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 $x = \pm\underbrace{(0.d_1 d_2 d_3 \dots d_t \dots)}_{\text{mantissa}} \cdot b^p$

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 mobiile* come:

$$fl^t(x) = \operatorname{sgn}(x) \cdot (0.d_1 d_2 \dots \tilde{d}_t) \cdot b^p.$$

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:

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

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

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:

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

2.2.2. Errore Relativo

Possiamo pero' anche stimare un errore relativo, tra il numero e l'approssimazione fatta; il *massimo errore relativo*, espresso in percentuale, di arrotondamento a n cifre in base b e' detto *precisione di macchina* e si indica con

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

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

Notiamo subito che l'errore perciò dipende dall'ordine di grandezza del numero:
numeri grandi in modulo avranno errori grandi, numeri piccoli in modulo avranno errori piccoli; la precisione di macchina e' dunque il minimo errore relativo di arrotondamento a t cifre.
Questo pero', abbinato alla distribuzione della rappresentazione dei reali nel calcolatore, ci permette di dire che la precisione di macchina dipende solo da b e da t e non dall'ordine di grandezza del numero.