RAPPRESENTAZIONE IN VIRGOLA MOBILE E PRECISIONE FINITA

- esponente e frazione di un numero reale non nullo

segno di x — esponente di x — base — Frazione di x

$x = (-1)^s \beta^b f$ $\quad f = \frac{|x|}{\beta^b}$ $\quad \beta^{b-1} \leq |x| < \beta^b$

$x \neq 0 \in \mathbb{R}$, β intero ≥ 2

$\rightarrow \beta^{b-1} \leq |x| < \beta^b \rightarrow \beta^{b-1} \leq f \cdot \beta^b < \beta^b \rightarrow \beta^{-1} \leq f < 1$

Parte intera di $x \rightarrow \lfloor x \rfloor$ = il più grande numero intero $\leq x$

Scrittura posizionale di f in base $\beta \rightarrow 0.c_1 c_2 \ldots$ con $c_1 \neq 0$

→ Come ottengo le cifre $c_1, c_2, \ldots$

1) $i = 1 \quad t_i = f \quad (t_1 = 0, c_1 c_2 \ldots)$
2) $c_i = \lfloor \beta t_i \rfloor \quad (\beta t_i = c_i c_{i+1} c_{i+2} \ldots)$
3) $t_{i+1} = \beta t_i - \lfloor \beta t_i \rfloor \quad (t_{i+1} = 0, c_{i+1} c_{i+2} \ldots)$
4) Se $t_{i+1} = 0 \Rightarrow$ STOP, altrimenti $\rightarrow i = i+1$ e vado al 2)

- numeri in virgola mobile, precisione

→ β intero ≥ 2 e m intero > 0, def INSIEME DEI NUMERI IN VIRGOLA MOBILE (normalizzati) in base β e precisione m

$$F(\beta, m) = \{0\} \cup \left\{ x \in \mathbb{R} \text{ t.c. } x = (-1)^s \beta^b 0.c_1 \ldots c_m \right\}$$ ‖ spesso si indica con M

→ $s \in \{0, 1\}$; $b \in \mathbb{Z}$; $c_1, \ldots, c_m$ cifre in base β ‖ $c_1 \neq 0$

→ È simmetrico rispetto a 0 : se $x \in F(\beta, m) \Rightarrow -x \in F(\beta, m)$

→ è un sottoinsieme proprio di $\mathbb{Q}$

→ è numerabile ed ordinato

→ zero è un punto di accumulazione

Qualunque intorno di 0 contiene almeno un punto appartenente all'insieme

→ $\sup(F(\beta, m)) = +\infty$, $\inf(F(\beta, m)) = -\infty$

→ Insieme elementi positivi di $F(10,1)$: $\forall b$ intero $(b \in \mathbb{Z})$

$$B_b = \{ 10^b\, 0{,}1,\ 10^b\, 0{,}2,\ \dots,\ 10^b\, 0{,}9 \}$$

→ B_b sono ordinati se: $c, d \in \mathbb{R}$ t.c. $c<d \Rightarrow \min B_c < \min B_d$ e proporzionalmente i punti di B_c sono a Sx di $\min B_d$ e i punti di B_d sono a Dx di $\max B_c$

→ distanza tra due elementi consecutivi in B_b : 10^{b-1}

- funzioni successore e predecessore

successore di ξ → $\sigma(\xi)$ || primo elem. di $F(\beta,m)$ a Dx di ξ

predecessore di ξ → $\pi(\xi)$ || primo elem. di $F(\beta,m)$ a Sx di ξ

|| sono def. $\forall$ elem. non nullo di $F(\beta,m)$

→ una è l'inversa dell'altra

- TEO: distribuzione degli elementi di F(β,m)

$\xi = \beta^b f \Rightarrow \sigma(\xi) - \xi = \beta^{b-m}$

Dim → $\sigma(\xi) = \beta^b (f + \beta^{-m}) = \beta^b f + \beta^{b-m} = \xi + \beta^{b-m} \Rightarrow \sigma(\xi) - \xi = \beta^{b-m}$

- Insieme dei numeri in virgola mobile con esponente limitato ed elementi normalizzati

→ $F(\beta, m, b_{min}, b_{MAX})$

→ è costituito da $\emptyset$ e da tutti gli elem con esponente limitato

- Insieme dei numeri in virgola mobile con esponente limitato ed elementi denormalizzati

→ elem. denormalizzati : elem. non null, con esponente minore di b_{min}

→ $F_d(\beta, m, b_{min}, b_{MAX}) = F(\beta, m, b_{min}, b_{MAX})$ + elem denormalizzati

↳ è sottoinsieme di $F(\beta,m)$ costituito da elem. con esponente limitato e da tutti i num. reali x t.c.:
$x = (-1)^s \beta^{b_{min}}\, 0.0 c_2 \dots c_m$

Relazione: $F(\beta, m, b_{min}, b_{MAX}) \subset F_d(\beta, m, b_{min}, b_{MAX}) \subset F(\beta, m)$

Oss: $\xi \in F_d(\beta, m, b_{min}, b_{MAX})$, se ξ ha esponente $\geq b_{min} \Rightarrow c_1, \dots, c_m$ sono le cifre della frazione di ξ. Se ξ ha esponente $< b_{min}$ (elem. denormalizzato) $\Rightarrow c_1 = 0, c_2 \dots, c_m$ non sono le cifre della frazione di ξ

In generale : $n < m \Rightarrow F(\beta, n) \subset F(\beta, m)$

- Funzione arrotondamento: è una funzione che uso per approssimare i numeri reali
  - Def: elementi adiacenti a x-> sono due elementi consecutivi di M tra i quali è compreso x

$x = \beta^b\ 0.c_1 c_2 \ldots \rightarrow$ prendo $\xi_- = \beta^b\ 0.c_1 \ldots c_m$ e $\xi_+ = \delta(\xi_-) \Rightarrow \xi_- < x < \xi_+$

  - Def: arrotondamento di x in M-> rd(x), è definito come l'elemento di M più vicino a x
    - Se x è equidistante dai due elementi di M ad esso adiacenti->ambiguità-> slz: procedure RTTE/RTTA

RTTE: approssima all'elemento con la cifra c_m pari ||è quella di Scilab

→ β deve essere pari

→ M insieme con esponente non limitato oppure con esponente limitato ed elementi denormalizzati

RTTA: approssima all'elemento più lontano da ϕ ||è quello della calcolatrice

  - Proprietà rd(x)
    - Non è invertibile
    - È dispari -> rd(-x) = -rd(x)
    - Debolmente crescente -> x>y => rd(x)>=rd(y)
    - rd(x)=x SSE x€M
    - Se M=F(β,m) allora rd(x)=0 SSE x=0

- funzione errore
  - Errore assoluto δ(x)=rd(x) - x $\forall x \in \mathbb{R}$ ||dispari
  - Errore relativo
    - ε(x)=[rd(x)-x]/x $\forall x \neq \phi \in \mathbb{R}$
    - η(x)=[rd(x)-x]/rd(x) $\forall x \in \mathbb{R}$ t.c. $rd(x) \neq \phi$ ||pari



- TEO: stime delle funzioni errore in F(β,m)

$M = F(\beta, m)$, $x = \beta^b f$ $(x \neq \phi)$

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

→ $|\varepsilon(x)| \leq \frac{1}{2}\beta^{1-m}$

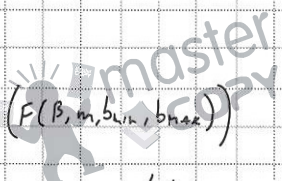
→ $|\eta(x)| \leq \frac{1}{2}\beta^{1-m}$

oss:

se $\xi_{min} = \inf\left(F(\beta, m, b_{min}, b_{max}) > \phi\right)$ e $\xi_{max} = \sup\left(F(\beta, m, b_{min}, b_{max})\right)$

$\Rightarrow$ se $x \in [\xi_{min}; \xi_{max}]$ le stime del teo sono valide

se $x \notin [\xi_{min}; \xi_{max}]$ le stime del teo possono <u>non</u> valere



- Precisione di macchina: serve per limitare superiormente il valore assunto dall'errore relativo. Mi dice con che precisione dell'insieme M posso approssimare i numeri reali

→ $u = \frac{1}{2}\beta^{1-m}$

quindi

→ $|\delta(x)| \leq u|x|$ oppure $|\delta(x)| \leq u|rd(x)|$

→ $|\varepsilon(x)| \leq u$

→ $|\eta(x)| \leq u$

  - Tanto più la precisione di macchina è piccola quanto più è stringente la limitazione dell'errore relativo commesso arrotondando i numeri reali

Note
- Errore relativo: è limitato
  - Per ogni x reale non nullo, l'errore relativo commesso approssimando x con rd(x) non supera la precisione di macchina
- Errore assoluto: non è limitato
  - Questa differenza con l'errore assoluto è dovuta alla struttura di F(β,m) e rende naturale misurare l'errore commesso approssimando un numero reale con un numero in virgola mobile e precisione finita con una funzione errore relativo

- TEO: arrotondamento e perturbazioni
  - Perturbazione additiva di x

$\exists d \in \mathbb{R}$ t.c. $rd(x) = x + d$ $(|d| \leq u|x|)$ $\quad |d| = |\delta(x)|$

  - Perturbazione additiva di rd(x)

$\exists d \in \mathbb{R}$ t.c. $x = rd(x) + d$ $(|d| \leq u|x|)$

$\begin{cases} e = \varepsilon(x) & \text{per } x \neq 0 \\ e = 0 & \text{per } x = 0 \end{cases}$

  - Perturbazione moltiplicativa di x

$\exists e \in \mathbb{R}$ t.c. $rd(x) = (1+e)x$ $(|e| \leq u)$

  - Perturbazione moltiplicativa di rd(x)

$\exists \varepsilon \in \mathbb{R}$ t.c. $x = (1+\varepsilon)rd(x)$ $(|\varepsilon| \leq u)$

$\begin{cases} \varepsilon = 0 & \text{per } rd(x) = 0 \\ \varepsilon = \eta(x) & \text{per } rd(x) \neq 0 \end{cases}$

- Funzioni predefinite
  - Operazioni aritmetiche

→ $\xi_1 \otimes \xi_2 = rd(\xi_1 \cdot \xi_2)$ → $\xi_1 \oplus \xi_2 = rd(\xi_1 + \xi_2)$

→ $\xi_1 \oslash \xi_2 = rd\left(\frac{\xi_1}{\xi_2}\right)$ → $\xi_1 \ominus \xi_2 = rd(\xi_1 - \xi_2)$

  - Funzioni elementari(trigonometrica, esponenziale, log..)

funzione predefinita → $\overline{F(\xi)} = rd(f(\xi))$ ← funzione vera

  - Confronti

$<, >, \leq, \geq, =, \neq$ → $\begin{cases} T = 1 & \text{// true} \\ F = 0 & \text{// false} \end{cases}$

- Proprietà pseudosomma
  - → Commutativa : $\xi_1 \oplus \xi_2 = \xi_2 \oplus \xi_1$
  - → <u>NON</u> è associativa : $(\xi_1 \oplus \xi_2) \oplus \xi_3 \neq \xi_1 \oplus (\xi_2 \oplus \xi_3)$
  - → Debolmente monotona : $\xi_1 > \xi_2 \Rightarrow \forall \alpha \in M$ s.t. $\xi_1 \oplus \alpha \geq \xi_2 \oplus \alpha$
  - → lo zero non è unico : $\xi \oplus \alpha = \xi \longrightarrow \alpha = \phi$ ma se $\xi \neq \phi$ α può anche non essere $= \phi$
  - → l'opposto è unico : $\xi \oplus (-\xi) = \phi$
- Proprietà pseudoprodotto
  - → Commutativa : $\xi_1 \otimes \xi_2 = \xi_2 \otimes \xi_1$
  - → <u>NON</u> è associativa : $(\xi_1 \otimes \xi_2) \otimes \xi_3 \neq \xi_1 \otimes (\xi_2 \otimes \xi_3)$
  - → Debolmente monotona : $\xi_1 > \xi_2 \Rightarrow \forall \alpha \in M$ s.t. $\xi_1 \otimes \alpha \geq \xi_2 \otimes \alpha$
  - → l'unità non è unica : $\xi \otimes \alpha = \xi \rightarrow \alpha = 1$ ma se $\xi \neq \phi$ α può anche non essere $= 1$
  - → l'inverso può $\exists$ o non essere unico : $\sigma \in M$ t.c. $\xi \otimes \sigma = 1$
- funzione SEN
  - → ha un unico ϕ in $\xi = \phi$: $SEN(\xi) = \phi = rd(sen(x)) = \phi \Rightarrow \xi = \phi$
- teorema esistenza zeri
  - → non è detto che sia verificato : $f: M \to M \Rightarrow f(\xi) > \phi$, $f(\sigma) < \phi$ ma non è detto che $\exists \alpha$ t.c. $f(\alpha) = \phi$
- errore relativo per le funzioni predefinite: il valore assoluto dell'errore relativo commesso approssimando il risultato di un operazione aritmetica con il valore ottenuto dalla funzione predefinita non supera la precisione di macchina
  - → Errore commesso approssimando $f(\xi)$ con $rd(f(\xi))$ è: $|\varepsilon| \leq u$
- Procedura di trasformazione: procedimento per trasformare una procedura che usa il tipo *numero reale* con una procedura che usa il tipo numero in *virgola mobile e precisione finita* e informazioni sul relativo errore commesso
  A. Sostituire a ciascuna costante a valore in R il suo arrotondato in M
  B. Sostituire a ciascuna operazione aritmetica o funzione elementare la corrispondente funzione predefinita aggiungendo, se è il caso, opportune precedenze tra operatori
- Esempio: a volte è utile introdurre F(x1, x2)=x1/x2

$$f(x_1, x_2) = \frac{x_1}{x_2}$$

→ $f(x_1, x_2) = f(rd(x_1), rd(x_2)) = f((1+\varepsilon_1)x_1, (1+\varepsilon_2)x_2) = (1+\varepsilon) f(x_1, x_2)$ (PERCHÈ? $x_1, x_2 \notin M$)

$\Rightarrow f(x_1, x_2) = f((1+\varepsilon_1)x_1, (1+\varepsilon_2)x_2) = \frac{(1+\varepsilon_1)x_1}{(1+\varepsilon_2)x_2} = (1+\varepsilon) f(x_1, x_2) = (1+\varepsilon)\frac{x_1}{x_2}$ e trovo "ε"