RAPPRESENTAZIONE DI UN NUMERO

Dato $x \in \mathbb{R}$, $x \neq 0$, esistono e sono univocamente determinati

① un intero $p \in \mathbb{Z}$ detto esponente della rappresentazione;

② una successione di numeri naturali $\{d_i\}_{i \geq 1}$, $d_1 \neq 0$, $0 \leq d_i \leq B-1$, non definitivamente $= B-1$ detta cifre della rappresentazione;

t.c.

$$x = \text{sign}(x) B^p \sum_{i=1}^{+\infty} d_i B^{-i}$$

} rappresentazione normalizzata in virgola mobile

$\downarrow$

perché l'esponente è determinato in modo che 1° CIFRA DOPO VIRGOLA $\neq 0$ e parte INTERA $= 0$

$d_1 \neq 0$
d_i non def. $= B-1$
$\Rightarrow$ UNICITÀ DELLA RAPPRESENTAZIONE

RAPPRESENTAZIONE NUMERI IN MACCHINA

$t, m, M \in \mathbb{N}$ con $t \geq 1$, $m, M > 0$ e

$$F(B,t,m,M) = \left\{ x \in \mathbb{R} \mid x = \text{sign}(x) \cdot B^p \cdot \sum_{i=1}^{t} d_i B^{-i} \right\} \cup \{0\}$$

con $0 \leq d_i \leq B-1$, $d_1 \neq 0$ $\quad -m \leq p \leq M$

$$\omega = B^{-(m+1)} \qquad \Omega = B^M (B-1) \cdot \sum_{i=1}^{t} B^{-i}$$

Errore assoluto $\tilde{x} - x$ è preferibile perché garantisce l'ordine di grandezza dell'ultima cifra rispetto all'ordine di grandezza di x

Errore di rappresentazione relativo $\varepsilon_x = \dfrac{\tilde{x} - x}{x} \rightarrow \tilde{x} = x(1 + \varepsilon_x)$ con $|\varepsilon_x| \leq u$

$\downarrow$

rappresentare $x \in \mathbb{R}$, $x \neq 0$ in macchina significa approssimare x con un numero $\tilde{x} \in F(B,t,m,M)$ commettendo un errore di rappresentazione

TEOREMA (maggiorazione dell'errore di rappresentazione)

$x \in \mathbb{R}$ con $\omega \le |x| \le \Omega$ Si ha

$$|\varepsilon_x| = \left| \frac{fl(x) - x}{x} \right| \le u = B^{1-t}$$

DIM $x = (-1)^s B^p \cdot$. $|fl(x) - x| \ge B^{p-t}$

↓ distanza tra 2 n. di macchina consecutivi

$$\left| \frac{fl(x) - x}{x} \right| \le B^{p-t}$$

Inoltre $|x| \ge B^{p-1}$ e quindi

$$|\varepsilon_x| = \left| \frac{fl(x) - x}{x} \right| \le \frac{B^{p-t}}{B^{p-1}} = u$$

oppure

$$|\varepsilon_x| = \left| \frac{\tilde{x} - x}{x} \right| = \frac{x - \tilde{x}}{x} \le$$

$$\le \frac{B^p \sum_{i=1}^{\infty} d_i B^{-i} - B^p \sum_{i=1}^{t} d_i B^{-i}}{B^p \cdot B^{-1}} = \frac{\sum_{i=t+1}^{+\infty} d_i B^{-i}}{B^{-1}} <$$

$0 \le d_i \le B-1$

$$< \frac{(B-1) B^{-(t+1)} \sum_{i=0}^{\infty} B^{-i}}{B^{-1}} = (B-1) B^{-(t+1)} \frac{B}{B-1} \cdot B =$$

$$= B^{1-t}$$

se $\omega \leq |x| \leq \Omega$

Allora

- ARROTONDAMENTO: x viene approssimato ~~rappresentato~~ con il numero rappresentabile $\tilde{x}$ più vicino
- TRONCAMENTO: x viene approssimato col maggior numero $\tilde{x}$ il cui valore assoluto risulti $\leq$ al valore assoluto di x
- con il maggior numero rappresentabile minore del dato
- con il minor numero rappresentabile maggiore del dato

$$\varepsilon_{in} = \frac{f'(x)}{f(x)}(\tilde{x}-x)\frac{x}{x} + \frac{f''(c)}{2f(x)}\frac{(\tilde{x}-x)^2 x^2}{x^2}$$

$$\varepsilon_{in} \doteq \frac{f'(x)}{f(x)} \cdot \varepsilon_x = c_x \varepsilon_x \qquad \varepsilon_{in} = \frac{f'(x)}{f(x)} \varepsilon_x \cdot x = c_x \varepsilon_x$$ (lo trascuro)

dove $c_x = \dfrac{f'(x)}{f(x)}\, x$ → COEFFICIENTE DI AMPLIFICAZIONE DELL'ERRORE fornisce una misura del condizionamento del problema

$$\boxed{f : \mathbb{R} \to \mathbb{R},\ f \in C^2(\mathbb{R})}$$

sviluppo di Taylor in un intorno di $\hat{x}$

$$f(\tilde{x}) = f(x) + f'(x)(\tilde{x}-x) + \frac{f''(\varepsilon)}{2}(\tilde{x}-x)^2 \qquad |\varepsilon - x| \leq |\tilde{x}-x|$$

$$\frac{f(\tilde{x})-f(x)}{f(x)} = \varepsilon_{in} = \frac{f'(x)}{f(x)}(\tilde{x}-x) + \frac{f''(\varepsilon)}{2f(x)}(\tilde{x}-x)^2$$

CARDINALITÀ FINITA $N = 2B^{t-1}(B-1)(M+m+1)+1$

Se $x \in F(B,t,m,M)$ $x \neq 0$

$$w = B^{-m-1} \leq |x| \leq B^M(1-B^{-t}) = \Omega$$

⇓

Non è possibile rappresentare numeri di modulo minore w

⇓

Rappresentazione denormalizzata:

- quando $p = -m$ NO $d_1 \neq 0$ e

$$B^{m-t} \leq |x| \leq B^{-m}(B^{-1}-B^{-t})$$

- quando $p = M$ $x = \pm\infty$, NaN

L'insieme dei numeri di macchina $F(B,t,m,M)$ è simmetrico rispetto all'origine quindi

se $x = (-1)^s B^p d$ il successivo sarà

$$y = (-1)^s B^p (d + B^{-t})$$

Quindi $|y - x| = B^{p-t}$

distanza fra due numeri in macchina

e quindi in generale

$$\varepsilon_{in} = \frac{1}{f(x)} \sum_{i=1}^{n} \frac{df(x)}{dx_i} x_i \varepsilon_{x_i}$$

$x = [x_1 \dots x_n]$

$\forall a, b \in F(B,t,m,M) \quad a \oplus b = fl(a+b)$

vale che

$$\left| \frac{fl(a+b) - (a+b)}{a+b} \right| = \frac{a}{a+b}\varepsilon_a + \frac{b}{a+b}\varepsilon_b + \varepsilon_1$$

e si ottiene perché

$$a \oplus b = \big(a(1+\varepsilon_a) + b(1+\varepsilon_b)\big)(1+\varepsilon_1)$$

($1+\varepsilon_1$): errore locale dell'operazione

$$\doteq a + a\varepsilon_a + b + b\varepsilon_b + a\varepsilon_1 + b\varepsilon_1$$

ho trascurato $a\varepsilon_a\varepsilon_1$ e $b\varepsilon_b\varepsilon_1$

$$= \frac{a}{a+b}\varepsilon_a + \frac{b}{a+b}\varepsilon_b + \varepsilon_1$$

esprime la dipendenza dell'errore totale complessivo nel calcolo della somma tra due numeri reali rispetto agli errori generati dall'approssimazione dei dati iniziali (errore inerente) e agli errori generati dall'algoritmo di calcolo (errore algoritmico)

$\doteq$ considero solo le componenti lineari dell'errore e trascuro le componenti di ordine superiore al primo

master copy