

# Errore di rappresentazione (virgola mobile)

Indichiamo con  $\tilde{x}$  il numero di macchina che approssima il valore reale  $x \neq 0$ . L'errore commesso nella rappresentazione è:

$$\begin{aligned}\eta_x &= \tilde{x} - x && \text{(assoluto)} \\ \epsilon_x &= \frac{\tilde{x} - x}{x} && \text{(relativo)}\end{aligned}$$

Possiamo scrivere la relazione tra  $x$  e  $\bar{x}$  come:

$$\bar{x} = x(1 + \epsilon_x).$$

## Lemma

Tolti casi di overflow e underflow,

- $|\text{trunc}(x) - x| < \beta^{p-t}$ ;
- $|\text{arr}(x) - x| \leq \frac{1}{2}\beta^{p-t}$ ;

## Dimostrazione

$x = \Omega$ :  $\text{trunc}(x) = \text{arr}(x) = \Omega \implies \text{trunc}(x) - x = \text{arr}(x) - x = 0$ ;

$0 \leq x < \Omega$ : sia  $x = \beta^p \sum_{i=1}^{+\infty} d_i \beta^{-i}$ .

Indichiamo con  $a$  e  $b$  rispettivamente i numeri di macchina più vicini a  $x$  a sinistra e destra:

$$a = \text{trunc}(x) \quad b = \text{trunc}(x) + \beta^{p-t}.$$

Visto che  $b - a = \beta^{p-t}$  e  $a \leq x < b$ ,

$$|\text{trunc}(x) - x| = |a - x| < |a - b| = \beta^{p-t}$$

e

$$\text{arr}(x) = \begin{cases} a & x < \frac{a+b}{2} \\ b & x \geq \frac{a+b}{2} \end{cases}$$

quindi

$$|\text{arr}(x) - x| \leq \frac{b - a}{2} = \frac{1}{2}\beta^{p-t}.$$

## Teorema

Se  $|x| \in [\omega, \Omega]$  (e  $x \neq 0$ ),

$$|\epsilon_x| < u$$

dove  $u$  è la *precisione di macchina*, che vale  $\beta^{1-t}$  con troncamento e  $\frac{1}{2}\beta^{1-t}$  con arrotondamento. L'errore relativo quindi è limitato da una costante del sistema.

## Dimostrazione

**troncamento** si osserva che:

- per il lemma,  $|\bar{x} - x| < \beta^{p-t}$ , e
- $x = \beta^p(0.d_1d_2 \dots d_{53})_\beta \geq \beta^p 0.1_\beta = \beta^{p-1}$ ,

quindi:

$$|\epsilon_x| = \left| \frac{\bar{x} - x}{x} \right| < \frac{\beta^{p-t}}{\beta^{p-1}} = \beta^{1-t}.$$

**arrotondamento** analogamente troviamo che

$$|\epsilon_x| < \frac{\frac{1}{2}\beta^{p-t}}{\beta^{p-1}} = \frac{1}{2}\beta^{1-t}.$$

Abbiamo un minore stretto nonostante nel lemma ci sia un  $\leq$  perché numeratore e denominatore non possono essere contemporaneamente uguali alle loro maggiorazioni.