CALCOLO NUMERICO

Fenomeno naturale → Modello matematico continuo → Modello matematico discreto → Metodo di risoluzione → Soluzione

Si possono produrre errori perché il calcolatore opera su numeri DATI NUMERICI rappresentati per mezzo di una sequenza finita di cifre.
I numeri irrazionali hanno problemi → troncamento → errore

- ERRORE DI RAPPRESENTAZIONE (INERENTE) (sorgente dall'omissione dei dati)
- ERRORE ALGORITMICO (sorgente dalle operazioni)

MASTER COPY PISA

Rappresentazione in base di un numero

L'insieme dei numeri rappresentabili con un calcolatore è finito

TEOREMA DI RAPPRESENTAZIONE IN BASE

- Sia $x \in \mathbb{R}$, $x \neq 0$
- Siano unici il numero intero p e la successione $\{d_i\}_{i=1,2,\dots}$ di numeri interi, $0 \leq d_i < \beta$, $d_1 \neq 0$, non definitivamente uguali a $\beta - 1$
- $p \in \mathbb{Z}$
- $\text{sgn}(x) = 1$ se $x > 0$; $\text{sgn}(x) = -1$ se $x < 0$

$$x = \text{sgn}(x)\, \beta^{p} \sum_{i=1}^{\infty} d_i \beta^{-i} \qquad \beta > 1$$

p esponente

d_i cifre della rappresentazione

$\sum_{i=1}^{\infty} d_i \beta^{-i}$ mantissa

MASTER COPY
PISA

master COPY
COPISTERIA
050/8312126 388/9837745

MASTER COPY
PISA

$$x = \text{sgn}(x)\, \beta^p \sum_{i=1}^{\infty} d_i \beta^{-i}$$

32 bit $\begin{cases} 1 \text{ bit per } \text{sgn}(x) \\ 8 \text{ bit per esponente} \\ 23 \text{ bit per cifre della mantissa} \end{cases}$

Es: $\beta = 2$

$$\pm 2^p \sum_{i=1}^{23} d_i 2^{-i} \qquad 1\cdot 2^0 + 1\cdot 2^1 + 1\cdot 2^2 + \dots + 1\cdot 2^7 = 127 \to 8 \text{ BIT}$$

$$-127 \le p \le 127$$

NUMERI DI MACCHINA, in base β, con t cifre significative

$t \ge 1, m, M > 0$

$$F(\beta, t, m, M) = \left\{ x \in \mathbb{R} : x = \text{sgn}(x)\, \beta^p \sum_{i=1}^{t} d_i \beta^{-i} \text{ con } 0 \le d_i < \beta \text{ per } i = 1, 2, \dots, t,\ d_1 \neq 0,\ -m \le p \le M \right\} \cup \{0\}$$

Tutti i numeri di macchina hanno una rappresentazione normalizzata

Proprietà:

- insieme di numeri finito
- simmetria rispetto all'origine
- esistono massimo e minimo

Ricorda $\boxed{-m \le p \le M}$ $\qquad$ $\boxed{0 < d_i \le \beta - 1}$

$$w = \beta^{-m} \cdot \beta^{-1} = \beta^{-m-1}$$

$$\Omega = \beta^M (\beta - 1) \sum_{i=1}^{t} \beta^{-i}$$

DISTRIBUZIONE NON UNIFORME



Es: $F(2, 3, 1, 1)$ $\quad m = 1 \quad M = 1 \quad \beta = 2 \quad t = 3$

$$w = \beta^{-m-1} = 2^{-2} = \frac{1}{4}$$

$$\Omega = \beta^M (\beta - 1) \sum_{i=1}^{t} \beta^{-i} = 2^1 (2-1) \sum_{i=1}^{3} 2^{-i} =$$

$$= 2 \cdot \left(2^{-1} + 2^{-2} + 2^{-3} \right) = 2 \cdot \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} \right) = 1 + \frac{1}{2} + \frac{1}{4} = \frac{4+2+1}{4} = \frac{7}{4}$$

$$-\frac{7}{4} < x < -\frac{1}{4} \quad \vee \quad \frac{1}{4} < x < \frac{7}{4}$$

$x > \Omega$ overflow

$x \in (0; \omega)$ underflow

Errori di rappresentazione

$$x \in \mathbb{R} \rightsquigarrow \tilde{x} \in F$$

$\tilde{x} - x$ errore assoluto

$\epsilon_x = \frac{\tilde{x} - x}{x}$ errore relativo, di rappresentazione

Condizioni standard $\quad \omega < x < \Omega$

TEOREMA

$$|\epsilon_x| = \left|\frac{\tilde{x} - x}{x}\right| < u = \beta^{1-t}$$

u = precisione di macchina, fornisce l'ordine dell'errore

Dimostrazione per maggiorazione

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

Ricordo: $\quad x = \beta^p \sum_{i=1}^{+\infty} d_i \beta^{-i} \qquad \tilde{x} = \beta^p \sum_{i=1}^{t} d_i \beta^{-i}$

$x > \tilde{x} \qquad x - \tilde{x} > 0 \qquad \tilde{x} - x < 0$

in generale $= \frac{x - \tilde{x}}{x} \leq \beta^p \sum_{i=t+1}^{+\infty} d_i \beta^{-i} \leq$ [illegible]

$$\leq \frac{\beta^p \sum_{i=t+1}^{\infty} d_i \beta^{-i}}{\beta^p \cdot \beta^{-1}} \leq$$

CONVERGENZA SERIE

$$\leq \frac{(\beta - 1) \cdot \beta^{-(t+1)} \sum_{i=0} \beta^{-i}}{\beta^{-1}} \cdot \frac{1}{1 - \frac{1}{\beta}} = (\beta - 1)\, \beta^{-t-1} \frac{\beta}{(\beta - 1)}\, \beta = \beta^{-t+1}$$

MASTER COPY PISA

$\tilde{x} = x(1+\varepsilon_x) \quad |\varepsilon_x| < u$

$\tilde{x} = x + \varepsilon_x x$

$\frac{\tilde{x} - x}{x} = \varepsilon_x$

ARITMETICA DI MACCHINA

$\tilde{x} \in F \qquad \tilde{x} + \tilde{y} \notin F$

$\tilde{y} \in F \qquad \tilde{x} \oplus \tilde{y} \in F$

$\tilde{x} \oplus \tilde{y} = \text{troncamento}(\tilde{x} + \tilde{y}) = (\tilde{x} + \tilde{y})(1+\varepsilon) \quad |\varepsilon| \le u$

Es: $f(x) = x^2 - 1 = (x+1)(x-1)$

□ $x^2 - 1$ Algoritmo 1

1. $z = \tilde{x} \odot \tilde{x}$
2. $w = z - 1$

$g_1(\tilde{x}) = (\tilde{x} \otimes \tilde{x}) \ominus 1$

$w = (z-1)(1+\varepsilon_1) = \left[\tilde{x}^2(1+\varepsilon_2) - 1\right](1+\varepsilon_1) =$

$= \left[x^2(1+\varepsilon_x)^2(1+\varepsilon_2) - 1\right](1+\varepsilon_1) \doteq \left[x^2(1+\varepsilon_x^2+2\varepsilon_x)(1+\varepsilon_2) - 1\right](1+\varepsilon_1) \doteq$

$\doteq \left[x^2(1+2\varepsilon_x)(1+\varepsilon_2) - 1\right](1+\varepsilon_1) \doteq \left[x^2(1+\varepsilon_2+2\varepsilon_x) - 1\right](1+\varepsilon_1) =$

$= \left[x^2 + \varepsilon_2 x^2 + 2\varepsilon_x x^2 - 1\right](1+\varepsilon_1) =$

$\doteq x^2 + \varepsilon_1 x^2 + \varepsilon_2 x^2 + 2\varepsilon_x x^2 - 1 - \varepsilon_1 =$

$= x^2(1+\varepsilon_1) + x^2(\varepsilon_2 + 2\varepsilon_x) - (1+\varepsilon_1)$

$\frac{\tilde{x} - x}{x} = \frac{w - (x^2-1)}{x^2-1} = \frac{x^2 + \varepsilon_1 x^2 + x^2\varepsilon_2 + 2\varepsilon_x x^2 - 1 - \varepsilon_1 - x^2 + 1}{x^2-1} =$

$= \frac{x^2(\varepsilon_1 + \varepsilon_2)}{x^2-1} + \frac{2x^2}{x^2-1}\cdot\varepsilon_x - \frac{\varepsilon_1}{x^2-1}$ → SO

□ $(x+1)(x-1)$ Algoritmo 2

1. $z = \tilde{x} \oplus 1$
2. $y = \tilde{x} \ominus 1$
3. $w = z \otimes y$

□ Le soluzioni mantengono solo gli errori in fondo!