## Insieme dei numeri rappresentabili

Scelta una particolare struttura dati per gli interi, l'insieme dei numeri con essa rappresentabili risulta essere l'intervallo

$$[-(b^{(t-1)}-1),(b^{(t-1)}-1)]$$

avendo indicato con $b$ la base e con $t$ il numero di cifre disponibili.

Scelta una particolare struttura dati per la rappresentazione *floating point* l'insieme $\mathcal{F}$ dei numeri con essa rappresentabili risulta essere un sottoinsieme finito dell'insieme dei razionali. Infatti ciascun numero risulta essere il rapporto o il prodotto di due interi tali che il primo ha un numero di cifre limitato e il secondo è una potenza intera di $b$ con esponente costituito da un numero di cifre limitato.

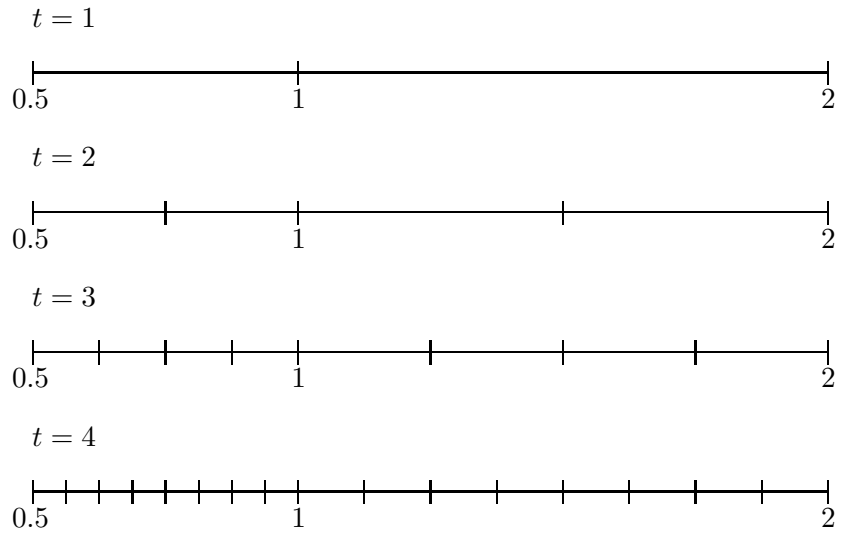
Una proprietà di $\mathcal{F}$ è che la distanza tra due numeri consecutivi che abbiano lo stesso esponente $f$ risulta

$$b^{-t}b^{f}$$

essendo $t$ il numero di cifre della mantissa.

Il più grande e il più piccolo numero in $\mathcal{F}$ dipendono, oltre che da $t$, dal numero di cifre $p$ riservate all'esponente.

Nel caso in cui la base $b$ sia 2 l'insieme $[0.5,2]\cap\mathcal{F}$ si può rappresentare, per diversi valori di $t$, nel modo seguente




avendo indicato ciascun numero con trattino. (*Provare a descrivere* $[0,0.5]\cap\mathcal{F}$ .)

Poichè $\mathcal{F}$ è contenuto nell'insieme $\mathbb{R}$ dei reali è utile definire una funzione

$$fl:\mathcal{R}\to\mathcal{F}$$

dove $\mathcal{R}$ è l'intervallo di $\mathbb{R}$ tra il più piccolo e il più grande numero in $\mathcal{F}$. Nella definizione usuale tale funzione (detta *arrotondamento*) trasforma un elemento di $\mathcal{R}$ nel più vicino elemento di $\mathcal{F}$ scegliendo, in caso di ambiguità, il più grande in valore assoluto.

Le operazioni di *addizione, moltiplicazione, sottrazione, divisione* sono definite su $\mathcal{F}$ come la composizione tra la corrispondente operazione in $\mathcal{R}$ e la funzione *arrotondamento*. Ad esempio la somma di $x, y \in \mathcal{F}$ è definita come

$$fl(x + y)$$

Da questo deriva che l'addizione non risulta associativa e la moltiplicazione non risulta distributiva sulla addizione.

Ad esempio, usando una rappresentazione in base 10 con una mantissa di 5 cifre le espressioni

$$(10^5 + 0.135) - 10^5$$
$$(10^5 - 10^5) + 0.135$$

in $\mathcal{F}$ hanno valori diversi. Infatti

$$fl(fl(10^5 + 0.135) - 10^5) = fl(10^5 - 10^5) = 0$$
$$fl(fl(10^5 - 10^5) + 0.135) = fl(0 + 0.135) = 0.135$$

Anche le espressioni

$$1100 \times (5.0001 - 5)$$
$$(1100 \times 5.0001) - 1100 \times 5$$

in $\mathcal{F}$ hanno valori diversi. Infatti

$$fl(1100 \times fl(5.0001 - 5)) = fl(1100 \times 0.0001) = 0.11$$
$$fl(fl(1100 \times 5.0001) - fl(1100 \times 5)) = fl(5500.1 - 5500) = 0.1$$

Più in generale il valore di una espressione numerica in $\mathcal{F}$ dipende dalla successione in cui sono eseguite le operazioni, ovvero dall'*algoritmo* usato.

Un esempio diverso consiste nel calcolo della soluzione del seguente sistema di equazioni

$$10^{-5}x_1 + x_2 = 1$$
$$x_1 + x_2 = 2$$

Sommando alla seconda equazione la prima moltiplicata per $-10^5$, con una mantissa di 5 cifre in base 10 si ottiene la soluzione

$$x_1 = 0$$
$$x_2 = 1$$

Sommando invece alla prima equazione la seconda moltiplicata per $-10^{-5}$, con la stessa mantissa si ottiene la soluzione

$$x_1 = 1$$
$$x_2 = 1$$

Notare che in questo caso sono stati usati due algoritmi diversi per calcolare il valore di una funzione

$$f : \mathcal{F}^2 \to \mathcal{F}^2$$