RAPPRESENTAZIONE DEI NUMERI

Sia $x \in \mathbb{R} \setminus \{0\}$ e $\beta \in \mathbb{N} : \beta > 1$ x ammette una rappresentazione in base β

$$x = \text{sign}(x)\, \beta^{b} \sum_{i=1}^{\infty} d_i \beta^{-i} \quad \text{con } b \in \mathbb{Z}$$

$$d_i \in \mathbb{N} : 0 \leq d_i \leq \beta - 1 \quad i = 0, 1, \dots$$

β = BASE DI RAPPRESENTAZIONE
b = ESPONENTE DEL NUMERO
d_i = CIFRE DI RAPPRESENTAZIONE
β^b = ORDINE DI GRANDEZZA DEL NUMERO

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

TH. UNICITÀ DELLA RAPPRESENTAZIONE

Data una base intera $\beta > 1$ e un qualunque numero reale diverso da zero, esiste un'unica rappresentazione in base β t.c.

1) $d_1 \neq 0$ 2) $\nexists j : d_i = \beta - 1 \quad i > j$ → serve ad evitare di avere numeri periodici

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

$\sum_{i=1}^{\infty} d_i \beta^{-i}$ è limitata sia inferiormente che superiormente

- basso $= \frac{1}{\beta}$
- alto $= 1$

Essa viene detta MANTISSA DEL NUMERO x

Il valore più alto che d_i può raggiungere, per ipotesi, è $\beta - 1$

$$\sum_{i=1}^{\infty} d_i \beta^{-i} < \frac{(\beta-1)}{\beta}\left(1 + \beta^{-1} + \beta^{-2} \dots\right) = \frac{\beta-1}{\beta} \cdot \frac{1}{1-\frac{1}{\beta}} = \frac{\beta-1}{\beta} \cdot \frac{\beta}{\beta-1} = 1$$

perché va contro l'ipotesi ② l'uguaglianza

RICONDUCIBILE ALLA PROGRESSIONE GEOMETRICA $a_0 \cdot \frac{1}{1-q}$

IL NUMERO VERRÀ RAPPRESENTATO COSÌ $= 0,127 \cdot 10^{-3}$

Ma perché NON si vogliono i periodici?

$1.1\overline{9} = \frac{119-11}{90} = 1.2$ Si avrebbero 2 rappresentazioni uguali per 2 numeri diversi

↳ tanti 9 quante sono le cifre del periodo
↳ tanti 0 quante cifre dopo la virgola, ma prima del periodo

Ma perché si scarta lo zero?

Perché avrebbe più rappresentazioni

RAPPRESENTAZIONE NUMERI SUL CALCOLATORE

1) $fd(x) = \text{sign}(x)\, \beta^{b'} \sum_{i=1}^{m} d_i' \beta^{-i}$

Perché la sommatoria arriva fino a m?

La sommatoria arriva fino a m perché non si può rappresentare, su un calcolatore, un numero infinito di cifre!
Si parla appunto di NUMERO DI MACCHINA, ottenuto grazie ad un ROUNDING (rd(x)), arrotondamento.

Perché si usano gli apici a β e di?

Si usano gli apici perché l'arrotondamento può produrre grandi cambiamenti

ES. → $d_i' \Rightarrow 0,1997$ arr alla 3ª diventa $0,2$

$b' \Rightarrow 0,9997 \times 10^0 = 0,1 \times 10^1$

$$rd(x) = \begin{cases} tr(x) & \text{SE } d_{m+1} < \frac{\beta}{2} \\ \text{sign}(x)\,\beta^b \left(\sum_{i=1}^{m} d_i \beta^{-i} + \beta^{-m} \right) & \text{SE } d_{m+1} \geq \frac{\beta}{2} \end{cases}$$

2) $tr(x) = \text{sign}(x)\,\beta^b \sum_{i=1}^{m} d_i \beta^{-i}$ ERRORE TRONCAMENTO: $|tr(x) - x| \leq \beta^{b-m}$

Si parla di rappresentazione per troncamento

Si indichi con F l'insieme dei numeri rappresentabili all'interno di un calcolatore, comunemente chiamati NUMERI DI MACCHINA
I NUMERI di macchina però sono in numero finito perciò poniamo $L \leq b \leq U$ con $L, U \in \mathbb{Z}$. Quindi F è dato da:

$F(\beta, m, L, U)$ da cui è possibile notare che più m è grande più l'approssimazione sarà preciso e più $(U-L)$ è grande maggiore sarà l'ordine di grandezza rappresentabile

PRECISIONE DI MACCHINA: $(\beta^{(1-m)})/2$

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

- 2 → segno
- $(U-L+1)$ → b: $2 \leq b \leq 5$, $5-3+1=3$
- $(\beta-1)$ → d_1
- β^{m-1} → restanti d
- $+1$ → per lo zero la sua configurazione non è unica, ogni calcolatore ha la sua maniera di rappresentarlo

ESERCIZIO ARROTONDAMENTO + OVERFLOW:

$F(10, 2, -2, 2)$ $\quad \frac{1}{10} < \text{mantissa} < 1$

$a = 11,50$ $\quad rd(a) = 0,12 \times 10^2$
$b = 0,01$ $\quad rd(b) = 0,1 \times 10^{-1}$ ⇒ d_1 deve essere $\neq 0$
$c = 9,27$ $\quad rd(c) = 0,93 \times 10^1$

$x = rd(a) \cdot rd(b) = 0,012 \times 10^1$
$y = rd(a) \cdot rd(c) = 0,1116 \times 10^3$

$rd(x) = 0,12 \times 10^0$ ⇒ $d_1 \neq 0$

$rd(y) = 0,11 \times 10^3$ XNOX si genera un OVERFLOW, il numero è troppo grande e non appartiene a F Voleva 10^2 al massimo, non 10^3

All'incontrario, se il numero è troppo piccolo e non appartiene a F si genera un UNDERFLOW

Ma qual'è l'errore che si commette ad approssimare?

ERRORE ASSOLUTO $\quad \delta x = \tilde{x}d(x) - x$

ERRORE RELATIVO $\quad \varepsilon x = \dfrac{\tilde{x}d(x) - x}{x} = \dfrac{\delta x}{x}$ è il più interessante, perchè ci indica l'errore rispetto all'ordine di grandezza.
es. errore 1 cm su 20 cm
errore 1 cm su 20000 cm

ES. $0,11|5378 = 0,12 \to 0,5 \cdot 10^{-2}$
$0,11|4975 = 0,11 \to 0,5 \cdot 10^{-2}$ / da queste considerazioni possiamo notare che

$$|\tilde{x}d(x) - x| = |\delta x| < \frac{1}{2}\beta^{b-m}$$

$$\frac{|\tilde{x}d(x) - x|}{|x|} = |\varepsilon x| < \left(\frac{\frac{1}{2}\beta^{b-m}}{\beta^{b-1}}\right) = \frac{1}{2}\beta^{1-m} \quad = \text{PRECISIONE DI MACCHINA}$$

↳ il + piccolo valore rappresentabile (RIFLETTI)

Consideriamo adesso $P_0 \in \mathbb{R}^n \quad P_0 = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, \dots, x_n^{(0)})$

$f : \mathbb{R}^n \to \mathbb{R} \quad : f(x_1, x_2) = x_1 x_2 \quad f(P_0)$?

Sia $D = \{ P \in \mathbb{R}^n \mid a_i \le x_i^{(0)} \le b_i \quad i = 1 \dots n \}$ l'INSIEME DI INDETERMINAZIONE

problema 1

$f_0(\sqrt{3}; \pi)$ ma $\sqrt{3}$ e π non posso rappresentarle sul calcolatore perchè hanno tantissime cifre. Decido allora di prendere solo le prime 2 cifre decimali

$P_1(1,74 ; 3,14)$ è comunque $\neq$ da P_0

problema 2

Sia ad esempio $f(x_1, x_2) = e^{\sqrt{x_1}\cos(x_2)}$. La funzione esponenziale viene ottenuta calcolando l'espansione dei primi n termini. Facendo ciò qualcosa viene perso. Si otterrà quindi una f_a = FUNZIONE ALGORITMICA

Il risultato di questi 2 problemi è la generazione di un errore detto

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

$$\delta f = \underbrace{f_a(P_1) - f(P_0)}_{\delta a} + \underbrace{f(P_1) - f(P_0)}_{\delta d} = \delta a + \delta d$$

δa ERRORE GENERATO DALL'ALGORITMO ⇓ definito e stimabile

δd ERRORE TRASMESSO DAI DATI ⇓ può essere rappresentabile

$$\delta d = f(P_1) - f(P_0) = \sum_{i=1}^{n} \underbrace{\frac{\partial f}{\partial x_i}}_{f_i} \underbrace{\left(x_i^{(1)} - x_i^{(0)}\right)}_{\delta x_i}$$

FORMULA DI TAYLOR ARRESTATA AL PRIMO TERMINE

COEFFICIENTE DI AMPLIFICAZIONE

$\delta d = f_i \cdot \delta x_i$

$$|\delta d| = |f(P_1) - f(P_0)| = \left|\sum_{i=1}^{n} f_i \delta x_i\right| \le \sum_{i=1}^{n} |f_i \delta x_i| = \sum_{i=1}^{n} |f_i|\,|\delta x_i|$$

$$\text{ma } |f_i| \le \sup_{P \in D} \left|\frac{\partial f}{\partial x_i}\right| = A_{x_i} \quad \to \quad |\delta d| \le \sum_{i=1}^{n} A_{x_i} |\delta x_i|$$

MASTER COPY
Tel. 050 831 2126
Cell. 340 9837745

ESERCIZIO PROBLEMA DIRETTO

$f(x_1, x_2) = x_1 x_2 \quad P_0 = \left(\pi, \frac{4}{3}\right) \quad |\delta f| = |\delta a + \delta d| \leq |\delta a| + |\delta d|$ ERRORE ASSOLUTO?

Scelgo un algoritmo che arrotondi il risultato alla 2ª cifra decimale

$|\delta a| \leq \frac{1}{2} 10^{-2}$

$|\delta x_1|, |\delta x_2| \leq \frac{1}{2} 10^{-2}$ mi indica con quale precisione devo prendere i dati

$$|\delta d| \leq \sum_{i=1}^{2} A_{xi} |\delta x_i| = A_{x1} \cdot \frac{1}{2} 10^{-2} + A_{x2} \cdot \frac{1}{2} \cdot 10^{-2} = 3 \cdot 10^{-2}$$

Fissiamo adesso un intervallo di indeterminazione $P_0 \in [3,4] \times [1,2]$

$$A_{x1} \geq \sup_{x \in D} \left|\frac{\partial f}{\partial x_1}\right| = x_2 \quad A_{x1} = 2$$

$$A_{x2} \geq \sup_{x \in D} \left|\frac{\partial f}{\partial x_2}\right| = x_1 \quad A_{x2} = 4$$

$|\delta f| \leq \frac{1}{2} \cdot 10^{-2} + 3 \cdot 10^{-2} = \frac{7}{2} \cdot 10^{-2}$ MASSIMO ERRORE ASSOLUTO

$x_1^{(1)} = 3,14$ $\quad x_2^{(1)} = 1,33$ $\quad x_1^{(1)} \cdot x_2^{(1)} = 4,1762 = 4,18$ perché cade $|\delta a| \leq \frac{1}{2} \cdot 10^{-2}$

$f(P_0) = 4,18879$

$|\delta f| = f_a(P_1) - f(P_0) = 0,00879 \leq \frac{7}{2} \cdot 10^{-2}$ ⇒ MASSIMO ERRORE ASSOLUTO

ESERCIZIO PROBLEMA INVERSO

$f(x_1, x_2) = x_1 \cdot x_2 \quad P_0 = \left(\pi, \frac{4}{3}\right)$ e voglio che $|f_a(P_1) - f(P_0)| < 10^{-2}$

Scelgo come intervallo di indeterminazione $P_0 \in [3,4] \times [1,2]$

$|\delta f| \leq |\delta a| + |\delta d| = |f_a(P_1) - f(P_0)| < 10^{-2}$ visto che il contributo è dato da 2 elementi lo divido a metà

$|\delta a| \leq \frac{1}{2} \cdot 10^{-2} \quad |\delta d| \leq \frac{1}{2} \cdot 10^{-2}$

$$|\delta d| \leq \sum_{i=1}^{2} A_{xi} |\delta x_i| \leq \frac{1}{2} \cdot 10^{-2}$$ visto che il contributo è dato da 2 elementi lo divido a metà

$A_{x1} |\delta x_1| \leq \frac{1}{4} \cdot 10^{-2} \quad |\delta x_1| \leq \frac{1}{4} \cdot \frac{1}{2} \cdot 10^{-2}$

$|\delta x_2| \leq \frac{1}{4} \cdot \frac{1}{4} \cdot 10^{-2}$

$x_1^{(1)} = 3,141$

$x_2^{(1)} = 1,333$ visto che $|fl(x) - x| \leq \beta^{b-m}$

$x_1^{(4)} \cdot x_2^{(2)} = 4,1869$ ma $|\delta a| \leq \frac{1}{2} \cdot 10^{-2}$ quindi $= 4,19$

$f(P_1) = 4,19$
$f(P_0) = 4,18879$
$\Big\}$ $1 \cdot 10^{-3} < 1 \cdot 10^{-2}$

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

ERRORE RELATIVO $\varepsilon_f = \frac{f_a(P_1) - f(P_0)}{f(P_0)}$

$\varepsilon_a = \frac{f_a(P_1) - f(P_1)}{f(P_1)}$ ⇒ errore che si commette sostituendo alla funzione l'algoritmo di calcolo

$\varepsilon_d = \frac{f(P_1) - f(P_0)}{f(P_0)}$ ⇒ errore che si commette sostituendo a P_0 P_1

Da qui risulta facile notare che $\varepsilon_f = \varepsilon_a + \varepsilon_d + \underbrace{\varepsilon_a \cdot \varepsilon_d}$ è solitamente trascurabile

ES. $\varepsilon_a = 10^{-2}$, $\varepsilon_d = 10^{-2}$
$\varepsilon_a \cdot \varepsilon_d = 10^{-4} << 10^{-2}$

$\varepsilon_d = \frac{\delta d}{f(P_0)} = \sum_{i=1}^{n} \frac{1}{f(P_0)} \cdot \frac{\partial f}{\partial x_i} \delta x_i$ moltiplicando e dividendo per $x_i^{(0)}$

$\varepsilon_d = \sum_{i=1}^{n} \underbrace{\frac{x_i^{(0)}}{f(P_0)} \frac{\partial f}{\partial x_i}} \cdot \frac{\delta x_i}{x_i^{(0)}} = \sum_{i=1}^{n} \gamma_i \varepsilon_{x_i}$

γ_{x_i} = COEFFICIENTI DI AMPLIFICAZIONE DEGLI ERRORI RELATIVI

PROBLEMA BEN CONDIZIONATO = Se i coefficienti γ_i sono tali che ε_d risulta dello stesso ordine di ε_{x_i}
(Errore dei dati ≈ a quello con cui sono stati introdotti)

PROBLEMA MAL CONDIZIONATO = Se a piccoli errori relativi ε_{x_i} corrisponde un ε_d rilevante
(Errore dei dati molto amplificato rispetto a quello con cui sono stati introdotti)

ALGORITMO STABILE = produce errori accettabilmente limitati nella sua applicazione

ALGORITMO INSTABILE = produce errori notevoli nella sua applicazione

Prendendo come $\delta d = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \delta x_i$ e $\varepsilon_d = \sum_{i=1}^{n} \frac{x_i^{(0)}}{f(P_0)} \frac{\partial f}{\partial x_i} \varepsilon_{x_i}$

$f(x_1, x_2)$	δd	ε_d
$x_1 + x_2$	$\delta x_1 + \delta x_2$	$\frac{x_1}{x_1+x_2}\varepsilon_{x_1} + \frac{x_2}{x_1+x_2}\varepsilon_{x_2}$
$x_1 - x_2$	$\delta x_1 - \delta x_2$	$\frac{x_1}{x_1-x_2}\varepsilon_{x_1} - \frac{x_2}{x_1-x_2}\varepsilon_{x_2}$
$x_1 \cdot x_2$	$x_2\delta x_1 + x_1\delta x_2$	$\varepsilon_{x_1} + \varepsilon_{x_2}$
$\frac{x_1}{x_2}$	$\frac{\delta x_1}{x_2} - \frac{x_1 \delta x_2}{x_2^2}$	$\varepsilon_{x_1} - \varepsilon_{x_2}$