Rappresentazione dei numeri (base β)

Dato un valore $x \in \mathbb{R} \setminus \{0\}$ esso può essere rappresentato come (è la notazione esponenziale):

$$x = \text{sign}(x)\, \beta^b \underbrace{\sum_{i=1}^{\infty} d_i \beta^{-i}}_{\text{MANTISSA}} \qquad \text{sign}(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$$

con $\underbrace{\beta > 1, \beta \in \mathbb{N}}_{\text{BASE}}$ e $b \in \mathbb{Z}$, con:

$$0 \leq d_i \leq \beta - 1 \qquad d_i \in \mathbb{N}_0$$

(d_i: CIFRE DELLA RAPPRESENTAZIONE)

MASTER COPY

Tel. 388/9837745

Prendiamo per esempio il numero $87{,}22 = 10^2 \cdot 0{,}8722 = 10^3 \cdot 0{,}08722$

$b = 2$ $\quad \beta = 10$

$d_1 = 8 \quad d_2 = 7 \quad d_3 = d_4 = 2$ $\quad \Rightarrow \quad x = 10^2 \cdot (8 \cdot 10^{-1} + 7 \cdot 10^{-2} + \dots)$

Per avere un'unicità della rappresentazione:

- $d_1 \neq 0$ (primo termine decimale non nullo)
- $\nexists\, k \in \mathbb{N}$ tale che $d_i = \beta - 1 \ \forall i > k$ ovvero si escludono numeri periodici con PERIODO UGUALE ALLA BASE -1.

Prendiamo $1{,}1\overline{3} = \dfrac{113 - 11}{90} = \dfrac{102}{90} = 1{,}2$ Vediamo che sono uguali, ecco perché si esclude!

(tante quante cifre periodo; tante quante le cifre antiperiodo)

Per quanto riguarda la mantissa (per questo si dice NORMALIZZATA) vediamo che può assumere:

$$\beta^{-1} \leq \sum d_i \beta^{-i} < 1$$

$d_1 = 1$ è il minimo

$$\sum_{i=1}^{\infty} d_i \beta^{-i} \leq \sum_{i=1}^{\infty} (\beta - 1)\beta^{-i} = (\beta - 1)\sum_{i=1}^{\infty} \beta^{-i} = (\beta - 1)\beta^{-1}\frac{1}{1 - \frac{1}{\beta}} =$$

(al massimo; progressione geometrica)

$$= (\beta - 1)\beta^{-1}\frac{\beta}{\beta - 1} = 1$$

In realtà in un calcolatore le cifre RAPPRESENTABILI sono FINITE per cui la sommatoria sarà finita.

Si parla in particolare di approssimazione per TRONCAMENTO se il numero viene troncato e quindi dopo la m-esima cifra si butta tutto via. Si ha invece l'ARROTONDAMENTO quando il numero si tronca e l'ultima cifra resta invariata se d_{m+1} è compresa tra 0 e $\beta/2$ (inclusi) oppure si somma al numero β^{-m} (la quantità più piccola) se $\frac{\beta}{2} < d_{m+1} \leq \beta - 1$:

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

$$\text{arr}(x) = \text{sign}(x)\, \beta^b \begin{cases} \sum_{i=1}^{m} d_i \beta^{-i} & \text{se } 0 \leq d_{m+1} \leq \beta/2 \\ \sum_{i=1}^{m} d_i \beta^{-i} + \beta^{-m} & \text{se } \beta/2 < d_{m+1} \leq \beta - 1 \end{cases}$$

In particolare se d_m è proprio uguale a $\beta - 1$ e dovrò che "si porta il riporto" e così via.

CARDINALITÀ DI F

Indichiamo ora con $F^{(o\,N)}$ l'insieme dei **NUMERI MACCHINA** che sono funzione di:

$$F(\beta, m, \angle, U) \qquad \text{dove } \angle \le b \le U \text{ cioè min/max possibile di } b$$

(m: numero di cifre decimali)

Il numero di elementi che fanno parte di F (CARDINALITÀ) e quindi i numeri rappresentabili si ricavano in questo modo:

① Quante combinazioni danno le cifre decimali?

$$(\beta-1)\underbrace{\beta\beta\ldots\beta}_{m-1} \Rightarrow (\beta-1)\,\beta^{m-1} \Rightarrow \beta^m - \beta^{m-1}$$

(prima cifra diversa da zero)

② Il numero di esponenti possibili sono: $U-\angle+1$

③ Poi vi è il segno e quindi per ogni singolo numero ve ne saranno in realtà due.

Da tutto ciò si evince che la cardinalità sarà:

$$\boxed{\text{card}(N) = 2\cdot(\beta^m - \beta^{m-1})(U-\angle+1) + 1}$$

(+1 lo zero)

ERRORE NEL $tr(x)$ E NELL'$arr(x)$

- Consideriamo ora la quantità δ_x ('sigma x') data da: $\delta_x = tr(x) - x$. Non è altro che l'errore **ASSOLUTO** l'approssimazione che si fa andando a troncare:

$$\boxed{|\delta_x| < \beta^{b-m}}$$

occorre considerare anche b

Nel caso invece dell'arrotondamento:

$$\boxed{|\delta_x| < \frac{1}{2}\beta^{b-m}}$$

ESEMPIO: $11{,}2735 = 10^2 \cdot 0{,}112735$

Con $m=5$ troncato sarà $0{,}11273 \cdot 10^2 \Rightarrow |\delta_x| < 10^{2-5} = 10^{-3}$ si vede ad occhio

- Vediamo ora l'**ERRORE RELATIVO** dato da: $E_x = \dfrac{tr(x)-x}{x}$ $\qquad E_x = \dfrac{arr(x)-x}{x}$

al denominatore ci va il più piccolo x possibile

TRONCAM.

$$\boxed{|E_x| < \frac{\beta^{b-m}}{\beta^{b-1}} = \beta^{1-m}}$$

ARROTONDAMENTO

$$\boxed{|E_x| < \frac{1}{2}\frac{\beta^{b-m}}{\beta^{b-1}} = \frac{1}{2}\beta^{1-m}}$$

Se passo da un numero alla sua rappresentazione in macchina quello è il massimo errore possibile (**PRECISIONE DI MACCHINA**).

ERRORE DI UNA FUNZIONE (ASSOLUTO)

Prendiamo $f(p): \mathbb{R}^m \to \mathbb{R}$ e un punto $p_0 = (x_1, x_2 \dots x_m)$

Ora calcolando proprio $f(p_0)$ voglio conoscere qual è l'errore che vado a commettere (vedi dopo perché c'è l'errore).
Innanzitutto stabilisco un insieme di INDETERMINAZIONE, cioè il PRODOTTO CARTESIANO DEGLI intervalli entro cui è compresa ogni componente di p_0:

$$D = \{ P \in \mathbb{R}^m \mid a_i \le x_i \le b_i \}$$

Se per esempio $p_0 = (\sqrt{2}, \pi, e)$ avremo: $1 < \sqrt{2} < 2$, $3 < \pi < 4$, $2 < e < 3$

L'errore nasce dal fatto che di quelle componenti noi diamo delle approssimazioni (sono valori infiniti)
Effettivamente si va a considerare il punto p_1 le cui componenti (anche non tutte) saranno delle approssimazioni di quelle di p_0 sempre presenti però nell'insieme di indeterminazione!
In realtà poi anche la funzione può in generale presentare delle approssimazioni (sempre per lo stesso motivo) e quindi in realtà la si sostituisce con un ALGORITMO DI CALCOLO (indicato con f_a) che introduce errori.
Quindi si passa:

$$\text{da } f(p_0) \Rightarrow f_a(p_1)$$

Si vuole studiare quindi l'errore!

ERR. ASSOLUTO

$$\delta p = f(p_0) - f_a(p_1) = \underbrace{-f_a(p_1) + f(p_1)}_{\text{contributo dato dai dati}} \underbrace{- f(p_1) + f(p_0)}_{\text{ERR. ASS. TRASMESSO DAI DATI } \delta_d}$$

Il contributo di questa differenza è dato da f_a (il punto è lo stesso)
ERR. ASS. ALGORITMICO δ_a

Calcoliamo:

$$\delta_d = f(p_1) - f(p_0) \qquad f(p_1) = f(p_0) + \sum_{i=1}^{m} \frac{\partial f}{\partial x_i} (x_i^{(1)} - x_i^{(0)})$$

SVILUPPO DI TAYLOR

$(x_i^{(1)} - x_i^{(0)})$ → differenza tra componenti (errore assoluto δx_i)

diventa: $f(p_1) = f(p_0) + \sum_{i=1}^{m} \frac{\partial f}{\partial x_i} \delta x_i$

Quindi: $\delta_d = \sum_{i=1}^{m} \frac{\partial f}{\partial x_i} \delta x_i$

$\frac{\partial f}{\partial x_i}$ → COEFFIC. DI AMPLIFICAZIONE p_i

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

$|\delta x_i|$ → DIFFERENZA COMPONENTI (È L'ER. AS. A PAG. 2)
p_i → DERIVATE PARZIALI CALCOLATE IN x_i

Noi sappiamo che le componenti di p_1 stanno però nell'insieme di indeterminazione. Quindi prendo l'estremo superiore di $|\frac{\partial f}{\partial x_i}|$ che sarà calcolato in uno dei punti di questo insieme. $P \in D$.
A_i

$$|\delta_d| \le \sum_{i=1}^{n} A_i |\delta x_i|$$

ESEMPIO: $f(x_1, x_2) = x_1 \cdot x_2$ $\quad p_0 = (\sqrt{2}, \pi)$ $\quad D = [1,2] \times [3,4]$

$\frac{\partial f}{\partial x_1} = x_2$ → MAX IN D È 4 $\quad \frac{\partial f}{\partial x_2} = x_1$ → MAX IN D È 2

$p_1 = (1{,}41, 3{,}14)$ (ERRORE $\le \frac{1}{2} 10^{-2}$ ARROTONDAMENTO)

$\delta x_1 \le \frac{1}{2} 10^{-2}$ $\quad \delta x_2 \le \frac{1}{2} 10^{-2}$

$$|\delta_d| \le 4 \cdot \frac{1}{2} \cdot 10^{-2} + 2 \frac{1}{2} \cdot 10^{-2} = 3 \cdot 10^{-2}$$

→ continua

L'errore algoritmico dipende dall'algoritmo scelto. Si supponga di prendere il risultato arrotondato alla seconda cifra decimale quindi:

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

$$\text{ERRORE} \leq \frac{1}{2}10^{-2} + 3\cdot 10^{-2} = \frac{7}{2}\cdot 10^{-2}$$ FINE!

Caso più interessante è invece quando a partire dall'errore, si vuole capire quale approssimazione usare per stare sotto tale limite. Vediamo un esempio chiarificatore:

$$f(x_1,x_2) = \frac{x_1}{x_2} \qquad P_0 = (\sqrt{2},\pi) \quad \Rightarrow \quad D = [1,2]\times[3,4]$$

Si vuole $|f_a(P_1) - f(P_0)| \leq 10^{-3} \quad \Rightarrow \quad \delta_a + \delta_d \leq 10^{-3}$

$$\frac{\partial f}{\partial x_1} = \frac{1}{x_2} \qquad \frac{\partial f}{\partial x_2} = -\frac{x_1}{x_2^2}$$

$$\downarrow \qquad \downarrow$$

$$A_{x_1} = \frac{1}{3} \qquad A_{x_2} = \frac{2}{3}$$

Generalmente l'errore si spezza a metà tra i due contributi: $\delta_a \leq \frac{1}{2}\cdot 10^{-3}$ e $\delta_d \leq \frac{1}{2}10^{-3}$:

- Per avere quel valore di δ_a basta arrotondare sulla terza cifra decimale!
- $$\sum_{i=1}^{2} A_{x_i}|\delta_{x_i}| \leq \frac{1}{2}10^{-3} \quad \Rightarrow \quad A_{x_1}|\delta_{x_1}| + A_{x_2}|\delta_{x_2}| \leq \frac{1}{2}10^{-3}$$

  due incognite

Spezziamo ancora l'errore in due: $A_{x_1}|\delta_{x_1}| \leq \frac{1}{4}10^{-3} \Rightarrow |\delta_{x_1}| \leq \frac{3}{4}10^{-3}$ — x_1 ARROTONDATO ALLA 3ª CIFRA (errore max $\frac{1}{2}10^{-3}$ che è min di $\frac{3}{4}10^{-3}$)

$A_{x_2}|\delta_{x_2}| \leq \frac{1}{4}10^{-3} \Rightarrow |\delta_{x_2}| \leq \frac{9}{8}10^{-3}$ — x_2 TRONCATO ALLA 3ª CIFRA

ERRORE DI UNA FUNZIONE (RELATIVO)

L'errore RELATIVO sarà: $\varepsilon_f = \frac{f_a(P_1)-f(P_0)}{f(P_0)} \overset{\text{andando a sostituire}}{=} \varepsilon_a + \varepsilon_d + \varepsilon_a\varepsilon_d$

ERRORE DELL'ALGORITMO $\varepsilon_a = \frac{f_a(P_1)-f(P_1)}{f(P_1)}$ — ERR. RELATIVO DEI DATI $\varepsilon_d = \frac{f(P_1)-f(P_0)}{f(P_0)}$

Poiché in realtà $\varepsilon_a\varepsilon_d$ ha come ordine di grandezza nettamente minore rispetto ai due separati (regola del prodotto):

$$\boxed{\varepsilon_f = \varepsilon_a + \varepsilon_d}$$

Vediamo ora nel dettaglio:

$$\varepsilon_d = \frac{f(p_1) - f(p_0)}{f(p_0)} \rightarrow \text{al numeratore } \partial d \qquad \sum_{i=1}^{m} \frac{\partial f}{\partial x_i} \cdot \frac{\delta_{x_i}}{f(p_0)}$$

moltiplico e divido $\Rightarrow$ per $x_i^{(0)}$

$$\sum_{i=1}^{m} \frac{\partial f}{\partial x_i} \frac{x_i^{(0)}}{f(p_0)} \cdot \frac{\delta_{x_i}}{x_i^{(0)}} = \sum_{i=1}^{m} \frac{\partial f}{\partial x_i} \frac{x_i^{(0)}}{f(p_0)} \varepsilon_{x_i}$$

ε l'errore relativo sulle singole componenti

Indichiamo con $\gamma_i = \frac{\partial f}{\partial x_i} \frac{x_i^{(0)}}{f(p_0)}$ detti <u>COEFF. DI AMPLIFICAZIONE</u> dell'errore relativo

$$\boxed{\varepsilon_d = \sum_{i=1}^{m} \gamma_i \varepsilon_{x_i}}$$

MASTER COPY

Tel. 388/9837745

ERRORI NELLE QUATTRO OPERAZIONI

	δd	ε_d
$x+y$	$\delta_x + \delta_y$	$\frac{x}{x+y}\varepsilon_x + \frac{y}{x+y}\varepsilon_y$
$x-y$	$\delta_x - \delta_y$	$\frac{x}{x-y}\varepsilon_x - \frac{y}{x-y}\varepsilon_y$
$x \cdot y$	$y\delta_x + x\delta_y$	$\varepsilon_x + \varepsilon_y$
x/y	$\frac{\delta_x}{y} - \frac{x}{y^2}\delta_y$	$\varepsilon_x - \varepsilon_y$

(vedi formula sopra) $\frac{\partial}{\partial f}(x+y) \cdot \frac{x}{x+y} \cdot \varepsilon_x + \frac{\partial}{\partial y}(x+y) \cdot \frac{y}{x+y} \cdot \varepsilon_y$

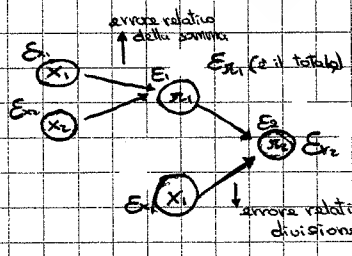
$\rightarrow y \cdot \frac{x}{xy}\varepsilon_x + x \cdot \frac{y}{xy}\varepsilon_y$

(vedi formula pag. 31) $\frac{\partial\left(\frac{x}{y}\right)}{\partial x} \cdot \delta_x + \frac{\partial\left(\frac{x}{y}\right)}{\partial y}\delta_y$

Moltiplicazione e divisione possono dare molti problemi in quanto l'errore assoluto dipende dall'ordine di grandezza dei fattori. Per l'errore relativo sono invece pericolose le operazioni di somma e differenza.

COME CALCOLARE ε_f?

- $f(x_1, x_2) = \frac{x_1 + x_2}{x_1}$ possiamo risolverlo in due parti riconducendoci alle quattro operazioni. ① $\pi_1 = x_1 + x_2$ ② $x_2 = \frac{\pi_1}{x_1}$



ε_{π_1} (è il totale) $\Rightarrow$ $\varepsilon_{\pi_1} = \varepsilon_1 + \frac{x_1}{x_1+x_2}\varepsilon_{x_1} + \frac{x_2}{x_1+x_2}\varepsilon_{x_2}$

errore relativo algoritmico

L'errore relativo si ottiene a partire <u>DAL FONDO RISALENDO</u>.

È L'ER. REL. ALGORITMICO (dovuto per esempio al troncamento del risultato)

$\varepsilon_f = \varepsilon_2 + \varepsilon_{x_{\pi_1}} - \varepsilon_{x_1}$

ε_f vedi pag. 4

ER. REL. DEI DATI (vedi tabellina per la differenza)

Ricaviamo $\varepsilon_{x_{\pi_1}}$ $\quad \varepsilon_f = \varepsilon_2 - \varepsilon_{x_1} + \varepsilon_1 + \frac{x_1}{x_1+x_2}\varepsilon_{x_1} + \frac{x_2}{x_1+x_2}\varepsilon_{x_2}$

// semplificando

$$\varepsilon_f = \underbrace{\varepsilon_1 + \varepsilon_2}_{\varepsilon_a} + \frac{x_2}{x_1+x_2}\varepsilon_{x_1} + \underbrace{\frac{x_2}{x_1+x_2}\varepsilon_{x_2}}_{\varepsilon_d}$$