ARITHETICA DI MACCHINA

1) Notazione posizionale
2) Notazione scientifica

1) $x = 12\,3456.789 = \sum_{i=-3}^{6} a_i \cdot 10^i$

(a_i · 100 · 1 · $\frac{1}{10}$ · $\frac{1}{100}$)

$x = (-1)^s \cdot \sum_{i=-m}^{n} a_i \beta^i$ $\quad \beta$ = base di rappresent.

con $s \in \{0,1\}$ $\quad a_i$ cifre

va bene per sistemi bancari $\quad 0 \le a_i \le \beta - 1$

ma non nel CS.

2) $x = 0.\underbrace{123456789}_{\text{mantissa}} \cdot \underbrace{10}_{\beta}^{\,6}$ (6 = esponente)

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

$s \in \{0,1\}$ $\quad 0 \le a_i \le \beta - 1$

$m = a_1 a_2 \dots a_t \in \mathbb{N}$ mantissa $\quad a_1 \neq 0$

β = base

e = esponente della base

t = n° cifre della mantissa

Sistema floating point (= virgola mobile)

$$\mathbb{F}(\beta, t, L, U) = \{ x = (-1)^s \cdot 0.a_1 \dots a_t \cdot \beta^e \text{ con } s \in \{0,1\},\ 0 \le a_i \le \beta - 1,\ a_1 \neq 0,\ L < 0,\ U > 0,\ L \le e \le U \} \cup \{0\}$$

(β: base; t: n° cifre mantissa >0; L: <0; U: >0)

- IEEE semplice precisione float 32

32 bit
4 Byte



$\beta = 2$

$2^8 = 256$ (-2 per le eccez.)

$-125 \le e \le 128$

L U

Se $\beta = 2$, $0 \le a_i \le 1$, $a_1 \neq 0 \Rightarrow a_1 = 1$ sempre

23 bit $\longrightarrow$ 24 cifre in base 2 $\Rightarrow t = 24$

24 cifre in base 2 $\sim$ 6/7 cifre in base 10

- IEEE doppia precisione 64 bit = 8 Byte float64



$\beta = 2$

52 bit $\Rightarrow$ 53 cifre in base 2 $\Rightarrow t = 53$

$\sim$ 15/16 cifre in base 10

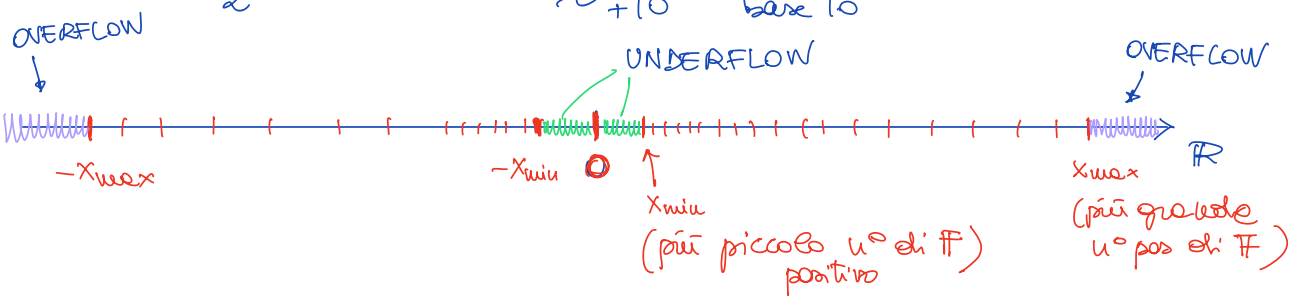
$2^{11} - 2 = 2046$

$-1021 \le e \le +1024$

L U

2^{-1021} in base 2 $\sim 10^{-308}$ base 10

$2^{1024} \sim +10^{+308}$ base 10



Esempio

$F(2, 3, -2, 3)$

β t L U

$x_{min} = 0.\underbrace{1}_{\frac{1}{2}}\underbrace{00}_{\text{base 2}} \cdot \underbrace{\beta_2^{-2}}_{\text{base 10}} = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8} = \frac{4}{32}$

not. posiz

$0.101 \cdot 2^{-2} = \left(\frac{1}{2} + \frac{1}{8}\right) \cdot \frac{1}{4} = \frac{5}{8} \cdot \frac{1}{4} = \frac{5}{32}$

$0.110 \cdot 2^{-2} = \left(\frac{1}{2} + \frac{1}{4}\right) \cdot \frac{1}{4} = \frac{3}{16} = \frac{6}{32}$

$0.111 \cdot 2^{-2} = \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8}\right) \cdot \frac{1}{4} = \frac{7}{8} \cdot \frac{1}{4} = \frac{7}{32}$

21/09/26

$0.100 \cdot 2^{-1} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = \frac{4}{16}$

$0.101 \cdot 2^{-1} = 5/16$

$0.110 \cdot 2^{-1} = 6/16$

$0.111 \cdot 2^{-1} = 7/16$

$0.100 \cdot 2^{0} = 1/2 = 4/8$

$0.101 \cdot 2^{0} = 5/8$

$0.110 \cdot 2^{0} = 6/8$

$0.111 \cdot 2^{0} = 7/8$

$0.100 \cdot 2^{1} = 1 = 4/4$

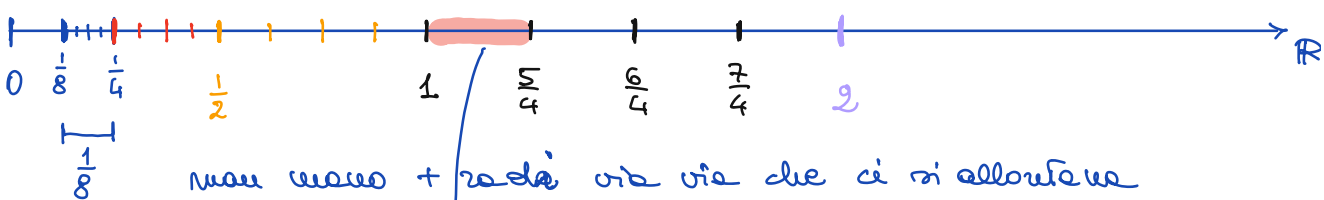
$0.101 \cdot 2 = 5/4$

$0.110 \cdot 2 = 6/4$

$0.111 \cdot 2 = 7/4$

$0.100 \cdot 2^{2} = 2$

$0.100 \cdot 2^{3} = \ldots$



man mano + radi via via che ci si allontana da zero

ε_M = precisione di macchina = è la distanza fra il n° 1 e il n° macchina successivo a 1.

$\varepsilon_M = \frac{1}{4}$



$\bullet \in \mathbb{F}$

$x \notin \mathbb{F}, x \in \mathbb{R} \quad x \in (1, 1+\varepsilon_M)$

$x \in \mathbb{R} \mapsto fl_t(x) \in \mathbb{F}$ è il n° macchina più vicino a x (float di x a t cifre)

? $|x - fl_t(x)| \leq \frac{\varepsilon_M}{2}$ $\quad u = \frac{\varepsilon_M}{2}$ = unità di arrotondamento

se x è generico $\in \mathbb{R}$, $x \in [-x_{max}; -x_{min}] \cup [x_{min}, x_{max}]$

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

se $x \neq 0$ $\quad \boxed{\frac{|x - fl_t(x)|}{|x|} \leq \frac{\varepsilon_M}{2}}$

errore relativo tra un n° $x \in \mathbb{R}$ e il suo float

"numero macchina" ≡ "n° floating point"

Se ho un sistema $\mathbb{F}$ a t cifre (della mantissa) $\Rightarrow$

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

$$\varepsilon_M = (1+\varepsilon_M) - 1 \Rightarrow \begin{array}{ll} 0.10\ldots\ldots01 \cdot \beta^1 & (1+\varepsilon_M) \\ - \; 0.10\ldots\ldots00 \cdot \beta^1 & (1) \\ \hline 0.0\ldots\ldots01 \cdot \beta^1 = \beta^{-t} \cdot \beta^1 = \beta^{1-t} & \end{array}$$

(t: posizione dell'ultima cifra)

$0.001 = 1 \cdot 10^{-3}$ (3: posizione della cifra 1)

maggiore è t e minore è ε_M

in $\mathbb{F}$ doppia precisione (float64) $\quad \varepsilon_M = 2^{-52} \sim 2.22 \cdot 10^{-16}$

unità di arrot. $\quad u = \frac{\varepsilon_M}{2} \sim 1.11 \cdot 10^{-16}$

In matlab `eps` contiene ε_M

Proprietà di $\mathbb{F}$: è simmetrico rispetto a 0,
è limitato e $[-x_{max}, +x_{max}]$
è finito (ha un n° finito di elementi)

In $\mathbb{F}$ non valgono + alcune proprietà fondamentali dell'aritmetica.

1) non vale + la proprietà associativa di + e ·

(es) $a = 10^{308}$, $b = 10^{308}$, $c = -4 \cdot 10^{307}$, $\mathbb{F}$ doppia prec.

$x_{max} \sim 1.8 \cdot 10^{308}$

$(a+b)+c = a+(b+c)$ prop. ass. in $\mathbb{R}$

- $a+b = 2 \cdot 10^{308} > x_{max}$ cioè $a+b \notin \mathbb{F}$

  Matlab dice NaN o Inf

  e $(a+b)+c$ non è calcolabile $\neq$

- $b+c = 6 \cdot 10^{307} \in \mathbb{F}$

  $a+(b+c) = 10^{308} + 6 \cdot 10^{307} = 1.6 \cdot 10^{308} \in \mathbb{F}$

(es) $\dfrac{1+(-1+x)}{x} = \dfrac{(1-1)+x}{x} = 1$ in carta

$x = 10^{-14}$ in $\mathbb{F}$ dp.

$= 0.9992$

$= 1$

Lo zero (elemento neutro della somma) non è più unico in $\mathbb{F}$

In $\mathbb{F}$ esistono + numeri che sommati ad $x \in \mathbb{F}$ restituiscono x.

$a + x = x$ anche se $a \neq 0$

(es) $\mathbb{F}(10, 3, -3, 3)$

$x = 0.123 \cdot 10^{0}$ $\quad y = 0.1 \cdot 10^{-3} = 0.0001 \cdot 10^{0}$

$$\begin{array}{lll} x & 0.123 & \cdot 10^{0} \\ y & 0.0001 & \cdot 10^{0} \\ \hline & 0.1231 & \cdot 10^{0} \end{array}$$

peso

$x + y = 0.123 \cdot 10^{0} = x$

anche se $y \neq 0$
$x + y = x$ in $\mathbb{F}$

(es)
$$\begin{array}{lll} x = & 0.2 & \cdot 10^{-3} \\ y = + & 0.1 & \cdot 10^{-3} \\ \hline & 0.3 & \cdot 10^{-3} \end{array}$$

$x + y \neq x$
$\neq y$

Propagazione degli errori di arrotondamento

(it) $x, y \in \mathbb{R}$ $\quad \bar{x} = fl_t(x)$ $\quad \bar{y} = fl_t(y)$

? $x + y$

al calcolatore trovo $fl_t(\bar{x} + \bar{y})$, non è garantito che $\bar{x} + \bar{y} \in \mathbb{F}$

Che errore si commette nell'eseguire la + al calcol.

$$e_+ = \frac{|(x+y) - fl_t(\bar{x} + \bar{y})|}{|x+y|} \leq \left(\frac{|x| + |y|}{|x+y|} + 1 \right) u$$

risultato dimostrabile; C_{xy}

(ip. $x + y \neq 0$)

C_{xy} può essere molto grande quando $|x| \sim |y|$ ma x e y sono discordi

(es) $x = 1 + 10^{-15}$ $\quad y = -1$, $|x| \sim |y|$, sono discordi

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

$$\Rightarrow \quad e_+ \leq (2\cdot 10^{15}+1)\cdot 10^{-16} \sim 10^{-1} \sim 10\,\%.$$

la somma è un'operazione instabile:
a piccoli errori sui dati (rappresentati da u)
possono corrispondere grossi errori nella soluzione

- prodotto

$$e_\cdot = \frac{|x\cdot y - fl_t(\bar{x}\cdot\bar{y})|}{|x\cdot y|} \leq 3\cdot u$$

$x\cdot y \neq 0$

l'operazione di prodotto è stabile