TEORIA degli ERRORI

RAPPRESENTAZIONE dei NUMERI

Scelto un numero qualunque $\beta \in \mathbb{N} : \beta > 1$, ogni numero $x \in \mathbb{R} \setminus \{0\}$ ammette una rappresentazione in base β

$$x = \text{sign}(x)\, \beta^{b} \sum_{i=1}^{\infty} \alpha_i \beta^{-i}$$

$\beta \in \mathbb{N} : \beta > 1$ BASE

$b \in \mathbb{Z}$ ESPONENTE

$\alpha_i \in \mathbb{N} : 0 \leq \alpha_i \leq \beta - 1 \quad i = 1, 2, 3, \ldots$ CIFRE

$$\text{sign}(x) = \begin{cases} -1 & \text{se } x < 0 \\ 1 & \text{se } x > 0 \end{cases}$$

MASTER COPY

Tel. 388/9837745

(TH) di UNICITÀ della RAPPRESENTAZIONE

Data una BASE $\beta \in \mathbb{N} : \beta > 1$ e un qualunque numero $x \in \mathbb{R} \setminus \{0\}$ esiste un'unica rappresentazione in base β tale che:

(1) $\alpha_1 \neq 0$

la 1ª cifra dopo la virgola non può essere 0.

(2) $\nexists k \in \mathbb{N} : \alpha_j = \beta - 1, \forall j > k$

l'ultima cifra non può essere $(\beta - 1)$; ciò serve a fare in modo che numeri come $1,1\overline{9}$ e $1,2$ non abbiano la stessa rappresentazione.

La rappresentazione definita dal Teorema si dice RAPPRESENTAZIONE IN VIRGOLA MOBILE NORMALIZZATA del NUMERO REALE x.

INSIEME di VARIAZIONE della MANTISSA

Fissato β, sono univocamente determinati i numeri b e α_i. La serie $\sum_{i=1}^{\infty} \alpha_i \beta^{-i}$ è detta MANTISSA del NUMERO REALE x e il suo insieme di variazione è:

$$\frac{1}{\beta} \leq \sum_{i=1}^{\infty} \alpha_i \beta^{-i} < 1$$

MASTER COPY

Tel. 388/9837745

Dimostrazione

Se tutti gli $\alpha_i = 0$ e $\alpha_1 \neq 0$ poniamo $\alpha_1 = 1$; quindi il valore minimo è $\alpha_1 \beta^{-1} = \frac{1}{\beta}$ → VALORE MINIMO

Altrimenti:

$$\sum_{i=1}^{\infty} \alpha_i \beta^{-i} < \sum_{i=1}^{\infty} (\beta-1)\beta^{-i} = (\beta-1)\sum_{i=1}^{\infty} \frac{1}{\beta^i} = (\beta-1)\frac{1}{1-\frac{1}{\beta}}\beta^{-1} = 1$$

- $(\beta-1)$: valore più grande che una cifra può assumere
- $\sum_{i=1}^{\infty} \frac{1}{\beta^i}$: progressione geometrica di ragione $\left(\frac{1}{\beta}\right)$
- $= 1$: VALORE MASSIMO

RAPPRESENTAZIONE nel CALCOLATORE

Nel calcolatore si possono rappresentare solo un certo numero di cifre m della mantissa di x; la rappresentazione nel calcolatore corrisponde perciò ad uno fra i 2 numeri seguenti:

TRONCAMENTO

$$y = tr(x) = sign(x)\,\beta^b \sum_{i=1}^{m} \alpha_i \beta^{-i}$$

ARROTONDAMENTO

$$y = rd(x) = \begin{cases} tr(x) & \text{se } 0 \leq \alpha_{m+1} < \frac{\beta}{2} \\ sign(x)\,\beta^b \left(\sum_{i=1}^{m} \alpha_i \beta^{-i} + \beta^m\right) & \text{se } \frac{\beta}{2} \leq \alpha_{m+1} < \beta. \end{cases}$$

Nel caso particolare che $\alpha_1 = \alpha_2 = \dots = \alpha_m = \beta - 1$ e $\alpha_{m+1} > \frac{\beta}{2}$ con l'arrotondamento si deve aumentare di una unità l'esponente b.

$\beta = 10 \quad m = 3 \qquad x = 0.997 \times 10^5$

$rd(x) = 0.100 \times 10^6$

In generale l'arrotondamento è un'approssimazione migliore del numero x

$$|tr(x) - x| < \beta^{(b-m)} \qquad |rd(x) - x| < \frac{1}{2}\beta^{(b-m)}$$

RAPPRESENTAZIONE dello ZERO

Lo 0 è l'unico numero non rappresentabile, ma esiste la funzione $sign_0(x)$ che lo aggiunge, tale che:

$$\text{sign}_0(x) = \begin{cases} \text{sign}(x) & \text{se } x \neq 0 \\ 0 & \text{se } x = 0 \end{cases}$$

INSIEME dei NUMERI di MACCHINA

M = insieme finito dei numeri di Macchina

Fissati β, m e supposto $L \leq b \leq U$ con:

$L \in \mathbb{Z}$ = minimo esponente

$U \in \mathbb{Z}$ = massimo esponente

si ha che la cardinalità di M è

$$\text{card}(M) = 2(\beta^m - \beta^{m-1})(U - L + 1) + 1$$

- 2 → SEGN.
- $(\beta^m - \beta^{m-1})$ → LA 1ª CIFRA DELLA MANTISSA NON PUÒ ESSERE 0. $\Rightarrow (\beta - 1)\beta^{(m-1)}$
- $(U - L + 1)$ → POSSIBILI ESPONENTI
- $+1$ → RAPPRESENTAZIONE DELLO ZERO

La rappresentazione di M è data da

$F(\beta, m, L, U)$ che rappresenta le caratteristiche di macchina

- β → BASE
- m → PRECISIONE
- L → MIN. ESP.
- U → MAX ESP.

OVERFLOW → (sempre segnalato) Il numero da rappresentare è "troppo grande" e non appartiene a M.

Alcuni calcolatori arrestano l'esecuzione del programma, mentre altri proseguono con $\text{rd}(x) = \text{sign}(x) \max_{y \in M} |y|$

UNDERFLOW → (non sempre segnalato) Il numero da rappresentare è "troppo piccolo" e non appartiene a M.

Alcuni calcolatori arrestano l'esecuzione del programma, mentre altri proseguono con $\text{rd}(x) = 0$

MASTER COPY
Tel. 388/9837745

ERRORE della RAPPRESENTAZIONE del NUMERO REALE x.

$$\delta_x = \begin{cases} rd(x) - x & \text{ERRORE ASSOLUTO per ARROTONDAMENTO} \\ tr(x) - x & \text{ERRORE ASSOLUTO per TRONCAMENTO} \end{cases}$$

$$\varepsilon_x = \frac{\delta_x}{x} = \begin{cases} \frac{rd(x)-x}{x} & \text{ERRORE RELATIVO per ARROTONDAMENTO} \\ \frac{tr(x)-x}{x} & \text{ERRORE RELATIVO per TRONCAMENTO} \end{cases}$$

Dalle precedenti considerazioni si hanno le limitazioni:

- $|\delta_{x_{rd}}| \leq \frac{1}{2}\beta^{b-m}$
- $|\delta_{x_{tr}}| < \beta^{b-m}$ → SI USA IL MINORE STRETTO PERCHÉ IL PIÙ PICCOLO NUMERO SI RAPPRESENTA SENZA ERRORI
- $|\varepsilon_{x_{rd}}| \leq \frac{1}{2}\beta^{1-m}$
- $|\varepsilon_{x_{tr}}| < \beta^{1-m}$ ↗

Indichiamo con $U = \frac{1}{2}\beta^{1-m}$ la PRECISIONE di MACCHINA

In generale, quando l'errore relativo ε_x di un'approssimazione non supera il valore $\frac{1}{2}\beta^{1-k}$, si dice che l'approssimazione è corretta almeno fino alla k-esima cifra significativa. Quindi $rd(x)$ approssima x almeno fino la m-esima cifra significativa.

MASTER COPY
Tel. 388/9837745

SISTEMA BINARIO → $\beta = 2$

Ogni cifra della rappresentazione ha valore 0 o 1 e si dice BIT. L'insieme dei t bit che rappresentano un numero, si chiama parola di lunghezza t

Per ragioni tecniche si usa spesso una base $\beta = 2^n$ $(n > 1)$ in cui ciascuna cifra tradotta in binario richiede n bit.

OPERAZIONI di MACCHINA

Nell'insieme N non sono definite tutte le proprietà delle 4 operazioni elementari, perché il risultato deve appartenere a N; quindi le operazioni elementari all'interno della macchina sono diverse da quelle ordinarie.

In particolare non sono definite, la proprietà associativa dell'addizione e la proprietà distributiva della molti pli=

cazione rispetto alla somma.

Quando si sottraggono 2 numeri di macchina con uguale segno e esponente b e diverse mantisse, si perdono cifre significative.

ERRORE NEL CALCOLO di una FUNZIONE

Una funzione NON RAZIONALE φ viene sempre sostituita all'interno del calcolatore, da una funzione RAZIONALE f, il cui uso comporta un errore di troncamento $f - \varphi$.

Nel calcolo di una funzione RAZIONALE $f(x_1, \dots, x_n)$ in un punto assegnato $P_0 = (x_1^0, \dots, x_n^0)$, in generale non si ottiene il valore $f(P_0)$ cercato, a causa delle approssimazioni.

ERRORE ALGORITMICO → le operazione aritmetiche che compaiono in $f(P)$ devono essere sostituite con le corrispondenti operazioni di macchina e organizzate in un certo algoritmo; quindi viene sostituita $f(P)$ con $f_a(P)$

ERRORE TRASMESSO dai DATI → le coordinate del punto P_0 non possono essere rappresentate esattamente e si approssimano con numeri di macchina; quindi viene sostituito P_0 con P_1.

$\Rightarrow$ sostituiamo $f(P_0)$ con $f_a(P_1)$

ERRORE ASSOLUTO

Si vuole calcolare la funzione $f(P)$ nel punto $P_0 \in \mathbb{R}^n : P_0 = (x_1^0, \dots, x_n^0)$

Consideriamo l'insieme D di INDETERMINAZIONE di P_0

$$D = \{P_0 \in \mathbb{R}^n : a_i \le x_i^0 \le b_i \text{ con } i = 1, \dots, n \quad a_i, b_i \in \mathbb{R}\}$$

Il valore cercato $f(P_0)$ viene sostituito dal valore calcolato $f_a(P_1)$ con $P_1 \in \mathbb{R}^n : P_1 = (x_1', \dots, x_n')$ e $P_1 \in D$.

$\delta_f = f_a(P_1) - f(P_0)$ ERRORE ASSOLUTO TOTALE

$\delta_a = f_a(P_1) - f(P_1)$ ERRORE ASSOLUTO ALGORITMICO

$\delta_d = f(P_1) - f(P_0)$ ERRORE ASSOLUTO TRASMESSO dai DATI