Análisis y Modelamiento Numérico I

Sucesiones: Definición y propiedades. Espacio normado de las matrices

Los profesores[1]

[1]Facultad de Ciencias
Universidad Nacional de Ingeniería

2023-1

Contenidos

Definición

Definición 1

Una sucesión de números reales es una aplicación:

$$\begin{array}{rcl} f : \mathbb{N} & \to & \mathbb{R} \\ n & \mapsto & f(n) = a_n \end{array}$$

Se denota por $\{a_n\}_{n\in\mathbb{N}}$. Al número a_n se le llama el término $n-$ésimo de la sucesión.
En algunas ocasiones el valor de cada término a_n se puede expresar a partir del índice " n ".
En ese caso, a esa expresión se le llama **término general de la sucesión**.

Límite de una sucesión

Definición 2

Se dice que el límite de una sucesión $\{a_n\}$ *es un valor* $L \in \mathbb{R}$ *y se denota por:*

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

cuando:

$$\forall\, \varepsilon > 0\, \exists n_0 \in \mathbb{N} \textit{ tal que para todo } n \geq n_0 \,:\, |a_n - L| < \varepsilon.$$

Es decir, a partir de un valor de n suficientemente grande, todos los elementos de la sucesión se **acumulan** cerca de L.

Tipos de sucesiones

Dada una sucesión $\{a_n\}_{n\in\mathbb{N}}$, ésta puede ser:

1. **Acotada**, es decir,existe un número real c tal que $|a_n| \leq c$ para todo $n \in \mathbb{N}$. Caso contrario se dice que es **no acotada**.
2. **Acotada superiormente** cuando existe un número real c tal que $a_n \leq c$ para todo $n \in \mathbb{N}$.
3. **Acotada inferiormente** cuando existe un número real c tal que $c \leq a_n$ para todo $n \in \mathbb{N}$.
4. **Creciente** cuando $x_n < x_{n+1}$ para todo $n \in \mathbb{N}$. Si se cumple $x_n \leq x_{n+1}$ para todo $n \in \mathbb{N}$ entonces es llamada **no decreciente**.
5. **Decreciente** cuando $x_n > x_{n+1}$ para todo $n \in \mathbb{N}$. Si se cumple $x_n \geq x_{n+1}$ para todo $n \in \mathbb{N}$ entonces es llamada **no creciente**.
6. Las sucesiones crecientes, no decrecientes, decrecientes y no crecientes son también llamadas **sucesiones monótonas**.

Ejemplos: I

Analice las siguientes sucesiones:

$$a_n = \frac{2n-1}{n}, \quad b_n = \frac{1}{n^3}$$

Solución:
Observe lo siguiente:

$$a_{n+1} - a_n = \frac{2(n+1)-1}{n+1} - \frac{2n-1}{n} = \frac{1}{(n+1)n} > 0 \Rightarrow a_{n+1} > a_n$$

es decir, es una sucesión creciente.
Por otro lado:

$$(n+1)^3 > n^3 \Leftrightarrow \frac{1}{n^3} > \frac{1}{(n+1)^3} \Leftrightarrow b_n > b_{n+1}$$

es decir, es una sucesión decreciente.

Ejemplo I

Aplique la definición de límite para demostrar los siguientes límites:

$$\lim_{n\to\infty} \frac{n+1}{n-2} = 1$$

Solución: Se debe probar lo siguiente:

$$\forall\, \varepsilon > 0\, \exists n_0 \in \mathbb{N} \text{ tal que } : \left|\frac{n+1}{n+2} - 1\right| < \varepsilon \quad \text{para todo} \quad n > n_0.$$

Observe que:

$$\left|\frac{n+1}{n-2} - 1\right| < \varepsilon \Leftrightarrow \frac{3}{n-2} < \varepsilon \Leftrightarrow \frac{3}{\varepsilon} < n-2 \Leftrightarrow \frac{1}{\varepsilon} + 2 < n$$

por tanto, fijado $\varepsilon > 0$ basta elegir $n_0 = \left[\!\left[\frac{1}{\varepsilon} + 2\right]\!\right]$ para que se cumpla la definición de límite.

Algunos resultados importantes

1. Toda sucesión convergente es acotada.
2. **Teorema de Weierstrass**: Toda sucesión monótona acotada es convergente.
3. Para que una función $f : X \to \mathbb{R}$ sea continua en un punto $a \in X$ es necesario y suficiente que se cumpla $\lim_{n\to\infty} f(x_n) = f(a)$ para toda sucesión de puntos $\{a_n\}_{n\in\mathbb{N}} \subset X$ tal que $\lim_{n\to\infty} a_n = a$.

Ejemplo:

Dada la sucesión $a_1 = 1$ y $a_n = \dfrac{1}{3 - a_{n-1}}$, demuestre que:

1. $(a_n)^2 - 3a_n + 1 \leq 0$ para todo número natural.
2. La sucesión $\{a_n\}_{n\in\mathbb{N}}$ es convergente y calcule su límite.

Solución: I

1. Demostremos la desigualdad usando inducción:
   - Para $n = 1$ se observa:
   $$1^3 - 3(1) + 1 = -1 \leq 0$$
   y por tanto es verdadera.
   - Hipótesis inductiva: Supongamos que $(a_n)^2 - 3a_n + 1 \leq 0$ y se debe probar que:
   $$(a_{n+1})^2 - 3a_{n+1} + 1 \leq 0.$$
   Se observa:
   $$(a_{n+1})^2 - 3a_{n+1} + 1 = \left(\frac{1}{3 - a_n}\right)^2 - 3\left(\frac{1}{3 - a_n}\right) + 1 = \frac{1 - 3a_n + a_n^2}{(3 - a_n)^2} \leq 0$$
   de lo anterior podemos concluir que:
   $$(a_{n+1})^2 - 3a_{n+1} + 1 \leq 0.$$

Solución: II

2. Para demostrar la convergencia, un modo de hacerlo es probar que sea monótona y acotada. Dando valores a n sugiere que en efecto es decreciente. Afirmamos que $0 < a_n < 1$, lo cual se demuestra usando inducción.
   - Por dato se tiene: $0 < a_1 \leq 1$.
   - Hipótesis inductiva: Supongamos que es verdad $0 < a_n < 1$, veamos que se cumple: $0 < a_{n+1} < 1$.
     Observe:
     $$-1 \leq -a_n < 0 \Rightarrow 2 < 3 - a_n < 3 \Rightarrow \frac{1}{3} < \frac{1}{3 - a_n} < \frac{1}{2}$$
     por tanto: $0 < a_{n+1} \leq 1$.

Solución: III

Ahora probemos que es decreciente, es decir: $a_{n+1} \leq a_n$. Para nuestro caso basta demostrar:

$$\frac{1}{3-a_n} \leq a_n$$

En efecto, como:

$$(a_n)^2 - 3a_n + 1 \leq 0 \Rightarrow 1 \leq a_n(3-a_n) \Rightarrow \frac{1}{3-a_n} \leq a_n \Rightarrow a_{n+1} \leq a_n$$

Por el teorema de Weierstrass, por ser una sucesión monótona y acotada podemos concluir que es una sucesión convergente. Sea L el valor de este límite, es decir:

$$\lim_{n\to\infty} a_n = L \Rightarrow \lim_{n\to\infty} a_n = \lim_{n\to\infty} \frac{1}{3-a_{n-1}} \Rightarrow L = \frac{1}{3-L}$$

Solución: IV

entonces:

$$L(3-L)=1 \Rightarrow L^2-3L+1=0 \Rightarrow L=\frac{3\pm\sqrt{5}}{2}$$

como la sucesión es monótona decreciente y $a_1=1$ resulta que:

$$L=\frac{3-\sqrt{5}}{2}.$$

Cálculos de límites

Sean $\{a_n\}_{n\in\mathbb{N}}$ y $\{b_n\}_{n\in\mathbb{N}}$ dos sucesiones convergentes y sean L_1 y L_2 sus límites respectivos. Se cumple:

1. $\lim\limits_{n\to\infty}(a_n \pm b_n) = L_1 \pm L_2$.
2. $\lim\limits_{n\to\infty}(a_n b_n) = L_1 L_2$.
3. Si $b_n \neq 0$ para todo $n \in \mathbb{N}$ y $L_2 \neq 0$ se cumple:

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{L_1}{L_2}.$$

Razón de convergencia

Definición 3

Sea $\{x_k\}$ una sucesión que converge a L. Si existe una sucesión $\{y_k\}$ que converge a cero y una constante positiva λ (independiente de " k "), tal que:

$$|x_k - L| \leq \lambda |y_k|$$

para todo " k " suficientemente grande. Luego, decimos que $\{x_k\}$ converge a L con tasa de convergencia $O(y_k)$.

Observaciones

1. Cuando $\{x_k\}_{k\in\mathbb{N}}$ converge a " L " con razón de convergencia $O(y_k)$ es comúnmente escrito en forma escrita de la forma $x_k = L + O(y_k)$, de aquí que el término O da una referencia de la rapidez con que el error se aproxima a cero.
2. La sucesión $\{y_k\}_{k\in\mathbb{N}}$, el cual es usual ser de la forma $1/n^a$ o $1/a^n$ para alguna constante positiva a, sirve como punto de referencia que permite la comparación entre dos sucesiones diferentes. Por ejemplo, una sucesión con razón de convergencia $O(1/n^2)$ converge más lentamente que uno que tiene razón de convergencia $O(1/n^{10})$, el cual a la vez converge más lentamente que una sucesión con razón de convergencia $O(1/2^n)$.

Ejemplo: Comparación de tasas de convergencia

Considere las sucesiones:

$$a_n = \left\{ \frac{n+3}{n+7} \right\} \quad \text{y} \quad \left\{ b_n = \frac{2^n+3}{2^n+7} \right\}$$

Observe que:

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = 1$$

Aunque ambas sucesiones convergen a 1, en la Tabla 1 se puede ver que los términos de la sucesión b_n se aproximan a 1 de forma más rápida que los términos de la sucesión a_n.

Resultados

Se obtienen los siguientes resultados:

n	a_n	b_n
0	0.428571428571	0.500000000000
1	0.500000000000	0.555555555556
2	0.555555555556	0.636363636364
3	0.600000000000	0.733333333333
4	0.636363636364	0.826086956522
5	0.666666666667	0.897435897436
6	0.692307692308	0.943661971831
7	0.714285714286	0.970370370370
8	0.733333333333	0.984790874525
9	0.750000000000	0.992292870906

Cuadro 1: Comparación a_n y b_n

Solución (cont.)

Determinemos las tasas de convergencia de cada sucesión:

$$\left|\frac{n+3}{n+7}-1\right| = \frac{4}{n+7} < 4\left(\frac{1}{n}\right)$$

de aquí, podemos considerar $\lambda = 4$ e $y_n = 1/n$ en la definición de tasa de convergencia. Por tanto, la sucesión $\{a_n\}_{n\in\mathbb{N}}$ converge a 1 con tasa de convergencia $O(1/n)$.
Análogamente:

$$\left|\frac{2^n+3}{2^n+7}-1\right| = \frac{4}{2^n+7} < 4\left(\frac{1}{2^n}\right)$$

de aquí, podemos considerar $\lambda = 4$ e $y_n = 1/2^n$ en la definición de tasa de convergencia. Por tanto, la sucesión $\{b_n\}_{n\in\mathbb{N}}$ converge a 1 con tasa de convergencia $O(1/2^n)$.
Estos resultados confirman nuestra evidencia numérica desde que $1/2^n$ se aproxima de cero más rápido que $1/n$ cuando $n \to \infty$.

Ejemplo

Calcule el límite de la sucesión $\{x_k\}$ y su correspondiente tasa de convergencia, donde:

$$x_k = \frac{k-1}{k^3+2}, \quad k \in \mathbb{N}.$$

Solución:

Se observa:

$$\lim_{k\to\infty} x_k = \lim_{k\to\infty} \frac{k-1}{k^3+2} = \lim_{k\to\infty} \frac{\frac{1}{k^2} - \frac{1}{k^3}}{1+\frac{2}{k^3}} = 0.$$

Por otro lado:

$$|x_k - 0| = \left|\frac{k-1}{k^3+2}\right| \leq \left|\frac{k}{k^3+2}\right| \leq \left|\frac{k}{k^3}\right| = \frac{1}{k^2} \Rightarrow x_k = O(1/k^2).$$

Orden de convergencia

Definición 4

Sea $\{x_k\}$ *una sucesión que converge a* L. *Considere la sucesión* $e_k = x_k - L$ *para todo* $k \geq 0$. *Si existen constantes* λ *y* α *tal que:*

$$\lim_{k\to\infty} \frac{|x_{k+1} - L|}{|x_k - L|^\alpha} = \lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^\alpha} = \lambda,$$

entonces decimos que $\{x_k\}$ *converge a* L *con* **orden de convergencia** α *y* **constante de error asintótico** λ.

Un **método iterativo** es llamado de **orden** α si la sucesión que genera es de orden α.

Observaciones

Observe que cuando una sucesión $\{x_k\}$ tiene orden de convergencia α, de la Definición 4 se tiene que la sucesión $\{e_k\}$ satisface la siguiente **relación asintótica**:

$$|e_{k+1}| \approx \lambda |e_k|^{\alpha}.$$

Los valores más usuales para α son:

- $\alpha = 1$ y decimos que la sucesión $\{x_k\}$ **converge linealmente** a L.
- $\alpha = 2$ y decimos que la sucesión $\{x_k\}$ **converge cuadráticamente** a L.
- $\alpha = 3$ y decimos que la sucesión $\{x_k\}$ **converge cúbicamente** a L.

Ejemplo: Comparación órdenes de convergencia

Suponga tres métodos, uno lineal, uno cuadrático y uno cúbico, todos ellos aplicados al mismo problema. Cada método tiene una constante de error asintótica $\lambda = 0{,}5$ y el error en la aproximación inicial $|e_0| = 1$. Luego se tiene la siguiente tabla:

	LINEAL	CUADRÁTICO	CÚBICO
n	$\lvert e_{n+1}\rvert \approx 0{,}5\lvert e_n\rvert$	$\lvert e_{n+1}\rvert \approx 0{,}5\lvert e_n\rvert^2$	$\lvert e_{n+1}\rvert \approx 0{,}5\lvert e_n\rvert^3$
1	0.500000000000	0.500000000000	0.5
2	0.250000000000	0.125000000000	0.0625
3	0.125000000000	0.007812500000	0.0001220703125
4	0.062500000000	0.000030517578	9.094947017729282e-13
5	0.031250000000	0.000000000466	3.76158192263132e-37
6	0.015625	1.0842×10^{-19}	
7	7.8125×10^{-3}	5.8775×10^{-39}	

Cuadro 2: Tabla de comparación

Solución (cont.)

Observe la clara diferencia entre los métodos lineal y cuadrático. A menos que en cada iteración del método cuadrático se requiera más trabajo que en el método lineal, el método lineal nunca competirá con un método cuadrático.
Por otro lado, se observa una ligera diferencia (2 o 3 iteraciones) entre los métodos cuadrático y cúbico. En la práctica, el trabajo extra para alcanzar convergencia cúbica no estaría justificado.

Definición 5

Decimos que una sucesión $\{x_k\}$ **converge superlinealmente** *a* L *cuando:*

$$\lim_{k\to\infty} \frac{|x_{k+1} - L|}{|x_k - L|} = 0.$$

Ejemplo:

Determine el orden de convergencia de la sucesión:

$$x_{k+1} = \frac{x_k^3 + 3x_k a}{3x_k^2 + a}$$

si se sabe que converge a $\sqrt{a}$. ¿Cuál es la constante de error asintótico?.

Solución: I

Se observa:

$$|x_{k+1} - \sqrt{a}| = \left| \frac{x_k^3 + 3x_k a}{3x_k^2 + a} - \sqrt{a} \right| = |(x_k - \sqrt{a})^3|,$$

por tanto:

$$\lim_{k\to\infty} \frac{|x_{k+1} - \sqrt{a}|}{|x_k - \sqrt{a}|^3} = \lim_{k\to\infty} \left| \frac{(x_k - \sqrt{a})^3}{(x_k - \sqrt{a})^3(3x_k^2 + a)} \right| = \lim_{k\to\infty} \left| \frac{1}{3x_k^2 + a} \right| = \frac{1}{4a},$$

entonces la sucesión $\{x_k\}$ converge a $\sqrt{a}$ con orden de convergencia $\alpha = 3$ y constante de error asintótico $\lambda = \dfrac{1}{4a}$.

Determinar el orden de convergencia

Considere el siguiente esquema iterativo:

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right) \tag{1}$$

que se usa para aproximar la raíz cuadrada de un número real positivo a. Se desea determinar el orden de convergencia del esquema iterativo dado. Por tanto:

$$x_{n+1} - \sqrt{a} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right) - \sqrt{a} = \frac{x_n^2 - 2x_n\sqrt{a} + a}{2x_n} = \frac{(x_n - \sqrt{a})^2}{2x_n}$$

$$\Rightarrow \lim_{n\to\infty} \frac{|x_{n+1} - \sqrt{a}|}{|x_n - \sqrt{a}|^2} = \lim_{n\to\infty} \frac{1}{2x_n} = \frac{1}{2\sqrt{a}}.$$

de lo anterior, el esquema generado por este método iterativo tiene orden de convergencia igual a 2 y constante de error asintótico igual a $1/(2\sqrt{a})$.

Verificación numérica de orden de convergencia.

Una tarea común es confirmar un orden de convergencia teórico usando datos numéricos. Por ejemplo, se ha visto que el esquema 1 debe converger con orden de convergencia igual a 2. En la práctica, ¿en verdad el esquema tiene convergencia cuadrática? Para responder esta pregunta, primero seleccionamos un valor particular de " a " y generamos la secuencia respectiva para luego examinar la razón $|e_n|/|e_{n+1}|^2$. Si esta razón se aproxima a una constante cuando " n " aumenta (la razón, en particular, debe aproximarse a la constante de error asintótico), entonces se tiene evidencia numérica de convergencia cuadrática.

Ejemplo:

Considere $a = 5$ y $x_0 = 6$ y los cinco primeros términos de la sucesión generada por 1, resultados que son mostrados en la siguiente tabla:

n	x_n	$\lvert e_n\rvert = \lvert x_n - \sqrt{5}\rvert$	$\lvert e_n\rvert / \lvert e_{n-1}\rvert^2$
0	4.312500000000	3.687500000000	
1	2.735960144928	1.576539855072	4.790034734699
2	2.281736073126	0.454224071802	0.530302270304
3	2.236524992442	0.045211080684	0.374000598904
4	2.236068024193	0.000456968248	0.361923070508
5	2.236067977500	0.000000046694	0.361803411100
6	2.236067977500	0.000000000000	0.361803398875
7	2.236067977500	0.000000000000	0.361803398875

Cuadro 3: Aproximación $\sqrt{5}$

Método de Aitken I

Suponga que $\{p_n\}_{n\in\mathbb{N}}$ es una sucesión linealmente convergente con límite p. La idea es construir una sucesión $\{\hat{p}_n\}_{n\in\mathbb{N}}$ que converja más rápido a p que la sucesión original. Para esto, primero considere que los signos de:

$$p_n - p, p_{n+1} - p, p_{n+2} - p$$

son los mismos y que para n suficientemente grande se cumple:

$$\frac{p_{n+1} - p}{p_n - p} \approx \frac{p_{n+2} - p}{p_{n+1} - p}$$

luego:

$$(p_{n+1} - p)^2 \approx (p_{n+2} - p)(p_n - p)$$

Método de Aitken II

resolviendo y despejando p resulta:

$$p \approx \frac{p_{n+2}p_n - p_{n+1}^2}{p_{n+2} - 2p_{n+1} + p_n}.$$

Sumando y restando los términos p_n^2 y $2p_np_{n+1}$ en el numerador y agrupando términos adecuadamente se obtiene:

$$\begin{aligned} p &\approx \frac{p_np_{n+2} - 2p_np_{n+1} + p_n^2 - p_{n+1}^2 + 2p_np_{n+1} - p_n^2}{p_{n+2} - 2p_{n+1} + p_n} \\ &= \frac{p_n(p_{n+2} - 2p_{n+1} + p_n) - (p_{n+1}^2 - 2p_np_{n+1} + p_n^2)}{p_{n+2} - 2p_{n+1} + p_n} \\ &= p_n - \frac{(p_{n+1} - p_n)^2}{p_{n+2} - 2p_{n+1} + p_n}. \end{aligned}$$

Método de Aitken III

El **método de Aitken** Δ^2 está basado en la hipótesis que la sucesión $\{\hat{p}_n\}_{n\in\mathbb{N}}$ definida por:

$$\hat{p_n} = p_n - \frac{(p_{n+1} - p_n)^2}{p_{n+2} - 2p_{n+1} + p_n}.$$

converge más rápido a p que la sucesión original $\{p_n\}_{n\in\mathbb{N}}$.

Ejemplo: I

La sucesión $\{p_n\}_{n\in\mathbb{N}}$ donde:

$$p_n = \frac{2 - e^{p_{n-1}} + p_{n-1}^2}{3},$$

converge linealmente. Determine los 5 primeros términos de la sucesión que se obtiene por el método de Aitken.

Solución:

Con el fin de calcular $\hat{p}_n$ se deben de calcular los términos p_n, p_{n+1} y p_{n+2}. Por tanto, para calcular $\hat{p}_5$ se deben de calcular los primeros 7 términos de $\{p_n\}_{n\in\mathbb{N}}$. Los resultados son mostrados en la siguiente tabla:

Ejemplo: II

n	p_n	$\hat{p}_n$
1	0.200426243100	0.257613211870
2	0.272749065098	0.257535845041
3	0.253607156584	0.257530659437
4	0.258550376265	0.257530301809
5	0.257265636335	0.257530301809
6	0.257598985162	0.257530272007
7	0.257512454515	0.257530272007
8	0.257534913615	0.257530272007
9	0.257529084168	0.257530272007

Cuadro 4: Sucesión generada por el método de Aitken

Diferencia progresiva

Definición 6

Dada una sucesión $\{p_n\}_{n\in\mathbb{N}}$, *la* **diferencia progresiva** Δp_n *se define por:*

$$\Delta p_n = p_{n+1} - p_n \quad \textit{para} \quad n \in \mathbb{N}.$$

Las potencias de mayor orden del operador Δ *se definen recursivamente:*

$$\Delta^k p_n = \Delta(\Delta^{k-1} p_n) \quad \textit{para todo} \quad k \geq 2.$$

Esta definición implica que la sucesión que se obtiene por el método de Aitken pueda ser escrito como sigue:

$$\hat{p}_n = p_n - \frac{(\Delta p_n)^2}{\Delta^2 p_n}, \quad \text{para todo} \quad n \in \mathbb{N}.$$

Teorema

El siguiente teorema explica el significado de ser la sucesión generada por el método de Aikten más rápida a la convergencia en relación a la sucesión original.

Teorema 1

Suponga que $\{p_n\}_{n\in\mathbb{N}}$ *es una sucesión que converge linealmente al límite* p *y que además:*

$$\lim_{n\to\infty} \frac{p_{n+1} - p}{p_n - p} < 1.$$

Entonces la sucesión generada por el método de Aitken converge a p *más rápido que la sucesión* $\{p_n\}_{n\in\mathbb{N}}$ *en el sentido siguiente:*

$$\lim_{n\to\infty} \frac{\hat{p}_n - p}{p_n - p} = 0.$$

Normas vectoriales

Definición 7

Dado un espacio vectorial V, una norma es una función $\|\cdot\|$ de V en $[0, +\infty\rangle$, que satisface:

1. $\|\boldsymbol{x}\| > 0,\ \boldsymbol{x} \neq 0, \boldsymbol{x} \in V$
2. $\|\lambda \boldsymbol{x}\| = |\lambda| \|\boldsymbol{x}\|,\ \lambda \in \mathbb{R}, \boldsymbol{x} \in V$
3. $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|,\ \boldsymbol{x}, \boldsymbol{y} \in V$

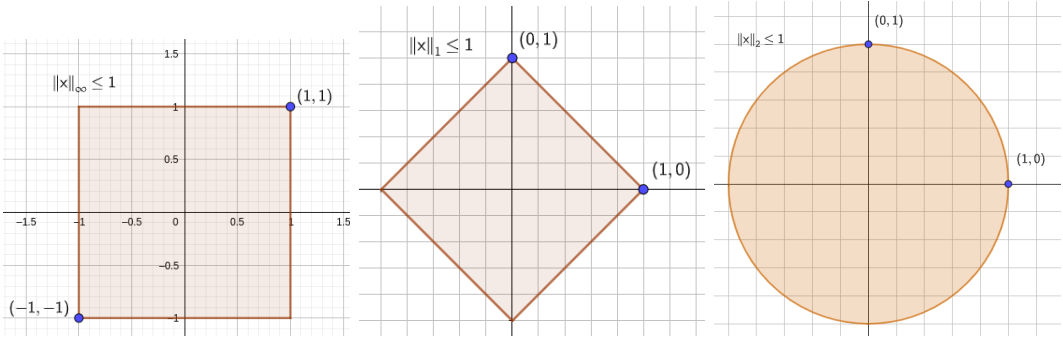
Ejemplo 1

Ejemplos comunes de normas:

- $\|\boldsymbol{x}\|_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}$
- $\|\boldsymbol{x}\|_\infty = \max\limits_{1 \leq i \leq n} |x_i|$
- $\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|$
- $\|\boldsymbol{x}\|_p = \left(\sum^{n} |x_i|^p \right)^{1/p}, \geq 1$

Ejemplo 2

Bola unitaria en $\mathbb{R}^2$ *El conjunto* $\{\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^2, \|x\| \leq 1\}$ *se llama bola unitaria y se muestra en las tres normas dadas anteriormente*



Definición 8

Una sucesión $\{\boldsymbol{x}^{(k)}\}$ *de vectores en* $\mathbb{R}^n$ *se dice convergente a* $\boldsymbol{x}$ *respecto a la norma* $\|\cdot\|$ *si para* $\epsilon > 0$ *existe* N *tal que:*

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| < \epsilon, \quad \forall k \geq N$$

Teorema 2

La sucesión $\{\boldsymbol{x}^{(k)}\}$ *converge a* $\boldsymbol{x}$ *en* $\mathbb{R}^n$ *respecto a* $\|\cdot\|_\infty$ *si y solo si* $\lim_{k\to\infty} x_i^{(k)} = x_i$ *para todo* $i = 1, 2, \ldots, n$

Demostración.

Si $\{\boldsymbol{x}^{(k)}\}$ converge a $\boldsymbol{x}$ en $\mathbb{R}^n$ entonces para $\epsilon > 0$ existe N tal que si $k \geq N$

$$\max_{i=1,2,\ldots,n} |x_i^{(k)} - x_i| = \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|_\infty < \epsilon$$

por lo tanto $|x_i^{(k)} - x_i| < \epsilon$ para $i = 1, 2, \ldots, n$. □

De igual manera si $\lim_{k\to\infty} x_i^{(k)} = x_i$ para todo $i = 1, 2, \ldots, n$, entonces para $\epsilon > 0$ existen N_i tales que si $k \geq N_i$ entonces $|x_i^{(k)} - x_i| < \epsilon$.
Definimos $N = \max_{1\leq i\leq n} N_i$, luego si $k \geq N$

$$\max_{i=1,2,\ldots,n} |x_i^{(k)} - x_i| = \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|_\infty < \epsilon$$

Ejemplo 3

Muestre que

$$\boldsymbol{x}^{(k)} = \left(1, 2 + \frac{1}{k}, \frac{3}{k^2}, e^{-k} \operatorname{sen} k\right)$$

converge a $\boldsymbol{x} = (1, 2, 0, 0)^T$ *respecto a la norma* $\|\cdot\|_\infty$
En efecto basta comprobar que $\lim_{k\to\infty} 1 = 1$, $\lim_{k\to\infty}(2 + 1/k) = 2$, $\lim_{k\to\infty} 3/k^2 = 0$, $\lim_{k\to\infty} e^{-k} \operatorname{sen} k = 0$.

Teorema 3

Si $\boldsymbol{x} \in \mathbb{R}^n$

$$\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_2 \leq \sqrt{n}\|\boldsymbol{x}\|_\infty$$

$$\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_1 \leq \sqrt{n}\|\boldsymbol{x}\|_2$$

$$\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_1 \leq \sqrt{n}\|\boldsymbol{x}\|_\infty$$

Prueba

Sea x_j tal que $\|\boldsymbol{x}\|_\infty = |x_j|$, entonces

$$\|\boldsymbol{x}\|_\infty^2 = |x_j|^2 = x_j^2 \leq \sum_{i=1}^{n} x_i^2 = \|\boldsymbol{x}\|_2^2 \implies \|\boldsymbol{x}\|_\infty^2 \leq \|\boldsymbol{x}\|_2^2$$

por otro lado

$$\|\boldsymbol{x}\|_2^2 = \sum_{i=1}^{n} x_i^2 \leq \sum_{i=1}^{n} x_j^2 = nx_j^2 = n\|\boldsymbol{x}\|_\infty^2$$

Norma matricial subordinada

Puede demostrase que todas las normas en $\mathbb{R}^n$ son equivalentes respecto a la convergencia, es decir si $\{\boldsymbol{x}^{(k)}\}$ converge a $\boldsymbol{x}$ respeto a la norma $\|\cdot\|'$ entonces también converge a $\boldsymbol{x}$ respecto a la norma $\|\cdot\|''$.

Definición 9

Denotamos por $\mathcal{M}_{m\times n}$ el conjunto de matrices reales de m filas y n columnas. Dada una norma $\|\cdot\|$ en $\mathbb{R}^n$ es posible definir para cada matriz $A \in \mathcal{M}_{m\times n}$, la cantidad $\|A\|$ como

$$\|A\| = \sup\{\|Au\| : u \in \mathbb{R}^n, \|u\| = 1\}$$

de tal manera que la aplicación $A \mapsto \|A\|$ es una norma en el espacio vectorial de matrices reales $m \times n$.

En efecto:

- Si $A \neq 0$ entonces al menos una columna $A^{(j)} \neq 0$, si escogemos el vector $x = e_j$ entonces para $v = \frac{1}{\|x\|}x$:

$$\|A\| \geq \|Av\| = \frac{\|Ax\|}{\|x\|} = \frac{\|A^{(j)}\|}{\|x\|} > 0$$

- $\|\lambda A\| = \sup\{\|\lambda Au\| : \|u\| = 1\} = |\lambda| \sup\{\|Au\| : \|u\| = 1\} = |\lambda|\|A\|$
- $\|A + B\| = \sup\{\|(A + B)u\| : \|u\| = 1\} \leq \sup\{\|Au\| + \|Bu\| : \|u\| = 1\} \leq \sup\{\|Au\| : \|u\| = 1\} + \sup\{\|Bu\| : \|u\| = 1\} = \|A\| + \|B\|$

Como consecuencia de la definición de norma matricial subordinada, tenemos que

$$\|Ax\| \leq \|A\|\|x\|$$

ademas si $A = I$ entonces

de igual manera

$$\|AB\| = \sup\{\|ABx\| : \|x\| = 1\} \leq \sup\{\|A\|\|Bx\| : \|x\| = 1\} = \|A\|\|B\|$$

Ejemplo 4

La norma Frobenius

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} |a_{i,j}|^2}$$

no es una norma matricial subordinada, pues $\|I\|_F = \sqrt{n} \neq 1$

La norma ∞

$$\|A\|_\infty = \sup_{\|x\|_\infty=1} \|Ax\|_\infty = \sup_{\|x\|_\infty=1} \max_i |(Ax)_i| = \max_i \sup_{\|x\|_\infty=1} |(Ax)_i|$$
$$= \max_i \sup_{\|x\|_\infty=1} |\sum_{j=1} a_{ij}x_j|$$

Como $|\sum_{j=1} a_{ij}x_j| \leq \sum_{j=1} |a_{ij}|$ entonces para $x_j = \operatorname{sgn}(a_{ij})$, $|\sum_{j=1} a_{ij}x_j| = \sum_{j=1} |a_{ij}|$ entonces $\sup_{\|x\|_\infty=1} |\sum_{j=1} a_{ij}x_j| = \sum_{j=1} |a_{ij}|$, luego

$$\|A\|_\infty = \max_i \sum_{j=1} |a_{ij}|$$

La norma 1

$$\begin{aligned}
\|A\|_1 &= \sup_{\|x\|_1=1} \|Ax\|_1 = \sup_{\|x\|_1=1} \left\| \sum_{j=1} x_j A^{(j)} \right\|_1 \\
&\leq \sup_{\|x\|_1=1} \sum_{j=1} |x_j| \|A^{(j)}\|_1 \\
&\leq \left(\sup_{\|x\|_1=1} \sum_{j=1} |x_j| \right) \max_j \|A^{(j)}\|_1
\end{aligned}$$

basta tomar $x = e_j$ donde $\|A^{(j)}\|_1 = \max_k \|A^{(k)}\|_1$,

$$\|Ax\|_1 = \left\| \sum_{k=1} x_k A^{(k)} \right\|_1 = \left\| \sum_{k=1} x_k A^{(k)} \right\|_1 = \left\| A^{(j)} \right\|_1 = \max_k \|A^{(k)}\|_1$$

El radio espectral

Recordemos que si M es una matriz simétrica real $n \times n$ entonces es diagonalizable con valores propios no negativos y sus vectores propios forman una base de $\mathbb{R}^n$. Definimos el radio espectral $\rho(M)$, como

$$\rho(M) = \max \{|\lambda| : Mv = \lambda v, v \neq 0\}$$

La norma 2

Consideremos A una matriz $m \times n$, entonces

$$\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2 = \sup_{\|x\|_2=1} (x^T A^T Ax)^{1/2}$$

Sea $v_1, \ldots v_n$ una base ortonormal formada por vectores propios de A^TA, ordenada de manera que

$$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$$

entonces

$$\begin{aligned} x &= \alpha_1 v_1 + \ldots + \alpha_n v_n \\ \implies Ax &= \alpha_1 Av_1 + \ldots + \alpha_n Av_n \\ \implies A^TAx &= \alpha_1 A^TAv_1 + \ldots + \alpha_n A^TAv_n \\ &= \alpha_1\lambda_1 v_1 + \ldots + \alpha_n\lambda_n v_n \\ \implies x^TA^TAx &= \alpha_1^2\lambda_1 + \ldots + \alpha_n^2\lambda_n \end{aligned}$$

si $\|x\|_2 = 1$ entonces $\alpha_1^2 + \ldots + \alpha_n^2 = 1$ y $(x^TA^TAx)^{1/2} \leq \sqrt{\lambda_n}$, basta escoger $x = v_n$ para establecer que

$$\|A\|_2 = \sqrt{\lambda_n} = \sqrt{\rho(A^TA)}$$

El radio espectral tiene importantes consecuencias teóricas, pero es de difícil aplicación computacional.

La norma 2 tiene las siguientes propiedades

- Si A es simétrica entonces $\|A\|_2 = \rho(A)$
- $\|A\|_2 = \|A^T\|_2$
- $\|A^T A\|_2 = \|AA^T\|_2 = \|A\|_2^2$
- $\|A^{-1}\|_2 = \frac{1}{\delta_{\min}}$ donde $\delta_{\min}$ es el mínimo valor propio de $A^T A$

Lema 1

Si $\|\cdot\|$ es una norma matricial entonces

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

Demostración.

Sea λ un valor propio de A, y $v \neq 0$ su vector propio correspondiente, definimos $B = [v|v|\dots|v] \in \mathcal{M}_n \setminus \{0\}$, luego

$$AB = \lambda B \implies |\lambda|\|B\| = \|AB\| \leq \|A\|\|B\| \implies |\lambda| \leq \|A\|$$

Lema 2

Dado $A \in \mathcal{M}_n$ y $\epsilon > 0$ entonces existe una norma matricial $\|\cdot\|$ tal que

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

Prueba

Toda matriz cuadrada es semejante a una matriz triangular superior T compleja. Si $0 < \delta < 1$ y $D = \text{diag}(\delta, \delta^2, \ldots, \delta^n)$.

$$(D^{-1}TD)_{ij} = t_{ij}\delta^{j-i}.$$

Como T es una matriz triangular superior los elementos debajo de la diagonal en $D^{-1}TD$ son 0, y elementos sobre la diagonal son $|t_{ij}|\delta^{j-i} \leq |t_{ij}|\delta$

$$D^{-1}TD = \text{diag}(T) + S$$

como existe P no singular tal que $P^{-1}AP = T$ entonces

$$\underbrace{D^{-1}P^{-1}}_{Q^{-1}} A \underbrace{PD}_{Q} = \text{diag}(T) + S$$

$$\|Q^{-1}AQ\|_\infty = \|\text{diag}(T) + S\|_\infty \leq \|\text{diag}(T)\|_\infty + \|S\|_\infty$$

Los elementos de $\text{diag}(T)$ contienen los valores propios de A, por lo tanto $\|\text{diag}(T)\|_\infty = \rho(A)$.
S es una matriz estrictamente triangular superior con $s_{ij} = t_{ij}\delta^{j-i}$, luego $\|S\|_\infty \leq \delta\|T\|_\infty$.
Podemos escoger δ suficientemente pequeño tal que $\delta\|T\|_\infty \leq \epsilon$, luego

$$\|Q^{-1}AQ\|_\infty \leq \rho(A) + \epsilon$$

Vemos que $\|Q^{-1}AQ\|_\infty$ define una norma matricial subordinada, por lo tanto tomando ínfimo sobre las normas subordinadas

$$\inf_{\|\cdot\|} \|A\| \leq \rho(A) + \epsilon$$

Teorema 4

El radio espectral satisface que

$$\rho(A) = \text{ínf}\{\|A\| : \|\cdot\| \text{ es una norma matrical subordinada}\}$$

Demostración.

Se sigue del Lema 1 que

$$\rho(A) \leq \inf_{\|\cdot\|} \|A\|$$

y del Lema 2.

$$\inf_{\|\cdot\|} \|A\| \leq \rho(A)$$

Ejemplo 5

Para la matriz

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

$$\|A\|_\infty = \text{máx}\{1+2+1, 0+3+1, 5+1+1\} = 7$$

$$\|A\|_1 = \text{máx}\{1+0+5, 2+3+1, 1+1+1\} = 6$$

como

$$A^T A = \begin{pmatrix} 26 & -3 & 4 \\ -3 & 14 & -6 \\ 4 & -6 & 3 \end{pmatrix} \implies \begin{array}{l} \lambda_1 = 0{,}1172, \\ \lambda_2 = 14{,}9792, \\ \lambda_3 = 27{,}9036, \end{array} \implies \|A\|_2 = \text{máx}\{1+0+5, 2+3+1, 1+1+1$$

El cálculo de los valores propios es un problema que será abordado en un capítulo posterior.

Teorema 5

Si $A \in \mathcal{M}_{m\times n}$ entonces

1. $\frac{1}{\sqrt{n}}\|A\|_\infty \leq \|A\|_2 \leq \sqrt{m}\|A\|_\infty$
2. $\|A\|_2 \leq \|A\|_F \leq \sqrt{m}\|A\|_2$
3. $\frac{1}{\sqrt{m}}\|A\|_1 \leq \|A\|_2 \leq \sqrt{n}\|A\|_1$
4. $\|A\|_2 \leq \sqrt{\|A\|_1\|A\|_\infty}$

Demostración.

$$\|Ax\|_\infty \leq \|Ax\|_2 \text{ y } \|x\|_2\sqrt{n}\|x\|_\infty$$

luego

$$\frac{\|Ax\|_\infty}{\|x\|_\infty} \leq \sqrt{n}\frac{\|Ax\|_2}{\|x\|_2}$$

como $\frac{1}{\|x\|_\infty}x$ y $\frac{1}{\|x\|_2}x$ son de norma unitaria, entonces

$$\sup \|Au\|_\infty \leq \sqrt{n} \sup \|Au\|_2 \implies \frac{1}{\sqrt{n}}\|A\|_\infty \leq \|A\|_2 \leq \sqrt{m}\|A\|_\infty$$

Ejemplos en Python

Ejecute los siguientes comandos y observe los resultados:

- $1e20 + 1 - 1e20$.
- ```
  import math
  x = [0.1]*10
  sum(x)
  math.fsum(x)
  ```
  ```

Representación en una base β

Seguiendo [1], un número x expresado en base β es denotado por x_β y tiene la forma siguiente:

$$x_\beta = (a_j a_{j-1} \ldots a_1 a_1 a_0)_\beta, \quad 0 \leq a_k \leq (\beta - 1)$$

el cual al expresarse en forma polinomial resulta:

$$x_{10} = a_j \beta^j + a_{j-1} \beta^{j-1} + \ldots + a_2 \beta^2 + a_1 \beta^1 + a_0 \beta^0 \tag{2}$$

es decir, se obtiene la representación de x en base 10.

Cambiar de base 2 a base 10

La representación del número $x_2 = (a_j a_{j-1} \dots a_2 a_1 a_0)_2$ en base 10, denotada por b_0, se obtiene a través del siguiente proceso:

$$\begin{aligned} b_j &= a_j \\ b_{j-1} &= a_{j-1} + 2b_j \\ b_{j-2} &= a_{j-2} + 2b_{j-1} \\ &\vdots \\ b_1 &= a_1 + 2b_2 \\ b_0 &= a_0 + 2b_1 \end{aligned}$$

Ejemplo:
Use el proceso descrito para calcular la representación en base 10 del siguiente número: $(10111)_2$. **Rpta: 23**.

Cambiar de base β a base 10

Dado $N = (a_n a_{n-1} \ldots a_1 a_0)_\beta$, para hallar su equivalente en base 10, resulta a partir de (2) que se reduce en evaluar el polinomio:

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

en $x = \beta$. Para esto se usa el siguiente algoritmo recursivo:

$$\begin{aligned} b_0 &= a_n \\ b_k &= a_{n-k} + \beta b_{k-1} \quad k = 1, 2, \ldots, n \end{aligned}$$

y por último N en base diez resulta: $N_{10} = b_n = P(2)$.

Ejemplo: Convierta $N = 11111001111_{(2)}$ a su equivalente a base 10.

Seguimos la siguiente tabla:

k	0	1	2	3	4	5	6	7	8	9	10
a_{n-k}	1	1	1	1	1	0	0	1	1	1	1
b_k	1	3	7	15	31	62	124	249	499	999	1999

Cambiar de base 10 a base 2

Considere un número entero x_{10} y su representación binaria $x_2 = (a_j a_{j-1} \dots a_2 a_1 a_0)_2$. El algoritmo calcula el dígito a_k para cada k.

Paso 0: $k = 0$.
$N_k = N$

Paso 1: Determine q_k y r_k tales que:
$N_k = 2q_k + r_k$
Asigne $a_k = r_k$

Paso 2: Si $q_k = 0$, pare.
Caso contrario, asigne $N_{k+1} = q_k$.
$k = k + 1$ y vuelva al Paso 1.

Ejemplo:
Use el proceso descrito para calcular la representación en base 2 del siguiente número: 81.
Rpta: 101001.

Cambiar de base 10 a base 2

Para encontrar la representación binaria de un entero positivo N (en base 10),se tiene el siguiente algoritmo: Hacer $N_0 = N$ y repetir hasta $N_k = 0$:

$$N_{k+1} = \frac{N_k - a_k}{2}, \quad k = 0, 1, 2, \ldots$$

donde:

$$a_k = \begin{cases} 1, & \text{si} \quad N_k \quad \text{es impar} \\ 0, & \text{si} \quad N_k \quad \text{es par} \end{cases}$$

Ejemplo: Determine la representación binaria de $N = 1999$.

k	0	1	2	3	4	5	6	7	8	9	10
N_k	1999	999	499	249	124	62	31	15	7	3	1
a_k	1	1	1	1	0	0	1	1	1	1	1

por tanto:

$$N = 11111001111_2.$$

Convertir un número fraccionario de base 10 a base 2

Considere los siguientes números:

$$r = 0{,}125, \quad s = 0{,}6666\ldots, \quad \pi = 3{,}141592653\ldots$$

Decimos que r tiene representación finita mientras que s y π tienen representación infinita.
Sea $r \in \mathbb{R}$ tal que $0 < r < 1$ en el sistema decimal y sea $r_2 = (0.d_1 d_2 \ldots d_j \ldots)_2$ su respectiva representación en base 2.
¿Cómo calcular los dígitos d_j para todo $j \geq 1$?

Los dígitos binarios $d_1, \ldots, d_j, \ldots$ se calculan a través del siguiente algoritmo:

Paso 0: $k = 1$, $r_1 = r$.
Paso 1: Determine $2r_k$.
Si $2r_k \geq 1$, asigne $d_k = 1$,
caso contrario, asigne $d_k = 0$.
Paso 2: Asigne $r_{k+1} = 2r_k - d_k$.
Si $r_{k+1} = 0$, pare.
Caso contrario, ir al Paso 3.
Paso 3: $k = k + 1$.
Regrese al Paso 1.

Ejemplo: Calcule la representación binaria de 0,125 y de 0,1.

Veamos el caso de $r = 0{,}125$.

$$
\begin{array}{lll}
k = 1: \quad 2r_1 = 0{,}25 & \Rightarrow & d_1 = 0 \\
 & & r_2 = 0{,}25 - 0 = 0{,}25 \\
k = 2: \quad 2r_2 = 0{,}5 & \Rightarrow & d_2 = 0 \\
 & & r_3 = 0{,}5 - 0 = 0{,}5 \\
k = 3: \quad 2r_3 = 1 & \Rightarrow & d_3 = 1 \\
 & & r_4 = 1 - 1 = 0
\end{array}
$$

por tanto:

$$r = (0{,}001)_2$$

Veamos el caso de $r = 0{,}1$.

$$
\begin{array}{lll}
k = 1: \quad 2r_1 = 0{,}2 & \Rightarrow & d_1 = 0 \\
 & & r_2 = 0{,}2 - 0 = 0{,}2 \\
k = 2: \quad 2r_2 = 0{,}4 & \Rightarrow & d_2 = 0 \\
 & & r_3 = 0{,}4 - 0 = 0{,}4 \\
k = 3: \quad 2r_3 = 0{,}8 & \Rightarrow & d_3 = 0 \\
 & & r_4 = 0{,}8 - 0 = 0{,}8 \\
k = 4: \quad 2r_4 = 1{,}6 & \Rightarrow & d_4 = 1 \\
 & & r_5 = 1{,}6 - 1 = 0{,}6 \\
k = 5: \quad 2r_5 = 1{,}2 & \Rightarrow & d_5 = 1 \\
 & & r_6 = 1{,}2 - 1 = 0{,}2 = r_2 \\
k = 6: \quad 2r_6 = 0{,}4 & \Rightarrow & d_6 = 0 \\
 & & r_7 = 0{,}4 - 0 = 0{,}4 = r_3 \\
\vdots & & \vdots
\end{array}
$$

observamos que los resultados para k de 2 a 5 se repetirán y así $r_{10} = r_6 = r_2 = 0$ y así indefinidamente, por tanto:

$$r = (0{,}0001100110011\overline{0011}\ldots)_2$$

resultando que $(0{,}1)_{10}$ no tiene representación binaria finita.
El hecho de que un número no tiene representación finita en el sistema binario provoca la ocurrencia de errores, dado que un computador sólo almacena un cantidad fija de dígitos y es esta aproximación la que es usada para realizar los cálculos y así es natural esperar un resultado no exacto.

Representación en complemento a dos I

- Es una forma eficiente de representar números enteros con signo.
- Se utiliza el bit más significativo de la palabra binaria para representar el signo (0 si es positivo y 1 si es negativo).
- Para una palabra de n bits, la capacidad de representación es de hasta $2^{n-1}-1$ para números positivos y de 2^{n-1} negativos. La figura siguiente muestra el esquema para n bits:

$n-1$	$n-2$	$\ldots$	2	1	0
s	b	$\ldots$	b	b	b

  donde n es el tamaño de la palabra, s bit para el signo y b bits que representan un número.
- Un número positivo x tal que $0 \leq x \leq 2^{n-1}-1$ es representado por la representación binaria, con bit de signo " 0 ".

Representación en complemento a dos II

- Un número negativo $-y$ donde $1 \leq y \leq 2^{n-1}$ es representado por la representación binaria de $2^n - y$.

Ejemplo: Representar 13 en un computador con longitud de palabra de 8 bits.
Convertimos 13 a base 2:

$$13 = 1101_2$$

luego, la representación pedida es:

7	6	5	4	3	2	1	0
0	0	0	0	1	1	0	1

Representación en complemento a dos III

donde 7 representa el bit más significativo.
Ejemplo: Representar -15 en un computador con longitud de palabra de 8 bits.
Luego $y = 15$ y convertimos $2^8 - 15 = 256 - 15 = 241$ a base 2:

$$241 = 11110001_2$$

luego, la representación pedida es:

7	6	5	4	3	2	1	0
1	1	1	1	0	0	0	1

donde 7 representa el bit más significativo.
Observación: Para negar en complemento a dos, use el siguiente algoritmo:

- Invierta todos los bits del número $(0 \leftarrow 1)$ y $(1 \leftarrow 0)$.

Representación en complemento a dos IV

- Sume 1 al resultado obtenido anteriormente.

Ejemplo: Use el algoritmo anterior para obtener la representación de -15 en una palabra de 8 bits.

$$15 = 00001111_2$$

luego:

- Invirtiendo los bits: 11110000
- Sumando 1:

$$\begin{array}{r} 11110000\ + \\ 00000001 \\ \hline 11110001 \end{array}$$

luego, la representación es:

Representación en complemento a dos V

7	6	5	4	3	2	1	0
1	1	1	1	0	0	0	1

Ejemplo:

Represente todos los números que pueden representarse en un computador que trabaja con longitud de palabra igual a 4.

Solución:

Los números que podemos representar son de 0 a 7 y de -8 a -1. Los resultados se muestran en la siguiente tabla.

bits	número	bits	número	$2^n - y$
0000	0	1000	-8	$2^4 - 8 = 8$
0001	1	1001	-7	$2^4 - 7 = 9$
0010	2	1010	-6	$2^4 - 6 = 10$
0011	3	1011	-5	$2^4 - 5 = 11$
0100	4	1100	-4	$2^4 - 4 = 12$
0101	5	1101	-3	$2^4 - 3 = 13$
0110	6	1110	-2	$2^4 - 2 = 14$
0111	7	1111	-1	$2^4 - 1 = 15$

Representación de números reales I

Fijado un número natural $\beta \geq 2$ que usado como base, todo número real admite una **representación posicional** en la base β como sigue:

$$(-1)^s(a_n\beta^n + a_{n-1}\beta^{n-1} + \ldots + a_1\beta^1 + a_0\beta^0 + a_{-1}\beta^{-1} + a_{-2}\beta^{-2} + \ldots)$$

donde los coeficientes a_i son los **dígitos** en el sistema β, es decir, enteros positivos tal que $0 \leq a_i \leq \beta - 1$.
Los coeficientes $a_{i\geq 0}$ se consideran como los dígitos de la **parte entera**.
Los coeficientes $a_{i<0}$ se consideran como los dígitos de la **parte fraccionaria**.
El número es representado en base β como sigue:

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

donde hemos usado **un punto** para separar la parte entera de la fraccionaria.

Representación de Punto Fijo

Supongamos un computador de longitud de palabra N, es decir, se tiene N posiciones para almacenar un número real respecto de cierta base β. Dado un número real r, una forma de representarlo es utilizar el bit más significativo para el signo, las $N-k-1$ posiciones siguientes para los dígitos de la parte entera y las k posiciones restantes para la parte fraccionaria. De esta forma, la secuencia de N dígitos

$$a_{N-1}\underbrace{a_{N-2}a_{N-3}\ldots a_k}_{N-k-1}\underbrace{a_{k-1}\ldots a_0}_{k}$$

donde $a_{N-1} = s(0 \text{ ó } 1)$ corresponde al número:

$$(-1)^s(a_{N-2}a_{N-3}\ldots a_k\,.\,a_{k-1}\ldots a_1a_0)_\beta = (-1)^s\beta^{-k}\sum_{j=0}^{N-2}a_j\beta^j$$

Ejemplos:

Supongamos $\beta = 10$, $N = 11$ y $k = 6$. Es decir, disponemos de $k = 6$ dígitos para la parte fraccionaria y $N - k - 1 = 4$ dígitos para la parte entera.

1. Represente 30,412:

$$\Rightarrow \boxed{0}\boxed{0030}\boxed{412000}$$

2. Represente $-0{,}0437$:

$$\Rightarrow \boxed{1}\boxed{0000}\boxed{000437}$$

Observación:
Para una longitud de palabra N y con k dígitos para la parte fraccionaria, el rango de valores de los números reales que pueden representarse se encuentra dentro del intervalo $[-\beta^{N-k}, \beta^{N-k}]$.

Representación de números reales en punto flotante

Todo número real no nulo puede ser escrito en forma única, respecto a la base β, en la **notación científica normalizada**:

$$(-1)^s 0.a_1 a_2 a_3 \ldots a_t a_{t+1} a_{t+2} \ldots \times \beta^e,$$

donde los dígitos a_i respecto a la base β son enteros positivos tales que $1 \leq a_1 \leq \beta - 1$, $0 \leq a_1 \leq \beta - 1$ para $i = 1, 2, \ldots$ y constituyen la parte fraccionaria o **mantisa** $(m,\ \beta^{-1} \leq m < 1)$ del número, en tanto que e es el exponente o **característica** indicando la posición del punto correspondiente a la base β.

Sistema de números reales en punto flotante

En todo dispositivo de cálculo, el número de dígitos a representar la mantisa es **finito**, digamos t dígitos en la base β y el exponente sólo puede variar dentro de un rango finito $L \leq e \leq U$ (con $L < 0$ y $U > 0$). Esto implica que sólo un conjunto finito de números reales pueden ser representados, los cuales tienen la forma:

$$(-1)^s 0.a_1 a_2 a_3 \ldots a_t \times \beta^e$$

Tales números se denominan **números de punto flotante** con t dígitos (o de precisión t) en la base β y rango $[L, U]$.
El conjunto de los números de punto flotante es denotado por $\mathbb{F}(\beta, t, L, U)$.

Observaciones:

1. $\mathbb{F}(\beta, t, L, U)$ es discreto y finito. El número de elementos de $\mathbb{F}$ es:

$$2(\beta - 1)\beta^{t-1}(U - L + 1)$$

2. De la definición de normalización resulta que el **cero** no puede ser representado como un número en punto flotante, es decir, $0 \notin \mathbb{F}(\beta, t, L, U)$. Por tanto, definimos $\mathbb{F} = \mathbb{F}(\beta, t, L, U) \cup \{0\}$.
3. Si $x \in \mathbb{F}$ entonces $-x \in \mathbb{F}$.
4. $\mathbb{F}$ es acotado, pues se cumple:

$$x_{min} = \beta^{L-1} \leq |x| \leq x_{max} = \beta^{U}(1 - \beta^{-t}) \quad \forall x \in \mathbb{F}.$$

donde x_{min} y x_{max} son el menor y mayor número en $\mathbb{F}$ positivos.

Overflow y underflow

De las observaciones anteriores, podemos diferenciar 5 regiones en $\mathbb{F}$:

1. Los números negativos menores que $-x_{max}$, región denominada **overflow negativo**.
2. Los números negativos mayores que $-x_{min}$, región denominada **underflow negativo**.
3. El cero.
4. Los números positivos menores que x_{min}, región denominada **underflow positivo**.
5. Los números positivos mayores que x_{max}, región denominada **overflow positivo**.

Representación gráfica de las regiones en $\mathbb{F}$.

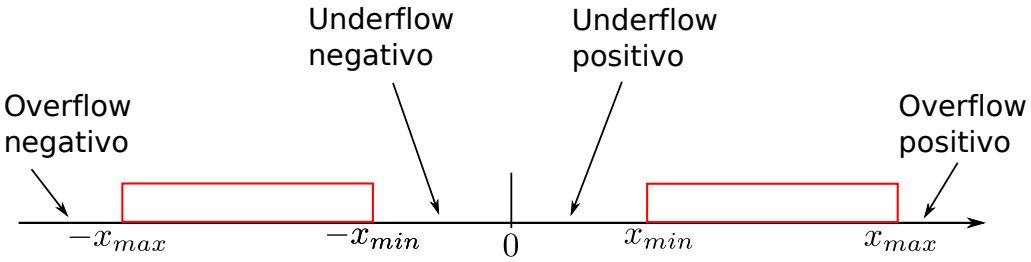


Figura 1: Underflow - Overflow

Ejemplo:

Determine los números del conjunto $\mathbb{F}(2, 3, -1, 2)$.

Solución:

Son los siguientes:

$$(0{,}100)_2 \times 2^{-1} = \frac{1}{4}, \quad (0{,}101)_2 \times 2^{-1} = \frac{5}{16}, \quad (0{,}110)_2 \times 2^{-1} = \frac{3}{8}, \quad (0{,}111)_2 \times 2^{-1} = \frac{7}{16},$$

$$(0{,}100)_2 \times 2^{0} = \frac{1}{2}, \quad (0{,}101)_2 \times 2^{0} = \frac{5}{8}, \quad (0{,}110)_2 \times 2^{0} = \frac{3}{4}, \quad (0{,}111)_2 \times 2^{0} = \frac{7}{8},$$

$$(0{,}100)_2 \times 2^{1} = 1, \quad (0{,}101)_2 \times 2^{1} = \frac{5}{4}, \quad (0{,}110)_2 \times 2^{1} = \frac{3}{2}, \quad (0{,}111)_2 \times 2^{1} = \frac{7}{4},$$

$$(0{,}100)_2 \times 2^{2} = 2, \quad (0{,}101)_2 \times 2^{2} = \frac{5}{2}, \quad (0{,}110)_2 \times 2^{2} = 3, \quad (0{,}111)_2 \times 2^{2} = \frac{7}{2}.$$

Sistema de números reales en punto flotante

Todo $x \in \mathbb{F}$ puede ser escrito en la forma:

$$x = (-1)^s(0.a_1a_2\ldots a_t) \times \beta^e = (-1)^s m\beta^{e-t}$$

donde $m = a_1a_2\ldots a_t$ es un entero positivo tal que $0 \leq m \leq \beta^t - 1$.
Suponiendo un computador con longitud de palabra N, distribuimos 1 bit para el signo, t dígitos para la mantisa (m) y los $N - t - 1$ bits restantes son para el exponente. El número **cero** tiene una representación separada.

Representación de números del sistema en punto flotante

Dado $x \in \mathbb{F}$ entonces: $x = (-1)^s(0.a_1a_2 \ldots a_t) \times \beta^e$

Usualmente un computador tiene dos formatos disponibles para números del sistema en punto flotante $\mathbb{F}$: **precisión simple** y **precisión doble**. En representación binaria, los N bits de la palabra son divididos como sigue:

Formato	N	s	$N-t-1$ bits para " e "	t bits para " m "
Precisión simple	32	1	8	23
Precisión doble	64	1	11	52

Épsilon de la máquina

Los números en punto flotante no están uniformemente distribuidos sobre la recta real, sino que están más próximos cerca del origen y más separados a medida que nos alejamos de él. Con mayor precisión, en un intervalo fijo $[\beta^e, \beta^{e+1}]$ los números de punto flotante presentes están igualmente espaciados con una separación igual a β^{e-t}.
Conforme e se incrementa, el espaciado entre los mismos crece también. Una medida de este espaciamento es dado por el llamado **epsilon de la máquina**

$$\varepsilon_M = \beta^{1-t} \tag{3}$$

el cual representa la distancia entre el número 1 y el número de punto flotante siguiente más próximo, es decir, es el número más pequeño en $\mathbb{F}(\beta, t, L, U)$ tal que $1 + \varepsilon_M > 1$, [2].

Bibliografía

M. A. G. Ruggiero and V. L. d. R. Lopes, *Cálculo numérico: aspectos teóricos e computacionais*.
Makron Books do Brasil, 1997.

A. Quarteroni, R. Sacco, and F. Saleri, *Numerical mathematics*, vol. 37.
Springer Science & Business Media, 2010.