Punto Flotante

Forma de codificar números reales, basada en la notación científica $\pm 7{,}13 \times 10^{-6}$ pero usando base 2. $\pm 1{,}\ldots \times 2^{e}$

Representamos los números con una cantidad Fija de bits. 32 bits precisión simple
64 bits precisión doble
usa octave por defecto

$\Rightarrow$ Solo podemos representar finitos números. n bits $\rightarrow 2^n$ combinaciones.

(2) usamos precisión simple. Mantisa

$$x = \pm (1+f) 2^{e}$$

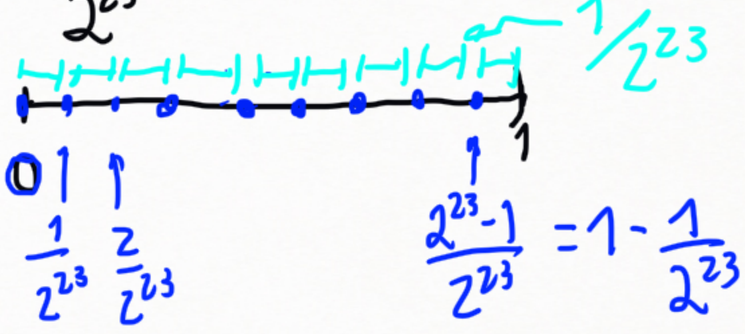
$0 \le f < 1$

$f = \frac{j}{2^{23}}$ con $j \in \{0, 1, \ldots, 2^{23}-1\}$ son 2^{23}

1 bit signo
23 bits mantisa
8 bits exponente
1 bit signo, 52 bits mantisa, 11 bits exponente
precisión doble

2^{23} posibles valores equiespaciados

$0, \frac{1}{2^{23}}, \frac{2}{2^{23}}, \ldots \frac{j}{2^{23}}, \ldots \frac{2^{23}-1}{2^{23}}$

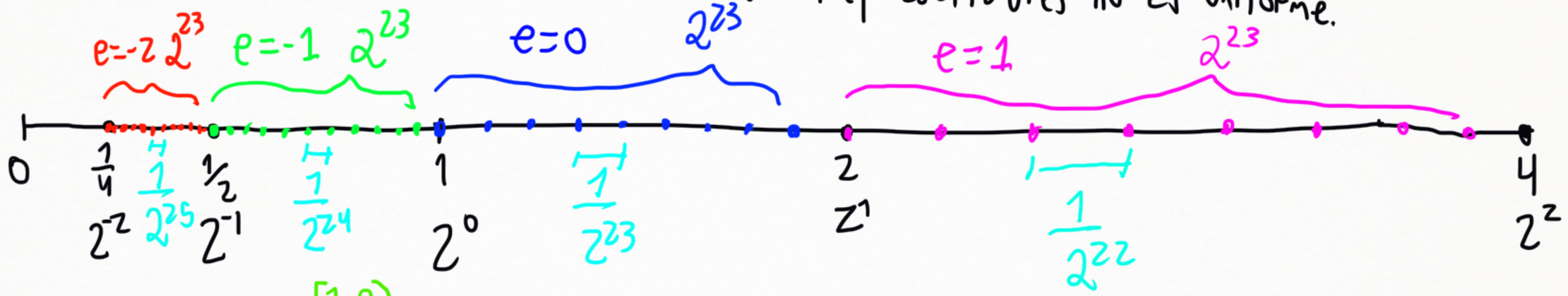


Exponente

$-126 \le e \le 127$

+ los valores $128, -127$ se reservan para ciertos casos
0, InF, NaN

La distribución de los números representables no es uniforme.



$[1,2)$

exponente = e: $x = (1+F)2^e$

$0 \le F < 1 \Rightarrow 2^e \le x < 2^{e+1}$

$F=0 \to x = 2^e$

$F=1 \to x = (1+1)2^e$ (cota sup)
$= 2 \cdot 2^e = 2^{e+1}$

$e \to x \in [2^e, 2^{e+1})$

por los distintos valores de F, tenemos 2^{23} números equiespaciados.

separación entre uno y el siguiente: 2^{e-23}

$F=0 \to 2^e$

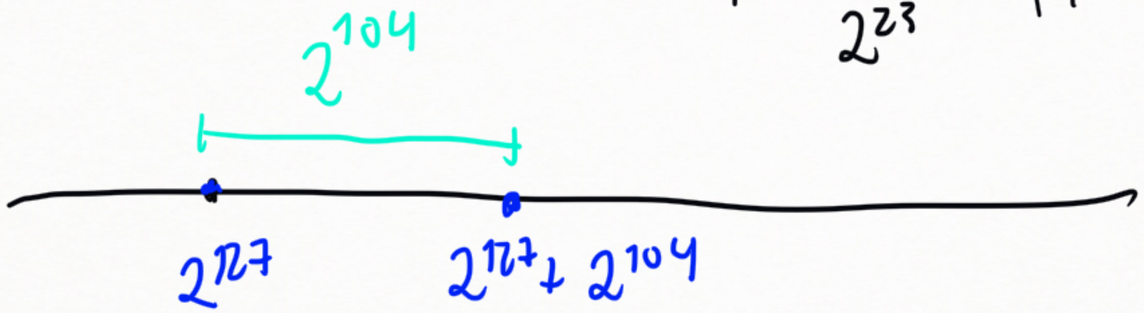
$F=\frac{1}{2^{23}} \to \left(1+\frac{1}{2^{23}}\right)2^e = 2^e + 2^{e-23}$

$F=\frac{2}{2^{23}} \to \left(1+\frac{2}{2^{23}}\right)2^e = 2^e + 2\times 2^{e-23}$

La máxima separación es con el exponente más grande.

$e=127$: $x = (1+F)\,2^{127}$ $\quad F=0: \; 2^{127}$

$F=\frac{1}{2^{23}}: \; (1+\frac{1}{2^{23}})\,2^{127} = 2^{127} + \frac{2^{127}}{2^{23}} = 2^{127}+2^{104}$



Normalmente nos importa más el error relativo que el absoluto.

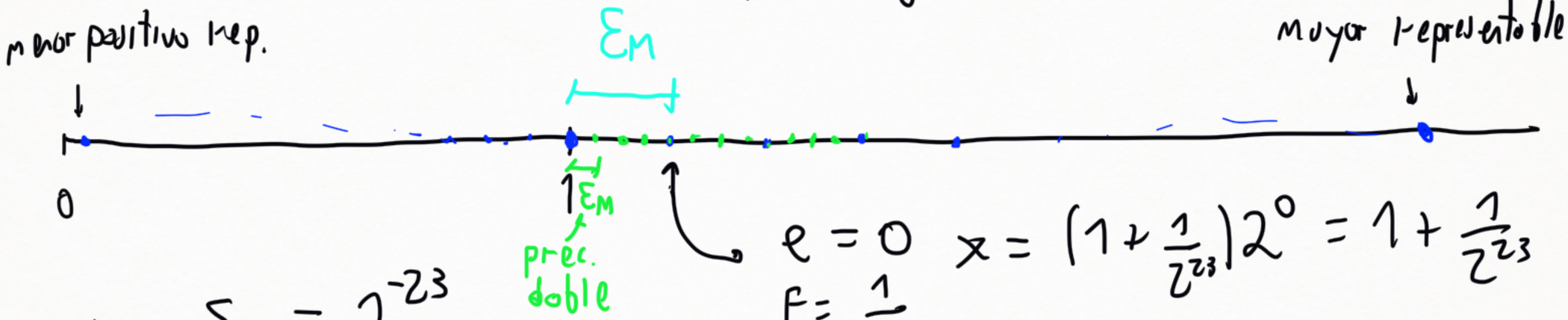
$x \in \mathbb{R} \quad x \neq 0$ $\qquad e_x := \hat{x} - x$ error absoluto

$\hat{x}$ aproximación. $\qquad \varepsilon_x := \frac{\hat{x}-x}{x}$ error relativo.

$x=30 \quad \hat{x}=32 \quad \varepsilon_x = \frac{32-30}{30} = \frac{2}{30} = \frac{1}{15}$

$x=100000000 \quad \hat{x}=100000002 \quad \varepsilon_x = \frac{2}{100000000}$ chico

ε_M: la separación entre 1 y el siguiente número representable.



$e = 0 \quad x = \left(1 + \frac{1}{2^{23}}\right) 2^0 = 1 + \frac{1}{2^{23}}$

$F = \frac{1}{2^{23}}$

en este caso $\varepsilon_M = 2^{-23}$

en precisión doble, 52 bits para mantiso, $\varepsilon_M = 2^{-52}$

Mayor representable

$(1+F)2^e$ F, e lo más grandes posible

$e = 127$

$F = \frac{2^{23}-1}{2^{23}} = 1 - \frac{1}{2^{23}}$

$x = \left(1 + 1 - \frac{1}{2^{23}}\right) 2^{127}$

$= \left(2 - \frac{1}{2^{23}}\right) 2^{127} = 2^{128} - 2^{104}$

Menor representable

$(1+F)2^e$

$F = 0, \quad e = -126$

$x = 2^{-126}$

$\varepsilon_M \simeq 1{,}2 \times 10^{-7}$

$M_{ax} \simeq 3{,}4 \times 10^{38}$

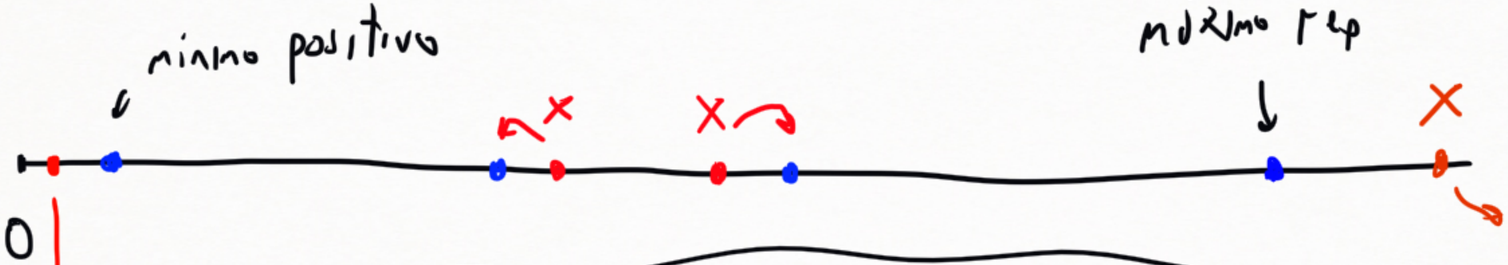
$M_{in} \simeq 1{,}2 \times 10^{-38}$

e	sep. 2^{e-23}
e=23	sep = 1
e=24	sep = 2

$(1+F)2^e$

$\left(1 + \frac{j}{2^{23}}\right) 2^e$

$= 2^e + 2^{e-23} j$



0

0

usar subnormales

$$g(x) = \frac{x^3}{x - \operatorname{sen}(x)}$$

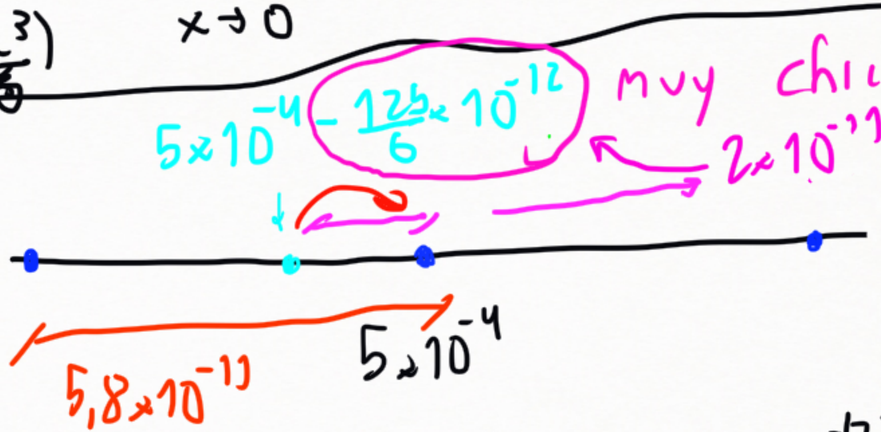
$$\operatorname{sen}(x) = x - \frac{x^3}{6} + H(x) \quad \text{(Taylor orden 3)}$$

$$\lim_{x\to 0} g(x) = \lim_{x\to 0} \frac{x^3}{x - \operatorname{sen}(x)} = \lim_{x\to 0} \frac{x^3}{x - \left(x - \frac{x^3}{6}\right)} = \lim_{x\to 0} 6 = 6$$

$$x - \operatorname{sen}(x) \approx x - \left(x - \frac{x^3}{6}\right)$$

$$5\times10^{-4} - \left[5\times10^{-4} - \frac{125}{6}\times10^{-12}\right]$$

5×10^{-4}



e

sep 2^{e-23}

$5\times10^{-4} \in [2^{-11}, 2^{-10})$ separado $2^{-(23+11)} = 2^{-34}$

$\simeq 5,8\times10^{-11}$