Teoria degli Errori - Rappresentazione dei numeri

Calcolo Numerico - Ing. Inf. - Lezione 1

Outline

Outline

Fissato un numero intero $\beta > 1$, ogni numero non nullo $x \in \mathbb{R} - \{0\}$ ammette una **rappresentazione in base β** data da

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

Nella precedente formula si ha

1. $b \in \mathbb{Z}$
2. $\alpha_i \in \mathbb{N}_0$ con $0 \leq \alpha_i \leq \beta - 1$, $i = 1, 2, \ldots$
3. $sign(x) = \begin{cases} 1 & \text{se} \quad x > 0 \\ -1 & \text{se} \quad x < 0 \end{cases}$

base

Il numero β si chiama la **base** della rappresentazione

esponente

b è l'**esponente** (determina l'ordine di grandezza del numero)

cifre

I numeri α_i sono le **cifre** della rappresentazione.

Teorema di rappresentazione

Data una base intera $\beta > 1$ e un qualunque numero reale x diverso da zero, esiste un'unica rappresentazione in base β

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

tale che

1. $\alpha_1 \neq 0$,
2. $\nexists \, k \in \mathbb{N}$ per cui si abbia $\alpha_j = \beta - 1, \; \forall j > k$.

La rappresentazione ora stabilita si dice **rappresentazione in virgola mobile normalizzata** del numero reale x

Fissato β, sono quindi univocamente determinati i numeri b e α_i, $i = 1, 2, \ldots$, della rappresentazione normalizzata e la serie $\sum_{i=1}^{\infty} \alpha_i \beta^{-i}$ è detta **mantissa** del numero x

È immediato verificare che

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

(motivo per cui si parla di rappresentazione normalizzata)

Mantissa

Il minimo valore assunto dalla mantissa si ottiene dando a α_1 il valore minimo accettabile e quindi $\alpha_1 = 1$ e ponendo $\alpha_i = 0$, $i = 2, 3, \ldots,$ per cui

$$\sum_{i=1}^{\infty} \alpha_i \, \beta^{-i} \; = \; 1 \cdot \beta^{-1} \; = \; \frac{1}{\beta}$$

Per determinare l'estremo superiore dei valori assunti dalla mantissa, ipotiziamo che risulti $\alpha_i = \beta - 1$, $i = 1, 2, 3, \ldots$ (ricordiamo che abbiamo escluso questa possibilità e quindi stiamo maggiorando la mantissa)

Mantissa

In questo caso risulta

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

La serie a secondo membro è una serie geometrica di ragione $\frac{1}{\beta}$ $(0 < \frac{1}{\beta} < 1)$ con primo termine $\frac{1}{\beta}$

Ricordando l'espressione della somma di una serie geometrica si ha

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

Osservazione

Perchè sono escluse le rappresentazioni periodiche di periodo $\beta - 1$?

Consideriamo $\beta = 10$ ed il numero $1.2\overline{9}$

Di un numero periodico sappiamo calcolare la **frazione generatrice** che in questo caso è

$$1.2\overline{9} = \frac{129 - 12}{90} = \frac{117}{90} = 1.3$$

Quindi $1.2\overline{9}$ e 1.3 sono **DUE** rappresentazioni diverse dello stesso numero!!!!

All'interno di un calcolatore si possono rappresentare solo m ($m \in \mathbb{N}$) cifre della mantissa di x.

Si hanno due modi di passare dalla rappresentazione infinita (infinte cifre della mantissa) alla rappresentazione finita (m cifre della mantissa)

Esiste la **rappresentazione per troncamento**

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

E la **rappresentazione per arrotondamento**

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

Nella rappresentazione per arrotondamento, in particolare nel secondo caso, si potrebbe dover ricorrere ad una nuova normalizzazione della mantissa con eventuale modifica anche dell'esponente b

Si può dimostrare che

$$| \, tr(x) - x \, | < \beta^{b-m}$$

e

$$| \, rd(x) - x \, | \leq \frac{1}{2}\beta^{b-m}$$

Segue che la rappresentazione per arrotondamento è, in generale, una migliore approssimazione del numero reale x

Si indichi con M l'insieme dei numeri z rappresentabili all'interno di un calcolatore, comunemente chiamati **numeri di macchina**.

L'insieme dei numeri di macchina M è un insieme finito

Infatti, fissati β ed m e supposto $L \leq b \leq U$ $(L, U \in \mathbb{Z})$, la cardinalità di M risulta

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

L'insieme M viene indicato con il simbolo $F(\beta, m, L, U)$ per evidenziare le caratteristiche della macchina.

Overflow

Dato un qualunque numero reale $x \neq 0$, non è assicurata l'esistenza di $rd(x)$ fra i numeri di macchina

Sia, per esempio, $F(10, 3, -99, 99)$ e $x = 0.9998 \times 10^{99}$
si ha $rd(x) = 0.1 \times 10^{100}$ che non rientra nell'insieme dei numeri di macchina considerato

In questo caso si ha una situazione di **overflow** (il numero da rappresentare è troppo grande e non appartiene a M)

In generale, i calcolatori segnalano il presentarsi di un overflow, alcuni arrestano l'esecuzione del programma, altri proseguono ponendo $rd(x) = sign(x) \max_{y \in M} |y|$

Underflow

Analogamente, si consideri il numero $x = 0.01 \times 10^{-99}$
si ha $rd(x) = 0.1 \times 10^{-100}$ che non è un numero di macchina
In questo caso si presenta una situazione di **Underflow** (il numero da rappresentare è troppo piccolo e non appartiene a M)

Non tutti i calcolatori segnalano questa situazione e nel caso in cui proseguano l'esecuzione del programma pongono $rd(x) = 0$

È possibile dimostrare che $rd(x)$ soddisfa la relazione

$$| \, rd(x) - x \, | \leq | \, z - x \, |, \quad \forall z \in M$$

Se $rd(x) \in M$ allora, in valore assoluto, differisce da x meno di qualunque altro numero di macchina.

Errore Assoluto

Errore Assoluto

Si definisce **errore assoluto** della rappresentazione del numero

reale x il valore

$$\delta_x = rd(x) - x$$

È immediato ricavare la limitazione

$$| \delta_x | \leq \frac{1}{2}\beta^{b-m}$$

Errore Relativo

Errore Relativo

Si definisce **errore relativo** il valore

$$\epsilon_x = \frac{rd(x) - x}{x} = \frac{\delta_x}{x}.$$

Anche in questo caso si ha una limitazione data da

$$\mid \epsilon_x \mid < \frac{1}{2}\beta^{1-m}$$

Precisione di Macchina

Il numero

$$u = \frac{1}{2}\beta^{1-m}$$

è detto **precisione di macchina**

La precisione di macchina è il massimo errore relativo commesso nel passaggio da x a $rd(x)$

Se l'errore relativo di una approssimazione non supera $\frac{1}{2}\beta^{1-m}$ si dice che l'approssimazione è corretta almeno fino alla m-esima cifra significativa

Operazioni di Macchina

Nell'insieme M le quattro operazioni elementari non sono chiuse...cioè il risultato della operazione tra due numeri di macchina non è detto che sia un terzo numero di macchina

Nell'insieme M non tutte le proprietà delle quattro operazioni elementari risultano verificate, in quanto il risultato di una operazione deve essere ricondotto ad un numero di macchina

Quindi le quattro operazioni elementari all'interno di una macchina sono diverse dalle corrispondenti operazioni ordinarie
Per esempio, l'addizione tra numeri di macchina non gode della proprietà associativa

Sia $M = F(10, 3, -99, 99)$ e siano $x = 0.135 \times 10^{-4}$,
$y = 0.258 \times 10^{-2}$ e $z = -0.251 \times 10^{-2}$
Indicanco con $\oplus$ l'operazione di addizione tra elementi di M, si ha

$$\begin{aligned} x \oplus (y \oplus z) &= 0.135 \times 10^{-4} \oplus \left(0.258 \times 10^{-2} \oplus -0.251 \times 10^{-2}\right) \\ &= 0.135 \times 10^{-4} \oplus 0.700 \times 10^{-4} \\ &= 0.835 \times 10^{-4}, \end{aligned}$$

mentre

$$\begin{aligned} (x \oplus y) \oplus z &= \left(0.135 \times 10^{-4} \oplus 0.258 \times 10^{-2}\right) \oplus -0.251 \times 10^{-2} \\ &= 0.259 \times 10^{-2} \oplus -0.251 \times 10^{-2} \\ &= 0.800 \times 10^{-4}. \end{aligned}$$

Cancellazione

Cancellazione

Se si sottraggono due numeri di macchina dello stesso segno che hanno lo stesso esponente b e con le mantisse che differiscono di poco, si ha una perdita di cifre significative nel risultato

Questo fenomeno è detto **cancellazione** e produce una notevole amplificazione degli errori relativi.