1 Precisione di macchina come max errore relativo di arrotondamento nel sistema floating-point

Definiamo arrotondamento a t cifre di un numero reale scritto in notazione floating-point

$$x = sign(x)(0, d_1 d_2 \dots d_t \dots) \cdot b^p$$

il numero

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

dove la mantissa è stata arrotondata alla t-esima cifra

$$\tilde{d}_t = \begin{cases} d_t & \text{se } d_{(t+1)} < \frac{b}{2} \\ d_t + 1 & \text{se } d_{(t+1)} \geq \frac{b}{2} \end{cases}$$

Definiamo:

$$\text{Errore Relativo} \longleftarrow \frac{\overbrace{|\, x - fl^t(x) \,|}^{\text{Errore Assoluto}}}{|\, x \,|} \quad \text{per} \quad x \neq 0$$

Definiamo:

$$\text{Errore Relativo} \longleftarrow \frac{\overbrace{|\, x - fl^t(x) \,|}^{\text{Errore Assoluto}}}{|\, x \,|} \quad \text{per} \quad x \neq 0$$

Stimiamo il numeratore

$$\begin{aligned} |\, x - fl^t(x) \,| &= b^p \cdot \overbrace{|(0, d_1 d_2 \dots d_t \dots) - (0, d_1 d_2 \dots \tilde{d}_t)|}^{\text{Errore di arrotondamento a } t \text{ cifre dopo la virgola} \leq \frac{b^{-t}}{2}} \\ &\leq b^p \cdot \frac{b^{-t}}{2} = \frac{b^{p-t}}{2} \end{aligned}$$

Notiamo subito un aspetto: l'errore dipende da p, cioè dall'ordine di grandezza del numero (in base b).
Stimiamo da sopra $\frac{1}{|x|}$, ovvero da sotto $|\, x \,|$:

$$|\, x \,| = (0, d_1 d_2 \dots d_t \dots) \cdot b^p$$

Poiché $d_1 \neq 0$, p fissato, il minimo valore della mantissa è $0,100\dots = b^{-1}$. Quindi:

$$|\, x \,| \geq b^{-1} \cdot b^p = b^{p-1} \iff \frac{1}{|\, x \,|} \leq \frac{1}{b^{p-1}}$$

Otteniamo

$$\frac{|\, x - fl^t(x) \,|}{|\, x \,|} \leq \frac{\frac{b^{p-t}}{2}}{b^{p-1}} = \frac{b^{p-t+1-p}}{2} = \frac{b^{1-t}}{2} = \varepsilon_M$$