Algebra I

Paolo Bettelini

Contents

1 Floating points

L'insieme dei floating point è

$$f(\beta,t,m,M)=\{0,\text{NaN},\pm\infty\}\cup\left\{x=\text{sign}(x)\cdot\beta^e\sum_{i=1}^{t}y_i\beta^{-i}\;\middle|\;t,y_i,m,M\in\mathbb{N},y_1\neq 0,-m\leq e\leq M\right\}$$

Stimiamo ora l'errore relativo

$$\frac{|x-\tilde{x}|}{|x|}$$

dove $x\in\mathbb{R}$ e $\tilde{x}\in f(\beta,t,m,M)$ è la sua rappresentazione migliore in un calcolatore. Consideriamo $x>0$. Chiaramente, se $\tilde{x}\in\mathbb{R}$, allora $|x-\tilde{x}|=0$. Altrimenti, $x\in[a,b]$ dove $a,b\in f$ e sono consecutivi in f. Quindi

$$|x-\tilde{x}|\leq\frac{b-a}{2}$$

Abbiamo allora

$$a=\beta^e\sum_{i=1}^{t}y_i\beta^{-i}$$

e

$$b=\beta^e\left(\sum_{i=1}^{t}y_i\beta^{-i}+\beta^{-t}\right)=a+\beta^{e-t}$$

Quindi la differenza è data da

$$|x-\tilde{x}|\leq\frac{1}{2}\beta^{e-t}$$

Dobbiamo ora minorare l'elemento normalizzante

$$|x|=\beta^e\sum_{i=1}^{\infty}y_i\beta^{-i}\geq\beta^e\cdot y_1\beta^{-1}\geq\beta^{e-1}$$

Abbiamo quindi

$$\frac{1}{|x|}\leq\beta^{1-e}$$

Combinando i due risultati otteniamo

$$\frac{|x-\tilde{x}|}{|x|}\leq\frac{1}{2}\beta^{e-t}\beta^{1-e}=\frac{1}{2}\beta^{1-t}\triangleq u$$

Allora u è la precisione macchina.