Appunti per il corso di Laboratorio di Programmazione e Calcolo

I Numeri Macchina*

E.Carlini

*Segnalazioni di errori o refusi sono ben accette

NUMERI MACCHINA : RAPPRESENTAZIONE dei NUMERI al CALCOLATORE

Rif Cap 2 quarteroni

RICHIAMI SUI $\mathbb{R}$

NUMERI RAZIONALI $\mathbb{Q}$

- ottenibili come rapporto di interi
- hanno parte decimale finita o periodica

Es 2 ; $3,5$ $(7/2)$; $0.\bar{6}$ $(2/3)$

NUMERI IRRAZIONALI $\mathbb{R} \setminus \mathbb{Q}$

- non ottenibili come rapp di interi
- hanno parte decimale non finita e non periodica

Es $\sqrt{2}, \sqrt{3}, \pi, e, \dots$

DENSITÀ dei RAZIONALI in $\mathbb{R}$

Ogni numero reale può essere rappresentato con precisione arbitraria con un numero razionale

(Dati due numeri reali a, b esiste sempre un razionale r compreso tra a e b)

$\Rightarrow$ TUTTI i REALI si possono rappresentare bene con numeri aventi una rappresentazione finita

<u>CONTINUITÀ di $\mathbb{R}$</u> : I numeri reali possono essere messi in corrispondenza biunivoca con i numeri di una retta reale

(la retta reale non ha buchi)

7/

SISTEMA POSIZIONALE (RAPPRESENTAZIONE dei reali)

Fissata una base $\beta \in \mathbb{N}$, $\beta \geq 2$, e sia $x \in \mathbb{R}$, $x \neq 0$ allora

(*) $x = \pm\, x_n x_{n-1} \dots x_1 x_0 , x_{-1} x_{-2} \dots x_{-m} \dots$ con $x_n \neq 0$

detta RAPP. POSIZIONALE

(*) sta a significare $x = x_n \beta^n + x_{n-1} \beta^{n-1} + \dots + x_1 \beta + x_0 + x_{-1} \beta^{-1} + \dots \; x_{-n} \beta^{-n} + \dots$

ES $102{,}56$ $\beta = 10$

$x = 1 \cdot 10^2 + 0 \cdot 10^1 + 2 \cdot 10^0 + 5 \cdot 10^{-1} + 6 \cdot 10^{-2}$

Analogamente $\beta = 2$ $x = 101{,}11$

$x = 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 1 \cdot 2^{-2}$

Limitatezza del calcolatore $\Rightarrow$ solo un n. finito di numeri reali possono essere rappresentati

NUMERI MACCHINA sono i numeri reali che il calcolatore può rappresentare esattamente e li denotiamo con $\mathbb{F}$

La cardinalità di $\mathbb{F}$ è finita

...alogamente $\beta = 2$, $x = 101,11$

$x = 1 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 1 \cdot 2^{-2}$

OSS. Un numero (razionale) può avere una rappresentazione finita in una base e infinita in un'altra

ES $x = 1/3$ $\beta = 10$ infinito
$\beta = 3$ finito

Da un punto di vista astratto, le basi sono EQUIVALENTI

Nella pratica, $\beta = \begin{cases} 2 & \text{binaria} \\ 10 & \text{decimale} \\ 16 & \text{esadecimale} \end{cases}$

Quasi tutti i calcolatori moderni usano $\beta = 2$

BIT (BINARY DIGIT) unità elementare d'informazione
può assumere valore 0 o 1

STATO ACCESO / SPENTO
1 0

Supponiamo di avere N posizioni di memoria

$\underbrace{\;}_{1\text{ segno}}\underbrace{\quad\quad}_{N-k-1 \text{ per le cifre intere}}\underbrace{\quad\quad}_{k \text{ cifre dopo la virgola}}$ (1) $x = (-1)^s a_{N-1} \dots a_{k+1} a_k \dots a_1$

SISTEMA a VIRGOLA FISSA
limita molto la ~~possibilità~~ il valore del n° massimo e minimo ed è poco conveniente

ES $x = 0,0001234$ $k = 4$ ⟹ faccio ragionare la scalatura che compare in 1
$\hat{x} = 0,0001$ perdo 3 cifre

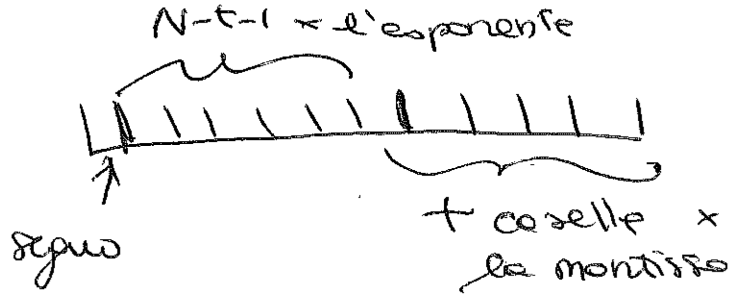
RAPPRESENTAZIONE in VIRGOLA MOBILE (o FLOATING-POINT)

$$x = \pm 0,a_1 \dots a_t \, \beta^{e} = \pm a_1 \dots a_t \, \beta^{e-t}$$

Un numero in virgola mobile è caratterizzato da

- $t \in \mathbb{N}$ numero di cifre significative
- $0 \le a_i \le \beta - 1$ CIFRE SIGNIFICATIVE
- $m = a_1 \dots a_t$ MANTISSA
- e esponente, tipica $L \le e \le U$ $\quad U > 0 \quad L < 0$

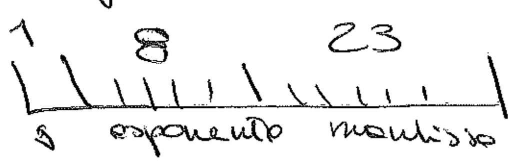
numeri piccoli $e < 0$
grandi $e > 0$



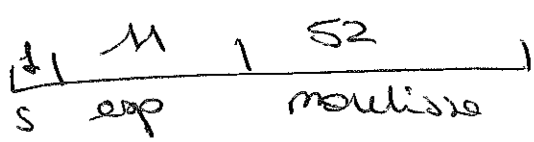
Lo standard più diffuso per definire il formato è lo standard IEEE (Institute of Electrical and Electronic Engineers (associazione professionale))
standard definito nel 198

due formati

SINGOLA PRECISIONE $N = 32$ bits



DOPPIA PRECISIONE $N = 64$ bits



DEF Si definisce insieme dei NUMERI MACCHINA (o numeri FLOATING POINT) con t cifre significative, base β ed esponente compreso tra L e U, l'insieme

$$\mathbb{F}(\beta, t, L, U) = \{x \in \mathbb{R}, x \neq 0, x = \pm 0{,}a_1 \ldots a_t \beta^e\} \cup \{0, Inf, NaN\}$$

dove

1. $t \geq 0$, $\beta \geq 1$
2. $0 \leq a_i \leq \beta - 1$
3. $a_1 \neq 0$ (garantisce unicità rappresentazione)
4. $L \leq e \leq U$ $\quad L < 0, U > 0$

OSS

0 ha una rappresentazione a parte

Inf sta per infinito

NaN " " Not a Number

$Inf - Inf$, Inf / Inf, $0/0$, $0 \times Inf$, 1^{Inf}

~~A~~LCUNE MAGGIORAZIONI

La mantissa $m = a_1 \dots a_t$ varia con $0 \le a_i \le \beta - 1$, $a_1 \neq 0$

varia tra $\beta^{t-1} \le m \le \beta^t - 1$

Es $\beta = 10$ $t = 3$ $\underbrace{100}_{\beta^2 = 10^2} \le m \le \underbrace{999}_{\beta^t - 1,\ \beta^3 = 1000}$

Se $x \in \mathbb{F}(\beta, t, L, U)$ allora

$$L \le e \le U$$

$$|x| = m\beta^{e-t} \ge \beta^{t-1}\beta^{L-t} = \beta^{L-\cancel{t}+\cancel{t}-1} = \beta^{L-1} := x_{min}$$

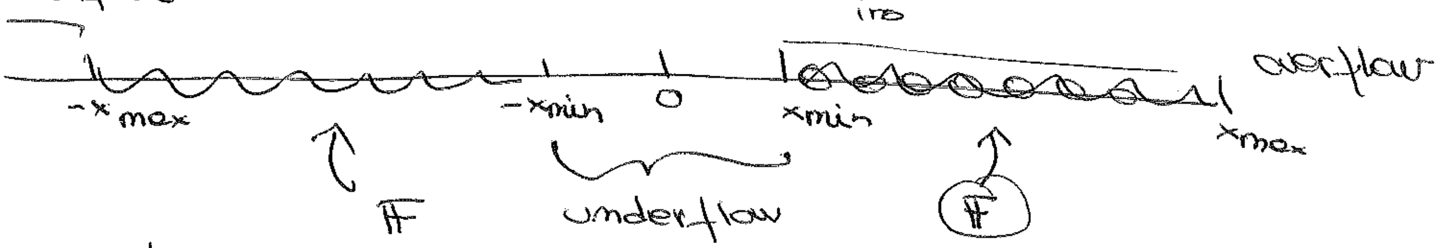
$$|x| = m\beta^{e-t} \le (\beta^t - 1)\beta^{U-t} = \beta^{U-\cancel{t}+\cancel{t}} - \beta^{U-t} = \beta^U(1-\beta^{-t}) := x_{max}$$

ESEMPIO Standard PRECISIONE SINGOLA $t = 23$

$U = 127$ $L = -128$

$$x_{min} \simeq (2)^{-128} \simeq 10^{-38}$$

$$x_{max} \simeq \beta^U = (2)^{127} \simeq 10^{38}$$



overflow: i numeri in modulo maggiore di x_{max}

underflow: " " minore di x_{min}

Quali sono i n. che non appartengono a $\mathbb{F}(\beta, t, L, U)$

- tutti i numeri irrazionali (decimali illimitati non periodici) $\sqrt{2}, \pi$
- i numeri (in base β) decimali finiti e infiniti con più di t cifre significative x la mantissa
- i numeri di modulo maggiore di x_{max}
  " " minore di x_{min}

In altre parole $\mathbb{F}$ è pieno di buchi e tutti i numeri che non sono in $\mathbb{F}$ vanno approssimati

Quali sono i numeri che possono essere rappresentati in maniera esatta

- I numeri macchina sono un sottoinsieme dei numeri razionali e si scrivono $\frac{m}{2^p}$ $m \in \mathbb{N}, p \in \mathbb{N}$

Come si rappresenta $x \notin \mathbb{F}$?

Definiamo l'applicazione

$$fl: \mathbb{R} \to \mathbb{F}$$
$$x \to \underbrace{fl(x)}_{\text{rappresentazione di } x \in \mathbb{R} \text{ in } \mathbb{F}}$$

nel seguente modo

$x = \pm 0, a_1 \ldots a_p \beta^e$ con $p \leq t$ ed $e \in [L, U]$

allora $fl(x) = x$

x è rappresentato esattamente come un numero macchina

[Es. $\beta = 10$, $t = 3$, $L = -5, U = 5$

$x = 0,0123$ $\quad fl(x) = 0.123 \cdot 10^{-1}$ $\quad x = fl(x)$]

2.) se x t. che $|x| > x_{max} \Rightarrow fl(x) = \pm Inf$

3) se x t. che $|x| < x_{min} \Rightarrow fl(x) = 0$

4) se $x = \pm 0, a_1 \ldots a_t a_{t+1} \ldots \beta^e$ $\quad e \in [L, U]$

allora si definisce

A) APPROSSIMAZIONE PER ARROTONDAMENTO $fl(x) = \pm 0, a_1 \ldots a_{t-1} \tilde{a}_t \beta^e$ dove $\tilde{a}_t = \begin{cases} a_t & \text{se } a_{t+1} < \frac{\beta}{2} \\ a_t + 1 & \text{se } a_{t+1} \geq \frac{\beta}{2} \end{cases}$

B) APP. PER TRONCAMENTO $fl(x) = \pm 0, a_1 \ldots a_t \beta^e$

Es $x = 0,1234\,56489$ $\quad t = 3$

$fl(x) = 0,123$ arr e tron coincidono

$x = 0,123\,6789$ $\quad t = 3$

$fl(x) = 0,124$ arr

$fl(x) = 0,123$ tr

OPERAZIONI di MACCHINA effettuate in $\mathbb{F}$

~~DEF~~ Chiamiamo OPERAZIONE MACCHINA, il risultato dell'operazione eseguita sui numeri macchina seguita da un'approssimazione in $\mathbb{F}$:

DEF ~~Definiamo~~ Indichiamo con $\oplus, \ominus, \oslash, \otimes$ operazioni macchina corrispondenti a $+, -, /, \times$, ~~definite~~ come

$$a \odot b := fl(fl(a) \cdot fl(b)) \qquad \cdot = +, -, /, \times$$

Per stimare un'approssimazione si usano le seguenti misure d'errore.

DEF Sia $x \in \mathbb{R}$ e $\tilde{x}$ una sua approssimazione. Si definiscono

ERRORE ASSOLUTO $e_a(x) := |x - \tilde{x}|$

ed ERRORE RELATIVO $e_r(x) := \dfrac{|x-\tilde{x}|}{|x|}$

PRECISIONE MACCHINA : costante di ogni aritmetica floating point

DEF Definiamo precisione macchina ε_M il più piccolo numero macchina t.c.

$$1 \oplus \varepsilon_M > 1$$

PROPRIETÀ Si può dimostrare che

1) $\mathbb{F}(\beta, t, L, U)$, $\varepsilon_M = \begin{cases} \frac{1}{2}\beta^{1-t} & \text{(ARROTONDAMENTO)} \\ \beta^{1-t} & \text{(TRONCAMENTO)} \end{cases}$

2) se $x \in \mathbb{R}$, $x \in [-x_{max}, -x_{min}] \cup [x_{min}, x_{max}]$ (x è nel range di $\mathbb{F}$) allora

ERRORE ROUNDOFF

$$\frac{|x - fl(x)|}{|x|} \leq \varepsilon_M$$

ε_M rappresenta la massima precisione relativa di calcolo raggiungibile sul calcolatore

...empi $\beta = 10$ $t = 3$ $\epsilon_M =$ 0.01 tronc. / $0.5 \cdot 10^{-2}$

caso troncamento $1 \oplus 0.01 = \underbrace{1.01}$ ok

$1 \oplus .009 = 1.00(9) = \underline{1}$

caso arr. $1 \oplus 0.5 \cdot 10^{-2} = 1,00(5) = \cancel{1,00} \ 1,01$ ok

$1 \oplus 0.4 \cdot 10^{-2} = 1,00(4) = \underline{1,00}$

~~ESEMPI~~
OSSERVAZIONI su l'ERRORE ASS. e RELATIVO

Es $x = 1000$ $\tilde{x} = 1000,5$

$y = 0,01$ $\tilde{y} = 0,51$

$E_a^x = |x - \tilde{x}| = 0.5$ $\quad E_r = \frac{|x-\tilde{x}|}{|x|} = \frac{0.5}{10^3} = 5 \cdot 10^{-4}$

$E_a^y = |y - \tilde{y}| = 0.5$

$E_r = \frac{|y-\tilde{y}|}{|y|} = \frac{0.5}{10^{-2}} = 0.5 \cdot 10^2 = 50$!!

Occhio alle differenze

$x = 1,23451 \cdot 10^{-6}$ $\quad e_a = |x - \tilde{x}| = 0,1 \cdot 10^{-5} \cdot 10^{-6} = 10^{-11}$

$\tilde{x} = 1,2345 \cdot 10^{-6}$ $\quad e_r = \frac{|x-\tilde{x}|}{|x|} \simeq \frac{10^{-11}}{10^{-6}} = 10^{-5}$

$x = 123451$ $\quad e_a = |x - \tilde{x}| = 1 >> 10^{-11}$

$\tilde{x} = 123450$ $\quad e_r = \frac{|x-\tilde{x}|}{|x|} \simeq \frac{1}{10^5} = 10^{-5}$

"Due numeri reali con prime k cifre corrette differiscono per un errore relativo dell'ordine di 10^{-k}"

Attenzione che il viceversa non è vero:

Es $x = 1000$ $\quad \frac{|x-y|}{|x|} = \frac{1}{10^3} = 10^{-3} \Rightarrow$ ma non hanno

$y = 999$ 3 cifre uguali!

Quindi dalla proprietà 2 della precisione macchina

PER ARROTONDAM. → $|x-fl(x)|/|x| \leq \epsilon_M$

osserviamo

PR. SINGOLA $t=23$ $\beta=2$

$$\epsilon_M \simeq 2^{-22} \simeq 2.38 \cdot 10^{-7}$$

Se un'approssimazione di x avesse 10 cifre corrette allora $\frac{|x-\tilde{x}|}{|x|} < 10^{-10}$

ma in questo ↑ non è garantito in precisione singola

Nell'app. con precisione singola posso ottenere al + 7 cifre significative

Attenzione / che / se / errore < 10^{-16}

In PRECISIONE DOPPIA $t=52$ $\epsilon_M \simeq 2^{-51} \simeq 10^{-16}$

16 cifre significative al più corrette

oss

- ϵ_M rappresenta la massima precisione relativa di calcolo raggiungibile sul calcolatore
- non ha senso determinare approssimazioni con precisione relativa inferiore a ϵ_M
- ϵ_M dipende solo da β e t e rappresenta la distanza tra 1 ed il succ numero macchina
- x_{min} dipende da β e da L ed è il + piccolo numero macchina positivo

ALGEBRA dei NUMERI MACCHINA

L'insieme $\mathbb{F}$ non è chiuso rispetto alle operazioni elementari

$$x, y \in \mathbb{F} \text{ non implica } \begin{matrix} x+y \\ x-y \\ x*y \\ x/y \end{matrix} \in \mathbb{F}$$

Es $x = 0,62378\ 10^7$

$y = 0,32881 \cdot 10^2$

$$x \oplus y = fl(10^7 (0,62378 + 0,32881 \cdot 10^{-5})) =$$
$$= fl(10^7 (0,62379 + 0,0000032881)) =$$
$$= fl(10^7 (0,6237932881)) =$$
$$= 10^7 (0,62379) = x$$

- L'elemento neutro della somma non è UNICO

$$x \oplus y = x \quad \text{non implica } y = 0$$

Es $x = \epsilon_M/2$

$$(1 \oplus x) \oplus x = 1 \oplus x = 1$$
$$1 \oplus (x \oplus x) = 1 \oplus 2x = 1 \oplus \epsilon_M > 1$$

- La PROP. ASSOCIATIVA NON VALE

In generale la pr. commutiva si conserva, ma non valgono più le seguenti proprietà:

$$a \oplus (b \oplus c) \neq (a \oplus b) \oplus c$$
$$a \otimes (b \otimes c) \neq (a \otimes b) \otimes c$$
$$(a \otimes b) \oslash b \neq a$$
$$(a \oslash b) \otimes b \neq a$$

CARDINALITÀ di $\mathbb{F}$

N bits $\quad$ a_1 fisso

$\beta = 2$

$$\underbrace{1}_{2}\ \underbrace{N-t-1}_{2^{N-t-1}}\ \underbrace{t}_{2^{t-1}}$$

$$\text{Card } \mathbb{F} = 2^{N-1} + \underbrace{3}_{0,\ Inf,\ NaN}$$

Come viene memorizzato il segno dell'esponente?
Supponiamo di avere 8 bits per l'esponente

$$2^8 = 256$$ combinazioni di cifre 1 e 0 ci permette di rappresentare 256 interi

Quindi $U = -128 \leq e \leq 127 = L$

Per risparmiare la doppia rappresentazione dello zero

$E = e + k$, $k = 129$ (detto BIAS)

$1 \leq E \leq 256 \Rightarrow e = E - k$

Se usassi $\boxed{\pm}\underbrace{\quad\quad\quad}_{2^7 = 128}$ $\quad 0 \leq e \leq 127$

Poi uso il bit del segno e $-127 \geq e \geq 0$, ma lo zero avrebbe doppia rappresentazione.

Con il bias recupero -128

CANCELLAZIONE NUMERICA

Si dice CANCELLAZIONE NUMERICA il fenomeno di perdita di cifre significative che si verifica quando si esegue una sottrazione tra due numeri macchina "quasi uguali" tra loro

→ ENORME AMPLIFICAZIONE degli ERRORI

<u>ES</u> $x_1 = 0,19101972\ 10^3$ $x_2 = 0,19101708\ 10^3$

$x_1 \ominus x_2$ $\beta = 10$ $t = 6$ con troncamento

$(\epsilon_M = 10^{-5})$

$fl(x_1) = 0.191019 \cdot 10^3$ $fl(x_2) = 0.191017 \cdot 10^3$

$e_a(x_1) = |x_1 - fl(x_1)| = 0.72 \cdot 10^{-6} \cdot 10^3 = 0.72\ 10^{-3}$

$e_r(x_1) = \frac{|x_1 - fl(x_1)|}{|x_1|} = \frac{0.72\ 10^{-3}}{0.19101972\ 10^3} \simeq 0.37692 \cdot 10^{-5} < \epsilon_M$

$e_a(x_2) = |x_2 - fl(x_2)| = \underline{0.8 \cdot 10^{-7}}\ 10^3 = 0.8\ 10^{-4}$

$e_r(x_2) = \frac{|x_2 - fl(x_2)|}{|x_2|} = \frac{0.8\ 10^{-4}}{0.19101708\ 10^3} \simeq 0.41881\ 10^{-6} < \epsilon_M$

come ci aspettavamo dal teorema sull'errore round-off

Calcoliamo $\ominus$

$fl(x_1) \ominus fl(x_2) = 0.2 \cdot 10^{-5} \cdot 10^3 = 0.2\ 10^{-2}$

Errore commesso? Differenza esatta $x_1 - x_2 = 0.264000\ 10^{-2}$

Errore relativo ↓ osserva che l'errore assoluto è dello stesso ordine degli errori assoluti $e_a(x_1)$, $e_a(x_2)$

$\frac{|fl(x_1) \ominus fl(x_2) - (x_1 - x_2)|}{|x_1 - x_2|} = \frac{|0.2\ 10^{-2} - 0.264\ 10^{-2}|}{|0.264\ 10^{-2}|} = \frac{|0.064\ 10^{-2}|}{|0.264\ 10^{-2}|}$

$= \frac{0.064}{0.264} = 0.2424$

$\Rightarrow$ perdita di cifre significative nella differenza

$fl(x_1) - fl(x_2) = 0.000002 \cdot 10^{-3}$

approssimando le mantisse di x_1 e x_2 si sono "buttate via" le cifre che seguono la sesta (con un errore di approssimazione ~~entro~~ i limiti della precisione di macchina). $fl(x_1) = 0.191019$ $fl(x_2) = 0.191017$

La differenza effettuata tra x_1 e x_2 ha però amplificato molto la perdita di informazione dovuta all'approssimazione fino a farla salire alla prima cifra significativa del risultato!

Con arrotondamento! $\quad e_r(x \ominus y) = 0.1363$

La sottrazione di macchina non introduce alcun errore di approssimazione e fornisce il risultato esatto tra i numeri macchina, ma AMPLIFICA errori di approssimazione già esistenti.

OSS È l'errore relativo ad essere ~~amplificato~~, STATO PROPAGATO!

$$e_a(x \ominus y) = 0.064 \cdot 10^{-2} = 0.6 \cdot 10^{-3} \simeq \frac{1}{2} 10^{-3}$$

$e_a(x_1) = 0.72 \cdot 10^{-3}$ $\quad (0.72 - 0.08) \cdot 10^{-3} = 0.64 \cdot 10^{-3}$

$e_a(x_2) = 0.8 \cdot 10^{-4}$ $\quad \Big] \longrightarrow e_a(x_1) - e_a(x_2) = 0.64 \cdot 10^{-3}$

$\simeq e_a(x_1 \ominus x_2)$

PROPAGAZIONE degli ERRORI nelle 4 OPERAZIONI

Dati $x, y \in \mathbb{R}$, siano $\tilde{x}, \tilde{y}$ la loro rappresentazione in aritmetica floating point

$$\tilde{x} := fl(x) \quad , \tilde{y} = fl(y)$$

e sia $\varepsilon_x := x - \tilde{x}$, $\varepsilon_y = y - \tilde{y}$.

Da $\frac{|x - fl(x)|}{|x|} \le \epsilon_\pi$ sappiamo che $|\varepsilon_x| \le \epsilon_\pi |x|$
e $|\varepsilon_y| \le \epsilon_\pi |y|$.

Supponiamo di ~~da~~ conoscere τ quantità piccola e positiva t. che

$$|\varepsilon_x| \le \tau \qquad e \qquad |\varepsilon_y| \le \tau$$

SOMMA $\quad x+y = \tilde{x} + \varepsilon_x + \tilde{y} + \varepsilon_y = \tilde{x} + \tilde{y} + \varepsilon_x + \varepsilon_y$

SOTTRAZIONE $\quad x-y = \tilde{x} + \varepsilon_x - (\tilde{y} + \varepsilon_y) = \tilde{x} - \tilde{y} + \varepsilon_x - \varepsilon_y$

PRODOTTO $\quad xy = (\tilde{x} + \varepsilon_x)(\tilde{y} + \varepsilon_y) = \tilde{x}\tilde{y} + \tilde{x}\varepsilon_y + \tilde{y}\varepsilon_x + \varepsilon_y\varepsilon_x$

DIVISIONE
$$\frac{x}{y} = \frac{\tilde{x} + \varepsilon_x}{\tilde{y} + \varepsilon_y} = \frac{\tilde{x}}{\tilde{y}} + \frac{\tilde{x} + \varepsilon_x}{\tilde{y} + \varepsilon_y} - \frac{\tilde{x}}{\tilde{y}} = \frac{\tilde{x}}{\tilde{y}} + \frac{(\tilde{x} + \varepsilon_x)\tilde{y} - \tilde{x}(\tilde{y} + \varepsilon_y)}{(\tilde{y} + \varepsilon_y)\tilde{y}}$$
$$= \frac{\tilde{x}}{\tilde{y}} + \frac{\varepsilon_x \tilde{y} - \varepsilon_y \tilde{x}}{y\tilde{y}}$$

CONCLUSIONI

$$\mathcal{E}_a(x+y) = |x+y - (\tilde{x} + \tilde{y})| = |\varepsilon_x + \varepsilon_y| \le |\varepsilon_x| + |\varepsilon_y| \le 2\tau$$
$$\mathcal{E}_a(x-y) = |x-y - (\tilde{x} - \tilde{y})| = |\varepsilon_x - \varepsilon_y| \le |\varepsilon_x| + |\varepsilon_y| \le 2\tau$$
$$\mathcal{E}_a(xy) \le |xy - \tilde{x}\tilde{y}| = |\tilde{x}\varepsilon_y + \tilde{y}\varepsilon_x + \varepsilon_y\varepsilon_x| \le \tau(|\tilde{x}| + |\tilde{y}| + \tau)$$
$$\mathcal{E}_a(x/y) \le \left|\frac{x}{y} - \frac{\tilde{x}}{\tilde{y}}\right| = \left|\frac{\varepsilon_x \tilde{y} - \varepsilon_y \tilde{x}}{y\tilde{y}}\right| \le \tau \frac{|\tilde{x}| + |\tilde{y}|}{|y\tilde{y}|}$$

L'errore ASSOLUTO non subisce effetti di PROPAGAZIONE dall'operazioni $\oplus, \ominus, \odot, \oslash$

ERRORE RELATIVO

SOMMA $\quad \varepsilon_r(x+y) = \dfrac{\varepsilon_a(x+y)}{|x+y|} = \dfrac{|\varepsilon_x + \varepsilon_y|}{|x+y|}$

SOTTRAZIONE $\quad \varepsilon_r(x-y) = \dfrac{\varepsilon_a(x-y)}{|x-y|} = \dfrac{|\varepsilon_x - \varepsilon_y|}{|x-y|}$

PRODOTTO $\quad \varepsilon_r(xy) = \dfrac{\varepsilon_a(xy)}{|xy|} = \dfrac{|\varepsilon_y \tilde{x} + \varepsilon_x \tilde{y} + \varepsilon_x \varepsilon_y|}{|xy|}$

Conclusioni

Somma $\quad \varepsilon_r(x+y) \leq \dfrac{2\tau}{|x+y|} \quad \Leftarrow$ GRANDE se $|x+y| \simeq 0$

SOTTRAZIONE $\quad \varepsilon_r(x-y) \leq \dfrac{2\tau}{|x-y|} \quad \Leftarrow$ GRANDE se $|x-y| \simeq 0$

PRODOTTO

$$\underset{\substack{\uparrow \\ \text{Supponendo } \frac{|\tilde{x}|}{|x|} \leq 1 \text{ e } \left|\frac{\tilde{y}}{y}\right| \leq 1}}{\varepsilon_r(xy) \leq} \varepsilon_r(x) + \varepsilon_r(y) + \varepsilon_r(y)\varepsilon_r(x)$$

QUOZIENTE $\quad \varepsilon_r(x/y) \leq \dfrac{\varepsilon_r(x) + \varepsilon_r(y)}{1+\varepsilon_r(y)}$

Operazioni di prodotto e quoziente si comportano bene per quanto riguarda la propagazione dell' ERR. RELATIVO, infatti in entrambi le operazioni l' errore relativo sul risultato è al massimo l' errore relativo sui dati.