

2. Rappresentazione Floating Point (IEEE)

2.1. Rappresentazione nel calcolatore

La rappresentazione *floating point* prevede 4 componenti differenti: il segno, la mantissa, la base b e l'esponente della base p , ad esempio $x = \pm \underbrace{(0.d_1d_2d_3 \dots d_t \dots)}_{\text{mantissa}} \cdot b^p$

dove $d_j \in \{0, 1, \dots, b-1\}$, $d_1 \neq 0$ e $p \in \mathbb{Z}$.

Il vincolo $d_1 \neq 0$ e' imposto per impedire di avere infinite rappresentazioni di un certo numero, cosi' invece ne sono disponibili solo 2 (quella canonica e quella con cifre periodiche) e il calcolatore ne ha disponibile solamente una, quella canonica del numero, visto che non puo' rappresentare infinite cifre periodiche.

Nei calcolatori viene usato l'arrotondamento come metodo per limitare le cifre frazionarie; si ha infatti che un numero x e' definito nel calcolatore in *virgola mobile* come:

$$fl^t(x) = \text{sgn}(x) \cdot (0.d_1d_2 \dots \tilde{d}_t) \cdot b^p$$

dove la mantissa e' stata arrotondata alla t -esima cifra.

Ricordiamo che comunque anche p e' finito all'interno del calcolatore e, di conseguenza, non posso rappresentare tutto \mathbb{R} .

I numeri rappresentabili dai calcolatori si chiamano *numeri macchina* e sono definiti nel seguente modo:

$$\mathbb{F}(b, t, L, U) =$$

$$\{\mu \in \mathbb{Q}, \mu = \text{sgn}(\mu)(0.\mu_1\mu_2 \dots \mu_t)b^p : \mu \in \{0, 1, \dots, b-1\}, \mu_1 \neq 0, p \in [L, U] \subset \mathbb{Z}\}$$

tipicamente $L < 0$ e $U > 0$.

2.2. Stima dell'errore

L'errore si puo' descrivere in due modi: *assoluto* e *relativo*.

2.2.1. Errore Assoluto

L'errore assoluto e' quello che siamo stati abituati a calcolare fino ad ora:

$$|x - fl^t(x)| = b^p |(0.d_1 \dots d_t) - (0.d_1 \dots \tilde{d}_t)| \leq b^p \cdot \frac{b^{-t}}{2} = \frac{b^{p-t}}{2}.$$

Notiamo subito un aspetto: l'errore dipende da p , cioe' dall'ordine di grandezza del numero (in base b).

Notiamo subito che l'errore percio' dipende dall'ordine di grandezza del numero: numeri grandi in modulo avranno errori grandi, numeri piccoli in modulo avranno errori

piccoli; ma e' accettabile una situazione del genere? In generale la risposta e' SI, basta spostarsi dall'errore assoluto a quello relativo.

2.2.2. Errore Relativo

L'errore relativo non e' altro che l'errore assoluto su una quantita' *pesato dalla grandezza* della quantita'.

Data a una quantita' e \tilde{a} la sua approssimazione, si ha

$$\text{errore relativo} = \frac{a - \tilde{a}}{|a|}, a \neq 0.$$

Esso e' l'errore piu' importante in campo sperimentale e nelle applicazioni pratiche e di solito si esprime in percentuale. Scopriremo che l'errore relativo non dipende piu' da p .

Il *massimo errore relativo*, espresso in percentuale, di arrotondamento a n cifre in base b e' detto *precisione di macchina* e si indica con

$$\epsilon_M = \frac{|x - \text{fl}^t(x)|}{|x|} \leq \frac{b^{p-t}}{2} \cdot b^{1-p} = \frac{b^{1-t}}{2}.$$

Per farlo abbiamo usato il fatto che $|x| \geq b^{-1} \cdot b^p = b^{p-1}$.

Come vediamo esso dipende solamente da b e da t , rispettivamente la base e il numero di cifre di mantissa.

Per esempio, secondo lo standard *IEEE* per la rappresentazione dei numeri floating point a 64 bit, 53 bit sono dedicati alla mantissa quindi $\epsilon_M = 2^{-53}$.