Numeración de Punto Flotante

01

Mario R. Rosenberger

2020
Facultad de Ciencias Exactas Químicas y Naturales
Universidad Nacional de Misiones

Sistema de numeración posicional

Algoritmo: combinación de operaciones fundamentales realizada con números cualesquiera que dan origen a otros números.

Números Romanos…

BCD, código binario decimal

0000 = 0 0001 = 1 0010 = 2 0011 = 3 0100 = 4

0101 = 5 0110 = 6 0111 = 7 1000 = 8 1001 = 9

0101 0000 = 50 0001 0001 0010 = 112

Teorema Fundamental de la Numeración.

Considérese un sistema de numeración posicional de base (natural) x, $x>1$, entonces, cualquier otro natural N puede expresarse, de manera única, en la forma:

$N = a_n x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \ldots + a_2 x^2 + a_1 x^1 + a_0 x^0$ siendo, $a_n, a_{n-1}, a_{n-2}, \ldots, a_2, a_1, a_0$ símbolos del sistema.

la expresión del número N es $a_n\, a_{n-1}\, a_{n-2} \ldots a_2 a_1 a_0$

Los números decimales.

$$d_2\, d_1\, d_0 = d_2 * 10_{(10)}{}^2 + d_1 * 10_{(10)}{}^1 + d_0 * 10_{(10)}{}^0$$

$$735 = 7 * 10_{(10)}{}^2 + 3 * 10_{(10)}{}^1 + 5 * 10_{(10)}{}^0$$

Los números binarios (diádicos)

$$b_n\, b_{n-1} \ldots b_1\, b_0 = b_n * 10_{(2)}{}^n + b_{n-1} * 10_{(2)}{}^{n-1} + \ldots + b_1 * 10_{(2)}{}^1 + b_0 * 10_{(2)}{}^0$$

$$b_n\, b_{n-1} \ldots b_1\, b_0 = b_n * 2_{(10)}{}^n + b_{n-1} * 2_{(10)}{}^{n-1} + \ldots + b_1 * 2_{(10)}{}^1 + b_0 * 2_{(10)}{}^0$$

$$1011011111_{(2)} = 1 * 2_{(10)}{}^9 + 0 * 2_{(10)}{}^8 + 1 * 2_{(10)}{}^7 + 1 * 2_{(10)}{}^6 + 0 * 2_{(10)}{}^5 + 1 * 2_{(10)}{}^4 + 1 * 2_{(10)}{}^3 + 1 * 2_{(10)}{}^2 + 1 * 2_{(10)}{}^1 + 1 * 2_{(10)}{}^0$$

```
BaseForm[735, 2]
```

1011011111_2

DECIMAL	BINARIO
1	1
2	10
3	11
4	100
5	101
6	110
7	111
8	1000
9	1001
10	1010
11	1011
12	1100
13	1101
14	1110
15	1111
16	10000
17	10001
18	10010
19	10011
20	10100

n	2^n
0	1
1	2
2	4
3	8
4	16
5	32
6	64

Sistema de numeración posicional

Números no enteros.

Los números

$$d_0\, d_{-1}\, d_{-2} = d_0 * 10_{(10)}{}^{0} + d_{-1} * 10_{(10)}{}^{-1} + d_{-2} * 10_{(10)}{}^{-2}$$

$$3.25 = 3 * 10_{(10)}{}^{0} + 2 * 10_{(10)}{}^{-1} + 5 * 10_{(10)}{}^{-2}$$

n	2^n	2^n
0	1	1.
-1	$\frac{1}{10}$	0.1
-2	$\frac{1}{100}$	0.01
-3	$\frac{1}{1000}$	0.001
-4	$\frac{1}{10\,000}$	0.0001
-5	$\frac{1}{100\,000}$	0.00001
-6	$\frac{1}{1\,000\,000}$	$1. \times 10^{-6}$

Los números binarios (diádicos)

$$b_1 * 2_{(10)}{}^{1} + b_0 * 2_{(10)}{}^{0} + b_{-1} * 2_{(10)}{}^{-1}$$

$$11.01_{(2)} = 1 * 2_{(10)}{}^{1} + 1 * 2_{(10)}{}^{0} + 0 * 2_{(10)}{}^{-1} + 1 * 2_{(10)}{}^{-2}$$

n	2^n	2^n
0	1	1.
-1	$\frac{1}{2}$	0.5
-2	$\frac{1}{4}$	0.25
-3	$\frac{1}{8}$	0.125
-4	$\frac{1}{16}$	0.0625
-5	$\frac{1}{32}$	0.03125
-6	$\frac{1}{64}$	0.015625

n	2^n
0	1
1	2
2	4
3	8
4	16
5	32
6	64

Representación de números en la Computadora

Números enteros.

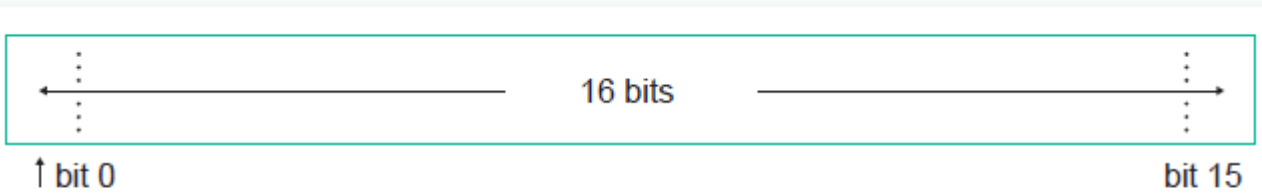


Figura 1.6 Esquema de una palabra de 16 bits para un número entero.

$525_{10} = 1015_{8} = 1000001101_{2}$, y su almacenamiento quedaría de la siguiente forma:

0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	1

$-26_{10} = -11010_{2}$ y su almacenamiento en una palabra de 16 bits quedaría así:

1	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0

Representación de números en la Computadora

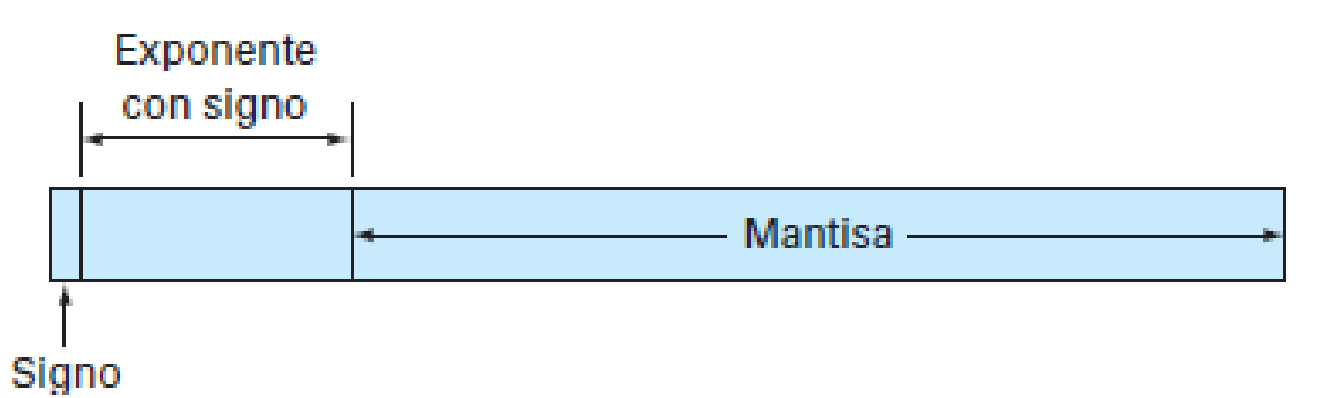
Números de coma flotante.

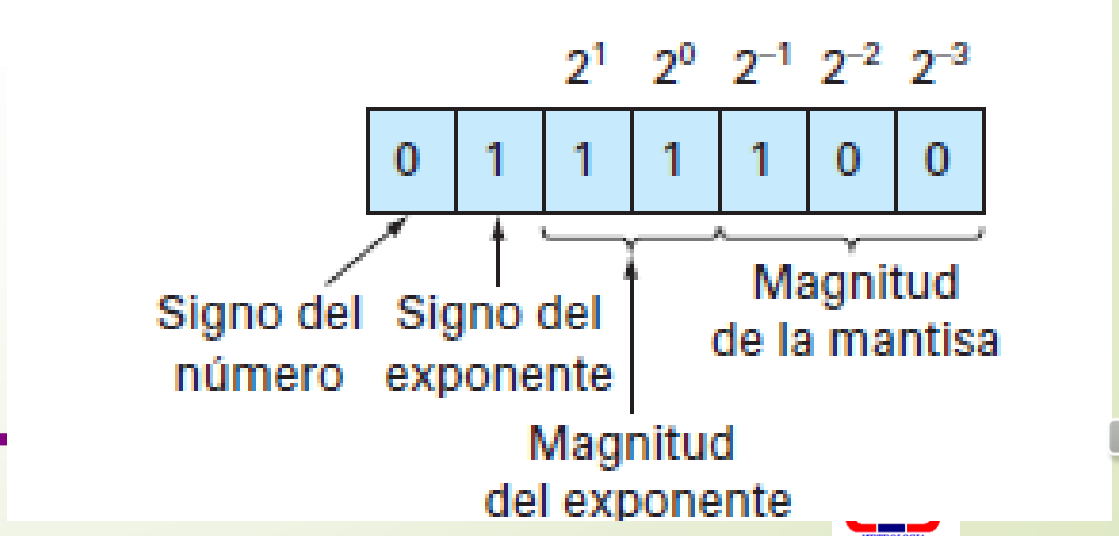
$$x = m * b^e$$

m: mantisa
b: base
e: exponente, o característica

$$12{,}57 = 0{,}1257 * 10^2$$

$$-0{,}0486 = -0{,}486 * 10^{-1} \qquad \rightarrow -0{,}0486 * 10^0$$





Representación de números en la Computadora

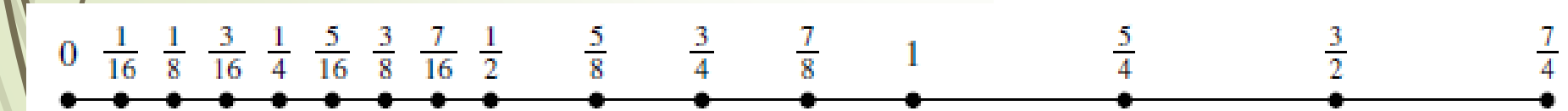
Números de coma flotante.

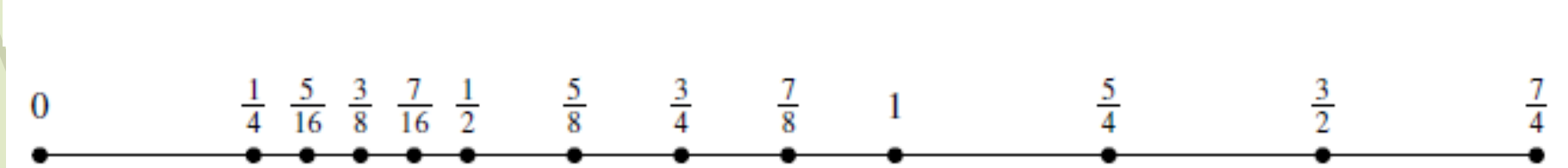
m: mantisa

$$x = m * b^e$$

$$x = \pm(0.1b_2b_3)_2 \times 2^{\pm k}$$

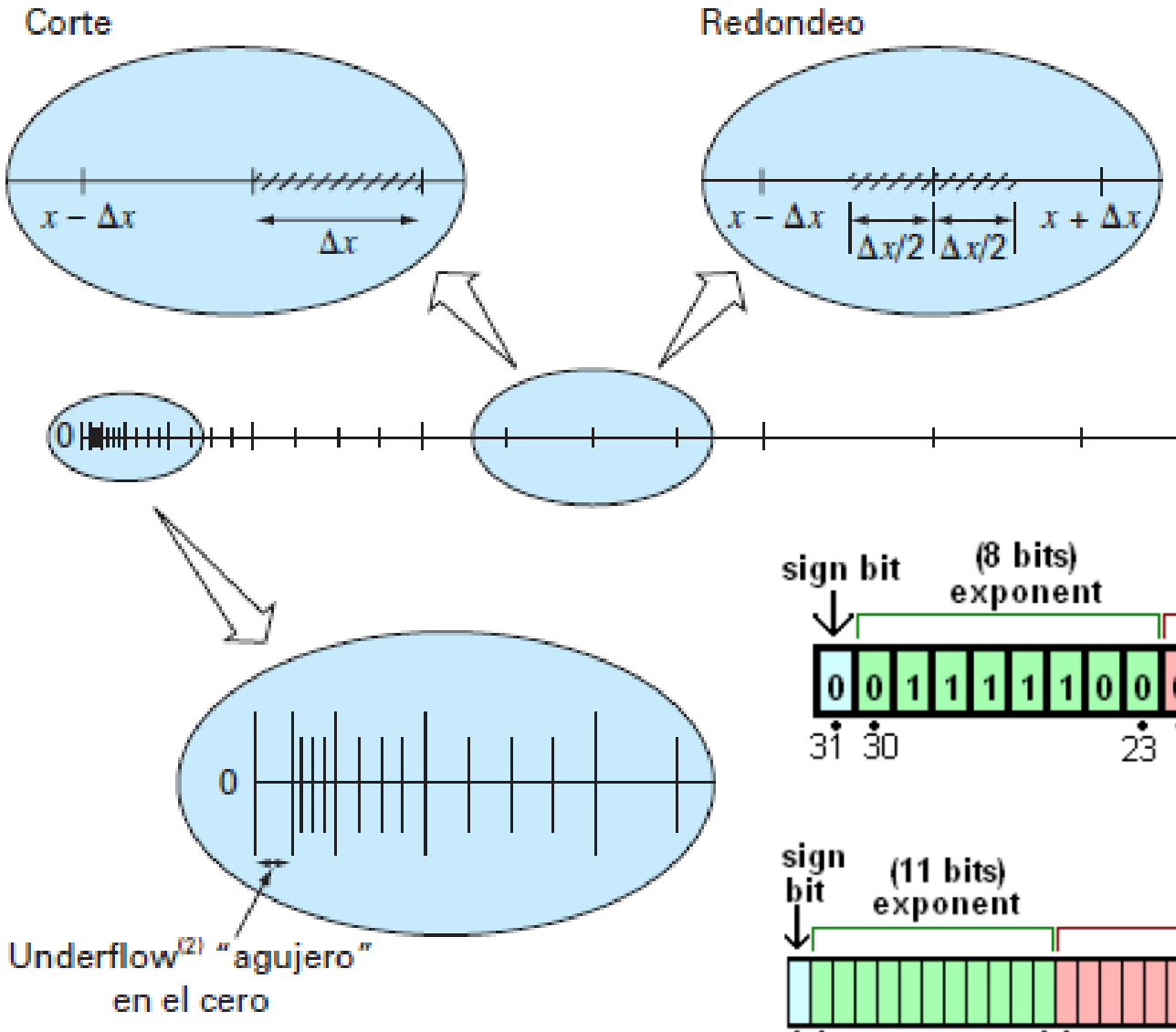
$0.000 \times 2^0 = 0$	$0.000 \times 2^1 = 0$	$0.000 \times 2^{-1} = 0$
$0.001 \times 2^0 = \frac{1}{8}$	$0.001 \times 2^1 = \frac{1}{4}$	$0.001 \times 2^{-1} = \frac{1}{16}$
$0.010 \times 2^0 = \frac{2}{8}$	$0.010 \times 2^1 = \frac{2}{4}$	$0.010 \times 2^{-1} = \frac{2}{16}$
$0.011 \times 2^0 = \frac{3}{8}$	$0.011 \times 2^1 = \frac{3}{4}$	$0.011 \times 2^{-1} = \frac{3}{16}$
$0.100 \times 2^0 = \frac{4}{8}$	$0.100 \times 2^1 = \frac{4}{4}$	$0.100 \times 2^{-1} = \frac{4}{16}$
$0.101 \times 2^0 = \frac{5}{8}$	$0.101 \times 2^1 = \frac{5}{4}$	$0.101 \times 2^{-1} = \frac{5}{16}$
$0.110 \times 2^0 = \frac{6}{8}$	$0.110 \times 2^1 = \frac{6}{4}$	$0.110 \times 2^{-1} = \frac{6}{16}$
$0.111 \times 2^0 = \frac{7}{8}$	$0.111 \times 2^1 = \frac{7}{4}$	$0.111 \times 2^{-1} = \frac{7}{16}$





Representación de números en la Computadora

Números de coma flotante.



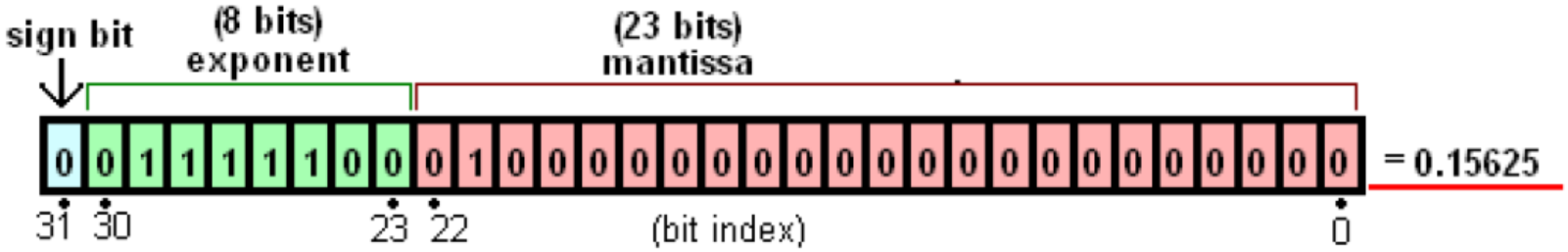
IEEE Standard 754 for Binary Floating-Point Arithmetic

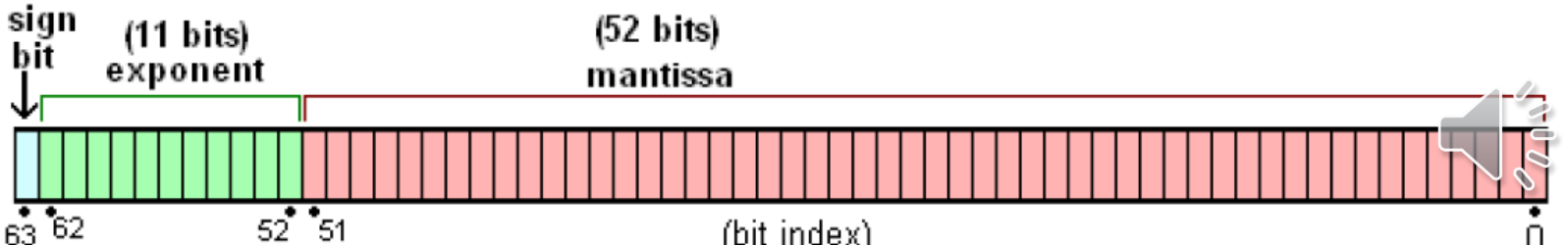
double	0 $2.23 \times 10^{-308} \leqslant x \leqslant 1.80 \times 10^{308}$	$2^{-1022} \rightarrow 2^{1024}$

Listing 1: Matlab Tools

```
>> realmax
ans = 1.7977e+308

>> realmin
ans = 2.2251e-308
```





Causas graves de errores en la computación con NPF. $m * b^e$

$$3.0 = .3000 \times 10^{1}$$
$$7956000 = .7956 \times 10^{7}$$
$$-0.0000025211 = -.2521 \times 10^{-5}$$

si se tiene una computadora
Decimal que tiene 4 cifras en la mantisa y 2 lugar para el exponente con signo

Sumas de números muy distintos en magnitud

sumar 0.002 a 600 en la computadora decimal imaginaria.

$$0.002 = .2000 \times 10^{-2}$$
$$600 = .6000 \times 10^{3}$$

$$\begin{array}{rr} & .000002 \times 10^{3} \\ + & .600000 \times 10^{3} \\ \hline & .600002 \times 10^{3} \end{array}$$

Resta de números casi iguales

a a restar 0.2144 de 0.2145.

$$\begin{array}{rr} & .2145 \times 10^{0} \\ - & .2144 \times 10^{0} \\ \hline & .0001 \times 10^{0} \end{array}$$

el resultado se almacena como $.1000 \times 10^{-3}$.

computació

Causas graves de errores en la computación con NPF. $m * b^e$

Overflow

al multiplicar 0.5000×10^8 por 0.2000×10^9, se tiene

$$\begin{array}{r} 0.5000 \times 10^8 \\ \times \quad 0.2000 \times 10^9 \\ \hline 0.1000 \times 10^{17} \end{array}$$

si se tiene una computadora Decimal que tiene 4 cifras en la mantisa y 2 lugar para el exponente con signo.

Underflow

$$(0.3000 \times 10^{-5}) \times (0.02000 \times 10^{-3}) = 0.006 \times 10^{-8} = 0.6000 \times 10^{-10}$$

Algunas veces es salvable.

$A = 0.3000 \times 10^{-5}$, $B = 0.0200 \times 10^{-3}$, $C = 0.4000 \times 10^{7}$,

$X = A * B * C$

$X = A * C * B$

A por C y se obtiene 0.1200×10^{2}.

0.2400×10^{-5}.

Causas graves de errores en la computación con NPF. $m * b^e$

Division entre un numero muy pequeño

$$X = A - B / C$$

$A = 0.1120 \times 10^9 = 112000000$

$B = 0.1000 \times 10^6 = 100000$

$C = 0.900 \times 10^{-3} = 0.0009$

si se tiene una computadora Decimal que tiene 4 cifras en la mantisa y 2 lugar para el exponente con signo.

Si el cálculo se realiza en la computadora decimal de cuatro dígitos, el cociente B / C es 0.1111×10^9, y X es 0.0009×10^9 o, después de ser normalizado, $X = 0.9000 \times 10^6$. Nótese que sólo hay un dígito significativo.

Vamos a imaginar ahora que se cometió un pequeño error de redondeo al calcular C en algún paso previo y resultó un valor $C^* = 0.9001 \times 10^{-3}$ ($EA = 0.0001 \times 10^{-3}$; $ER = 10^{-4}$ y $ERP = 0.01\%$).

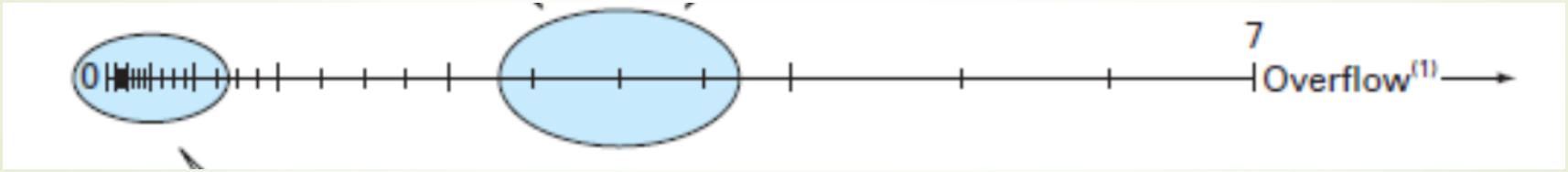
Si se calcula B / C^* se obtiene como cociente 0.1110×10^9 y $X^* = 0.1000 \times 10^7$. El valor correcto de X es 0.9000×10^6.

Causas graves de errores en la computación con NPF. $m * b^e$

Error de discretización

si se tiene una computadora Decimal que tiene 4 cifras en la mantisa y 2 lugar para el exponente con signo.

Dado que un número específico no se puede almacenar exactamente como número binario de punto flotante, el error generado se conoce como error de discretización (error de cuantificación), ya que los números expresados exactamente por la máquina (números máquina) no forman un conjunto continuo sino discreto.



Errores de salida

Tiene que ver con la cantidad de dígitos que usa el programa para dar un resultado en la pantalla.

Fin del bloque

1.- Chapra, S. C. y Canale, R. P. *Métodos numéricos para ingenieros.* 5ta. edición. McGraw-Hill. (2002).

2.- D. Kincaid, N. Cheney, Métodos numéricos y computación, Addison-Wesley. (2015)

3.- Burden, R. L. Y Faires, J. D. *Análisis Numérico*. 6ta. Ed.Thomson International. Méjico.(1998).

4,- Nieves Hurtado y Dominguez Sanchez, Métodos numéricos Aplicados a la ingeniería. Cecsa. 2014.