1 Rappresentazione posizionale

Per rappresentare un qualsiasi numero $\boldsymbol{N} \geq 1$ in notazione posizionale usando una base β (dove β = numero di simboli $S-1$) si può stabilire la relazione $\boldsymbol{N}= (c_p c_{p-1}, \cdots c_0)_\beta = (c_p \beta^p c_{p-1} \beta^{p-1} ... c_0)$.
Per un numero $\alpha < 1$ si ha $\alpha = (0.c_p c_{p-1}, \cdots c_0)_\beta$ rapresentabile anche come la sommatoria $\sum_{i=1}^{\infty} a_i \beta^{-i}$ quindi si ha che un qualsiasi numero $\alpha \in \Re$ è rapresentabile come

$$\alpha = \sum_{i=1}^{\infty} a_i \beta^{-i} \beta^p$$

dove p è il numero di simboli dopo il punto radice $\neq 0$.
Riassumendo: un qualsiasi numero $\alpha \in \Re$ è rappresentabile in notazione **Posizionale** come $segno(\alpha) m \beta^p$ dove $m = \sum_{i=1}^{\infty} a \beta^{-i}$

2 Rappresentazione floating point

Per memorizzare un qualsiasi numero reale ma in uno spazio finito di memorie viene uitilizzata la rappresentazione a **virgola mobile** che a differenza di altre rapresentazioni (es. virgola fissa) permette grande flessibilità. il formato di rappresentazione è cartterizzato da 4 elementi:

$$\beta, t, L, U$$

.

- β è la base usata, di norma base 2.
- t è il numero massimo di numeri dedicato alla rapresentazione della mantissa, di norma **single precision** = 23 **double precision** = 52.
- L, U sono rispettivamente il **L**ower e **U**pper bound per la rappresentazione di un esponente r quindi il numero minimo e massimo che si può rappresentare, di norma $[-127, +128]$.

altre caratteristiche del insieme dei numeri floating point

- è un insieme **discreto** e **finito**.
- è **simmetrico** al origine, quindi per ogni numero esiste l'opposto.
- la sua cardinalità è data da $card(\beta, t, L, U) = 2(\beta - 1)\beta^{t-1}(U - L + 1)$ dove
  - $(U - L + 1)$ è il numero di tutti i possibili esponenti
  - β^{t-1} è il numero di tutti i numero di valori che ogni posizione dedicata alla mantissa può prendere escluso un elemento.
  - $\beta - 1$ è il numero che il primo elemento può prendere (-1 perchè non può essere zero).
  - 2 perché un valore può essere sia positivo che negativo.

3 Errori di macchina e operazioni

Dato che l'insieme dei **Floating point** è un insieme discreto e finito spesso si fa uso di **troncamento** e **arrotondamento** per rapprensentare i numeri non appartenenti all'insieme di riferimento ottenendo una perdita di informazione chiamata ***errore***.
L'**errore assoluto** nella rappresentazione di un numero α è la differenza tra il numero è la sua rappresentazione, quindi si otttiene che: $E_a = \mid \alpha - \alpha^* \mid$ dove α^* è la rapresentazione macchiana.
L'errore relativo invece è dato da $E_r = \frac{E_a}{|\alpha|}$. Inoltre si ha che l'errore relativo ottenuto è sempre minore o uguale alla **precisione di macchina** rappresentata come $k\beta^{t-1}$ dove

- $k = 1$ nel caso si utilizzi logiche di approsimazione tramite toroncamento.
- $k = \frac{1}{2}$ nel caso si usi arrotondamento.

questo risultato è chiamato **teorema dell'errore di rappresentazione dei numeri reali**.
Dato che l'insieme dei floating point F non è chiuso alle operazioni aritmetiche si possono verificare, oltre ai casi di **overflow** e **underflow**, dei casi in cui il numero non è rappresentabile pur essendo compreso tra il numero massimo e il numero minimo rapresentabile del insieme, difatti si ha che un operazione tra numeri floating point restituisce un numero reale che andrà trasformato seguendo la logica vista nel capitolo Rappresentazione floating point