Osservazione:

$$(0.1)_{10} = (\,?\,)_2$$

$0.1 \times 2 = 0.2$ periodo

$0.2 \times 2 = 0.4$

$0.4 \times 2 = 0.8$

$0.8 \times 2 = 1.6$

$0.6 \times 2 = 1.2$

allora

$$(0.1)_{10} = (0.\overline{00011})_2 =$$

$$= \left(1.\overline{1001} \times 2^{-4}\right)_2$$

sviluppo finito in base 10 ma ∞ in base 2!

INSIEME DEI NUMERI DI MACCHINA

L'insieme dei numeri di Macchine in base β a $t+1$ cifre significative e range per l'esponente (M_1, M_2) è costituito dai seguenti elementi

$$\left\{ x \in \mathbb{R} : x = \pm \sum_{i=0}^{t} d_i \beta^{-i} \times \beta^{p} \right.$$

dove $d_i \in \mathbb{N}$, $0 \leq d_i \leq \beta - 1$ per $i = 0, 1, \dots, t$, $d_0 \neq 0$ e $p \in \mathbb{Z}$ con $M_1 + 1 \leq p \leq M_2 - 1$ $\left.\right\}$

Esso è denotato con

$$\mathbb{F}(\beta, t, M_1, M_2)$$

Osservazioni:

(1) $0 \notin \mathbb{F}$

(2) informalmente i numeri macchina hanno la seguente forma:

$$x = \pm \underbrace{d_0 . d_1 d_2 \dots d_t}_{\text{cifre}} \times \beta^{p}$$

β: base; p: esponente

ESEMPIO

$$3.14 \in \mathbb{F}(10,2,-1,2) =$$

$$= \{\pm d_0.d_1d_2 \times 10^p,\ p \in \{0,1\}\}$$

$$3.14 \notin \mathbb{F}(10,1,-1,2)$$

$$3.14 \notin \mathbb{F}(10,2,0,2)$$

---

più piccolo numero macchina strett. positivo:

Matlab → realmin $:= 1.00\ldots0 \times \beta^{M_1+1} = \beta^{M_1+1}$

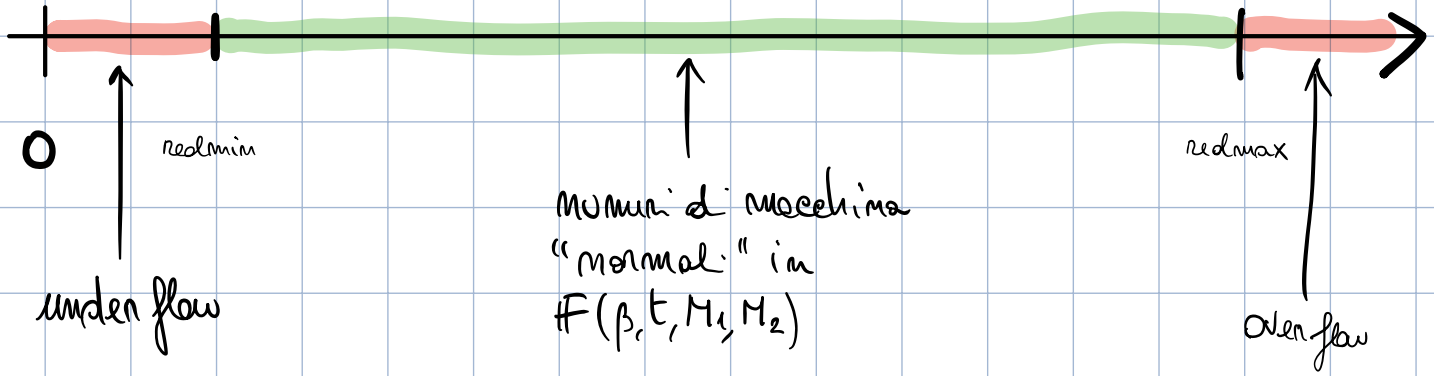
più grande numero di macchina:

realmax $:= \beta-1, \beta-1\ \beta-1 \ldots \beta-1 \times \beta^{M_2-1}$

aggiungo e sottraggo $0.00\ldots01 \times \beta^{M_2-1}$

$$
\begin{array}{l}
\left\{\begin{array}{llll}
\beta-1, & \beta-1 & \beta-1 \ \ldots & \beta-1 \ \beta-1 \times \beta^{M_2-1} + \\
0.0 & 0 \ \ldots & 0 \ 1 \times \beta^{M_2-1} - \\
0.0 & 0 \ \ldots & 0 \ 1 \times \beta^{M_2-1} =
\end{array}\right. \\
\hline
\begin{array}{llll}
1 & 0.0 & 0 \ \ldots & 0 \ 0 \times \beta^{M_2-1} - \\
 & 1.0 & 0 \ \ldots & 0 \ 0 \times \beta^{M_2-1-t} =
\end{array}
\end{array}
$$

$$\beta^{M_2} - \beta^{M_2-1-t} = \beta^{M_2-1}\left(\beta - \beta^{-t}\right)$$



Come vengono gestiti underflow e overflow?

Gestione "graduale" dell'underflow:

numeri denormali $= \left\{ \pm 0.d_1 d_2 \dots d_t \times \beta^{M_1+1}, \; d_i \neq 0 \right.$
per almeno un $\left. i = 1, 2, \dots, t \right\}$

Viene "rilassata" la condizione di normalizzazione $(d_0 \neq 0)$

più piccolo numero denormale strett. positivo:

$$\underbrace{0.0\dots01}_{t \text{ zeri}} \times \beta^{M_1+1} = \beta^{M_1+1-t}$$

ESEMPIO: $F(2, 2, -2, 3) =$

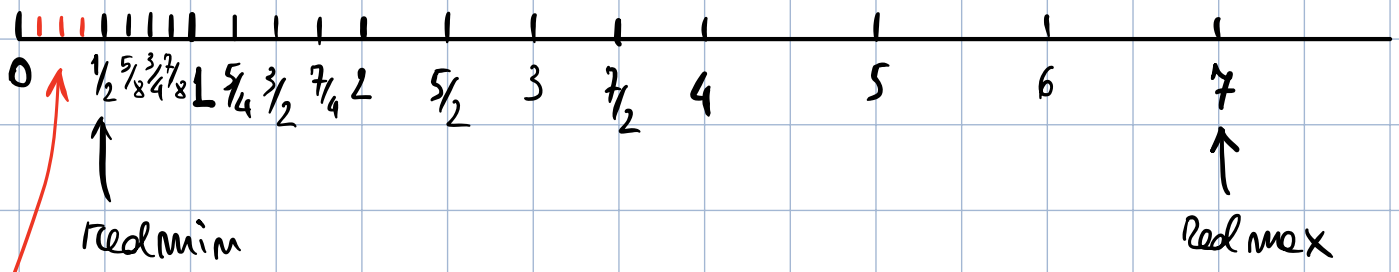
$$= \left\{ \pm 1.d_1 d_2 \times 2^p, \; p = -1, 0, 1, 2 \right\}$$

Sono 32 numeri (16 positivi). I positivi sono:

$p = -1$	1.00×2^{-1}	1.01×2^{-1}	1.10×2^{-1}	1.11×2^{-1}
$p = 0$	1.00×2^{0}	1.01×2^{0}	1.10×2^{0}	1.11×2^{0}

$p=1$	1.00×2^1	1.01×2^1	1.10×2^1	1.11×2^1
$p=2$	1.00×2^2	1.01×2^2	1.10×2^2	1.11×2^2

Rappresentiamoli sull'asse reale:



<u>OSSERV.</u>: maggiore densità per numeri reali

<u>OSSERV.</u>: denormali $= \left\{ \pm 0.d_1d_2 \times 2^{-1} \right\} =$

$= \left\{ \pm 1/8, \pm 1/4, \pm 3/8 \right\}$

Distanza tra numeri macchine consecutivi:

Sia $x = d_0.d_1d_2 \dots d_t \times \beta^p \in F(\beta, t, M_1, M_2)$.

Allora y, numero macchina successivo, è dato da

$$y = x + 0.0\dots01 \times \beta^p \Rightarrow$$

la distanza di x da y è $d = 0.0\dots01 \times \beta^p = \beta^{p-t}$

Conclusione: se $x < y \in F(\beta, t, M_1, M_2)$ sono numeri macchine consecutivi e $x, y \in [\beta^p, \beta^{p+1}]$, allora la loro distanza è

$$y - x = \beta^{p-t}$$

Esempio: distanza di $5/4$ dal numero macchine consecutivo in $F(2, 2, -2, 3)$:

$$5/4 \in [1, 2] = [2^0, 2^1] \Rightarrow$$

$$\text{distanza} = 2^{0-2} = 1/4$$

Standard IEEE 754 single/double precision

Single prec.: il numero è memorizzato sotto forma di stringa a 32 bit:

	S	q	f
# bit:	1	8	23
cosa:	segno	esponente	mantissa

Sono i numeri $(-1)^S \times 1.f \times 2^p$, dove

positivo se $s=0$
negativo se $s=1$

$$s = 0, 1, \quad p = q - \underbrace{\text{"bias"}}_{?}$$

$$q = (00000000)_2, (00000001)_2, \dots,$$

$$(11111110)_2, (11111111)_2 =$$

$$= 0, 1, \dots, 254, 255$$

sottraggo 127 per distribuirli meglio attorno a zero

$$p = q - 127 = -126, \dots, 127$$

↑ bias

Si tratta dell'insieme $\mathbb{F}(2, 23, -127, 128)$

Esempi:

$$\text{realmin} = \beta^{M_1+1} = 2^{-126} \approx 1.2 \times 10^{-38}$$

$$\text{realmax} = \beta^{M_2-1}(\beta - \beta^{-t}) = 2^{127}(2 - 2^{-23}) \approx$$

$$\approx 2^{128} \approx 1.7 \times 10^{38}$$

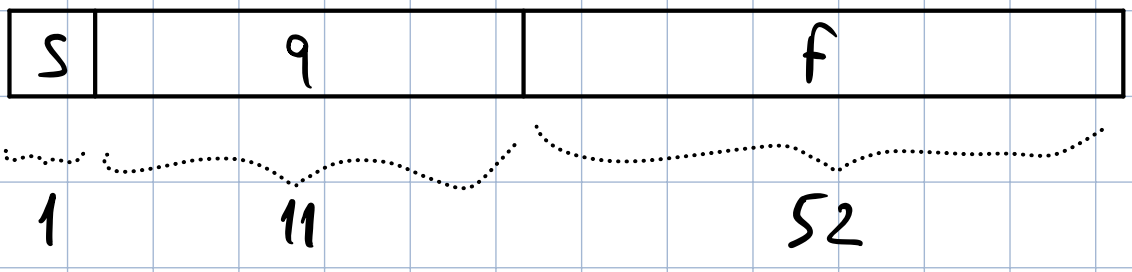
$$\text{più piccolo denormale positivo} = \beta^{M_1+1-t} = 2^{-149} \approx 1.4 \times 10^{-45}$$

Le stringhe q = 00000000, 1111 1111 non contribuiscono all'esponente perché rivestono un ruolo speciale:

q	mantissa	uso
00000000	0	zero macchina
00000000	$\neq 0$	denormali
1111 1111	0	$\pm$ Inf
1111 1111	$\neq 0$	NaN

ESEMPI (Matlab): $1/0 = \text{Inf}$, $\text{Inf} + \text{Inf} = \text{Inf}$, $\text{Inf} - \text{Inf} = \text{NaN}$

Double prec.: (default in Matlab)



$$q = 0, 1, \dots, 2046, 2047$$

bias: 1023

$$p = q - \text{"bias"} = -1022, \dots, 1023$$

si tratta di $\mathbb{F}(2, 52, -1023, 1024)$

Determinare realmin, realmax e "realmin denormale".