17/09/24

ARITMETICA di MACCHINA

$x = 123456.789$ (n: 6, m: 3)

$10^3 \quad 10^0 \quad 10^{-1}$

notazione posizionale

$$x = \sum_{k=-m}^{n} a_k \cdot \beta^k$$

β = base (10)

a_k sono le cifre : $a_k \in \{0, 1, \dots, \beta-1\}$

---

notazione scientifica o esponenziale

$x = 0.123456789 \cdot 10^6$ ≠ FLOATING POINT

$$x = (-1)^s \cdot 0.a_1 a_2 \dots a_t \cdot \beta^e$$

$s \in \{0,1\}$ $\quad a_i \in \{0, 1, \dots, \beta-1\}$, $a_1 \neq 0$

t cifre della mantissa

mantissa $m = a_1 a_2 \cdots a_t$

β = base di rappresentazione

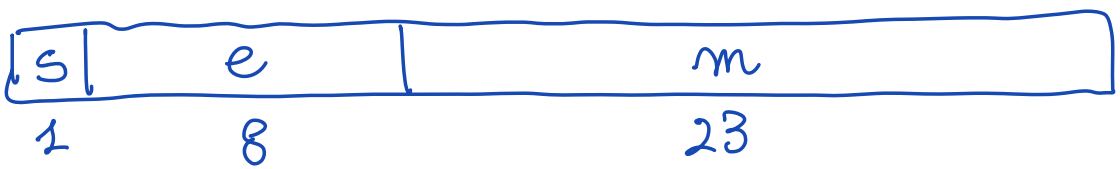
e = esponente della base (ordine di grandezza)

t = n° cifre della mantissa (precisione)

IEEE semplice precisione 4 Byte

doppia precisione 8 Byte

$\beta = 2$ 4 Byte (semplice)



es $1.f \neq 0$ con 23 bit per la mantissa, ho in realtà

$t = 24$ (n° cifre in base 2)

$\sim$ 6 o 7 cifre in base 10

$\beta = 2$, 8 Byte (doppia precisione)

$a_1 \neq 0 \Rightarrow a_1 = 1$, $t = 53$

$-1021 \leq e \leq 1024$ (2^e $\beta = 2$)

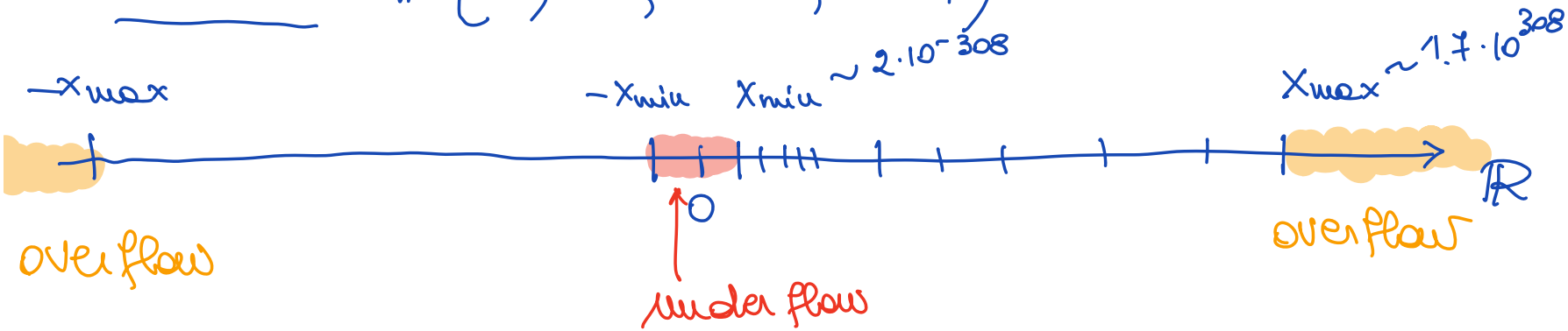
Sistema floating point

$$\mathbb{F}(\beta, t, L, U) = \left\{ x = (-1)^s \cdot 0.a_1 a_2 \cdots a_t \cdot \beta^e \text{, con:} \right.$$

$$s \in \{0,1\},\ a_i \in \{0, \ldots, \beta-1\},\ a_1 \neq 0,\ L \leq e \leq U$$

con $L < 0$, $U > 0$ $\} \cup \{0\}$

Matlab $\mathbb{F}(2, 53, -1021, 1024)$



$\mathbb{F}$ è simmetrico rispetto a 0
è limitato (non va a $+\infty$ $-\infty$)
è finito (ha un n° finito di elementi)

i numeri fp non sono equispaziati, ma sono + densi verso lo 0 e meno densi man mano che cresce l'ordine di grandezza.

Es: $F(2, 3, -2, 3)$
$\beta \quad t \quad L \quad U$

$x = (-1)^s \cdot 0.a_1a_2a_3 \cdot 2^e$

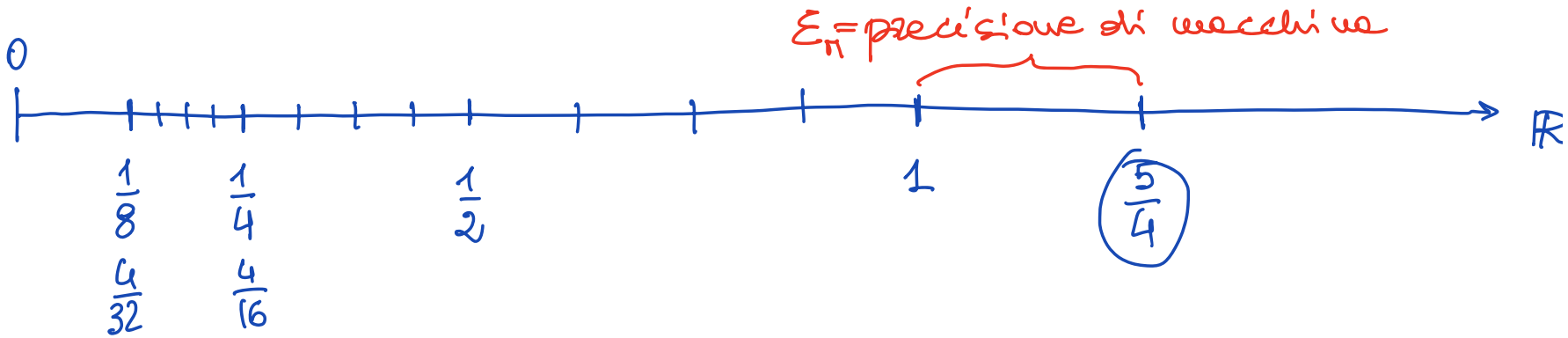
per simmetria prendo solo $s=0$

base 2 — base 10

$x_{min} = \begin{cases} 0.100 \cdot 2^{-2} = 2^{-1} \cdot 2^{-2} = \frac{1}{8} = \frac{4}{32} \\ 0.101 \cdot 2^{-2} = (2^{-1} + 2^{-3}) \cdot 2^{-2} = \left(\frac{1}{2} + \frac{1}{8}\right) \cdot \frac{1}{4} = \frac{5}{32} \\ 0.110 \cdot 2^{-2} = (2^{-1} + 2^{-2}) \cdot 2^{-2} = \left(\frac{1}{2} + \frac{1}{4}\right) \cdot \frac{1}{4} = \frac{3}{16} = \frac{6}{32} \\ 0.111 \cdot 2^{-2} = \ldots\ldots\ldots = \frac{7}{32} \end{cases}$

$\begin{cases} 0.100 \cdot 2^{-1} = 2^{-1} \cdot 2^{-1} = \frac{1}{4} \\ 0.101 \cdot 2^{-1} = (2^{-1} + 2^{-3}) \cdot 2^{-1} = \left(\frac{1}{2} + \frac{1}{8}\right) \cdot \frac{1}{2} = \frac{5}{16} \\ 0.110 \cdot 2^{-1} = \frac{6}{16} \\ 0.111 \cdot 2^{-1} = \frac{7}{16} \end{cases}$

$\frac{1}{4} = \frac{8}{32} = \frac{4}{16}$



ε_m = distanza tra il n° 1 e il più piccolo numero floating point maggiore di 1

$$\frac{5}{4} = 0.101 \cdot 2^1$$

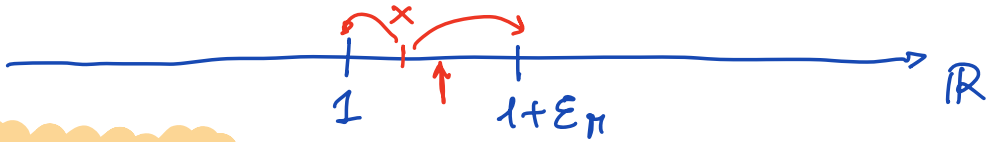
$$1 = 0.100 \cdot 2^1$$

$$\varepsilon_m = 0.001 \cdot 2^1 = \beta^{-t} \cdot \beta^1 = \beta^{1-t}$$

(t: β^{-t}; β)

$$\boxed{\varepsilon_m = \beta^{1-t}}$$

24/09



$x \in \mathbb{R}$: $1 \leq x \leq 1+\varepsilon_M$

Se $x \notin \mathbb{F} \Rightarrow x$ sarà approx dal n° fp più vicino a x

$fl_t(x)$ = rappresentante di macchina di x

Si dimostra $|x - fl_t(x)| \leq \frac{1}{2}\varepsilon_M$

Se $x \notin (1, 1+\varepsilon_M)$ allora si ha

errore relativo tra x e il suo float

$$\frac{|x - fl_t(x)|}{|x|} \leq \frac{1}{2}\varepsilon_M$$

$u = \frac{1}{2}\varepsilon_M$ unità di roundoff

MATLAB (doppia prec) $\beta = 2, t = 53 \Rightarrow \varepsilon_M \sim 2.22 \cdot 10^{-16}$

$u \sim 1.11 \cdot 10^{-16}$

Proprietà che cadono in $\mathbb{F}$

1) non vale più la prop associativa di +e.

es: doppia prec, $x_{max} \sim 1.8 \cdot 10^{308}$

$a = 10^{308}$ $b = 10^{308}$ $c = -4 \cdot 10^{307}$

$\underbrace{(a+b)}_{2 \cdot 10^{308} > x_{max}} + c$ $\neq$ $a + \underbrace{(b+c)}_{6 \cdot 10^{307}}$

Nan o Inf $\underbrace{\qquad\qquad}_{1.6 \cdot 10^{308} < x_{max}}$

Es2 $x = 10^{-14}$

$\frac{1 + (-1+x)}{x} = 0.9992$ in matlab

$\frac{(1-1)+x}{x} = 1$ in matlab

2) lo zero, in fatti come elemento neutro della somma, non è unico

$0 : \quad x + 0 = 0 + x = x \quad \forall x \in \mathbb{R}$

In $\mathbb{F}$, dato $x \in \mathbb{F}$ $\exists$ molti numeri vecchina $y \in \mathbb{F}$

t.c. $x + y = x$

Es $\mathbb{F}(10, 4, -5, 6)$ $\qquad \varepsilon_M = \beta^{1-t} = 10^{-3}$
$\beta \quad t \quad L \quad U$

$x = 1 \qquad y = 10^{-4} \qquad ? \; x + y$

$x = 1 = 0.1000 \cdot 10^1 = 0.1000 \cdot 10^1$

$y = 10^{-4} = 0.1000 \cdot 10^{-3} = 0.0000\,1 \cdot 10^1$

$0.1000\,(1) \cdot 10^1$

perso nel salvare il risultato in m

$\Rightarrow x + y = 0.1000 \cdot 10^1 = 1 = x$

Propagazione degli errori di arrotondamento

$x \longrightarrow fl_t(x)$

1) Somma

dati $x, y \in \mathbb{R}$? $x+y$, in realtà calcolo $\bar{x}+\bar{y}$
anzi calcolo
$fl_t(\bar{x}+\bar{y})$

$\bar{x} = fl_t(x)$ $fl_t(y) = \bar{y}$

la somma di 2 numeri F non è detto che stia in F

es: $F(10, 3, -2, 3)$

$\bar{x} = 0.123 \cdot 10^0 = 0.123 \cdot 10^0$
$\bar{y} = 0.456 \cdot 10^{-2} = 0.00456 \cdot 10^0$

$\bar{x}+\bar{y} = 0.12756 \cdot 10^0$

$fl_t(\bar{x}+\bar{y}) = 0.128 \cdot 10^0$

Che errore commetto nel prendere come soluzione
$fl_t(\bar{x}+\bar{y})$ al posto di $(x+y)$

risultato numerico

risultato esatto

$$\frac{\left|(x+y) - fl_t(\bar{x}+\bar{y})\right|}{|x+y|} \leq \underbrace{\left(\frac{|x|+|y|}{|x+y|} + 1\right)}_{C_{x,y}} \cdot u$$

u = unità di roundoff (dp $u \sim 1.11 \cdot 10^{-16}$)

se $x \sim -y \Rightarrow |x+y| \sim 0$, $|x|+|y| >> 0$

$\frac{|x|+|y|}{|x+y|}$ può diventare molto grande

es $x = 1 + 10^{-15}$ $\quad y = -1$

$|x|+|y| = 2 + 10^{-15}$

$|x+y| = 10^{-15}$

$$\frac{|x|+|y|}{|x+y|} \sim \frac{2}{10^{-15}} \sim 2 \cdot 10^{15}$$

$\Rightarrow C_{xy} \cdot u \simeq 2 \cdot 10^{15} \cdot 1.11 \cdot 10^{-16} \sim 2 \cdot 10^{-1}$

A fronte di una unità di roundoff di 10^{-16} in una data somma arrivo ad un errore nella sol di circa 10^{-1}.

La somma è un'operazione instabile perché a piccoli errori sui dati possono corrispondere grandi errori sulla soluzione.

Gli errori alla + sono detti errori di cancellazione

2) prodotto

$x, y \in \mathbb{R}$, $\bar{x} = fl_t(x)$, $\bar{y} = fl_t(y)$? $x \cdot y$ esatto

trovo $fl_t(\bar{x} \cdot \bar{y})$ (numerico) $x, y \neq 0$

$$err = \frac{|(x \cdot y) - fl_t(\bar{x} \cdot \bar{y})|}{|x \cdot y|} \leq C \cdot u$$

$\uparrow_{\sim 3}$ indip da x e y

il prodotto è un'op. stabile.