- CALCOLO NUMERICO
corso di Paolo Gallenzoni

RAPPRESENTAZIONE DEI NUMERI

- Vogliamo trovare una formalità per la rappresentazione dei numeri su ogni base.

- $x \in \mathbb{R} \setminus \{0\}$

- $\beta \in \mathbb{N}, \beta > 1$
  - ↳ BASE DELLA RAPPRESENTAZIONE

MASTER COPY
Tel. 050 8312126
Cell. 333 9837745

- $b \in \mathbb{Z}$
  - ↳ ESPONENTE DELLA BASE
- $d_i \in \mathbb{N}, 0 \le d_i \le \beta - 1$ (con $i = 0, 1, 2 \dots$)
  - ↳ CIFRE DELLA RAPPRESENTAZIONE

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

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

- Vogliamo inoltre ottenere l'UNICITÀ di questa rappresentazione, ponendo due condizioni:

1) $d_1 \neq 0$

2) $\nexists j : d_i = \beta - 1 \ \ \forall i \ge j$ → Esclude tutte le rappresentazioni di periodo $\beta - 1$ (come 1,999...)

- Im sostanza, tra $\cdot\ 1{,}27 \times 10^{-2}$, $\cdot\ 0{,}127 \times 10^{-1}$, $\cdot\ 12{,}7 \times 10^{-3}$, prendiamo la seconda rappresentazione, ossia $0{,}127 \times 10^{-2}$.

- Individuiamo il minimo/estremo inferiore e il massimo/estremo superiore della sommatoria scritta in precedenza.

$$\frac{1}{\beta} \le \underbrace{\sum_{i=1}^{\infty} d_i \beta^{-i}}_{\text{↳ È CHIAMATA MANTISSA DEL NUMERO}} < (\beta-1)\beta^{-1} + (\beta-1)\beta^{-2} + \dots =$$

$$= \frac{(\beta-1)}{\beta}\underbrace{\left(1 + \beta^{-1} + \beta^{-2} + \dots\right)}_{\text{SOMMA DI UNA SERIE GEOMETRICA}} =$$

$$= \frac{\beta-1}{\beta} \cdot \frac{1}{1-\frac{1}{\beta}} = \frac{\beta-1}{\beta} \cdot \frac{\beta}{\beta-1} = 1$$

$$\Rightarrow \underbrace{\frac{1}{\beta}}_{\text{↳ MINIMO}} \le \underbrace{\sum_{i=1}^{\infty} d_i \beta^{-i}}_{\text{↳ MANTISSA}} < \underbrace{1}_{\text{↳ ESTREMO SUPERIORE}}$$

- Tra le rappresentazioni di prima scegliamo $0{,}127 \times 10^{-2}$ perché
  - ha la mantissa minore di 1 ($0{,}127 < 1$)
  - ha la prima cifra dopo la virgola diversa da 0
- la seconda condizione che abbiamo posto esclude i numeri come

$$1{,}1\overline{9} = \frac{119 - 11}{90} = \frac{108}{90} = 1{,}2$$

parlando di ricità della rappresentazione.

- β^b dà quello che è chiamato l'ORDINE DI GRANDEZZA del numero (in $0{,}127 \times 10^{-3}$ è quel 10^{-3})
- In una macchina è impossibile rappresentare le cifre all'infinito: andiamo quindi a definire il NUMERO DI MACCHINA, ossia la rappresentazione fino a m cifre invece che ∞, attraverso due funzioni di approssimazione.

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

↳ TRONCAMENTO

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

↳ ARROTONDAMENTO

→ Cardinalità

$m \in \mathbb{N}$
$L, U \in \mathbb{Z}$
$L \le b \le U$
$F(\beta, m, L, U)$

- I segni possibili sono 2
- Le mantisse possibili sono $(\beta - 1)\beta^m$
- Gli esponenti possibili sono $(U - L + 1)$

$$\Rightarrow card(F) = \underbrace{2}_{\text{SEGNI}} \underbrace{(U - L + 1)}_{\text{ESPONENTI}} \underbrace{(\beta - 1)\beta^{m-1}}_{\text{MANTISSE}} + \underbrace{1}_{\text{RAPPRESENTAZIONE DELLO ZERO}}$$

- F indica una certa configurazione per la rappresentazione
- $card(F)$ è il numero di numeri che si possono rappresentare con la configurazione F.
- F non è un INSIEME CHIUSO rispetto alle operazioni possibili (ossia eseguendo operazioni sugli elementi che appartengono all'insieme, si possono ottenere risultati che non appartengono all'insieme.

MASTER COPY
Tel. 050 8312126
Cell. 388 9837745

Esercizio:

$F = (10, 2, -2, 2)$

$a = 11,50$
$b = 0,01$
$c = 9,27$

MASTER COPY
Tel. 050 09121 33
Cell. 388 9537745

Calcolare:

- $rd(a) = 0,12 \times 10^2$
- $rd(b) = 0,1 \times 10^{-1}$
- $rd(c) = 0,93 \times 10^1$
- $x = rd(a) \cdot rd(b) \rightarrow rd(x) = 0,12 \times 10^0$
- $y = rd(a) \cdot rd(c) \rightarrow rd(y) = 0,11 \times 10^3$ ma non appartiene ad F! (VA IN OVERFLOW)

→ Errore assoluto

$\sigma_x = rd(x) - x$
↳ ERRORE ASSOLUTO (conta anche il segno)

→ Errore relativo

$\varepsilon_x = \frac{rd(x) - x}{x}$
↳ ERRORE RELATIVO, ossia rapporto tra l'errore assoluto e x, per capire quanto incide rispetto alla dimensione di x

→ Massimo errore e precisione di macchina

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

$\Rightarrow |\varepsilon_x| \leq \frac{\frac{1}{2}\beta^{b-m}}{\beta^{b-1}} = \frac{1}{2}\beta^{1-m}$
↳ PRECISIONE DI MACCHINA

→ Errore nel calcolo di una funzione

$P_0 \in R^m$
$P_0 = (x_1^{(0)}, x_2^{(0)}, \ldots, x_m^{(0)})$ → VETTORE

$f : R^m \rightarrow R$ → FUNZIONE

Come si calcola $f(P_0)$?

Prendiamo come esempio $P_0 = (\sqrt{3}, \pi)$ e $f(x_1, x_2) = e^{\sqrt{x_1}\cos(x_2)}$
Dobbiamo mettere al posto di P_0 un punto P_1 rappresentabile, poiché $\sqrt{3}$ e π non lo sono (servono infinite cifre).

Stessa cosa vale per $f(x_1, x_2)$: deve essere trovato un algoritmo di calcolo finito (ad esempio, l'esponenziale si sostituisce con le polinomiali di Taylor). La nuova funzione la chiamiamo $f_a(x_1, x_2)$.

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

$$\delta f = \underbrace{f_a(P_1) - f(P_1)}_{\text{CONTRIBUTO DATO DAL CAMBIAMENTO DELLA FUNZIONE } \delta_a} + \underbrace{f(P_1) - f(P_0)}_{\text{CONTRIBUTO DATO DAL CAMBIAMENTO DEL PUNTO } \delta_d}$$

$$P_0 \in D = \{ p \in \mathbb{R}^2 \mid 1 \le x_1 \le 2,\ 3 \le x_2 \le 4 \}$$

↳ INSIEME DI INDETERMINAZIONE, ossia insieme dove possono trovarsi gli elementi.

$P_1 \in D$

- Su δ_a non è sempre possibile dare un'espressione precisa.
- Su $\delta_d = f(P_1) - f(P_0)$ possiamo applicare lo sviluppo di Taylor fino al primo termine.

$$\delta_d = f(P_1) - f(P_0) = \sum_{i=1}^{m} \frac{\partial f}{\partial x_i}(x_i^{(1)} - x_i^{(0)}) =$$

$$= \sum_{i=1}^{m} p_i \delta x_i$$

p_i: COEFFICIENTI DI AMPLIFICAZIONE

δx_i: ↳ ERRORE RELATIVO ALLA i-ESIMA COMPONENTE

$$|\delta_d| = |f(P_1) - f(P_0)| = \left| \sum_{i=1}^{m} p_i \delta x_i \right| \le \sum_{i=1}^{m} |p_i| |\delta x_i| \le$$

$$\le \underbrace{\sum_{i=1}^{m} A x_i |\delta x_i|}_{\text{↳ MAGGIORAZIONE DELL'ERRORE OTTENUTO DAL DATO}}$$

$$|p_i| \le \sup_{p \in D} \left| \frac{\partial f}{\partial x_i} \right| = A x_i$$

→ Problema diretto

- Consiste nel calcolare il massimo errore assoluto.

$$f(x_1, x_2) = x_1 \cdot x_2$$

$$P_0 = \left(2, \frac{9}{3}\right)$$

$$|\delta f| = |\delta_a + \delta_d| \le |\delta_a| + |\delta_d|$$

Calcoliamo la funzione fino alla terza cifra decimale e arrotondiamo alla seconda cifra decimale.

MASTER COPY
Tel. 050 831 2125
Cell. 388 9637745

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

$|\delta x_1|, |\delta x_2| \leq \frac{1}{2} 10^{-2}$ (Questo scelto è arbitrario. Generalmente si arrotonda alla cifra richiesta dalle $x_1, x_2, \dots, x_m$)

$|\delta d| \leq \sum_{i=1}^{2} Ax_i |\delta x_i| = Ax_1 |\delta x_1| + Ax_2 |\delta x_2|$

$P_0 \in [3*, 4] \times [1, 2]$ → INSIEME DI INDETERMINAZIONE

$\frac{\partial f}{\partial x_1} = x_2 \quad Ax_1 = 2 \qquad \frac{\partial f}{\partial x_2} = x_1 \quad Ax_2 = 4$

$\Rightarrow |\delta d| \leq 2 \cdot \frac{1}{2} \cdot 10^{-2} + 4 \cdot \frac{1}{2} \cdot 10^{-2} = 3 \cdot 10^{-2}$

$|\delta d| \leq \frac{7}{2} \cdot 10^{-2}$

$x_1^{(1)} = 3,14 \quad x_2^{(1)} = 1,33$

$f(P_0) = 4,18879$
$f(P_1) = 4,18$

→ Problema inverso

- Si vuole avere un valore che differisca dal valore esatto $f(P_0)$ meno di un errore prefissato. Ad esempio, che non superi 10^{-2}.

$|f_a(P_1) - f(P_0)| < 10^{-2}$

$|\delta f| \leq |\delta a| + |\delta d|$

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

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

Come dati del problema riprendiamo quelli che abbiamo usato anche per il problema diretto.

$|\delta d| \leq Ax_1 |\delta x_1| + Ax_2 |\delta x_2| \leq \frac{1}{2} 10^{-2}$

$Ax_1 |\delta x_1| \leq \frac{1}{4} \cdot 10^{-2}$

$|\delta x_1| \leq \frac{1}{8} \cdot 10^{-2}$

$Ax_2 |\delta x_2| \leq \frac{1}{4} 10^{-2}$

$|\delta x_2| \leq \frac{1}{16} \cdot 10^{-2}$

Con il troncamento $|tr(x) - x| \leq \beta^{b-m}$
Quindi tronchiamo x_1 allo stesso ordine decimale, mentre arrotondiamo x_2 allo stesso.