lineales - Métodos Directos

5.1. Eliminación de Gauss.

5.1.1. Explicación introductoria.

El método de **eliminación de Gauss** consiste en operar sobre la matriz ampliada del sistema hasta hallar la forma escalonada (una matriz triangular superior). Así, se obtiene un sistema fácil de resolver por sustitución regresiva.

Las dos etapas del método son:

1. Eliminación progresiva de incógnitas. Consiste en transformar el sistema en un sistema triangular superior usando operaciones elementales sobre las filas de la matriz ampliada.
2. Resolución del sistema triangular superior mediante sustitución regresiva.

Definición del sistema y matriz ampliada.

Denotamos el sistema lineal original como $A^{(1)}x = b^{(1)}$, donde el supraíndice indica la componente en el paso k y definimos la matriz ampliada:

$$[A \mid b] = [A^{(1)} \mid b^{(1)}] = \left[\begin{array}{ccc|c} a_{11}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ \vdots & & \vdots & \vdots \\ a_{n1}^{(1)} & \cdots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right]$$

Este sistema inicial se reduce en $n-1$ pasos a la forma:

$$[A^{(n)} \mid b^{(n)}] = \left[\begin{array}{cccc|c} a_{11}^{(1)} & \cdots & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & \ddots & & \vdots & \vdots \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & a_{nn}^{(n)} & b_n^{(n)} \end{array}\right]$$

Y luego el sistema trianuglar $A^{(n)}x = b^{(n)}$ se resuelve por sustitución regresiva.

5.1.2. Algoritmo de Eliminación de Gauss.

- **PASO 1:** Supongamos que $a_{11}^{(1)} \neq 0$. Sean los multiplicadores de cada fila $i = 2, 3, ...n$:

$$m_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}$$

Estos multiplicadores se usan para eliminar x_1 de las ecuaciones 2 a n. Definimos:

$$a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i1}a_{ij}^{(1)}, \quad i, j = 2, ...n$$

$$b_i^{(2)} = b_i^{(1)} - m_{i1}b_1^{(1)}, \quad i = 2, ...n$$

Luego, la primera fila de la matriz ampliada $[A^{(1)} \mid b^{(1)}]$ no se modifica, y la primera columna de $A^{(1)}$ debajo de la diagonal, se lleva a cero. La matriz ampliada $[A^{(2)} \mid b^{(2)}]$ tiene la forma:

$$[A^{(2)} \mid b^{(2)}] = \left[\begin{array}{cccc|c} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} & b_n^{(2)} \end{array}\right]$$

Donde a la primera fila la llamamos $E_1^{(1)}$, a la segunda $E_2^{(2)}$, y así hasta $E_n^{(2)}$, donde esta notación ($E_i^{(k)}$) se emplea para denotar la fila i de la matriz ampliada en el paso k. Esta fila representa la ecuación i del sistema $A^{(k)}x = b^{(k)}$.

- **PASO k:** Suponga que para $i = 1, ...k-1$ los x_i han sido eliminados de las ecuaciones $i+1, ...n$. Tenemos:

$$[A^{(k)} \mid b^{(k)}] = \left[\begin{array}{ccccccc} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & \cdots & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & \cdots & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ \vdots & \ddots & \ddots & & & \vdots & \vdots \\ \vdots & & 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} & b_k^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} & b_n^{(k)} \end{array}\right]$$

Y las filas son $E_1^{(1)}, E_2^{(2)}, E_k^{(k)}, E_n^{(k)}$. Supongamos que el elemento pivote $a_{kk}^{(k)} \neq 0$. Definimos los multiplicadores de fila $i = k+1, ...n$ como:

$$m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$$

Estos multiplicadores se usan para eliminar x_k de las ecuaciones $k+1, ...n$:

$$E_i^{(k+1)} = E_i^{(k)} - m_{ik}E_k^{(k)}, \quad i = k+1, ...n$$

Continuando de esta manera, después de $n-1$ pasos, obtenemos el sistema escalonado (triangular superior) $A^{(n)}x = b^{(n)}$.

- **SUSTITUCIÓN REGRESIVA:** Primero obtenemos x_n de la última ecuación del sistema escalonado $A^{(n)}x = b^{(n)}$:

$$x_n = \frac{b_n^{(n)}}{a_{nn}^{(n)}}$$

Este resultado se puede sustituir hacia atrás en la $(n-1)$-ésima ecuación y despejar x_{n-1} y así para las incógnitas restantes:

$$x_i = \frac{1}{a_{ii}^{(i)}}\left(b_i^{(i)} - \sum_{j=i+1}^{n} a_{ij}^{(i)}x_j\right), \qquad \text{para } i = n-1, n-2, ..,1$$

Esto completa el algoritmo de la eliminación de Gauss.

5.1.3. Ejemplo. Aplicación de la eliminación de Gauss.

Se quiere resolver el sistema lineal:

$$\begin{aligned} x_1 + 2x_2 + x_3 &= 0 \\ 2x_1 + 2x_2 + 3x_3 &= 3 \\ -x_1 + 3x_2 &= 2 \end{aligned} \tag{19}$$

Representamos este sistema lineal con la matriz ampliada:

$$[A \mid b] = \left[\begin{array}{ccc|c} 1 & 2 & 1 & 0 \\ 2 & 2 & 3 & 3 \\ -1 & -3 & 0 & 2 \end{array}\right]$$

En el siguiente diagrama, las flechas indican los pasos de eliminación, y los multiplicadores utilizados se indican al costado de las flechas:

$$\left[\begin{array}{ccc|c} 1 & 2 & 1 & 0 \\ 2 & 2 & 3 & 3 \\ -1 & -3 & 0 & 2 \end{array}\right] \xrightarrow[m_{31}=-1]{m_{21}=2} \left[\begin{array}{ccc|c} 1 & 2 & 1 & 0 \\ 0 & -2 & 1 & 3 \\ 0 & -1 & 1 & 2 \end{array}\right] \xrightarrow{m_{32}=\frac{1}{2}} \left[\begin{array}{ccc|c} 1 & 2 & 1 & 0 \\ 0 & -2 & 1 & 3 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{array}\right]$$

Resolviendo el sistema $A^{(3)}x = b^{(3)}$ con sustitución regresiva tenemos:

$$x_3 = 1, \quad x_2 = -1, \quad x_1 = 1$$

5.2. Pivoteo Parcial.

5.2.1. Definición. Técnica de pivoteo parcial.

En cada paso del proceso de eliminación visto en la sección anterior, supusimos que los elementos pivote eran distintos de cero, $a_{kk}^{(k)} \neq 0$.

Para eliminar esta hipótesis, podemos emplear **pivoteo parcial:** en el caso de que $a_{kk}^{(k)} = 0$, examinamos los elementos $a_{ik}^{(k)}$ en las filas $E_i^{(k)}$ para $i = k+1, ...n$ (nos fijamos en la misma columna k, los elementos de las filas de abajo y buscamos alguno que no sea cero).

Siendo A **no singular,** se puede demostrar que al menos uno de dichos elementos es distinto de cero. Luego, la ecuación $E_i^{(k)}$ con $a_{ik}^{(k)} \neq 0$ se intercambia con $E_k^{(k)}$ y luego se continua el proceso de eliminación.

5.2.2. Ejemplo.

Se quiere resolver el sistema lineal:

$$\begin{aligned} 6x_1 + 2x_2 + 2x_3 &= -2 \\ 2x_1 + \frac{2}{3}x_2 + \frac{1}{3}x_3 &= 1 \\ x_1 + 2x_2 - x_3 &= 0 \end{aligned} \tag{20}$$

Utilizando una calculadora decimal con una mantisa de 4 dígitos, tenemos que la solución exacta es:

$$x_1 = 2,6 \quad x_2 = -3,8 \quad x_3 = -5,0$$

Resolución errónea.

Veamos la solución que se obtiene aplicando el proceso de eliminación de Gauss con la aritmética de la calculadora:

$$\left[\begin{array}{rrr|r} 6{,}000 & 2{,}000 & 2{,}000 & -2{,}000 \\ 2{,}000 & 0{,}6667 & 0{,}3333 & 1{,}000 \\ 1{,}000 & 2{,}000 & -1{,}000 & 0{,}000 \end{array}\right] \xrightarrow[m_{31}=0{,}1667]{m_{21}=0{,}3333} \left[\begin{array}{rrr|r} 6{,}000 & 2{,}000 & 2{,}000 & -2{,}000 \\ 0{,}000 & 0{,}0001000 & -0{,}3333 & 1{,}667 \\ 0{,}000 & 1{,}667 & -1{,}333 & 0{,}3334 \end{array}\right]$$

$$\xrightarrow[m_{32}=16670]{} \left[\begin{array}{rrr|r} 6{,}000 & 2{,}000 & 2{,}000 & -2{,}000 \\ 0{,}000 & 0{,}0001000 & -0{,}3333 & 1{,}667 \\ 0{,}000 & 0{,}000 & 5555 & -27790 \end{array}\right]$$

Resolviendo por sustitución regresiva obtenemos:

$$x_1 = 1,335 \quad x_2 = -0,000 \quad x_3 = -5,003$$

Vemos que la solución obtenida difiere mucho de la solución verdadera. Este error se debe a que el elemento pivote $a_{22}^{(2)} = 0,0001000$ debió haber sido igual a cero, pero no lo es debido a los errores de redondeo. Este elemento pivote tiene un error relativo infinito.

Solución.

Para evitar este problema podemos intercambiar las ecuaciones $E_2^{(2)}$ y $E_3^{(2)}$, y luego continuar la eliminación:

$$\left[\begin{array}{rrr|r} 6{,}000 & 2{,}000 & 2{,}000 & -2{,}000 \\ 0{,}000 & 1{,}667 & -1{,}333 & 0{,}3334 \\ 0{,}000 & 0{,}0001000 & -0{,}3333 & 1{,}667 \end{array}\right] \xrightarrow[m_{32}=0{,}00005999]{} \left[\begin{array}{rrr|r} 6{,}000 & 2{,}000 & 2{,}000 & -2{,}000 \\ 0{,}000 & 1{,}667 & -1{,}333 & 0{,}3334 \\ 0{,}000 & 0{,}000 & -0{,}3332 & 1{,}667 \end{array}\right]$$

Mediante sustitución regresiva obtenemos en este caso:

$$x_1 = -2,602 \quad x_2 = -3,801 \quad x_3 = -5,003$$

que se aproxima mucho más a la solución real del sistema.

5.2.3. Procedimiento de pivoteo parcial para evitar errores de redondeo.

Vemos en el ejemplo anterior que no es suficiente simplemente pedir que los elementos pivote sean distintos de cero. Puede ocurrir que un elemento sea distinto de cero debido a errores de redondeo, lo cual conduce a errores gruesos en los cálculos subsiguientes. Para evitar esto, se introduce el pivoteo parcial, donde en cada paso k se selecciona el mayor elemento en valor absoluto de la columna como pivote. Esto reduce la posibilidad de dividir entre un número muy pequeño y, por ende, disminuye el riesgo de errores:

En el paso k, calcular:

$$c = \text{máx}_{k \leq i \leq n} |a_{ik}^{(k)}|$$

Si el elemento $|a_{kk}^{(k)}| < c$, luego intercambiar la ecuación $E_k^{(k)}$ con la correspondiente ecuación tal que $|a_{kk}^{(k)}| = c$.

Con el pivoteo parcial, los multiplicadores m_{ik} satisfacen:

$$|m_{ik}| \leq 1, \quad i = k+1, ...n \quad k = 1, ...n-1$$

Además del pivoteo parcial, existe también el pivoteo completo, en el cual, tanto en las columnas como en las filas se busca el elemento de mayor valor absoluto y luego se intercambian. El pivoteo completo agrega complejidad al cambiar el orden de las variables x.

5.3. Número de Operaciones del Método de Gauss.

Para analizar el número de operaciones necesarias para resolver el sistema $Ax = b$ por eliminación de Gauss, consideraremos separadamente la generación de $A^{(n)}$ a partir de $A^{(1)}$, la modificación de $b^{(1)}$ a $b^{(n)}$, y finalmente la obtención de la solución x por sustitución regresiva.

5.3.1. Generación de $A^{(n)}$ a partir de $A^{(1)}$

En el paso 1, se necesitan $n-1$ divisiones para calcular los multiplicadores $m_{i1},\ 2 \leq i \leq n$ (pues se calcula un multiplicador para cada fila distinta de la actual).

Luego, se usan $(n-1)^2$ multiplicaciones y $(n-1)^2$ sumas para crear los nuevos elementos $a_{ij}^{(2)}$. (Esto es pues la primer columna es cero. Parandonos en una fila tenemos que calcular $n-1$ multiplicaciones y $n-1$ sumas. Luego, este proceso se repite en $n-1$ filas. Por lo tanto, esto es $(n-1)+(n-1)+(n-1)... = (n-1)\cdot(n-1) = (n-1)^2$.

Continuando de esta manera en cada paso, iterando en las filas, obtenemos el conteo de operaciones indicado en la siguiente tabla:

Tabla 1: Conteo de operaciones en la eliminación gaussiana.

Paso	Sumas	Multiplicaciones	Divisiones
1	$(n-1)^2$	$(n-1)^2$	$n-1$
2	$(n-2)^2$	$(n-2)^2$	$n-2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n-1$	1	1	1
Total	$\frac{n(n-1)(2n-1)}{6}$	$\frac{n(n-1)(2n-1)}{6}$	$\frac{n(n-1)}{2}$

Denotamos por $SR(\cdot)$ el número de sumas y restas, y por $MD(\cdot)$ el número de multiplicaciones y divisiones. Luego,

- $SR(A^{(1)} \to A^{(n)}) = \frac{n(n-1)(2n-1)}{6} \approx \frac{n^3}{3}$
- $MD(A^{(1)} \to A^{(n)}) = \frac{n(n-1)(2n-1)}{6} + \frac{n(n-1)}{2} = \frac{n(n^2-1)}{3} \approx \frac{n^3}{3}$

siendo las estimaciones finales válidas para valores grandes de n.

5.3.2. Modificación de $b^{(1)}$ a $b^{(n)}$

- $SR(b^{(1)} \to b^{(n)}) = (n-1)+(n-2)+...+2+1 = \frac{n(n-1)}{2}$
- $MD(b^{(1)} \to b^{(n)}) = (n-1)+(n-2)+...+2+1 = \frac{n(n-1)}{2}$

5.3.3. Solución regresiva de $A^{(n)}x = b^{(n)}$

- $SR(x) = \frac{n(n-1)}{2}$
- $MD(x) = \frac{n(n+1)}{2}$

Vemos que, para valores de n grandes, el principal costo computacional de la eliminación de Gauss se da en la generación de $A^{(n)}$ a partir de $A^{(1)}$. Puede entonces afirmarse que para sistemas de

grandes dimensiones, el número de operaciones de la eliminación de Gauss es del órden de $\frac{2n^3}{3}$, es decir $O(n^3)$.

5.4. Casos Especiales para el Método de Gauss.

5.4.1. Definición. Matriz estrictamente diagonal dominante.

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

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

Es decir, una matriz es estrictamente diagonal dominante si los elementos de la diagonal son predominantemente mayores que la suma de los valores absolutos de los elementos fuera de la diagonal en cada fila o columna.

5.4.2. Teorema. Matriz estrictamente diagonal dominante es no singular.

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.

5.4.3. Teorema. Eliminación de Gauss en matrices simétricas y definidas positivas.

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.

5.5. Método de Gauss-Jordan.

El método de Gauss-Jordan es una variante de la eliminación de Gauss en el que se eliminan las incógnitas tanto por encima como por debajo de la diagonal. En este método, la matriz ampliada $[A \mid b]$ se convierte luego de n pasos en la matriz $[I \mid x]$, obteniéndose de este modo la solución x.

5.5.1. Iteración en el paso k.

1. Normalizar $E_k^{(k)}$ dividiéndolo por $a_{kk}^{(k)}$:

$$a_{kj}^{(k+1)} = \frac{a_{kj}^{(k)}}{a_{kk}^{(k)}}, \quad j = k, ..., n$$
$$b_k^{(k+1)} = \frac{b_k^{(k)}}{a_{kk}^{(k)}}$$

2. Eliminación de x_k en las ecuaciones por encima y por debajo de la ecuación k.

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - a_{ik}^{(k)} a_{kj}^{(k+1)}$$
$$b_i^{(k+1)} = b_i^{(k)} - a_{ik}^{(k)} b_k^{(k+1)}$$

para $j = k, ..., n, \quad i = 1, ..., n, \quad i \neq k.$

El método de Gauss-Jordan usa del orden de $\frac{n^3}{2}$ multiplicaciones y divisiones y $\frac{n^3}{2}$ sumas y restas, por lo que requiere un mayor número de operaciones que la eliminación de Gauss.

5.6. Factorización LU.

5.6.1. Factorización LU a partir de la Eliminación Gaussiana.

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

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

Luego definimos la matriz triangular inferior L, basada en los multiplicadores m_{ik} de la eliminación gaussiana:

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

5.6.2. Teorema. Factorización LU.

Sea $A \in \mathbb{R}^{n \times n}$ una matriz no singular, y sean L y U las matrices triangular inferior y triangular superior, definidas anteriormente usando la eliminación de Gauss. Luego, si U es generada sin pivoteo se tiene:

$$A = LU$$

Resolver el sistema $Ax = b$ es equivalente a resolver $LUx = b$, lo cual es equivalente a resolver dos sistemas triangular:

- $Ly = b$: sistema triangular inferior. Se resuelve por sustitución progresiva (hacia abajo).
- $Ux = y$: sistema triangular superior. Se resuelve por sustitución regresiva (hacia arriba).

5.6.3. Cantidad de operaciones y cuestiones del algoritmo.

Para resolver la sustitución progresiva y regresiva se requieren en total alrededor de n^2 multiplicaciones y divisiones y n^2 sumas y restas. Para determinar las matrices L y U se requiere el mismo número de operaciones que las que se necesitan para generar la matriz $A^{(n)} = U$ en la eliminación de Gauss. Sin embargo, una vez que se tiene la factorización, los sistemas en los que intervenga la matriz A pueden resolverse fácilmente para cualquier número de vectores b.

En la programación de la factorización LU a partir de la eliminación gaussiana, los elementos $a_{ij}^{(k+1)}, j \geq i$, siempre se almacenan reemplazando los elementos $a_{ij}^{(k)}$. A medida que los elementos debajo de la diaonal se hacen iguales a cero, es conveniente almacenar en dicho espacio los multiplicadores m_{ij}, ocupando el espacio utilizado originalmente para almacenar los elementos $a_{ij} i > j$.

5.6.4. Unicidad de la factorización LU.

Si $A \in \mathbb{R}^{n\times n}$ es tal que la eliminación de Gauss puede realizarse sin pivoteo, luego 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 $l_{ii} = 1$ y $l_{ij} = m_{ij}$, $i > j$.

Luego, dicha factorización es única.

Demostración.

Demostraremos la unicidad de la factorización LU. Notar que los factores L y U son no singulares (invertibles) 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, luego:

$$\begin{aligned} L_1U_1 &= L_2U_2 \\ L_2^{-1}L_1U_1 &= U_2 \\ L_2^{-1}L_1 &= U_2U_1^{-1} \end{aligned}$$

Sabemos que:

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

Por lo tanto, $L_2^{-1}L_1$ es triangular inferior, mientras que $U_2U_1^{-1}$ es triangular superior.

Luego la ecuación de arriba implica $L_2^{-1}L_1 = U_2U_1^{-1} = D$, siendo D una matriz diagonal (matriz triangular inferior y superior a la vez).

Como $[L_2]_{ii} = [L_2^{-1}]_{ii} = [L_1]_{ii} = 1$, tenemos $D = I$, y por ende $L_1 = L_2$ y $U_1 = U_2$. Por lo tanto, la factorización LU es única.

5.6.5. Definición. 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.

Este tipo de matrices se utiliza para reordenar las filas o columnas de otra matriz. El patrón de unos y ceros asegura que la permutación mantiene la estructura de la matriz pero cambia el orden de las filas o columnas según sea necesario.

Ejemplo.

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

Observación.

En los razonamientos previos hemos supuesto que A es tal que el sistema $Ax = b$ puede resolverse mediante el método de eliminación de Gauss sin intercambios de filas (sin pivoteo). Cuando se requieren intercambios de filas se puede usar una matriz de permutación.

5.6.6. Teorema. Existencia de matriz de permutación.

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, $PA = LU$.

Observación.

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 $Pb = \bar{b}$. Luego se resuelven los sistemas triangulares $Ly = \bar{b}$ y $Ux = y$ por sustitución progresiva y regresiva, respectivamente.

5.6.7. Método de Doolittle.

El **método de Doolittle** es un método de factorización LU denominado compacto, porque es sencillo de programar y requiere poca memoria una vez implementado en el ordenador.

Pasos del método Doolittle.

1. **Inicializar L y U.** Crea L como una matriz de ceros, excepto unos en su diagonal principal y crea U como una matriz de ceros del mismo tamaño.
2. **Calcular los elementos de L y U.** Se utiliza un proceso iterativo para llenar las matrices L y U desde la primera fila y columna hasta la última. Al ser las matrices L y U triangulares, podemos calcular cada elemento de la matriz A como:

$$a_{ij} = \sum_{k=1}^{min(i,j)} l_{ik}u_{kj}$$

El límite superior de la suma se debe a que las matrices triangulares tienen ceros en su mitad inferior o superior, y por tanto, el producto matricial sumaría cero.

En el caso de que $i \leq j$, es decir, el número de fila es más pequeño que el número de la columna, en otras palabras, estamos por encima de la diaggonal (o en la diagonal), entonces $min(i, j) = i$ y tendremos que el último elemento de la sumatoria $l_{ii} = 1$, luego:

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

Es decir, hemos despejado de la sumatoria el sumando cuando $k = i$, u eso nos permite obtener una expresión genérica para calcular u_{ij}.

$$\boxed{\mathrm{u}_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj} \text{ para } i \leq j}$$

Si empezamos en $i = 1$, la ecuación anterior nos dice que $u_{1j} = a_{1j}$, por lo que ya hemos calculado la primera fila de la matriz U. Para seguir calculando filas de U es necesario conocer los valores de L.

Si nos movemos por debajo de la diagonal, $j \leq i$, tendremos que $min(i, j) = j$, y obtenemos:

$$\boxed{\mathrm{l}_{ij} = \frac{1}{u_{jj}}(a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj}), \text{ para } i > j}$$

Por lo tanto, podemos calcular $l_{i1} = \frac{a_{i1}}{u_{11}}$, es decir, hemos calculado ya la primera columna de L. Al conocer la primera columna de L podemos intentar calcular la siguiente fila de U. Al terminar una fila de U, ya conocemos los elementos para calcular la siguiente columna de L. Aplicando sucesivamente este método, obtendremos de manera completa las matrices L y U.

3. **Resolver el sistema usando la factorización LU.** Primero resolver $Ly = b$ para y usando sustitución progresiva y luego $Ux = y$ para x usando sustitución regresiva.

Ejemplo.

Supongamos que tenemos la matriz:

$$A = \begin{bmatrix} 2 & -1 & 1 \\ 4 & 1 & -1 \\ -2 & 2 & 5 \end{bmatrix}$$

Entonces inicializamos las matrices L y U:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \quad U = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

- Calcular $u_{11} = a_{11} = 2$
- Calcular $u_{12} = a_{12} = -1$
- Calcular $u_{13} = a_{13} = 1$
- Calcular $l_{21} = \frac{a_{21}}{u_{11}} = \frac{4}{2} = 2$
- Calcular $l_{31} = \frac{a_{31}}{u_{11}} = \frac{-2}{2} = -1$
- Calcular $u_{22} = a_{22} - l_{21}u_{12} = 1 - 2 \cdot (-1) = 3$
- Calcular $u_{23} = a_{23} - l_{21}u_{13} = -1 - 2 \cdot 1 = -3$
- Calcular $l_{32} = \frac{1}{u_{22}}(a_{32} - l_{31}u_{12}) = \frac{1}{3}(2 - (-1) \cdot (-1)) = \frac{1}{3}$
- Calcular $u_{33} = a_{33} - (l_{31}u_{13} + l_{32}u_{23}) = 5 - (-1 \cdot 1 + \frac{1}{3} \cdot (-3)) = 5 - (-1 + (-1)) = 5 - (-2) = 7$

Entonces obtenemos:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & \frac{1}{3} & 1 \end{bmatrix} \quad U = \begin{bmatrix} 2 & -1 & 1 \\ 0 & 3 & -3 \\ 0 & 0 & 7 \end{bmatrix}$$

5.6.8. 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.

Pasos de la descomposición de Crout.

1. **Inicializar L y U.** L es una matriz de ceros inicialmente. U es una matriz identidad, ya que su diagonal principal se debe llenar de unos.
2. **Calcular los elementos de L y U.** Para los elementos de L en la columna j:

$$\boxed{\mathrm{l}_{ij} = a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj}} \text{ para } i \geq j$$

Para los elementos de U en la fila j:

$$\boxed{\mathrm{u}_{jk} = \frac{1}{l_{jj}} \left(a_{jk} - \sum_{i=1}^{j-1} l_{ji} u_{ik} \right)} \text{ para } k > j$$

Dado que U tiene unos en su diagonal, el cálculo de u_{jk} es sencillo, y se divide por el valor l_{jj}.

Ejemplo.

Ejemplificamos para un sistema de 3×3.

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

1. $l_{11} = a_{11},\ l_{21} = a_{21},\ l_{31} = a_{31}$
2. $u_{12} = \frac{a_{12}}{l_{11}},\ u_{13} = \frac{a_{13}}{l_{11}}$
3. $l_{22} = a_{22} - l_{21} u_{12},\ l_{32} = a_{32} - l_{31} u_{12}$
4. $u_{23} = \frac{a_{23} - l_{21} u_{13}}{l_{22}},\ l_{33} = a_{33} - l_{31} u_{13} - l_{32} u_{23}$

5.7. Factorización de Cholesky.

5.7.1. Teorema. Factorización de Cholesky.

La matriz $A \in \mathbb{R}^{n \times n}$ es definida positiva (sus autovalores son positivos) si y sólo si existe una única matriz triangular superior U con elementos diagonales positivos tal que $A = U^t U$. Esta es la factorización de Cholesky de A.

Resolver el sistema $Ax = b$ es equivalente a resolver $U^t U x = b$, lo cual es equivalente a resolver dos sistemas triangulares:

- $U^t y = b$: Sistema triangular inferior. Se resuelve por sustitución progresiva.
- $Ux = y$: Sistema triangular superior. Se resuelve por sustitución regresiva.

5.7.2. Ejemplos.

- La matriz identidad se puede escribir como $I = I^t I$, siendo I triangular superior invertible. Luego, existe la factorización de Cholesky para la matriz identidad.
- Si existe la factorización de Cholesky de una matriz, al ser U y U^t invertibles, entonces A debe ser invertible. Luego, la matriz nula no tiene factorización de Cholesky.
- Sea $A = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix}$, entonces:

$$\begin{bmatrix} u_{11} & 0 \\ u_{12} & u_{22} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix}$$

De donde surge el siguiente sistema de ecuaciones:

$$\begin{aligned} u_{11}^2 &= 1 \\ u_{11}u_{12} &= 2 \\ u_{12}^2 + u_{22}^2 &= 5 \end{aligned}$$

Y se deduce que $u_{11} = 1,\ u_{12} = 2,\ u_{22} = 1$. Por lo tanto:

$$U = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$$

Es decir:

$$U^tU = A \Rightarrow \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix}$$

Cuando se calculó u_{11} se hubiera podido tomar $u_{11} = -1$ y se hubiera podido obtener otra matriz U. Se puede demostrar que si se escogen los elementos diagonales u_{ii} positivos, entonces la factorización, cuando existe, es única.

- Sea $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$, entonces:

$$\begin{bmatrix} u_{11} & 0 \\ u_{12} & u_{22} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

De donde surge el siguiente sistema de ecuaciones:

$$\begin{aligned} u_{11}^2 &= 1 \\ u_{11}u_{12} &= 2 \\ u_{12}^2 + u_{22}^2 &= 4 \end{aligned}$$

Y se deduce que $u_{11} = 1,\ u_{12} = 2,\ u_{22} = 0$. Por lo tanto:

$$U = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}$$

Entonces, aunque existe U tal que $A = U^tU$, sin embargo no existe la factorización de Cholesky de A ya que U no es invertible.

5.7.3. Algoritmo.

Dada una matriz A de dimensión $n \times n$, queremos encontrar una matriz U tal que $A = U^tU$. Los elementos de U se calculan de la siguiente manera:

1. **Elementos diagonales de U:** Para cada $k = 2, ...n$, la siguiente fórmula calcula el valor de cada elemento en la diagonal de U.

$$\boxed{\mathrm{U}_{kk} = \sqrt{A_{kk} - \sum_{i=1}^{k-1} U_{ik}^2}}$$

2. **Elementos fuera de la diagonal (encima de la diagonal principal):** para cada $k = 2, ...n, \quad j = k+1, ...n$, la siguiente fórmula calcula los elementos de L que están por encima de la diagonal principal.

$$\boxed{\mathrm{U}_{kj} = \frac{1}{U_{kk}} \left(A_{kj} - \sum_{i=1}^{k-1} U_{ik} U_{ij} \right)}$$

5.7.4. Ejemplo usando algoritmo.

Sea $A = \begin{bmatrix} 16 & -12 & 8 & -16 \\ -12 & 18 & -6 & 9 \\ 8 & -6 & 5 & -10 \\ -16 & 9 & -10 & 46 \end{bmatrix}$, entonces:

- $u_{11} = \sqrt{a_{11}} = \sqrt{16} = 4$
- $u_{12} = \dfrac{a_{12}}{u_{11}} = \dfrac{-12}{4} = -3$
- $u_{13} = \dfrac{a_{13}}{u_{11}} = \dfrac{8}{4} = 2$
- $u_{14} = \dfrac{a_{14}}{u_{11}} = \dfrac{-16}{4} = -4$
- $u_{22} = \sqrt{a_{22} - u_{12}^2} = \sqrt{18 - (-3)^2} = \sqrt{9} = 3$
- $u_{23} = \dfrac{a_{23} - u_{12}u_{13}}{u_{22}} = \dfrac{-6 - ((-3) \cdot 2)}{3} = \dfrac{-6+6}{3} = 0$
- $u_{24} = \dfrac{a_{24} - u_{12}u_{14}}{u_{22}} = \dfrac{9 - ((-3) \cdot (-4))}{3} = \dfrac{9-12}{3} = -1$
- $u_{33} = \sqrt{a_{33} - u_{13}^2 - u_{23}^2} = \sqrt{5 - 2^2 - 0^2} = \sqrt{1} = 1$
- $u_{34} = \dfrac{a_{34} - u_{13}u_{14} - u_{23}u_{24}}{u_{33}} = \dfrac{-10 - (2 \cdot (-4))}{1} = \dfrac{-10+8}{1} = -2$
- $u_{44} = \sqrt{a_{44} - u_{14}^2 - u_{24}^2 - u_{34}^2} = \sqrt{46 - (-4)^2 - (-1)^2 - (-2)^2} = \sqrt{46 - 16 - 1 - 4} = \sqrt{25} = 5$

Entonces:

$$U = \begin{bmatrix} 4 & -3 & 2 & -4 \\ 0 & 3 & 0 & -1 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 5 \end{bmatrix}$$

5.8. Factorización QR.

5.8.1. Definición. Factorización QR.

La **factorización QR** descompone una matriz A en el producto de una matriz ortogonal Q y una matriz triangular superior R, de forma que:

$$A = QR$$

- Q es una matriz ortogonal u ortonormal (sus columnas son vectores ortonormaleS).

- R es una matriz triangular superior.

Esta factorización es útil en varios algoritmos numéricos, como el cálculo de autovalores y la resolución de sistemas de ecuaciones lineales sobredeterminados (cuando hay más ecuaciones que incógnitas).

5.8.2. Explicación de la factorización.

Sea $A \in \mathbb{R}^{m\times n}$ tal que $A = [a_1 \mid a_2 \mid ... \mid a_n]$ una matriz con columnas $a_1, a_2, ...a_n$ linealmente independientes (esto implica $m \geq n$, es decir A puede tener más filas que columnas).

Aplicando Gram-Schmidt a las columnas de A, resulta una base ortogonal normalizada $\{q_1, q_2, ...q_n\}$ del espacio columna de A, donde:

- $q_1 = \dfrac{a_1}{v_1}$
- $q_k = \dfrac{a_k - \sum_{i=1}^{k-1}(a_k^t q_i)q_i}{v_k}, \quad k = 2, ...n$
- $v_1 = \|a_1\|$
- $v_k = \|a_k - \sum_{i=1}^{k-1}(a_k^t q_i)q_i\|$

Reordenando estas ecuaciones, tenemos:

- $a_1 = v_1 q_1$
- $a_k = (a_k^t q_1)q_1 + ... + (a_k^t q_{k-1})q_{k-1} + v_k q_k, \quad k = 2, ...n$

En forma matricial:

$$[a_1 \mid a_2 \mid ... \mid a_n] = QR = [q_1 \mid q_2 \mid ... \mid q_n] \begin{bmatrix} v_1 & a_2^t q_1 & a_3^t q_1 & \cdots & a_n^t q_1 \\ 0 & v_2 & a_3^t q_2 & \cdots & a_n^t q_2 \\ 0 & 0 & v_3 & \cdots & a_n^t q_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & v_n \end{bmatrix}$$

- $A \in \mathbb{R}^{m\times n}$ es una matriz con columnas linealmente independientes.
- $Q \in \mathbb{R}^{m\times n}$ tiene por columnas a los vectores de la base ortogonal normalizada del espacio columna de A.
- $R \in \mathbb{R}^{n\times n}$ es una matriz triangular superior con elementos diagonales positivos (la norma siempre es positiva).

5.8.3. Teorema. 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.

Luego, el sistema $Ax = QRx = b$ es equivalente al siguiente sistema triangular superior que se resuelve por sustitución regresiva:

$$Rx = Q^t b$$

6. Normas Vectoriales y Matriciales.

6.1. Normas Vectoriales.

6.1.1. Definición. Norma vectorial.

Dado un espacio vectorial $\mathbb{V}$, una función $\|\cdot\| : \mathbb{V} \to \mathbb{R}$ es una **norma vectorial** si satisface las siguientes propiedades:

1. $\|x\| \geq 0 \ \forall x \in \mathbb{V}$ y $\|x\| = 0 \Leftrightarrow x = 0$
2. $\|\lambda x\| = |\lambda|\|x\|, \quad \forall \lambda \in \mathbb{R}, x \in \mathbb{V}$
3. $\|x + y\| \leq \|x\| + \|y\|, \quad \forall x, y \in \mathbb{V}$

Además, definimos la **distancia** entre los vectores x e y como $\|x - y\|$.

6.1.2. Diferentes tipos de normas vectoriales.

- **Norma euclídea:** para un vector $x \in \mathbb{R}^n$ se define la norma euclídea como:

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}} = \sqrt{x^t x}$$

- **Norma infinito:** para un vector $x \in \mathbb{R}^n$ se define la norma infinito como:

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

- **Norma p:** para un vector $x \in \mathbb{R}^n$ se define la norma p como:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$$

- **Norma 1:** para un vector $x \in \mathbb{R}^n$ se define la norma 1 como:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

Observación.

Notar que la norma 2 (euclídea) y la norma 1 son dos tipos de norma p con $p = 2$ y $p = 1$, respectivamente.

Además, la norma euclídea representa la noción común de distancia respecto del origen en caso de que x pertenezca a $\mathbb{R}, \mathbb{R}^2, \mathbb{R}^3$. Por ejemplo, la norma euclídea del vector $x = [x_1 \ x_2 \ x_3]^t$ denota la longitud del segmento de recta que une los puntos $(0, 0, 0)$ y (x_1, x_2, x_3).

Ejemplo. Sea $x = [1 \ 2 \ 3]^t$. Luego:

- $\|x\|_2 = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{1 + 4 + 9} = \sqrt{14} \approx 3{,}74$
- $\|x\|_\infty = 3$
- $\|x\|_1 = |1| + |2| + |3| = 6$

Región de normas vectoriales habituales.

En la siguiente figuta se representa la región $R = \{x \in \mathbb{R}^2 : \|x\| \leq 1\}$ para la norma euclídea, infinito y norma 1.

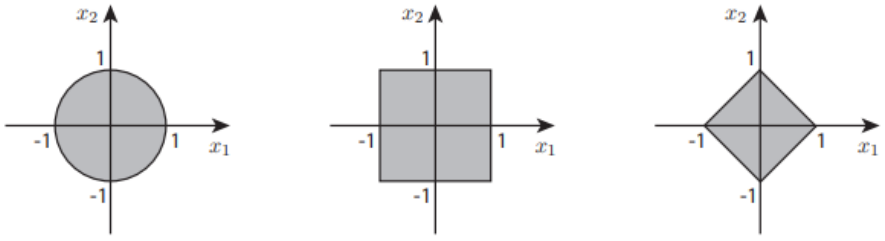


Figura 1: Normas vectoriales habituales.

6.1.3. Teorema. Equivalencia de normas.

Sean N y M normas vectoriales en $\mathbb{V} = \mathbb{R}^n$. Luego, existen constantes $c_1, c_2 > 0$ tales que:

$$c_1 M(x) \leq N(x) \leq c_2 M(x)$$

Es decir, no importa cuál norma elijamos en un espacio de dimensión finita, ya que siempre están acotadas entre sí por constantes positivas. Esto asegura que ninguna norma 'explota' o 'colapsa' más rápido que otra.

6.1.4. Definición. Sucesión de vectores convergente.

Se dice que una sucesión de vectores $\{x^{(1)}, x^{(2)}, ...x^{(m)}, ...\}$ en $\mathbb{R}^n$ **converge** a un vector x si y sólo si:

$$\|x - x^{(m)}\| \to 0 \text{ cuando } m \to \infty$$

Observación. Notar que no se especifica la elección de la norma. Para espacios vectoriales de dimensión finita no importa qué norma se usa.

6.2. Normas Matriciales.

El conjunto de todas las matrices de $m \times n$ con entradas reales puede considerarse equivalente al espacio vectorial $\mathbb{R}^{m\times n}$. Por lo tanto, las normas matriciales satisfacen las mismas propiedades que las normas vectoriales.

6.2.1. Definición. Norma matricial.

Sea $A \in \mathbb{R}^{m\times n}$. Dado un espacio vectorial $V = \mathbb{R}^{m\times n}$, una función $\|\cdot\| : \mathbb{V} \to \mathbb{R}$ es una **norma matricial** si satisface las siguientes propiedades:

1. $\|A\| \geq 0,\ \forall A \in \mathbb{V}$, y $\|A\| = 0 \Leftrightarrow A = 0$
2. $\|\lambda A\| = |\lambda|\|A\|, \quad \forall \lambda \in \mathbb{R}, A \in \mathbb{V}$
3. $\|A + B\| \leq \|A\| + \|B\|, \quad \forall A, B \in \mathbb{V}$

6.2.2. Definición. Norma matricial consistente.

Se dice que la norma matricial $\|\cdot\|$ es **consistente** con las normas vectoriales $\|\cdot\|_a$ en $\mathbb{R}^n$ y $\|\cdot\|_b$ en $\mathbb{R}^m$ si:

$$\|Ax\|_b \leq \|A\|\|x\|_a, \quad \forall x \in \mathbb{R}^n, A \in \mathbb{R}^{m\times n}$$

6.2.3. Definición. Norma matricial submultiplicativa.

Para una matriz cuadrada $A \in \mathbb{R}^{n\times n}$, se dice que la norma matricial $\|\cdot\|$ es **submultiplicativa** si:

$$\|AB\| \leq \|A\|\|B\|, \quad \forall A, B \in \mathbb{R}^{n\times n}$$

6.2.4. Definición. Norma matricial inducida.

Dada una norma vectorial, se define la norma matricial inducida para $A \in \mathbb{R}^{m\times n}$ como:

$$\|A\| = \text{máx}\{\|Ax\| : x \in \mathbb{R}^n, \|x\| = 1\}$$

$$\|A\| = \text{máx}\{\tfrac{\|Ax\|}{\|x\|} : x \in \mathbb{R}^n, x \neq 0\}$$

Se pueden estudiar las normas matriciales inducidas por las normas vectoriales 1, 2 e infinito:

- Si la norma vectorial es $\|\cdot\|_1$, luego la norma matricial inducida está dada por:

$$\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|, \quad A \in \mathbb{R}^{m\times n}$$

  Es decir, la norma 1 es el máximo de la suma de los valores absolutos de los elementos de cada columna.

- La expresión de la norma infinito de una matriz es la siguiente:

$$\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|, \quad A \in \mathbb{R}^{m\times n}$$

  Es decir, la norma infinito es el máximo de la suma de los valores absolutos de los elementos de cada fila.

- La expresión de la norma 2 o norma espectral de una matriz es:

$$\|A\| = \max_{\|x\|_2=1} \|Ax\|_2 = \sqrt{\rho(A^tA)}, \quad A \in \mathbb{R}^{m\times n}$$

  Donde $p(A^tA)$ denota el *radio espectral* de la matriz cuadrada A^tA.

Ejemplo. La norma matricial inducida de la matriz identidad es igual a uno:

- $\|I\|_1 = \max_{i\leq j\leq n} \sum_{i=1}^{m} |a_{ij}| = 1$
- $\|I\|_\infty = \max_{i\leq j\leq m} \sum_{i=1}^{n} |a_{ij}| = 1$

6.2.5. Definición. Espectro de una matriz.

Sea $B \in \mathbb{C}^{n\times n}$. Luego, el espectro de B, denotado por $\sigma(B)$, es igual al conjunto de autovalores de B:

$$\sigma(B) = \{\lambda_1, \lambda_2, ...\lambda_n\}$$

6.2.6. Definición. Radio espectral de una matriz.

El radio espectral de una matriz $B \in \mathbb{C}^{n\times n}$ es el máximo de los valores absolutos de los elementos de su espectro, indicándose como $\rho(B)$:

$$\rho(B) = \max_{\lambda\in\sigma(B)} |\lambda|$$

6.3. Teoremas de normas matriciales inducidas.

6.3.1. Teorema. Norma matricial inducida es submultiplicativa.

La norma matricial inducida es submultiplicativa. Es decir, $\|AB\| \leq \|A\|\|B\|$.

Demostración.

$$\|AB\| = \max_{x\neq 0} \frac{\|ABx\|}{\|x\|} = \max_{x\neq 0} \frac{\|A(Bx)\|}{\|Bx\|} \frac{\|Bx\|}{\|x\|}$$

Y sea $y = Bx$ entonces tenemos:

$$\|AB\| = \max_{x\neq 0} \frac{\|Ay\|}{\|y\|} \frac{\|Bx\|}{\|x\|} \leq \max_{x\neq 0} \frac{\|Ay\|}{\|y\|} \cdot \max_{x\neq 0} \frac{\|Bx\|}{\|x\|} = \|A\| \cdot \|B\|$$

Luego, $\|AB\| \leq \|A\|\|B\|$

6.3.2. Teorema. Desigualdad de norma vectorial.

Sea $A \in \mathbb{R}^{n\times n}$ y sea $\|\cdot\|$ una norma vectorial, luego:

$$\|Ax\| \leq \|A\|\|x\|, \quad \forall x \in \mathbb{R}^n$$

Demostración.

Para $x = 0$ la desigualdad se verifica trivialmente por igualdad: $0 = 0$.

Supongamos que $x \neq 0$, y sea $v = \dfrac{x}{\|x\|}$, luego $\|v\| = 1$ y se tiene:

$$\|Ax\| = \|Ax\frac{\|x\|}{\|x\|}\| = \|A\|x\|\frac{x}{\|x\|}\| = \|x\|\|Av\| \leq \|x\|\|A\|\|v\| = \|x\|\|A\|$$

Donde el $\leq$ vale por el teorema anterior. Luego, $\|Ax\| \leq \|x\|\|A\|$.

6.3.3. Teorema. Radio espectral menor o igual que norma matricial inducida.

Sea $A \in \mathbb{R}^{n\times n}$. Luego para cualquier norma matricial submultiplicativa vale:

$$\rho(A) \leq \|A\|$$

Es decir, el máximo de los autovalores de A es menor o igual que la norma matricial inducida de A.

Demostración.

Sea (λ, v) un par autovalor-autovector de A. Sea la matriz:

$$X = [\ v \mid 0 \mid ... \mid 0\ ]_{n\times n} \neq 0_{n\times n}$$

Sabiendo que $Av = \lambda v$, donde entonces:

$$\begin{aligned} \lambda X = AX \Rightarrow |\lambda|\|X\| &= \|\lambda X\| \\ &= \|AX\| \\ &\leq \|A\|\|X\| \quad (A\ submult.) \end{aligned}$$

Es decir, tenemos que:

$$|\lambda|\|X\| \leq \|A\|\|X\| \Rightarrow |\lambda| \leq \|A\|, \quad \forall \lambda \in \sigma(A)$$

Como $|\lambda| \leq \|A\|$, en particular máx $|\lambda| \leq \|A\|$ y por lo tanto $\rho(A) \leq \|A\|$.

6.3.4. Teorema. Norma matricial inducida por norma 1.

Sea $A \in \mathbb{R}^{m\times n}$. Entonces:

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|$$

Es decir, la norma 1 es el máximo de la suma de los valores absolutos de los elementos de cada columna.

Demostración.

Por definición, $\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1$ (norma matricial - norma vectorial). Luego:

$$\begin{aligned}
\|Ax\|_1 &= \sum_i |A_{i:}x| = \sum_i |\sum_j a_{ij}x_j| \\
&\leq \sum_i \sum_j |a_{ij}||x_j| \quad (\textit{des. triangular}) \\
&= \sum_j \left(|x_j| \sum_i |a_{ij}| \right) \\
&\leq \left(\sum_j |x_j| \right) \left(\max_j \sum_i |a_i j| \right)
\end{aligned}$$

Luego tenemos

$$\|Ax\|_1 \leq \max_j \sum_i |a_{ij}|$$

sujeto a $\|x\|_1 = 1$. Luego, la igualdad se logra ya que si $A_{:k}$ es la columna con mayor suma absoluta, y tomamos $x = e_k$ (notar que $\|e_k\| = 1$, obtenemos:

$$\|Ae_k\|_1 = \|A_{:k}\|_1 = \max_j \sum_i |a_{ij}|$$

Es decir, hallamos un valor de x, con $\|x\|_1 = 1$, tal que la desigualdad se cumple con igualdad.

6.3.5. Teorema. Norma matricial inducida por norma infinito.

Sea $A \in \mathbb{R}^{m \times n}$. Entonces:

$$\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|$$

Es decir, la norma infinito es el máximo de la suma de los valores absolutos de los elementos de cada fila.

Demostración.

Por definición, $\|A\|_\infty = \max_{\|x\|_1=1} \|Ax\|_\infty$. Luego:

$$\begin{aligned}
\|Ax\|_\infty &= \max_{1 \leq i \leq m} |(Ax)_i| \\
&= \max_{1 \leq i \leq m} |\sum_{j=1}^{n} a_{ij}x_j| \quad (\textit{producto matricial}) \\
&\leq \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}| \ \max_{1 \leq j \leq n} |x_j| \quad (\textit{prop. de maximos})
\end{aligned}$$

Dado que $\max_{i \leq j \leq n} |x_j| = \|x\|_\infty = 1$, tenemos que:

$$\|A\|_\infty \leq \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|$$

Ahora necesitamos demostrar la desigualdad opuesta para que valga la igualdad, es decir, debemos demostrar que $\|A\|_\infty \geq \max\limits_{1\leq i\leq m} \sum\limits_{j=1}^{n} |a_{ij}|$.

Sea p un entero tal que:

$$\sum_{j=1}^{n} |a_{pj}| = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|$$

Y sea x el vector con las componentes:

$$x_j = \begin{cases} 1 & si \quad a_{pj} \geq 0 \\ -1 & si \quad a_{pj} < 0 \end{cases}$$

Entonces $\|x\|_\infty = 1$ (maximo) y luego $a_{pj}x_j = |a_{pj}|$ para todo $j = 1, 2, ...n$ asi que:

$$\begin{aligned} \|Ax\|_\infty &= \max_{1\leq i\leq m} |\sum_{j=1}^{n} a_{ij}x_j| \\ &\geq |\sum_{j=1}^{n} a_{pj}x_j| \\ &= |\sum_{j=1}^{n} |a_{pj}|| \\ &= \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}| \end{aligned}$$

Este resultado implica que $\|A\|_\infty \geq \max\limits_{1\leq i\leq m} \sum\limits_{j=1}^{n} |a_{ij}|$.

Lo cual, junto con la otra desigualdad nos da:

$$\|A\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|$$

6.4. Estabilidad de Resolución de Sistemas de Ecuaciones Lineales.

6.4.1. Introducción y ejemplo de perturbación.

Al resolver un sistema de ecuaciones lineales $Ax = b$, es importante examinar la estabilidad de la solución x con respecto a perturbaciones en el sistema. Consideramos primero el sistema perturbado $A\tilde{x} = \tilde{b}$ que presenta perturbaciones pequeñas en el vector b.

Ejemplo. El siguiente sistema lineal tiene solución $x_1 = 0$ y $x_2 = 0{,}1$.

$$\begin{aligned} 5x_1 + 7x_2 &= 0{,}7 \\ 7x_1 + 10x_2 &= 1 \end{aligned}$$

Y el siguiente sistema perturbado tiene solución $\tilde{x}_1 = -0{,}17$ y $\tilde{x}_2 = 0{,}22$:

$$\begin{aligned} 5\tilde{x}_1 + 7\tilde{x}_2 &= 0{,}69 \\ 7\tilde{x}_1 + 10\tilde{x}_2 &= 1{,}01 \end{aligned}$$

Vemos que un cambio relativamente pequeño en b produce un cambio relativamente grande en la solución.

6.4.2. Teorema. Cota del error relativo de la solución del sistema perturbado.

Sea $A \in \mathbb{R}^{n\times n}$ no singular (es decir invertible). Luego, las soluciones de $Ax = b$ y del sistema perturbado $A\tilde{x} = \tilde{b}$ satisfacen:

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|b - \tilde{b}\|}{\|b\|}$$

Demostración.

$$\begin{aligned} Ax - A\tilde{x} &= b - \tilde{b} \\ A(x - \tilde{x}) &= b - \tilde{b} \\ x - \tilde{x} &= A^{-1}(b - \tilde{b}) \\ \|x - \tilde{x}\| &= \|A^{-1}(b - \tilde{b})\| \leq \|A^{-1}\|\|b - \tilde{b}\| \end{aligned}$$

Luego, dividimos ambos miembros de la desigualdad por $\|x\|$ y obtenemos:

$$\begin{aligned} \frac{\|x - \tilde{x}\|}{\|x\|} &\leq \frac{\|A^{-1}\|\|b - \tilde{b}\|}{\|x\|} \\ &= \frac{\|A^{-1}\|\|b - \tilde{b}\|}{\|x\|}\frac{\|A\|}{\|A\|} \\ &= \|A\|\|A^{-1}\|\frac{\|b - \tilde{b}\|}{\|A\|\|x\|} \end{aligned}$$

Pero $\|b\| = \|Ax\| \leq \|A\|\|x\|$. Luego se completa la demostración:

$$\frac{\|x-\tilde{x}\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|b-\tilde{b}\|}{\|b\|}$$

6.4.3. Definición. Número de condición de una matriz.

Sea $A \in \mathbb{R}^{n\times n}$ no singular. El número de condición de A, denotado $\kappa(A)$, está dado por:

$$\kappa(A) = \|A\|\|A\|^{-1}$$

donde $\|\cdot\|$ es una norma matricial submultiplicativa.

Observación.

El número de condición surge de la cota anterior de los sistemas perturbados. Este número indica cuán sensible es el resultado de un sistema de ecuaciones lineales respecto a los cambios o errores en los datos de entrada. Específicamente, mide la sensibilidad de la solución x de un sistema $Ax = b$ con respecto a variaciones en la matriz A o en el vector b.

- Si $\kappa(A) \approx 1$, entonces la matriz A se considera *bien condicionada*, lo que significa que pequeñas perturbaciones en los datos de entrada afectarán poco al resultado de la solución x.
- Si $\kappa(A) \gg 1$, entonces A es *mal condicionada*. Esto implica que pequeñas perturbaciones en los datos pueden producir grandes cambios en la solución x.
- Si $\kappa(A)$ es infinito, entonces A es singular y por lo tanto no tiene inversa. En este caso, el sistema $Ax = b$ no tiene una solución única (o no tiene solución en absoluto).

6.4.4. Lema. Característica del número de estabilidad de una matriz.

Sea $A \in \mathbb{R}^{n\times n}$ no singular. El número de condición de A satisface $\kappa(A) \geq 1$.

Demostración.

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\|$$

Luego, $\kappa(A) \geq 1$.

6.4.5. Perturbación de la matriz A.

La desigualdad del *Teorema de cota del error relativo* indica que el error relativo de la solución está acotado por el error relativo del vector b, multiplicado por el número de condición de la matriz.

Por otra parte, si se perturba la matriz A, tenemos el sistema perturbado $\tilde{A}\tilde{x} = b$.

Tomando $\tilde{A} = A + \delta A$ y $\tilde{x} = x + \Delta x$, podemos escribir:

$$
\begin{aligned}
(A + \Delta A)(x + \Delta x) &= b \\
Ax + A\Delta x + \Delta A(x + \Delta x) &= b \\
Ax + A\Delta x + \Delta A(x + \Delta x) &= Ax \\
A\Delta x + \Delta A(x + \Delta x) &= 0 \\
\Delta x &= -A^{-1}\Delta A(x + \Delta x) \\
\|\Delta x\| &= \|A^{-1}\|\|\Delta A\|\|(x + \Delta x)\|
\end{aligned}
$$

Luego, dividiendo primero por $\|x + \Delta x\|$ y luego multiplicando por $\frac{\|A\|}{\|A\|}$ obtenemos:

$$
\begin{aligned}
\frac{\|\Delta x\|}{\|x + \Delta x\|} &= \|A^{-1}\|\|\Delta A\|\frac{\|(x + \Delta x)\|}{\|x + \Delta x\|} \\
&= \|A^{-1}\|\|\Delta A\| \\
&= \|A^{-1}\|\|\Delta A\|\frac{\|A\|}{\|A\|} \\
&= \|A^{-1}\|\|A\|\frac{\|\Delta A\|}{\|A\|} \\
&= \kappa(A)\frac{\|\Delta A\|}{\|A\|}
\end{aligned}
$$

Notar que $\frac{\|\Delta A\|}{\|A\|}$ es el error relativo en la matriz, mientras que $\frac{\|\Delta x\|}{\|x + \Delta x\|}$ es similar al error relativo en la solución.

7. Resolución de Sistemas 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.

7.1. Método de Jacobi.

El **método de Jacobi** es un método de reemplazos simultáneos. Empezemos con el siguiente ejemplo:

7.1.1. Ejemplo introductorio.

Sea el sistema lineal:

$$\begin{aligned} 9x_1 + x_2 + x_3 &= b_1 \\ 2x_1 + 10x_2 + 3x_3 &= b_2 \\ 3x_1 + 4x_2 + 11x_3 &= b_3 \end{aligned}$$

Despejando x_j de la ecuación j obtenemos:

- $x_1 = \dfrac{1}{9}(b_1 - x_2 - x_3)$
- $x_2 = \dfrac{1}{10}(b_2 - 2x_1 - 3x_3)$
- $x_3 = \dfrac{1}{11}(b_3 - 3x_1 - 4x_2)$

Sea $x^{(0)} = [x_1^{(0)}, x_2^{(0)}, x_3^{(0)}]^t$ una estimación inicial de la solución x (que puede ser $x = 0$). El método de Jacobi define la iteración:

- $x_1^{(k+1)} = \dfrac{1}{9}\left(b_1 - x_2^{(k)} - x_3^{(k)}\right)$
- $x_2^{(k+1)} = \dfrac{1}{10}\left(b_2 - 2x_1^{(k)} - 3x_3^{(k)}\right)$
- $x_3^{(k+1)} = \dfrac{1}{11}\left(b_3 - 3x_1^{(k)} - 4x_2^{(k)}\right)$

7.1.2. Forma general del Método de Jacobi.

En forma general, el sistema a resolver es $Ax = b$. Luego, la ecuación i-ésima es:

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ii}x_i + \cdots + a_{in}x_n = b_i$$

de donde podemos despejar:

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

Luego, el método de Jacobi propone como iteración:

$$\boxed{\mathrm{x}_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1, j\neq i}^{n} a_{ij}x_j^{(k)}\right)} \quad i = 1, ..., n \quad k \geq 0 \tag{21}$$

siendo estas las ecuaciones que se emplean para programar el método. A continuación procederemos a reescribir el sistema de ecuaciones (21) en forma matricial:

$$\begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+2)} \\ \vdots \\ x_n^{(k+1)} \end{bmatrix} = \begin{bmatrix} \frac{1}{a_{11}} & 0 & \cdots & 0 \\ 0 & \frac{1}{a_{12}} & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \frac{1}{a_{nn}} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} + \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & \cdots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \cdots & -\frac{a_{2n}}{a_{22}} \\ \vdots & & \ddots & \vdots \\ -\frac{a_{n1}}{a_{nn}} & \cdots & \cdots & 0 \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \vdots \\ x_n^{(k)} \end{bmatrix} \tag{22}$$

Donde la primera matriz representa $\left(\frac{1}{a_{ii}}b_i\right)$ y la segunda a $\left(-\frac{1}{a_{ii}}\sum_{j=1, j\neq i}^{n} a_{ij}x_j^{(k)}\right)$ (distributiva).

Sea $D \in \mathbb{R}^{n\times n}$ la matriz diagonal de A. Luego:

$$D^{-1} = \begin{bmatrix} \frac{1}{a_{11}} & 0 & \cdots & 0 \\ 0 & \frac{1}{a_{22}} & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \frac{1}{a_{nn}} \end{bmatrix} \qquad (I - D^{-1}A) = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & \cdots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \cdots & -\frac{a_{2n}}{a_{22}} \\ \vdots & & \ddots & \vdots \\ -\frac{a_{n1}}{a_{nn}} & \cdots & \cdots & 0 \end{bmatrix}$$

Luego, el sistema de ecuaciones (22) nos queda:

$$\boxed{x^{(k+1)} = D^{-1}b + (I - D^{-1}A)x^{(k)}} \tag{23}$$

Se puede utilizar como criterio de parada el hecho de que $\|x^{(k+1)} - x^{(k)}\| < \epsilon$ donde ϵ es una tolerancia dada.

7.2. Método de Gauss-Seidel.

El **método de Gauss-Seidel** es un método de reemplazos sucesivos. Es más eficiente que el método de Jacobi en muchos casos porque usa las actualizaciones más recientes de las variables dentro de cada iteración, lo que puede acelerar la convergencia.

Ejemplo introductorio.

Consideremos nuevamente el sistema lineal del Ejemplo anterior. Esta vez utilizamos en forma inmediata la información de cada nuevo componente x_i calculado:

- $x_1^{(k+1)} = \frac{1}{9}\left(b_1 - x_2^{(k)} - x_3^{(k)}\right)$
- $x_2^{(k+1)} = \frac{1}{10}\left(b_2 - 2x_1^{(k+1)} - 3x_3^{(k)}\right)$

- $x_3^{(k+1)} = \dfrac{1}{11}\left(b_3 - 3x_1^{(k+1)} - 4x_2^{(k+1)}\right)$

7.2.1. Forma general del Método de Gauss-Seidel.

En forma general, el método de Gauss-Seidel propone como iteración:

$$\boxed{\begin{aligned} x_1^{(k+1)} &= \frac{1}{a_{ii}}\left(b_1 - \sum_{j=2}^{n} a_{1j}x_j^{(k)}\right) \\ x_i^{(k+1)} &= \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right) \quad i = 2, ..., n-1 \\ x_n^{(k+1)} &= \frac{1}{a_{nn}}\left(b_n - \sum_{j=1}^{n-1} a_{nj}x_j^{(k+1)}\right) \end{aligned}} \tag{24}$$

siendo estas las ecuaciones que se emplean para programar el método. Procederemos ahora a escribir el sistema de ecuaciones (24) en forma matricial. Primero, reescribimos (24) pasando el término $\frac{1}{a_{ii}}$:

$$a_{ii}x_i^{(k+1)} = b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} \quad i = 1, ...n$$

donde la primera sumatoria se anula para $i = 1$, y la segunda sumatoria se anula para $i = n$. A su vez, esta última ecuación se puede reescribir como:

$$\sum_{j=1}^{i} a_{ij}x_j^{(k+1)} = b_i - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)} \quad i = 1, ...n$$

Y en forma matricial esto es:

$$\begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{n1} & \cdots & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ \vdots \\ x_n^{(k+1)} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} - \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ 0 & 0 & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix} \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \vdots \\ x_n^{(k)} \end{bmatrix} \tag{25}$$

Por conveniencia, introducimos la descomposición $A = L + D + U$, donde L es la matriz triangular inferior de A que no incluye la diagonal, D es la diagonal de A, y U es la matriz triangular superior de A que no incluye la diagonal.

$$L = \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ a_{21} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix} \quad D = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix} \quad U = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{bmatrix}$$

Luego, el sistema de ecuaciones (25) se puede escribir como:

$$\boxed{(L + D)x^{(k+1)} = b - Ux^{(k)}}$$

Luego, podemos escribir:

$$\begin{aligned} x^{(k+1)} &= (L+D)^{-1}b - (L+D)^{-1}Ux^{(k)} \\ &= (L+D)^{-1}b - (L+D)^{-1}(A-(L+D))x^{(k)} \\ &= \boxed{(L+D)^{-1}b + (I-(L+D)^{-1}A)x^{(k)}} \end{aligned} \tag{26}$$

7.3. 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:

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

Por lo general, 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:

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

Y el método de Gauss-Seidel se define con:

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

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

7.4. Condiciones de Convergencia.

Vimos que los métodos iterativos se pueden escribir en forma vectorial como:

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

Luego:

$$\begin{aligned} x^{(k+1)} &= N^{-1}((N-A)x^{(k)} + b) \\ &= (I - N^{-1}A)x^{(k)} + N^{-1}b \end{aligned} \tag{27}$$

Por otra parte, la solución del sistema cumple:

$$x = (I - N^{-1}A)x + N^{-1}b \tag{28}$$

Introduciendo el error $e^{(k)} = x - x^{(k)}$, y restando (27) de (28), obtenemos:

$$\boxed{e^{(k+1)} = (I - N^{-1}A)e^{(k)}} \tag{29}$$

7.4.1. Teorema. Condición suficiente de convergencia a partir de matriz del método iterativo.

Si $\|I - N^{-1}A\| < 1$, entonces la sucesión $\{x^{(k)}\}$, definida por el proceso iterativo 27, converge a la solución del sistema $Ax = b$ para cualquier estimación inicial $x^{(0)} \in \mathbb{R}^n$.

Demostración.

Tomando la norma del error (29) tenemos:

$$\begin{aligned}
\|e^{(k+1}\| &= \|(I - N^{-1}A)e^{(k)}\| \leq \|I - N^{-1}A\|\|e^{(k)}\| \\
&= \|I - N^{-1}A\|\|(I - N^{-1}A)e^{(k-1)}\| \\
&\leq \|I - N^{-1}A\|^2\|e^{k-1}\| \cdots \\
&\leq \|I - N^{-1}A\|^{k+1}\|e^{(0)}\|
\end{aligned}$$

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

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

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

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

7.4.2. Teorema. 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 sólo si $\rho(B) < 1$ (radio espectral menor a 1).

7.4.3. Corolario. Condición suficiente y necesaria de convergencia.

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 sólo si $p(I - N^{-1}A) < 1$. Es decir, tomando $B = I - N^{-1}A$ para la B del teorema anterior.

Demostración.

La demostración surge de aplicar el teorema anterior al proceso iterativo dado por la ecuación 29.

7.5. Condiciones de Convergencia para Matrices Diagonalmente Dominantes.

7.5.1. Definición. Matriz diagonalmente dominante.

La matriz $A \in \mathbb{R}^{n\times n}$ es **diagonal dominante** si:

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

7.5.2. Teorema. Matriz diagonal dominante converge por el método de Jacobi.

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

Demostración.

El método de Jacobi usa $N = D = diag(A) = diag(a_{11}, a_{22}, ...a_{nn})$ que suponemos invertible. Luego, usando el esquema general $Nx = Nx - Ax + b$ con $N = D$ tenemos:

$$\begin{aligned} Dx^{(k+1)} &= (D - A)x^{(k)} + b \\ x^{(k+1)} &= (I - D^{-1}A)x^{(k)} + D^{-1}b \end{aligned}$$

Veamos la forma que tiene la matriz $(I - D^{-1}A)$:

$$\begin{array}{ccc|ccc} & \mathsf{D^{-1}A} & & a_{11} & \dots & a_{1n} \\ & & & \vdots & \ddots & \vdots \\ & & & a_{n1} & \dots & a_{nn} \\ \hline \frac{1}{a_{11}} & \dots & 0 & 1 & \dots & \frac{a_{1n}}{a_{11}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{a_{nn}} & \frac{a_{n1}}{a_{nn}} & \dots & 1 \end{array}$$

Luego:

$$\mathbf{I} - \mathsf{D^{-1}A} = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \dots & \dots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} & \dots & -\frac{a_{2n}}{a_{22}} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 & -\frac{a_{n-1,n}}{a_{n-1,n-1}} \\ -\frac{a_{n1}}{a_{nn}} & \dots & \dots & -\frac{a_{n,n-1}}{a_{nn}} & 0 \end{pmatrix}$$

Y se tiene entonces que:

$$\|I - D^{-1}A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1, j\neq i}^{n} |\frac{a_{ij}}{a_{ii}}| \tag{30}$$

Por otra parte, como A es diagonal dominante, entonces:

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

Luego vale:

$$\sum_{j=1, j\neq i}^{n} |\frac{a_{ij}}{a_{ii}}| < 1 \quad i = 1, ...n \tag{31}$$

Combinando (30) y (31), tenemos que:

$$\|I - D^{-1}A\|_\infty < 1$$

Luego, por el teorema de condición suficiente de convergencia a partir de matriz del método iterativo, el método de Jacobi converge a la solución del sistema $Ax = b$ para cualquier vector inicial $x^{(0)} \in \mathbb{R}^n$ siendo A diagonal dominante.

7.5.3. Teorema. Matriz diagonal dominante converge por el método de Gauss-Seidel.

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

Demostración.

Utilizaremos la descomposición $A = L + D + U$ definida anteriormente. Demostraremos que si A es diagonal dominante se cumple que $\rho(I - N^{-1}A) < 1$, con $N = L + D$, es decir, se cumple la condición necesaria y suficiente de convergencia para el método de Gauss-Seidel.

Sea λ un autovalor de $(I - N^{-1}A)$ y v el autovector asociado tal que $\|v\|_\infty = 1$. Nos preguntamos si $|\lambda| < 1$. Veamos que:

$$\begin{aligned} (I - N^{-1}A)v &= \lambda v \\ N(I - N^{-1}A)v &= \lambda Nv \\ Nv - Av &= \lambda Nv \\ -Uv &= \lambda(L + D)v \end{aligned} \tag{32}$$

Veamos la forma que tiene el vector Uv:

$\mathbf{Uv}$							v_1 v_2 $\vdots$ v_i $\vdots$ v_n
0	a_{12}	$\dots$	$\dots$	$\dots$	$\dots$	a_{1n}	$\sum_{j=2}^{n} a_{1j}v_j$
0	0	a_{23}	$\dots$	$\dots$	$\dots$	a_{2n}	$\vdots$
$\vdots$		$\ddots$	$\ddots$			$\vdots$	$\vdots$
$\vdots$			$\ddots$	$a_{i,i+1}$	$\dots$	a_{in}	$\sum_{j=i+1}^{n} a_{ij}v_j$
$\vdots$				$\ddots$	$\ddots$	$\vdots$	$\vdots$
$\vdots$					$\ddots$	$a_{n-1,n}$	$a_{n-1,n}v_n$
0	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	0	0

Veamos ahora la forma que tiene el vector $(L+D)v$:

$(\mathbf{L+D})\mathbf{v}$							v_1 v_2 $\vdots$ v_i $\vdots$ v_n
a_{11}	0	$\dots$	$\dots$	$\dots$	$\dots$	0	$a_{11}v_1$
a_{21}	a_{22}	0				$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\ddots$	$\ddots$			$\vdots$	$\vdots$
a_{i1}	a_{i2}	$\dots$	a_{ii}	0		$\vdots$	$\sum_{j=1}^{i} a_{ij}v_j$
$\vdots$				$\ddots$	$\ddots$	$\vdots$	$\vdots$
$\vdots$					$\ddots$	0	$\vdots$
a_{n1}	a_{n2}	$\dots$	$\dots$	$\dots$	$\dots$	a_{nn}	$\sum_{j=1}^{n} a_{ij}v_j$

Luego, el sistema de ecuaciones (32) se puede escribir como:

$$-\sum_{j=i+1}^{n} a_{ij}v_j = \lambda \sum_{j=1}^{i} a_{ij}v_j \quad i = 1, ...n$$

De donde surge que:

$$
\begin{aligned}
-\sum_{j=i+1}^{n} a_{ij}v_j &= \lambda\left(\sum_{j=1}^{i-1} a_{ij}v_j + a_{ii}v_i\right) \\
-\sum_{j=i+1}^{n} a_{ij}v_j &= \lambda\left(\sum_{j=1}^{i-1} a_{ij}v_j\right) + \lambda a_{ii}v_i \\
-\sum_{j=i+1}^{n} a_{ij}v_j - \lambda\left(\sum_{j=1}^{i-1} a_{ij}v_j\right) &= \lambda a_{ii}v_i
\end{aligned}
$$

Como $\|v\|_\infty = \max_i |v_i| = 1$, luego existe un índice m tal que $|v_m| = 1 \geq |v_j|, \forall j \neq m$. Luego:

$$
\lambda a_{mm}v_m = -\lambda \sum_{j=1}^{m-1} a_{mj}v_j - \sum_{j=m+1}^{n} a_{mj}v_j
$$

Tomando el valor absoluto de la expresión anterior:

$$
\begin{aligned}
|\lambda||a_{mm}| &\leq |\lambda| \sum_{j=1}^{m-1} |a_{mj}||v_j| + \sum_{j=m+1}^{n} |a_{mj}||v_j| \\
|\lambda||a_{mm}| &\leq |\lambda| \sum_{j=1}^{m-1} |a_{mj}| + \sum_{j=m+1}^{n} |a_{mj}| \\
|\lambda|\left(|a_{mm}| - \sum_{j=1}^{m-1} |a_{mj}|\right) &\leq \sum_{j=m+1}^{n} |a_{mj}|
\end{aligned}
\tag{33}
$$

Por otra parte, siendo A diagonal dominante, se tiene que:

$$
\begin{aligned}
|a_{mm}| &> \sum_{j=1, j\neq m}^{n} |a_{mj}| = \sum_{j=1}^{m-1} |a_{mj}| + \sum_{j=m+1}^{n} |a_{mj}| \\
|a_{mm}| - \sum_{j=1}^{m-1} |a_{mj}| &> \sum_{j=m+1}^{n} |a_{mj}|
\end{aligned}
\tag{34}
$$

Combinando (33) y (34) obtenemos:

$$
|\lambda| \leq \frac{\sum_{j=m+1}^{n} |a_{mj}|}{|a_{mm}| - \sum_{j=1}^{m-1} |a_{mj}|} < 1
$$

Con lo cual queda demostrado que para todo autovalor λ, se cumple que $|\lambda| < 1$, es decir, que el radio espectral de $(I-(L+D)^{-1}A)$ es menor que uno. Luego por el corolario de condición suficiente y necesaria de convergencia, el método de Gauss-Seidel converge a la solución del sistema $Ax = b$ para cualquier vector inicial $x^{(0)} \in \mathbb{R}^n$ y A siendo una matriz diagonal dominante.

7.6. Métodos de Relajación.

Los **métodos de relajación** son una clase de métodos iterativos utilizados para resolver sistemas de ecuaciones lineales. En estos métodos, en lugar de usar directamente las nuevas aproximaciones calculadas en cada iteración, se usa una combinación (una relajación) entre la solución anterior y la nueva solución calculada. Esto ayuda a estabilizar el método y puede mejorar la velocidad de convergencia o prevenir oscilaciones.

7.6.1. Método de SOR.

El **método de SOR** es un mejoramiento del método de Gauss-Seidel en el que se introduce un factor de relajación ω, que permite controlar cuánto de la nueva aproximación se usa en cada iteración.

El método de SOR modificando Gauss-Seidel propone la siguiente iteración:

$$\boxed{\begin{aligned} x_1^{(k+1)} &= (1-\omega)x_1^{(k)} + \frac{\omega}{a_{ii}}\left(b_1 - \sum_{j=2}^{n} a_{1j}x_j^{(k)}\right) \\ x_i^{(k+1)} &= (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right) \quad i = 2, ...n-1 \\ x_n^{(k+1)} &= (1-\omega)x_n^{(k)} + \frac{\omega}{a_{nn}}\left(b_n - \sum_{j=1}^{n-1} a_{nj}x_j^{(k+1)}\right) \end{aligned}} \tag{35}$$

donde ω es el **factor de escala.** Podemos distinguir los siguientes casos:

- 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.

Método de SOR en forma matricial.

Podemos reescribir la ecuación (35) como sigue:

$$\begin{aligned} x_i^{(k+1)} - (1-\omega)x_i^{(k)} &= \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right) \\ a_{ii}x_i^{(k+1)} - (1-\omega)a_{ii}x_i^{(k)} &= \omega\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right) \\ a_{ii}x_i^{(k+1)} - (1-\omega)a_{ii}x_i^{(k)} &= \omega b_i - \omega\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \omega\sum_{j=i+1}^{n} a_{ij}x_j^{(k)} \\ a_{ii}x_i^{(k+1)} + \omega\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} &= (1-\omega)a_{ii}x_i^{(k)} - \omega\sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + \omega b_i \end{aligned} \tag{36}$$

Utilizando la descomposición de A como $A = L + D + U$, reescribimos (36) en forma matricial:

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

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

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

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

Error del método de SOR.

El error del método de SOR está determinado por:

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

Luego, por el *Teorema de Estabilidad asintótica de un proceso iterativo lineal*, el método de SOR converge a la solución de $Ax = b$ para todo vector inicial $x^{(0)}$ si y sólo 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 consideramos el caso particular de las matrices definidas positivas y tridiagonales.

7.6.2. Teorema. Omega óptimo para el método de SOR.

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.

8. Aproximación de Autovalores.

8.1. Autovalores y Autovectores.

8.1.1. Definición. Autovalor y autovector.

Sea $A \in \mathbb{R}^{n\times n}$. Si λ es un escalar, y $v \neq 0$ es un vector tales que verifican:

$$Av = \lambda v$$

decimos que λ es un **autovalor** de A y v es un **autovector** asociado.

- Observación 1. Notar que si v es un autovector, luego cv también es un autovector asociado al mismo autovalor, con c escalar real distinto de cero. Pues $A(cv) = c(Av) = c(\lambda v) = \lambda(cv)$.

- Observación 2. Se pueden entender a los autovectores como aquellos vectores que no se salen de su subespacio generado. Es decir, son aquellos vectores v tal que aplicar Av el vector resultante es de la forma λv, es decir, un múltiplo de v.

- Observación 3. Reformulando la ecuación tenemos que:

  $$Av = \lambda v \Rightarrow (Av - \lambda v) = 0 \Rightarrow (A - \lambda I)v = 0$$

  Esta última igualdad muestra que los autovectores son vectores distintos de cero que pertenecen al espacio nulo $N(A - \lambda I)$. Pero $N(A - \lambda I)$ contiene vectores no nulos si y sólo si $A - \lambda I)$ es singular.

  Por lo tanto, los autovalores son precisamente los valores de λ que hacen que la matriz $(A-\lambda I)$ sea singular, o en forma equivalente, los valores de λ para los cuales $det(A - \lambda I) = 0$.

8.1.2. Teorema. Polinomio característico y ecuación característica.

Sea $A \in \mathbb{R}^{n\times n}$. Entonces:

- 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, en 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 $A \in \mathbb{R}^{n\times n}$, luego 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 $det(A) = det(A^t)$.

8.1.3. Definición. Matriz diagonalizable.

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.

8.1.4. Teorema de la diagonalización.

Una matriz $A \in \mathbb{R}^{n\times n}$ es diagonalizable si y sólo si A tiene n autovectores linealmente independientes, o en forma equivalente, los autovectores de A conforman una base de $\mathbb{R}^n$.

Observación. Si la matriz $A \in \mathbb{R}^{n\times n}$ es diagonalizable, por este teorema, los autovectores de A forman una base y luego 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:

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

8.1.5. Ejemplo de autovalores y autovectores.

- Sea la matriz $A = \begin{pmatrix} 1{,}25 & 0{,}75 \\ 0{,}75 & 1{,}25 \end{pmatrix}$

  cuyos autovalores y autovectores son:

$$\lambda_1 = 2{,}0 \qquad v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\lambda_1 = 0{,}5 \qquad v_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

  Los autovectores conforman una base para $\mathbb{R}^2$. Luego, para $x = [x_1, x_2]^t$, podemos escribir este vector genérico como combinación lineal de los autovectores de la siguiente forma:

$$\begin{aligned} x &= c_1 v_1 + c_2 v_2 \\ (x_1, x_2) &= c_1(1,1) + c_2(-1,1) \\ (x_1, x_2) &= (c_1, c_1) + (-c_2, c_2) \\ (x_1, x_2) &= (c_1 - c_2, c_1 + c_2) \end{aligned}$$

  De donde surge el sistema:

$$\begin{aligned} c_1 - c_2 &= x_1 \\ c_1 + c_2 &= x_2 \end{aligned}$$

  Y resolviendolo llegamos a que:

$$c_1 = \frac{x_1 + x_2}{2}, \qquad c_2 = \frac{x_2 - x_1}{2}$$

Aplicando estos c en la ecuación $Ax = \sum_{i=1}^{n} \lambda_i c_i v_i$ tenemos:

$$\begin{aligned} Ax &= \lambda_1 c_1 v_1 + \lambda_2 c_2 v_2 \\ &= 2c_1 v_1 + \frac{1}{2} c_2 v_2 \end{aligned}$$

La interpretación gráfica de la descomposición de x y de Ax es de la siguiente manera:

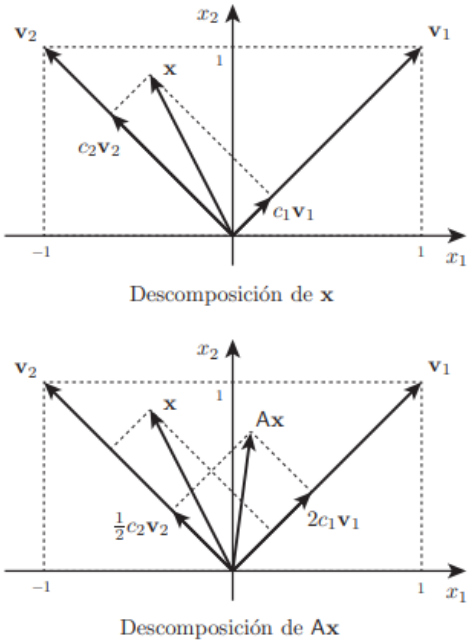


Figura 1: Descomposición en autovectores usando la matriz del Ejemplo 3.

Observación. Consideremos ahora el proceso iterativo $x^{(k+1)} = Ax^{(k)}$. En la siguiente figura puede verse que cuando $k \to \infty$, la componente de $x^(k)$ en la dirección del autovector v_1 (cuyo autovalor es $\lambda_1 = 2$) tiende a infinito. Debe cmplirse que el radio espectral de A sea menor que 1 (máximo autovalor menor que 1) para que ambas componentes tiendan a cero, y por ende $x^{(k)}$ tienda a cero.

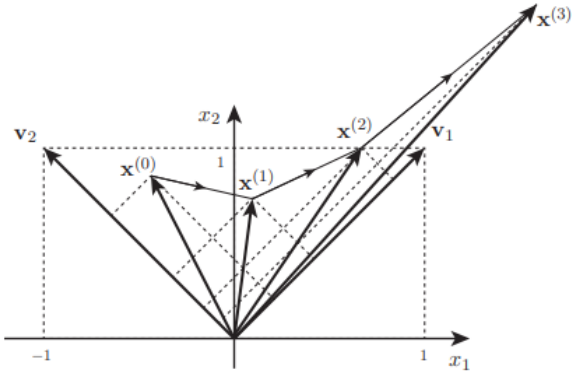


Figura 2: Iteración $\mathbf{x}^{(k+1)} = \mathbf{A}\mathbf{x}^{(k)}$ usando la matriz del Ejemplo 3.

8.2. Teorema de Gershgorin.

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

8.2.1. Definición. Círculos de Gershgorin.

Sea $A \in \mathbb{C}^{n\times n}$. Se definen los **círculos de Gershgorin** como:

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

Donde r_i es el radio del círculo C_i.

Observación. Los círculos de Gershgorin de una matriz A se definen para cada fila i de la matriz, con centro en el elemento diagonal de la fila y radio igual a la suma de los valores absolutos de los elementos no diagonales de esa fila.

8.2.2. Teorema de Gershgorin.

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

Es decir, todos los autovalores de una matriz A están contenidos dentro de la unión de los círculos de Gershgorin.

Demostración.

Queremos demostrar que la distancia entre un autovalor y un elemento diagonal de A es menor al radio del círculo de Gershgorin para esa componente: $|\lambda - a_i| \leq r_i$.

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 (hacemos Av en la fila k de A):

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

$$\sum_{j=1,j\neq k}^{n} a_{kj}v_j + a_{kk}v_k = \lambda v_k$$

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

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

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

$$|\lambda - a_{kk}||v_k| \leq \sum_{j=1,j\neq k}^{n} |a_{kj}||v_j| \quad (des.\ triangular)$$

$$|\lambda - a_{kk}||v_k| \leq |v_k| \sum_{j=1,j\neq k}^{n} |a_{kj}| \quad (|vk| \geq |v_j|)$$

$$|\lambda - a_{kk}| = \sum_{j=1,j\neq k}^{n} |a_{kj}| \qquad (dividir\ por\ |v_k|)$$

$$|\lambda - a_{kk}| = r_k$$

Como queríamos probar.

Observación. 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.

8.2.3. Ejemplos. Autovalores y teorema de Gershgorin.

- Sea la matriz $A = \begin{pmatrix} 1 & 2 \\ 1 & -1 \end{pmatrix}$ cuyos autovalores son $\lambda_1 = \sqrt{3}$ y $\lambda_2 = -\sqrt{3}$.

  De las filas de A obtenemos un círculo de radio 2 centrado en (1, 0) y un círculo de radio 1 centrado en (-1, 0):

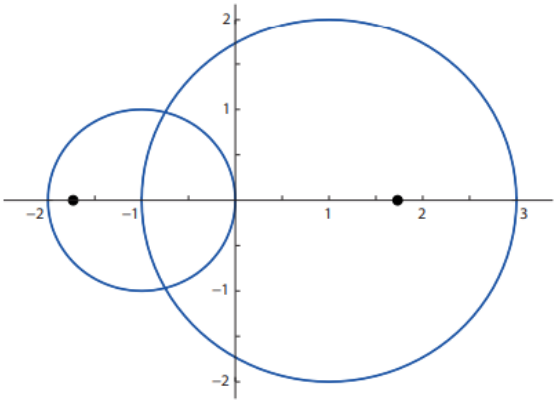


Figura 3: Círculos de Gershgorin de la matriz del Ejemplo 4.

A partir de la figura anterior, podemos concluir que $\rho(A) \leq 3$.

- El Teorema de Gershgorin nos dice que todo autovalor se encuentra en un círculo de Gershgorin. Sin embargo, el Teorema no dice que todo círculo contiene un autovalor.

  Sea la matriz $A = \begin{pmatrix} 1 & -1 \\ 2 & -1 \end{pmatrix}$ cuyos autovalores son $\lambda_1 = i$ y $\lambda_2 = -1$.

  De las filas de A obtenemos un círculo de radio 1 centrado en (1, 0) y un círculo de radio 2 centrado en (-1, 0):

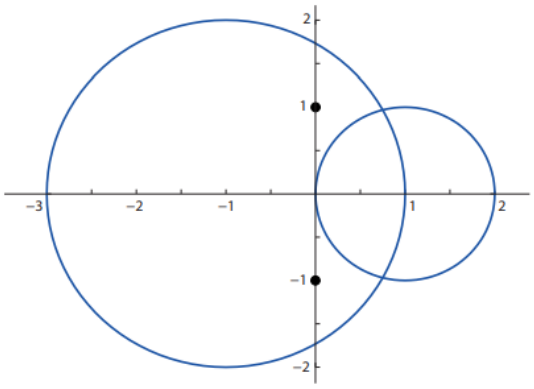


Figura 4: Círculos de Gershgorin de la matriz del Ejemplo 5.

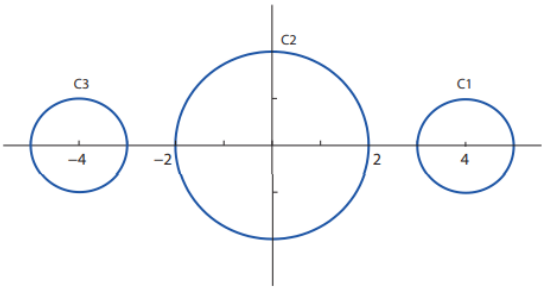
Se puede ver claramente que ambos autovalores se encuentran en el círculo definido por la segunda fila, mientras que ningún autovalor se encuentra en el círculo definido por la primera fila.

- En el caso de las matrices simétricas, sabemos que sus autovalores son reales, y por lo tanto para estas matrices el Teorema de Gershgorin nos da intervalos en el eje de los reales en los que se encuentran los autovalores.

  Sea la matriz simétrica $A = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & -4 \end{pmatrix}$

  Según el Teorema de Gershgorin, los autovalores deben estar contenidos en los intervalos

- $C_1 : |\lambda - 4| \leq 1$
- $C_2 : |\lambda| \leq 2$
- $C_3 : |\lambda + 4| \leq 1$



- Sea la matriz simétrica $A = \begin{pmatrix} 4 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 4 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 4 & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 & 4 & 1 \\ 0 & \cdots & \cdots & 0 & 1 & 4 \end{pmatrix}$

  Siendo una matriz simétrica, sabemos que sus autovalores son reales. Todos los radios r_i son iguales a 1 o 2. Los centros de todos los círculos son $a_{ii} = 4$. Luego, por el Teorema de Gershgorin, los autovalores deben estar en el intervalo [2, 6]. Esta información nos dice que A es invertible y definida positiva.

8.2.4. Aplicación de Teorema de Gershgorin para ver que matrices diagonalmente dominantes son invertibles.

Una matriz es invertible si y sólo si no posee ningún autovalor igual a cero. Por el Teorema de Gershgorin, podemos asegurar que ningún autovalor es igual a cero si el punto (0, 0) en el plano complejo no pertenece a ningún círculo de Gershgorin. Es fácil verifica que esto se cumple si:

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

es decir, si la matriz A es diagonalmente dominante (pues así el radio del círculo nunca superaría al centro, y por lo tanto nunca ocupa el punto (0, 0)).

Empleando el Teorema de Gershgorin, acabamos de demostrar que las matrices diagonalmente dominantes son invertibles.

8.2.5. Corolario. Teorema de Gershgorin en matrices no simétricas.

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

Demostración.

Por el *Teorema de Polinomio característico y ecuación característica* sabemos que A y A^t tienen los mismos autovalores.

Luego, por el Teorema de Gershgorin, los autovalores de A deben encontrarse en los círculos de Gershgorin de las filas de A^t, que son las columnas de A.

Observación. Este corolario nos dice que el Teorema de Gershgorin también es aplicable a las columnas de una matriz. Es decir, en vez de tomar la suma de los valores absolutos de una fila, tomamos los valores de la columna.

Esto nos permite acotar aún más la región en la que se encuentran los autovalores, considerando la intersección de los círculos obtenidos por filas y por columnas.

Para los primeros dos ejemplos de la sección de *Ejemplos de autovalores y teorema de Gershgorin.*, se puede asegurar que los autovalores se encuentran en la región sombrada de la siguiente figura:

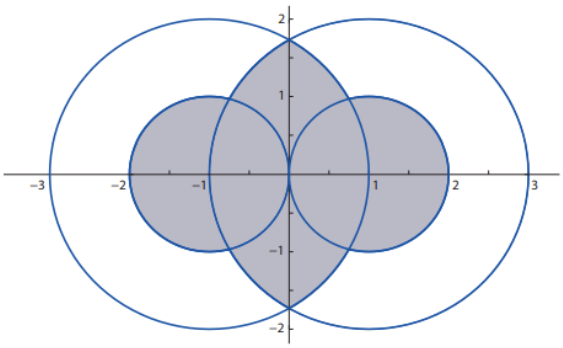


Figura 6: Intersección de los círculos obtenidos por filas y por columnas.

8.2.6. Definición. Grupo disjunto de círculos de Gershgorin.

Se dice que un subconjunto $\mathcal{G}$ de círculos de Gershgorin es un **grupo disjunto de círculos** si ningún círculo en $\mathcal{G}$ intersecta con un círculo que no pertenece a $\mathcal{G}$.

8.2.7. Teorema. Grupo disjunto de círculos y cantidad de autovalores.

Si un grupo disjunto de círculos de Gershgorin $\mathcal{G}$ contiene k círculos, luego $\mathcal{G}$ 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. Es decir, $\bar{A}(p) = p(A - D) + D$.

Luego, si $p = 0$, entonces $\bar{A}(p) = D$ y luego la matriz $\bar{A}(p)$ tendrá 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, los autovalore 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 de Gershgorin.

Ejemplo.

Sea la matriz $A = \begin{pmatrix} 5 & 0 & 0 & -1 \\ 1 & 0 & -1 & 1 \\ -1{,}5 & 1 & -2 & 1 \\ -1 & 1 & 3 & -3 \end{pmatrix}$

Cuyos autovalores son $\lambda_1 \approx 5{,}17$, $\lambda_2 \approx -4{,}15$, $\lambda_3 \approx -1{,}38$ y $\lambda_4 \approx 0{,}35$.

De A obtenemos un círculo de radio 1 centrado en (5, 0), un círculo de radio 3 centrado en (0, 0), un círculo de radio 3,5 centrado en (-2, 0) y un círculo de radio 5 centrado en (-3, 0). Los círculos y los autovalores se encuentran representados en la siguiente figura:

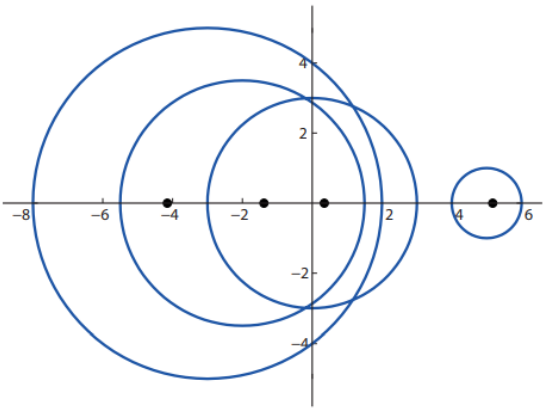


Figura 7: Grupos disjuntos de círculos de Gershgorín.

Como puede verse, tenemos dos grupos disjuntos de círculos. El grupo $\mathcal{G}_1$ conformado por por un círculo centrado en (5, 0) posee 1 autovalor en su interior, y el grupo $\mathcal{G}_2$, de mayor tamaño, conformado por 3 círculos, posee 3 autovalores.

8.3. Método de la Potencia.

El **método de la potencia** es un algoritmo iterativo utilizado para encontrar el autovalor de mayor módulo de una matriz cuadrada A y su correspondiente autovector. Es un método especialmente útil cuando solo se necesita el autovalor dominante de una matriz.

8.3.1. Teorema. Método de la Potencia.

Sea la matriz $A \in \mathbb{R}^{n\times n}$ y sean $\lambda_1, ..., \lambda_n$ 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 λ_1.

Sea $\{v_1, v_2, ...v_n\}$ la base de autovectores correspondiente asociados a los autovalores. Seleccionamos $z^{(0)}$ como una estimación inicial del autovector v_1, elegida como sea, incluso de forma aleatoria. Luego, definimos:

- El vector $w^{(n+1)}$ como el producto de la matriz A y el vector actual $z^{(n)}$:

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

- Luego, normalizamos este nuevo vector $w^{(n+1)}$ para obtener el siguiente vector $z^{(n+1)}$:

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

Entonces resulta que cuando $n \to \infty$, el vector $z^{(n)}$ tiende a aproximarse al autovector dominante v_1, normalizado por su componente de mayor valor absoluto:

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

Y para obtener una estimación del autovalor λ_1, seleccionamos una componente k que no sea nula en $w^{(n-1)}$. Haciendo el siguiente cálculo, conseguiremos una convergencia a λ_1 cuando $n \to \infty$:

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

Demostración.

Primero demostraremos por inducción que:

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

- Verificamos que se cumple para el caso base $n = 1$:

$$z^{(1)} = \frac{w^{(1)}}{\|w^{(1)}\|_\infty} = \frac{Az^{(0)}}{\|Az^{(0)}\|_\infty}$$

- Verificamos que se cumple para el caso base $n = 2$:

$$z^{(2)} = \frac{w^{(2)}}{\|w^{(2)}\|_\infty} = \frac{Az^{(1)}}{\|Az^{(1)}\|_\infty} = \frac{A\frac{Az^{(0)}}{\|Az^{(0)}\|_\infty}}{\left\|A\frac{Az^{(0)}}{\|Az^{(0)}\|_\infty}\right\|_\infty} = \frac{A^2 z^{(0)}}{\|A^2 z^{(0)}\|_\infty}$$

- Veririficamos el caso inductivo. Por inducción fuerte suponemos que vale para todos los números anteriores a n (HI) y probamos el caso para n:

$$z^{(n)} = \frac{w^{(n)}}{\|w^{(n)}\|_\infty} = \frac{Az^{(n-1)}}{\|Az^{(n-1)}\|} = \frac{A\frac{A^{n-1}z^{(0)}}{\|A^{n-1}z^{(0)}\|_\infty}}{\left\|A\frac{A^{n-1}z^{(0)}}{\|A^{n-1}z^{(0)}\|_\infty}\right\|_\infty} = \frac{A^n z^{(0)}}{\|A^n z^{(0)}\|_\infty}$$

  Como queríamos probar.

Siendo $\{v_1, ...v_n\}$ una base de $\mathbb{R}^n$, podemos escribir al vector $z^{(0)}$ como una combinación lineal de estos vectores:

$$z^{(0)} = \sum_{j=1}^{n} \alpha_j v_j$$

Supongamos que $\alpha_1 \neq 0$ (esto siempre es posible pues $z^{(0)}$ es arbitrario). Luego, podemos expresar Az con la definición de z expresado como combinación lineal:

$$\begin{aligned} Az^{(0)} &= A(\alpha_1 v_1 + \cdots + \alpha_n v_n) \\ &= \alpha_1 Av_1 + \cdots + \alpha_n Av_n \\ &= \alpha_1 \lambda_1 v_1 + \cdots + \alpha_n \lambda_n v_n \\ &= \lambda_1 \alpha_1 v_1 + \cdots + \lambda_n \alpha_n v_n \end{aligned}$$

Luego, utilizando esto último también podemos expresar $A^m z$:

$$\begin{aligned} A^m z^{(0)} &= A^m \sum_{j=1}^{n} \alpha_j v_j \\ &= \lambda_1^m \alpha_1 v_1 + \cdots + \lambda_n^m \alpha_n v_n \\ &= \lambda_1^m \left(\alpha_1 v_1 + \sum_{j=2}^{n} \alpha_j \left(\frac{\lambda_j}{\lambda_1}\right)^m v_j \right) \end{aligned}$$

Como $|\lambda_1| > |\lambda_j|$ para todo $j = 2, ...n$, luego $\left(\frac{\lambda_j}{\lambda_1}\right)^m \to 0$ cuando $m \to \infty$ (serie geométrica con $r < 1$). Luego, escribimos $z^{(m)}$ utilizando lo probado por inducción y este último resultado:

$$z^{(m)} = \frac{A^m z^{(0)}}{\|A^m z^{(0)}\|_\infty} = \frac{\lambda_1^m \left(\alpha_1 v_1 + \sum_{j=2}^{n} \alpha_j \left(\frac{\lambda_j}{\lambda_1}\right)^m v_j \right)}{|\lambda_1|^m \left\| \alpha_1 v_1 + \sum_{j=2}^{n} \alpha_j \left(\frac{\lambda_j}{\lambda_1}\right)^m v_j \right\|_\infty}$$

Tomando el límite cuando $m \to \infty$ vemos que la sumatoria se anula y luego obtenemos:

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

$$\beta = \left(\frac{\lambda_1}{|\lambda_1|}\right)^m \frac{\alpha_1}{|\alpha_1|}$$

Y podemos ver que β es una constante de módulo igual a 1, por lo tanto $z^{(m)} \to v_1$. Luego, eligiendo una componente k no nula de $w^{(m)}$ tenemos:

$$\lambda^{(m)} = \frac{w_k^{(m)}}{z_k^{(m-1)}} = \frac{(Az^{(m-1)})_k}{(z^{(m-1)})_k}$$

Y luego, como $z^{(m)} \to v_1$ cuando $m \to \infty$ tenemos:

$$\lambda^{(m)} \to \frac{(Av_1)_k}{(v_1)_k} = \lambda_1 \text{ cuando } m \to \infty$$

Como queríamos probar.

Observación.

Cabe resaltar que el método de la potencia es válido empleando cualquier norma vectorial.

Este método tiene la desventaja de que al inicio no sabe si la matriz tiene o no un solo autovalor dominante. La convergencia del método puede ser lenta, pero con una implementacióna adecuada es efectivo para matrices dispersas de grandes dimensiones.

9. Interpolación Polinómica.

9.1. Problema de Interpolación Polinómica.

9.1.1. Introducción a la interpolación.

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

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

Sea $x_{min} = min\{x_1, x_2, \cdots, x_{n+1}\}$ y $x_{max} = max\{x_1, x_2, \cdots x_{n+1}\}$. 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.**

9.1.2. Interpolación Polinómica.

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

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

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

$$p(x_i) = y_i, \quad i = 0, 1 \cdots n \tag{37}$$

¿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 $a_0, a_1, \cdots a_m$. Puesto que (37) impone $n+1$ condiciones sobre $p(x)$, es razonable considerar el caso en que $m = n$. Es decir, queremos encontrar $a_0, a_1, \cdots a_n$ tales que:

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

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

$$Xa = y$$

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

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

$$det(X) = \prod_{0 \leq j < i \leq n} (x_i - x_j) \tag{38}$$

Por ejemplo, para una matriz X con 4 nodos x_0, x_1, x_2, x_3, tenemos:

$$det(X) = (x_3 - x_0)(x_2 - x_0)(x_1 - x_0)(x_3 - x_1)(x_2 - x_1)(x_3 - x_2)$$

9.1.3. Teorema. Existencia y unicidad del polinomio interpolante.

Dados $n+1$ puntos distintos $(x_0, y_0), (x_1, y_1), \cdots, (x_n, y_n)$ con $x_0, x_1, \cdots x_n$ 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.

Las siguientes demostraciones concluyen que existe un único polinomio interpolante de grado menor o igual a n que pasa por los $n+1$ puntos dados. La primer demostración utiliza la propiedad de invertibilidad de la matriz de Vandermonde y la segunda utiliza el teorema fundamental del álgebra al suponer una contradicción.

Demostración A. Considerando la expresión del determinante de la matriz de Vandemonde, dado en la ecuación (38), se tiene que $det(X) \neq 0$, porque si $i \neq j$, entonces $x_i \neq x_j$. Luego, X es no singular (invertible) y el sistema $Xa = y$ tiene solución única.

El grado del polinomio interpolante puede ser menor o igual a n para todos los $n+1$ puntos dados, ya que algunos de los coeficientes $a_i, i = 0, 1 \cdots n$, pueden ser iguales a cero (si el coeficiente a_n es cero, entonces el polinomio tiene grado menor que n).

Demostración B. Por contradicción, supongamos que X no es invertible, es decir, que rango(X) $< n+1$. Luego X tiene un espacio nulo no trivial, es decir, $N(A) \neq \{0\}$.

En particular, existe un vector $a = (a_0, a_1, \cdots a_n)^t \in \mathbb{R}^{n+1}$ con $a \neq 0$, tal que $Xa = 0$. Escribiendo el polinomio para este sistema, tenemos:

$$p(x_i) = a_0 + a_1 x_i + \cdots + a_n x_i^n = 0, \quad i = 0, 1, \cdots n \tag{39}$$

Como a_n y otros coeficientes pueden ser iguales a cero, el grado m de $p(x)$ puede ser $m \leq n$. Las ecuaciones (39) indican que $p(x)$ se anula en los $n+1$ puntos $x_0, x_1, \cdots x_n$.

Sin embargo, por el Teorema Fundamental del Álgebra, un polinomio de grado m tiene como máximo m raíces distintas. Esto es una contradicción ya que $m < n+1$. Esta contradicción surge de suponer que X no es invertible. Luego, rango(X) $= n+1$ y esto garantiza que el sistema $Xa = y$ tiene solución única, que corresponde a los coeficientes del polinomio único $p(x)$ que interpola los $n+1$ puntos.

9.1.4. Limitaciones computacionales de la matriz de Vandermonde.

La matriz de Vandermonde X es no singular pero está mal condicionada. Por lo general, a medida que el número de puntos de interpolación aumenta, $det(X)$ tiende a cero.

Ejemplo. Considerar el caso de 4 nodos equiespaciados en el intervalo [0, 1]. Tenemos:

$$x_0 = 0, \quad x_1 = \frac{1}{3}, \quad x_2 = \frac{2}{3}, \quad x_3 = 1$$

Luego, utilizando la fórmula del determinante de la matriz de Vandermonde:

$$\begin{aligned} det(X) &= (x_3 - x_0)(x_2 - x0)(x_1 - x_0)(x_3 - x_1)(x_2 - x_1)(x_3 - x_2) \\ &= 1 \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} = 0{,}016 \end{aligned}$$

Es posible obtener el polinomio interpolante resolviendo el sistema $Xa = y$. Sin embargo, este sistema puede estar mal condicionado lo cual puede conducir a errores numéricos. Por otra parte, es más sencillo calcular el polinomio interpolante a partir de la fórmula de interpolación de Lagrange.

9.2. Interpolación de Lagrange.

9.2.1. Caso lineal.

Queremos 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:

- $L_0(x) = \dfrac{x - x_1}{x_0 - x_1}$
- $L_1(x) = \dfrac{x - x_0}{x_1 - x_0}$

Luego se define el polinomio interpolante p como:

$$\boxed{p(x) = L_0(x)f(x_0) + L_1(x)f(x_1)}$$

Evaluando las funciones L en x_0 y x_1 observamos que:

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

Y por lo tanto tenemos que:

$$p(x_0) = L_0(x_0)f(x_0) + L_1(x_0)f(x_1) = f(x_0) = y_0$$
$$p(x_1) = L_0(x_1)f(x_0) + L_1(x_1)f(x_1) = f(x_1) = y_1$$

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

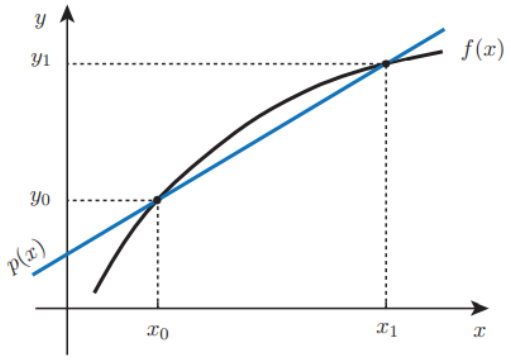


Figura 1: Interpolación lineal.

9.2.2. Caso general.

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

Para $k = 0, 1, \cdots n$, definimos:

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

Podemos ver que $L_k(x)$ satisface $L_k(x_i) = 0$ si $i \neq k$ y $L_k(x_k) = 1$.

Luego, el polinomio interpolador de Lagrange está dado por:

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

Notar que $p(x)$ interpola los datos, pues $p(x_i) = y_i$ para $i = 0, 1, \cdots n$.

Además, el grado de $p(x)$ es menor o igual a n ya que el grado de $L_k(x)$ es igual a n. Luego, $p(x)$ es el único en el conjunto de polinomios de grado menor o iual a n, de acuerdo al Teorema anterior.

9.2.3. Ejemplo de polinomio interpolador de Lagrange.

Queremos obtener $p_2(x)$ de grado 2, que pase por los puntos $(0, -1), (1, -1), (2, 7)$. Aplicando la fórmula de Lagrange obtenemos:

$$L_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} = \frac{(x - 1)(x - 2)}{(0 - 1)(0 - 2)} = \frac{(x - 1)(x - 2)}{2}$$

$$L_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} = \frac{(x - 0)(x - 2)}{(1 - 0)(1 - 2)} = \frac{x(x - 2)}{-1}$$

$$L_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \frac{x(x - 1)}{(2 - 0)(2 - 1)} = \frac{x(x - 1)}{2}$$

Y luego el polinomio $p_2(x)$ resulta:

$$p(x) = L_0(x)y_0 + L_1(x)y_1 + L_2(x)y_2 = \frac{(x - 1)(x - 2)}{2}(-1) + \frac{x(x - 2)}{-1}(-1) + \frac{x(x - 1)}{2}7$$

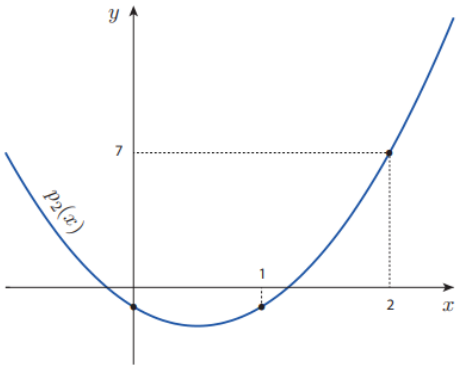


Figura 2: Interpolación cuadrática.

Desventajas de la interpolación de Lagrange.

- Requiere gran cantidad de cálculos.
- Para cada valor de x hay que 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.

9.3. Método de las Diferencias Divididas de Newton.

El método de las **diferencias divididas de Newton** es una técnica para construir el **polinomio interpolador de Newton,** que es una forma eficiente de encontrar un polinomio que pase exactamente por un conjunto de puntos dados $(x_0, y_0), (x_1, y_1), \cdots (x_n, y_n)$. Es especialmente útil cuando se añaden nuevos puntos, ya que permite actualizar el polinomio sin recalcular desde cero.

9.3.1. Idea general.

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

$$\boxed{p(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots + a_n(x - x_0)(x - x_1)\cdots(x - x_{n-1})} \quad (40)$$

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, se necesita conocer cómo calcular los coeficientes $a_i, i = 0, 1, \cdots n-1$. Imponiendo las condiciones de interpolación, de que $p_n(x_i) = y_i$, obtenemos:

$$\begin{aligned} p_n(x_0) &= a_0 = y_0 \\ p_n(x_1) &= y_0 + a_1(x_1 - x_0) = y_1 \end{aligned}$$

Y luego surge que:

$$a_1 = \frac{y_1 - y_0}{x_1 - x_0}$$

Luego, calculando $p_n(x_2)$ reemplazando adecuadamente obtenemos:

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

Y surge entonces el valor de a_2:

$$a_2 = \frac{\dfrac{y_2 - y_0}{x_2 - x_0} - \dfrac{y_1 - y_0}{x_1 - x_0}}{x_2 - x_1}$$

9.3.2. Diferencias divididas.

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 de 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.**

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

- **Diferencia dividida de orden k.**

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

9.3.3. Proposición. Permutación en diferencias divididas.

Sea $(i_0, i_1, ..., i_n)$ una permutación (o reubicación) de los enteros $(0, 1, ..., n)$. Entonces:

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

Demostración. La demostración es trivial para $n = 1$ y $n = 2$, pero no es trivial para el caso general. Veremos una demostración más adelante.

9.3.4. Teorema. Fórmula de interpolación por diferencias divididas de Newton.

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

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

Demostración.

La demostración es por inducción sobre el valor de k. Para $k = 1$ sabemos que el teorema es cierto para cualquier valor de i. Supondremos que es cierto para k y cualquier valor de i (HI). Luego, probaremos para $k + 1$:

$$p_{i,k+1} = p_{i,k}(x) + (x - x_i)(x - x_{i+1}) \cdots (x - x_{i+k})(a_{k+1})$$

Este polinomio es de grado $k + 1$ e interpola $f(x)$ en los puntos $\{x_i, x_{i+1}, ..., x_{i+k+1}\}$. Debemos elegir a_{k+1} tal que:

$$p_{i,k+1}(x_{i+k+1}) = f(x_{i+k+1})$$

Y demostrar que:

$$a_{k+1} = f[x_i, x_{i+1}, ..., x_{i+k+1}]$$

Notar que:

- El coeficiente de x^k en $p_{i,k}(x)$ es $f[x_i, x_{i+1}, ..., x_{i+k}]$
- El coeficiente de x^{k+1} en $p_{i,k}(x)$ es a_{k+1}

Por lo tanto, construiremos el siguiente polinomio:

$$q(x) = \frac{(x - x_i)p_{i+1,k}(x) - (x - x_{i+k+1})p_{i,k}(x)}{x_{i+k+1} - x_i}$$

Este polinomio satisface:

- $q(x_i) = \dfrac{-(x_i - x_{i+k+1})p_{i,k}(x_i)}{x_{i+k+1} - x_i} = p_{i,k}(x_i) = f(x_i)$

- $q(x_{i+k+1}) = \dfrac{(x_{i+k+1} - x_i)p_{i+1,k}(x_{i+k+1})}{x_{i+k+1} - x_i} = p_{i+1,k}(x_{i+k+1}) = f(x_{i+k+1})$

- $q(x_j) = \dfrac{(x_j - x_i)f(x_j) - (x_j - x_{i+k+1})f(x_j)}{x_{i+k+1} - x_i}, \quad \forall i < j < i + k + 1$

Luego, $q(x)$ es un polinomio de grado $\leq k+1$ que interpola $f(x)$ en los puntos $\{x_i, x_{i+1}, ..., x_{i+k+1}\}$. Por lo tanto, $q(x) = p_{i,k+1}(x)$.

Además, el coeficiente de x^{k+1} en $q(x)$ es:

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

Por lo que el teorema queda demostrado. □

Observación. Este teorema nos dice que los coeficientes $a_i, \quad i = 0, 1, ..., n$ en la ecuación (40) son iguales a las diferencias divididas, con lo cual, la fórmula de interpolación por diferencias divididas de Newton nos queda:

$$\boxed{p_n(x) = f(x_0) + (x - x_0)f[x_0, x_1] + \ \cdots \ + (x - x_0)(x - x_1)\cdots(x - x_{n-1})f[x_0, x_1, ..., x_n]} \quad (41)$$

9.3.5. Tabla de diferencias divididas.

Para obtener las diferencias que se necesitan en la ecuacion (41) construimos una tabla de diferencias divididas de la siguiente forma:

Tabla 1: Tabla de diferencias divididas.

x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$\cdots$
x_0	$f(x_0)$			
x_1	$f(x_1)$	$f[x_0, x_1]$	$f[x_0, x_1, x_2]$	
x_2	$f(x_2)$	$f[x_1, x_2]$	$f[x_1, x_2, x_3]$	$\cdots$
x_3	$f(x_3)$	$f[x_2, x_3]$	$f[x_2, x_3, x_4]$	$\cdots$
x_4	$f(x_4)$	$f[x_3, x_4]$	$\vdots$	
$\vdots$	$\vdots$	$\vdots$		

9.3.6. Fórmula de multiplicaciones encajadas.

Para evaluar el polinomio de interpolación por diferencias divididas de Newton (41) de manera eficiente, podemos usar multiplicaciones encajadas. Primero, escribimos (41) como:

$$p_n(x) = D_0 + (x - x_0)D_1 + (x - x_0)(x - x_1)D_2 + \ \cdots \ + (x - x_0)(x - x_1)\cdots(x - x_{n-1})D_n$$

Donde, las variables D_i son de la siguiente manera:

$$D_0 = f(x_0), \qquad D_i = f[x_0, ..., x_i]$$

Lo cual se puede escribir en forma encajada de la siguiente manera. Esta forma solo requiere de n multiplicaciones para evaluar $p_n(x)$:

$$\boxed{p_n(x) = D_0 + (x - x_0)[D_1 + (x - x_1)[D_2 + \cdot + (x - x_{n-2})[D_{n-1} + (x - x_{n-1})D_n]\cdots]]}$$

9.4. Error de la Interpolación Polinómica.

9.4.1. Teorema. Error de la interpolación polinómica.

Sean $x_0, x_1, ..., x_n$, $n+1$ números distintos en $[a,b]$, y sea $f(x) \in C^{n+1}$ en $[a,b]$ (continuamente diferenciable).

Luego, siendo $p(x)$ el polinomio interpolante de grado menor o igual a n, introducimos el error de interpolación. Para todo $x \in [a,b]$ existe $\xi(x) \in (a,b)$ tal que:

$$\boxed{f(x) - p(x) = \frac{(x-x_0)(x-x_1)...(x-x_n)}{(n+1)!} f^{(n+1)}(\xi(x))} \tag{42}$$

Demostración.

Se observa primero que si $x = x_k$ con $k = 0, 1, ...n$, entonces la expresión del error es 0 y por lo tanto $f(x_k) = p(x_k)$ para cualquier $\xi(x) \in (a,b)$.

Si $x \neq x_k$, definimos la siguiente función $g(t)$ para $t \in [a,b]$:

$$\begin{aligned} g(t) &= f(t) - p(t) - (f(x) - p(x)) \frac{(t-x_0)(t-x_1)...(t-x_n)}{(x-x_0)(x-x_1)...(x-x_n)} \\ &= f(t) - p(t) - (f(x) - p(x)) \prod_{i=0}^{n} \frac{(t-x_i)}{(x-x_i)} \end{aligned} \tag{43}$$

Puesto que $f \in C^{n+1}$ en $[a,b]$ (continuamente diferenciable $n+1$ veces en $[a,b]$), y $p \in C^{\infty}$ en $[a,b]$, se deduce que $g \in C^{n+1}$ en $[a,b]$.

Cuando evaluamos $g(t)$ con $t = x_k$ tendremos lo siguiente:

$$\begin{aligned} g(x_k) &= (f(x_k) - p(x_k)) - (f(x) - p(x)) \prod_{i=0}^{n} \frac{(x_k - x_i)}{(x - x_i)} \\ &= 0 - (f(x) - p(x)) \cdot 0 \\ &= 0 \end{aligned}$$

Donde $f(x_k) - p(x_k) = 0$ por la observación del principio de la demostración y la productoria es 0 pues en algún momento se evalúa $(x_k - x_k)$ y multiplica toda la expresión por cero. Además,

$$\begin{aligned} g(x) &= (f(x) - p(x)) - (f(x) - p(x)) \prod_{i=0}^{n} \frac{(x - x_i)}{(x - x_i)} \\ &= (f(x) - p(x)) - (f(x) - p(x)) \cdot 1 \\ &= 0 \end{aligned}$$

Por lo tanto, $g \in C^{n+1}$ en $[a,b]$, y g se anula en los $n+2$ números distintos $x, x_0, x_1, ..., x_n$. Luego, por el Teorema Generalizado de Rolle, existe $\xi \in (a,b)$ tal que $g^{(n+1)}(\xi) = 0$.

Diferenciando $n+1$ veces la ecuación (43), igualando a cero y evaluando en $t = \xi$ obtenemos:

$$0 = g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - p^{(n+1)}(\xi) - (f(x) - p(x)) \frac{d^{n+1}}{dt^{n+1}} \left(\prod_{i=0}^{n} \frac{(t-x_i)}{(x-x_i)} \right)_{t=\xi} \tag{44}$$

Notar que como el polinomio interpolante $p(x)$ es de grado $\leq n$, al derivarlo $n+1$ veces tenemos que $p^{(n+1)}(x) = 0$.

Por otra parte, la productoria es un polinomio mónico y de grado $n+1$ (al multiplicar los numeradores). Entonces podemos reescribir:

$$\prod_{i=0}^{n} \frac{t-x_i}{x-x_i} = \frac{1}{\prod_{i=0}^{n}(x-x_i)} t^{n+1} + \ terminos\ de\ grado\ \leq n \tag{45}$$

Diferenciando esta última ecuación (45) $n+1$ veces respecto de t obtenemos:

$$\frac{d^{n+1}}{dt^{n+1}} \left(\prod_{i=0}^{n} \frac{(t-x_i)}{(x-x_i)} \right) = \frac{(n+1)!}{\prod_{i=0}^{n}(x+x_i)}$$

Luego, la ecuación (44) de la derivada $g^{(n+1)}(\xi)$ se reduce a:

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

Despejando $f(x)$ tenemos:

$$f(x) = p(x) + \prod_{i=0}^{n}(x-x_i) \frac{f^{(n+1)}(\xi)}{(n+1)!}$$
$$f(x) - p(x) = \frac{(x-x_0)(x-x_1)...(x-x_n)}{(n+1)!} f^{(n+1)}(\xi)$$

Como queríamos probar.

9.4.2. Caso particular. Error de la interpolación lineal

Sean $x, x_0, x_1 \in [a,b]$, y sea $f(x) \in C^2$ en $[a,b]$ (continuamente diferenciable dos veces). Luego, el error de la interpolación lineal está dado de la siguiente manera:

$$\boxed{f(x) - p_1(x) = \frac{(x-x_0)(x-x_1)}{2} f''(c_x)}$$

para algún c_x entre el mínimo y el máximo de x, x_0 y x_1. (Es aplicar el teorema del error para 3 nodos).

9.4.3. Ejemplo de acotación del error.

Sea $f(x) = log_{10}(x) = \frac{log(x)}{log(10)}$. Notar que $log_{10}(e) = \frac{1}{log(10)}$.

Buscaremos el error del polinomio interpolante para 2 nodos. La derivida 2da de f está dada por:

$$f'(x) = \frac{1}{log(10)} \cdot \frac{1}{x} = log_{10}(e) \cdot \frac{1}{x} = log_{10}(e) x^{-1}$$
$$f''(x) = -\frac{1}{x^2} log_{10}(e)$$

Luego, usando el teorema de error de interpolación, tenemos el siguiente error:

$$log_{10}(x) - p_1(x) = \frac{(x-x_0)(x-x_1)}{2}\left(-\frac{log_{10}(e)}{c_x^2}\right) = \frac{(x-x_0)(x_1-x)}{2}\left(\frac{log_{10}(e)}{c_x^2}\right)$$

Si estamos interpolando (def) se cumple que $x_0 \leq x \leq x_1$, y en este caso tenemos:

$$(x-x_0)(x_1-x) \geq 0, \qquad x_0 \leq c_x \leq x_1$$

Por lo tanto, podemos acotar el error de interpolación lineal como:

$$(x-x_0)(x_1-x)\left(\frac{log_{10}(e)}{2x_1^2}\right) \leq log_{10}(x) - p_1(x) \leq (x-x_0)(x_1-x)\left(\frac{log_{10}(e)}{2x_0^2}\right)$$

Supongamos que queremos hallar una cota del error que sea válida para todo $x \in [x_0, x_1]$. Tenemos:

$$0 \leq log_{10}(x) - p_1(x) \leq \max_{x_0 \leq x \leq x_1} (x-x_0)(x_1-x)\left(\frac{log_{10}(e)}{2x_0^2}\right)$$

Siendo $q(x) = (x-x_0)(x_1-x)$, esta función es cuadrátrica con signo negativo, por lo tanto, es estrictamente cóncava. Luego, el valor de x que maximiza la función en el intervalo $[x_0, x_1]$ puede hallarse en el (único) punto estacionario, es decir, el punto $\bar{x}$ para el cual $q'(\bar{x}) = 0$, si $\bar{x} \in [x_0, x_1]$, o puede hallarse en un punto extremo del intervalo si $\bar{x} \notin [x_0, x_1]$.

$$\begin{aligned} q(x) &= (x-x_0)(x_1-x) = xx_1 - x^2 - x_0x_1 + x_0x = -x^2 + x_1x + x_0x + x_0x_1 \\ q'(x) &= -2x + x_1 + x_0 \end{aligned}$$

Igualando $q'(x) = 0$ y despejando x obtenemos:

$$\bar{x} = \frac{x_0 + x_1}{2} \in (x_0, x_1)$$

Luego, reemplazando el máximo de $q(x)$ por este valor encontrado tenemos:

$$\begin{aligned} \max_{x_0 \leq x \leq x_1} q(x) = q(\bar{x}) &= \left(\frac{x_0+x_1}{2} - x_0\right)\left(x_1 - \frac{x_0+x_1}{2}\right) \\ &= \left(\frac{x_0+x_1-2x_0}{2}\right)\left(\frac{2x_1-x_0-x_1}{2}\right) \\ &= \left(\frac{x_1-x_0}{2}\right)\left(\frac{x_1-x_0}{2}\right) \\ &= \left(\frac{x_1-x_0}{2}\right)^2 \\ &= \frac{(x_1-x_0)^2}{4} = \frac{h^2}{4} \end{aligned}$$

Donde $h = x_1 - x_0$. Luego, por definición de máximo podemos acotar:

$$0 \leq (x-x_0)(x_1-x) \leq \frac{h^2}{4}, \quad x \in [x_0, x_1]$$

Y por lo tanto, la acotación del error de interpolación polinomial lineal nos queda:

$$0 \leq log_{10}(x) - p_1(x) \leq \frac{h^2}{4}\left(\frac{log_{10}(e)}{2x_0^2}\right) \approx 0,0543\ \frac{h^2}{x_0^2}$$

Supongamos ahora que tenemos tabulados los valores de $log_{10}(x)$ en el intervalo $[1, 10]$, con un espaciado $h = 0,01$. Esta tabla se puede emplear para calcular por interpolación lineal el valor de $log_{10}(x)$ para cualquier $x \in [1, 10]$. Si queremos una cota uniforme para todo $x_0 \in [1, 10)$ tenemos:

$$0 \leq log_{10}(x) - p_1(x) \leq 0,0543h^2$$

Para $h = 0,01$ tenemos:

$$0 \leq log_{10}(x) - p_1(x) \leq 5,43 \times 10^{-6}$$

Típicamente, las tablas de $log_{10}(x)$ presentaban valores con 4 cifras decimales, por ejemplo, $log_{10}(5,41) = 0,7332$. El máximo error de redonde es $0,00005$, el cual es mayor que el máximo error de interpolación lineal.

9.4.4. Acotación del error. Caso general.

El procedimiento empleado en el ejemplo anterior para acotar el error de interpolación se puede generalizar. Para $x_0, x_1, ..., x_n$ distintos en $[a, b]$ y $x \in [a, b]$, el error de interpolación está dado por:

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

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

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

9.4.5. Fenómeno de Runge.

El **fenómeno de Runge** es un problema que sucede cuando se usa interpolación polinómica con polinomios de alto grado utilizando nodos equidistantes. Veamos el siguiente ejemplo:

Ejemplo.

Supongamos que tenemos $n + 1$ nodos equiespaciados en un intervalo $[a, b]$. Tenemos:

$$h = \frac{b - a}{n}, \quad x_i = a + ih$$

En particular, tomemos:

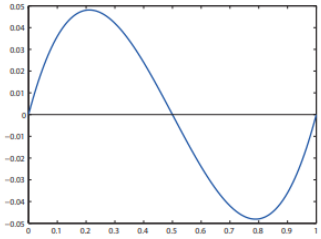
$$a = x_0 = 0, \quad x_1 = h, \quad x_2 = 2h, \quad ... \quad b = x_n = 1$$

Luego:

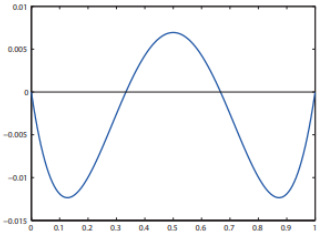
$$\Phi_n(x) = x(x - h)(x - 2h)...(x - 1)$$

Para valores de n grandes (por ejemplo, $n \geq 5$), los valores de $|\Phi_n(x)|$ varían mucho en el intervalo $[x_0, x_n]$. Los valores en los extremos del intervalo pueden ser mucho mayores que los valores en el medio del intervalo. Esta disparidad aumenta a medida que n aumenta.

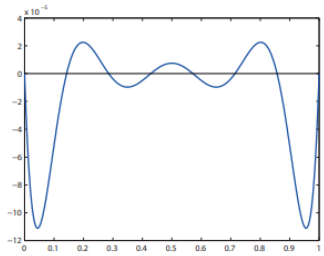
La oscilación se puede minimizar usando nodos de Chebyshev en lugar de equidistantes. En este caso, se garantiza que el error máximo disminuye al crecer el orden polinómico.



(a) $\Phi_2(x)$



(b) $\Phi_3(x)$



(c) $\Phi_7(x)$

Figura 3: Fenómeno de Runge.