→ RAPPRESENTAZIONE NORMALIZZATA IN VIRGOLA MOBILE

$$x = (-1)^s B^p \sum_{i=1}^{+\infty} d_i B^{-i} \longrightarrow x = \pm (0, d_1 d_2 d_3 d_4)_B B^p \quad d_1 \neq 0$$

dinon dep = a d-1

s	m	Valore del numero
[1, 254]	qualsiasi	$(-1)^s \cdot 2^{m-127} \cdot 1.m$ (numeri normali)
0	$\neq 0$	$(-1)^s \cdot 2^{-126} \cdot 0.m$ (numeri subnormali o denormalizzati)
0	0	$(-1)^s$ (zero con segno)
255	0	$(-1)^s \cdot \infty$ (infinito con segno)
255	$\neq 0$	NaN (not a number)

→ IEEE 754 $x = \pm 2^p \sum_{i=1}^{53} d_i 2^{-i}$ $d_1 = 1$

$p \in [-1021, 1024]$

$T = 53$

INSIEME NUMERI MACCHINA:

→ $F(B, T, m, M) = \{ x \in \mathbb{R} : x = \pm B^p \sum_{i=1}^{T} d_i B^{-i}$ $0 \leq d_i \leq B-1$, $d_1 \neq 0$, $p \in [-m, M] \}$

Base, ampiezza range — non è troppo restrittivo in macchina — $\forall$ IO4

→ dipende dalle celle di memoria (BITS)

Insieme N° di macchina, NUMERICO: $p = -1022$

numeri denormalizzati

$-\Omega \quad -\omega \quad 0 \quad \omega \quad \Omega$

$P = M \begin{cases} NAN \rightsquigarrow \text{Tutti a Tutti 0} \\ \pm\infty \end{cases}$

L'insieme è numerico — ogni posto può assumere $B-1$ valori

→ Cardinalità di F (insieme di numeri di macchina normalizzati) $= 2 \cdot (m + M + 1)(B-1) B^{T-1} + 1$ (numeri che può assumere $p(+1=0)$; 0.)

X appartiene ai numeri di macchina? $B^{m-1} \leq |x| \leq B^M (1 - B^{-T})$

ω — Ω

→ $x = B^p \sum_{i=1}^{T} d_i B^{-i}$ $\forall\, B^p \left(\sum_{i=1}^{T} d_i B^{-i} + B^{-T} \right)$

aggiungo un'unità all'ultima cifra disponibile → $\delta_{cons.} = B^{(p)} B^{-T}$

$-\Omega \quad -\omega \quad 0 \quad \omega \quad \Omega$ non a mono che all'altra + grandi

le cifre assumono tutti i valori $\neq 0$ — assumono tutte $= B-1$

→ $\Omega = B^M \cdot (B-1) \sum_{i=1}^{T} B^{-1} \longrightarrow = B^M (1 - B^{-T})$

$\omega = B^{-m} \cdot B^{-1}$

prima cifra uguale a 1 e tutte le altre uguali a 0.

$|x| > \Omega$ overflow $x \rightsquigarrow \pm \inf$ (ed x viene arrotondato $\pm$ inf.)

$|x| < \omega$ $x \neq 0$ underflow

$-\omega \quad 0 \quad \omega$

* TRONCAMENTO: Tronca le cifre alle prime T significative

ANALISI DELL'ERRORE

a) $x \in \mathbb{R} \rightarrow \tilde{x} \in F(B,T,m,M)$

$\tilde{x} = x(1+\epsilon_x)$ $\quad |\epsilon_x| \leq u \rightarrow$ precisione di macchina

2) $\tilde{x}, \tilde{y} \in F(B,T,m,M) \Rightarrow \begin{cases} \tilde{x}+\tilde{y} \\ \tilde{x}-\tilde{y} \\ \tilde{x}\cdot\tilde{y} \\ \tilde{x}/\tilde{y} \end{cases} \notin F(B,T,m,M)$

op. fatte bene non sono definite in macchina

- Tecniche di approssimazione:
- ARROTONDAMENTO: $\mathbb{R}$ numero x viene approssimato con $\mathbb{R}$ numero rappresentabile più vicino
- TRONCAMENTO: $\mathbb{R}$ numero x viene approssimato con $\mathbb{R}$ più grande numero rappresentabile $\leq \mathbb{R}$ con valore assoluto minore o uguale al valore assoluto di x
- ROUND TOWARD PLUS/MINUS INFINITY: x viene approssimato al più piccolo/grande numero appartenente maggiore/minore del dato

(*) ERRORE RELATIVO DI RAPPRESENTAZIONE: non dipende dal numero ma dall'aritmetica di macchina

$\rightarrow \epsilon_x = \frac{\tilde{x}-x}{x} \rightarrow \tilde{x} = x(1+\epsilon_x)$ $\quad |\epsilon_x| \leq u = \frac{B^{1-T}}{2} = u$ = PRECISIONE DI MACCHINA (non dipende da x)

$\tilde{x}$ numero di macchina, perciò rappresentabile

due numeri in numero di macchina

dim: $|\epsilon_x| = \left|\frac{\tilde{x}-x}{x}\right| \leq \left|\frac{\tilde{x}-\tilde{y}}{|x|}\right| \leq \frac{B^{p-T}}{|x|} \leq \frac{B^{p-T}}{B^{p-1}} \leq B^{1-T} = u$

u MODULO DELL'ERRORE DI RAPPRESENTAZIONE IN MACCHINA DEI NUMERI IN ENTRATA. $\rightarrow$ PRECISIONE DI MACCHINA

in IEE $u = B^{1-53} = 2^{-52} \rightarrow (2^{-53})$ perché ha num. $(ho \frac{1}{2} B^{p-T}$, lavoro x ARROTONDAMENTO)

↓ si LAVORO CON TRONCAMENTO

ho la stessa distanza



non è in nessun intervallo a sinistra

$\epsilon_x = \frac{\tilde{x}-x}{x} \Leftrightarrow \tilde{x} = x(1+\epsilon_x) \rightarrow |\epsilon_x| \leq u$ ERRORE DI RAPPRESENTAZIONE DEI NUMERI IN ENTRATA

(*) LE OPERAZIONI ARITMETICHE NON SONO INTERNE ALL'INSIEME DEI NUMERI DI MACCHINA

↓

(*) ERRORI LOCALI NELLE OPERAZIONI $\rightarrow \varepsilon_i$ $\quad |\varepsilon_i| \leq u$

1) x non è numero di macchina: rappresento $\tilde{x} = x(1+\epsilon_x)$

2) vado a calcolare la funzione, ho 2 casi diversi: $\rightarrow$ $q_1(\tilde{x}) = (\tilde{x} \ominus 1) \oslash \tilde{x}$

$q_2(\tilde{x}) = 1 \ominus (1 \oslash \tilde{x})$

che nell'implementazione in macchina non sono equivalenti

$\rightarrow$ in base a come organizzo il calcolo in macchina il risultato cambia

ANALISI ALL'INDIETRO degli ERRORI $\rightarrow$ trasformo tutti i termini rispetto ai x degli errori

$(\pm)$ = PRODOTTI DI ERRORI perché gli errori sono piccoli

✳ ERRORE INERENTE

ERRORE INEVITABILE la macchina non può rappresentare x ma $\tilde{x}$ ⟶ perturbazione dei dati in ingresso a cui si aggiunge l'errore di arrotondamento

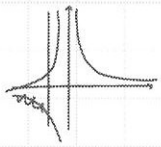
NON DIPENDE DAL MODO IN CUI CALCOLO LA FUNZIONE ⟶ è una caratteristica della funzione che sto calcolando

⟶ $\epsilon_m = \frac{f(\tilde{x}) - f(x)}{f(x)}$ ⟶ CONDIZIONAMENTO DEL PROBLEMA

(ben condizionato se non dipende dai dati in ingresso)

$$f(\tilde{x}) = \frac{\tilde{x}-1}{\tilde{x}} = \frac{x(1+\epsilon_x)-1}{x(1+\epsilon_x)} = \frac{[x(1+\epsilon_x)-1](1-\epsilon_x)}{x} = \frac{1-1}{x} + \frac{\epsilon_x}{x}$$

$f(x)$



calcolo mal condizionato quando $x \simeq 1$ (numeri vicini a 1)

$$\frac{f(\tilde{x}) - f(x)}{f(x)} = \frac{\epsilon_x}{x \cdot f(x)} = \frac{\epsilon_x}{x-1} \longrightarrow |\epsilon_m| = \left|\frac{\epsilon_x}{x-1}\right| \leq \frac{u}{|x-1|} \leftarrow \left(\frac{1}{|x-1|}\right) u$$

✳ ERRORE ALGORITMICO:

dipende da $q(\tilde{x})$ ⟶ DIPENDE DA COME CALCOLO LA FUNZIONE

Dipende dall'algoritmo di calcolo ⟶ dipende dopo ERRORI LOCALI DELLA FUNZIONE

⟶ $\epsilon_{alg} = \frac{q(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})}$ ⟶ STABILITÀ DELL'ALGORITMO

dati in ingresso : numeri di macchina

✳ ERRORE TOTALE:

$$\longrightarrow \epsilon_{TOT} = \frac{q(\tilde{x}) - f(x)}{f(x)} = \epsilon_m + \epsilon_{alg.}$$

quello che effettivamente calcolo in macchina

CALCOLO MEMANTE GRAFI

x funzioni razionali

$$\text{Infatti: } \frac{q(\tilde{x}) - f(\tilde{x}) + f(\tilde{x}) - f(x)}{f(x)} = \frac{q(\tilde{x}) - f(\tilde{x})}{f(x)} + \underbrace{\frac{f(\tilde{x}) - f(x)}{f(x)}}_{\epsilon_m} = \frac{q(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})} \cdot \frac{f(\tilde{x})}{f(x)} + \epsilon_m$$

$$= \epsilon_{alg.}(\epsilon_m + 1) + \epsilon_m \doteq \epsilon_{alg} + \epsilon_m$$

$$\epsilon_m = \frac{f(\tilde{x}) - f(x)}{f(x)} \qquad \epsilon_{m+1} = \frac{f(\tilde{x})}{f(x)}$$

COEFFICIENTE AMPLIFICAZIONE:

Amplifica l'Errore di Rappresentazione del dato iniziale sul risultato $\quad \epsilon_m = C_x \epsilon_x$

$\epsilon_m \leq C_x \cdot u$

$\longrightarrow C_x = \frac{f'(x)}{f(x)} \cdot x$

Thm:

$f: \mathbb{R} \to \mathbb{R} \qquad f \in C^2(\mathbb{R}) \longrightarrow f$ continua con derivate 1 e 2 continue $\to f$ di classe C^2

$f(\tilde{x}) = f(x) + f'(x)(\tilde{x}-x) + \frac{f''(\xi)\cdot(\tilde{x}-x)^2}{2}$ Sviluppo di TAYLOR di f con centro x e raggio $\tilde{x}-x$ al II ordine

ξ è un numero compreso tra $\tilde{x}$ e x

$\frac{f(\tilde{x})-f(x)}{f(x)} = \frac{f'(x)}{f(x)} \cdot \frac{\tilde{x}-x}{x} \cdot x + \frac{f''(\xi)}{f(x)} \left(\frac{\tilde{x}-x}{x}\right)^2 x^2$

ϵ_x (Errore di Rappresentazione) $\qquad \epsilon_x^2$ $\times$ Analisi al I ordine

$\epsilon_m = \frac{f(\tilde{x})-f(x)}{f(x)} \doteq \frac{f'(x)}{f(x)} x \cdot \epsilon_x$

C_x

generalizzazione formula x Funzioni REALI di più variabili reali:

$$\longrightarrow \epsilon_m = \frac{1}{f(x)} \left[\frac{\partial f}{\partial x_1}(x) \cdot x_1 \epsilon_1 + \frac{\partial f}{\partial x_2}(x)\, x_2 \epsilon_2 + \ldots + \frac{\partial f(x)}{\partial x_n}(x) \times m f_n \right]$$

$\partial f(x)$ = derivata parziale di f fatta rispetto ad x

$$f: \mathbb{R}^n \to \mathbb{R}^m = \begin{pmatrix} f_1(x_1, \ldots x_n) \\ \vdots \\ f_m(x_1, \ldots x_m) \end{pmatrix}$$

$$\begin{aligned}
&f(x,y) = x + y, \quad \epsilon_{in} \doteq c_x\epsilon_x + c_y\epsilon_y, \quad c_x = \frac{x}{x+y},\ c_y = \frac{y}{x+y};\\
&f(x,y) = x - y, \quad \epsilon_{in} \doteq c_x\epsilon_x + c_y\epsilon_y, \quad c_x = \frac{x}{x-y},\ c_y = -\frac{y}{x-y};\\
&f(x,y) = x \cdot y, \quad \epsilon_{in} \doteq c_x\epsilon_x + c_y\epsilon_y, \quad c_x = 1,\ c_y = 1;\\
&f(x,y) = x/y, \quad \epsilon_{in} \doteq c_x\epsilon_x + c_y\epsilon_y, \quad c_x = 1,\ c_y = -1.
\end{aligned}$$

$\longrightarrow \tilde{x} = \tilde{y}$ PROBLEMA DI CANCELLAZIONE NUMERICA gli errori di Rappresentazione non vanno ad aumentare anche molto più

$f: \mathbb{R} \to \mathbb{R}$

$x \rightsquigarrow f(x)$ commettiamo una serie di errori

E.m = Errore nella rappresentazione in macchina

$\varepsilon_x = \frac{\tilde{x}-x}{x}$ Errore di rappresentazione $\longrightarrow$ Analizzo l'impatto dell'errore di rappresentazione del dato iniziale sul risultato — X CONDIZIONAMENTO

$\varepsilon_m = \frac{f(\tilde{x})-f(x)}{f(x)} \doteq \frac{f'(x)}{f(x)} \cdot x \, \varepsilon_x$

ε_x amplificato

Quando implemento in macchina il calcolo della funzione f

implemento una funzione che è composizione delle corrispondenti operazioni di macchina

↓ compio un ε Alg. $\longrightarrow$ la f una f razionale: $\varepsilon_{tot} = \left(\frac{g(\tilde{x})-f(x)}{f(x)}\right) \doteq \varepsilon_m + \varepsilon_{Alg.}$

$\varepsilon_{Alg.} = \frac{g(\tilde{x})-f(\tilde{x})}{f(\tilde{x})}$ dati in ingresso: numeri di macchina

Alg. $\begin{cases} r = x+y \\ \downarrow \\ \tilde{r} = \tilde{x} \oplus \tilde{y} \end{cases}$ in macchina ho: $\tilde{x} = x(1+\varepsilon)$ ε = tutto l'errore che mi è accumulato nel calcolo del dato x

$\tilde{y} = y(1+\delta)$ δ = tutto l'errore che mi è accumulato nel calcolo del dato y

$\tilde{r} = \tilde{x} + \tilde{y}$

↳ valore che calcolo in macchina

$\tilde{r} = \tilde{x} \oplus \tilde{y} = (\tilde{x}+\tilde{y})(1+\varepsilon_1) \longrightarrow |\varepsilon_1| \le u$ errore locale dell'operazione

(precisione di macchina)

$= (x+\varepsilon x + y + \delta y)(1+\varepsilon_1)$

$= x + x\varepsilon_1 + x\varepsilon + x\varepsilon\varepsilon_1 + y + y\varepsilon_1 + \delta y + \delta y\varepsilon_1$

$= x(1+\varepsilon)+y(1+\delta)](1+\varepsilon_1)$

$\doteq x(1+\varepsilon+\varepsilon_1)+y(1+\delta+\varepsilon_1)$

$= \underbrace{(x+y)}_{r} + (x+y)\varepsilon_1 + x\varepsilon + y\delta$

$\frac{\tilde{r}-r}{r} = \varepsilon_1 + \frac{x}{x+y}\varepsilon + \frac{y}{x+y}\delta \doteq \mu$ (circa) $\longrightarrow$ Tutto l'errore che mi è accumulato nel calcolo di r $\tilde{r} = r(1+\mu)$

errore locale operazione

errori locali dagli operandi $\tilde{x}$ e $\tilde{y}$ → nel risultato sono amplificati dai coef. di amplif.

Esempio: coepp. di amplificazione errori corrispondente