1 Aritmetica computazionale

1.0.1 Rappresentazione dei numeri reali

I **numeri finiti** sono utilizzati dai calcolatori per rappresentare i numeri reali poiché questi ultimi possono avere un numero infinito di cifre, che i calcolatori, avendo una memoria limitata, non sono in grado di rappresentare.

Teorema (Rappresentazione in base). *Sia α un numero reale non nullo. Possiamo rappresentare tale numero con una base $\beta \geq 2$, un numero intero scelto da noi, nel seguente modo:*

$$\begin{aligned}\alpha &= \pm(\alpha_1\beta^{-1} + \alpha_2\beta^{-2} + \ldots)\beta^p \\ \alpha &= \pm(\sum_{i=1}^{\infty} \alpha_i\beta^{-i})\beta^p\end{aligned} \tag{1}$$

I vari termini dell'uguaglianza vengono detti:

β	*base*
p	*esponente*
α_i	*cifre del numero*
$\sum_{i=1}^{\infty} \alpha_i\beta^{-i}$	*mantissa*

Ogni cifra α_i è un numero intero che varia tra 0 e $\beta - 1$. Ad esempio, se lavoriamo in base 10, le cifre saranno numeri interi compresi tra 0 e 9.
Per garantire l'unicità della rappresentazione, è necessario che $\alpha_1 \neq 0$. Se così non fosse, il numero 13 potrebbe essere rappresentato come 13, 013, 0013, eccetera, il che va contro l'unicità della rappresentazione.

Possiamo scrivere un numero $\alpha \in \mathbb{R}$ con $\alpha \neq 0$ in due modi:

1. **forma mista**.
$$\alpha = \begin{cases} \pm(0.000\alpha_1\alpha_2\ldots)_\beta & p \leq 0 \\ \pm(\alpha_1\alpha_2\ldots)_\beta & p > 0 \end{cases}$$

2. **forma scientifica**. L'idea è quella di spostare il punto decimale al primo numero $\neq 0$ e poi moltiplicare il tutto per β^p per riportare il numero al suo valore originale.
$$\alpha = \pm 0.\alpha_1\alpha_2\ldots \cdot \beta^p$$

Esempio:

$$\begin{aligned}\alpha &= (12.37)_{10} \quad & \alpha &= 0.12237 \cdot 10^2 \\ \alpha &= (0.0045)_{10} \quad & \alpha &= 0.45 \cdot 10^{-2} \\ & & &= (4 \cdot 10^{-1} + 5 \cdot 10^{-2}) \cdot 10^{-2}\end{aligned}$$

Definizione (Numeri finiti). *L'insieme $\mathbb{F}$ dei numeri finiti è definito come l'insieme dei numeri espressi in base β (dove $\beta \geq 2$), utilizzando t cifre (con $t \geq 1$). Poiché anche l'esponente p potrebbe essere così grande da non poter essere rappresentato, è necessario limitare l'intervallo degli esponenti rappresentabili. Qui, λ indica il più piccolo esponente che può essere rappresentato e ω il più grande esponente rappresentabile.*

$$\begin{aligned}\mathbb{F}(\beta, t, \lambda, \omega) &= \{0\} \cup \{\alpha \in \mathbb{R} : \alpha = \pm 0.\alpha_1\alpha_2 \dots \alpha_t \cdot \beta^p\} \\ &= \{0\} \cup \{\alpha \in \mathbb{R} : \alpha = \pm(\sum_{i=1}^{t} \alpha_i \beta^{-i})\beta^p\}\end{aligned}$$

Esempio: