

CAPITOLO 0

il significato della virgola dipende dall'esponente

poiché $\beta^{-2} \leq x < \beta^0$

$x = \frac{|x|}{\beta^b} \Rightarrow \beta^{-1} \leq x < 1$

(NUMERI IN VIRGOLA MOBILE, PRECISIONE)

$F(\beta, m) = \{0\} \cup \{x \in \mathbb{R} \text{ t.c. } x = (-1)^s \beta^b \cdot 0, c_1 \dots c_m\}$ con $s \in \{0, 1\}, b \in \mathbb{Z}$,
 c_1, \dots, c_m cifre in base β , $c_1 \neq 0$
ELEMENTI NORMALIZZATI

SIMMETRICO RISPETTO A 0

Introduciamo questo TIPO "REALIZZABILE" perché abbiamo l'esigenza di ridurre il TIPO "IDEALE" (quindi numeri REALI) a un tipo elaborabile da parte del calcolatore (numeri in VIRGOLA MOBILE e PRECISIONE FINITA).

Scrivere $\frac{1}{10}$, utilizzando il tipo in virgola mobile, in base $\beta=10$ e $\beta=2$

$X = \frac{1}{10} \quad \beta^{b-1} \leq X < \beta^b \quad b=0 \quad \frac{1}{10} \leq \frac{1}{10} < 1 \quad s=0 \quad X = (-1)^0 \cdot 10^0 \cdot 0,1 \quad (\beta=10)$

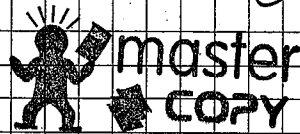
Se $\beta=2 \quad \frac{1}{16} \leq \frac{1}{10} < \frac{1}{8} \quad \frac{1}{2^4} \leq \frac{1}{10} < \frac{1}{2^3} \quad b=-3 \quad X = (-1)^0 \cdot 2^{-3} \cdot 0,1c_1c_2c_3 \dots$

$0,1c_1c_2c_3 \dots = \frac{0,1}{2^{-3}} = \frac{0,1}{1/8} = 0,8 = \frac{8}{10} = \frac{4}{5} \quad \frac{4}{5} = 0,1c_1c_2c_3 \dots \Rightarrow \frac{8}{5} = c_1c_2c_3 \dots$

$\frac{8}{5} = 1 + \frac{3}{5} \Rightarrow \underline{c_1=1} \quad \frac{3}{5} = 0,1c_2c_3c_4 \dots \Rightarrow \frac{6}{5} = c_2c_3c_4 \dots \Rightarrow \underline{c_2=1} \quad \frac{1}{5} = 0,1c_3c_4 \dots$

$\Rightarrow \frac{2}{5} = c_3c_4c_5 \dots \Rightarrow \underline{c_3=0} \quad \frac{4}{5} = c_4c_5c_6 \dots \Rightarrow \underline{c_4=0} \quad \left(\frac{8}{5}\right) \text{ ricomincia il periodo}$

$\Rightarrow \frac{1}{10} = (-1)^0 \cdot 2^{-3} \cdot 0,1100$



Se ne deduce che a seconda della scelta della base un numero può avere un n° di cifre dopo la virgola) una scrittura FINITA o NON FINITA.

$F(\beta, m)$ si può vedere come unione di insiemi PARTIZIONATI $\bigsqcup_{b \in \mathbb{Z}} (-\beta^b, \beta^b] \cup \{0\} \cup \bigsqcup_{b \in \mathbb{Z}} (-\beta^b, -\beta^{b+1}]$

Qual è la distanza tra il max della partizione precedente e il min della successiva? β^{b-1}

Fatti: $|10^{b-1} \cdot 0,9 - 10^{b+1} \cdot 0,1| = |10^b (0,9 - 1)| = 0,1 \cdot 10^0 = 10^{b-1}$

Proprietà di $F(\beta, m)$:
 • $F(\beta, m) \subset \mathbb{Q}$ infatti $(-1)^s \beta^{b-m} \cdot 0, c_1 \dots c_m$
 • $F(\beta, m)$ è numerabile e ordinata (infatti lo è \mathbb{Q})
 • $F(\beta, m)$ simmetrica rispetto a 0
 • 0 è l'unico PUNTO di ACCUMULAZIONE (in ogni suo intorno c'è un elemento che è a $F(\beta, m)$)
gli altri sono punti isolati

Es. $\inf_{m \in \mathbb{N}} \beta^m = 0$ $\lim_{m \rightarrow \infty} \beta^m = +\infty$ $\lim_{m \rightarrow -\infty} \beta^m = -\infty$
 • $\sup F(\beta, m) = +\infty$ $\inf F(\beta, m) = -\infty$
 $\xi \in F(\beta, m)$ con $\xi \neq 0$

Successore di ξ : $\sigma: F(\beta, m) - \{0\} \rightarrow F(\beta, m) - \{0\} \quad \sigma(\xi) = \min\{\theta \in F(\beta, m) \mid \theta > \xi\}$

Predecessore di ξ : $\pi: F(\beta, m) - \{0\} \rightarrow F(\beta, m) - \{0\} \quad \pi(\xi) = \max\{\theta \in F(\beta, m) \mid \theta < \xi\}$

$\sigma(\pi(\xi)) = \xi = \pi(\sigma(\xi))$

TEOREMA (distribuzione degli elementi di $F(\beta, m)$)
 Se $\xi = \max B_b \Rightarrow \sigma(\xi) = (\beta^{b+1} \cdot 0,1)$
 $\sigma(\xi) = \xi + \beta^{b-m} = \beta^b (0,1 + \beta^{-m})$
 $= (\beta^b \cdot 0, c_1 \dots c_m + 1)$

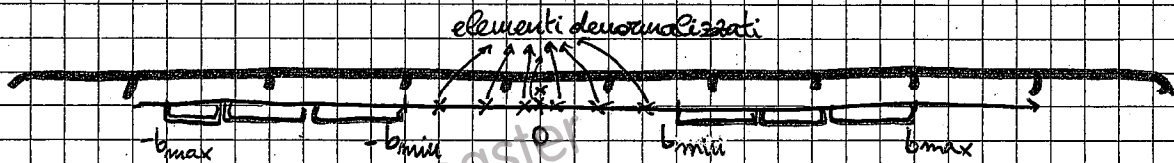
conseguenza del TEOREMA:

$$\sigma(\frac{b}{\beta}) - \frac{b}{\beta} = \beta^{b-m}$$

DISTANZA tra $\frac{b}{\beta}$ e $\sigma(\frac{b}{\beta}) = \beta^{b-m}$

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

La distanza tra elementi consecutivi di $F(\beta, m)$ è PROPORZIONALE a β^b (ORDINE di GRANDEZZA di $\frac{b}{\beta}$). Tanto + siamo lontani dall'origine, tanto più gli elementi di $F(\beta, m)$ sono distanti tra loro.



$F(\beta, m, b_{\min}, b_{\max})$ NUMERI IN VIRGOLA MOBILE con ESPONENTE LIMITATO con precisione m
 $F_d(\beta, m, b_{\min}, b_{\max})$ NUMERI IN VIRGOLA MOBILE con ESPONENTE LIMITATO e ELEMENTI DENORMALIZZATI con precisione m
 $F(\beta, m)$ NUMERI IN VIRGOLA MOBILE con precisione m

$$F_d(\beta, m, b_{\min}, b_{\max}) = F(\beta, m, b_{\min}, b_{\max}) + \{x = (-1)^i \beta^{b_{\min}} a_1 a_2 \dots a_m\}$$

$\exists \sigma(b) \wedge \exists \tau(0)$

Si avrà che $F(\beta, m, b_{\min}, b_{\max}) \subset F_d(\beta, m, b_{\min}, b_{\max}) \subset F(\beta, m)$.

n° finito di elementi

n° ∞ di elementi



Esercizi

E1) Determinare l'esponente e la frazione di $\frac{2}{5}$ in base 3.

$$\frac{1}{3} < \frac{2}{5} < \frac{2}{3} \quad b=0 \quad q = \frac{2}{5} = \frac{2}{5} \quad (\frac{2}{5} = 0, \overline{10})$$

E2) Giudicare quali dei seguenti numeri reali appartengono a $F(2,3)$: $1, \frac{1}{3}, -\frac{1}{16}, \frac{3}{16}, \pi$

$1 \quad 2^0 \leq 1 < 2^1 \quad (-1)^0 \cdot 2^0 \cdot \frac{1}{2} \quad (\frac{1}{2} \text{ in base } 2 = 0,1) \Rightarrow (-1)^0 \cdot 2^0 \cdot 0,100 \in F(2,3)$

$\frac{1}{3} \quad 2^{-2} \leq \frac{1}{3} < 2^{-1} \quad (-1)^0 \cdot 2^{-1} \cdot \frac{1}{3} \quad (\frac{1}{3} \text{ in base } 2 = 0, \overline{101}) \notin F(2,3)$

$-\frac{1}{16} \quad 2^{-4} \leq -\frac{1}{16} < 2^{-3} \quad (-1)^1 \cdot 2^{-3} \cdot \frac{1}{2} \quad (\frac{1}{2} = 0,1) = -1 \cdot 2^{-4} \in F(2,3)$

$\frac{3}{16} \quad 2^{-3} \leq \frac{3}{16} < 2^{-2} \quad (-1)^0 \cdot 2^{-2} \cdot \frac{3}{4} \quad (\frac{3}{4} = 0,101) \Rightarrow (-1)^0 \cdot 2^{-2} \cdot 0,101 \in F(2,3)$

$0 = 0 \in F(2,3)$

$\pi \notin \mathbb{Q} \Rightarrow \pi \notin F(2,3)$

E3) Dimostrare che $F(2,2) \subset F(2,3)$ $F(2,2) = \{x \in F(2,3) : x = (-1)^s 2^b a_1 a_2, a_1, a_2 = 0,1\}$

E4) Sia $x = 3,7$ (in base 10). Decidere se $x \in F(2,8)$.

$$2^1 \leq 3,7 \leq 2^2 \Rightarrow x = (-1)^0 \cdot 2^1 \cdot \frac{37}{20} \rightarrow 0,111010100 \quad x \notin F(2,8)$$

E5) Mostrare che tutti gli elementi positivi di $F(2,4)$ con esponente ≥ 4 sono interi, poi determinare:

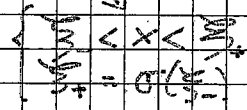
$$\max\{\xi \in F(2,4) \text{ t.c. } \xi \geq 0 \text{ e } \xi \notin \mathbb{Z}\} \text{ e } \min\{\xi \in \mathbb{N} \text{ t.c. } \xi \notin F(2,4)\}$$

$$|x| = 2^b \cdot 0,1a_1a_2a_3a_4 = 2^{b-4} c_1c_2c_3c_4 \quad 2^3 \cdot 0,1111 \quad 2^5 \cdot 0,10001 \Rightarrow 10001$$

FUNZIONE ARROTONDAMENTO in M

$rd: \mathbb{R} \rightarrow M$ " $rd(x)$ " è l'elemento di M più vicino a x

Se x è equidistante dai 2 elementi di M adiacenti si risolve adottando la tecnica RTTE o RTTA \rightarrow away (più lontano da zero) o even (con cifre pari)



Arrotondare $x = 1/10$ in $M = (2, 2)$

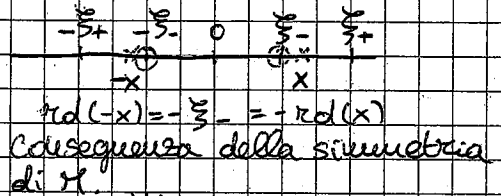
$x = 2^{-3} \cdot 0,100$ $\xi_- = 2^{-3} \cdot 0,11$ $\xi_+ = 2^{-2} \cdot 0,10$

$m = 2^{-3} \cdot 0,111 > x$ $rd(x) = \xi_- = \frac{1}{8} \left(\frac{1}{2} + \frac{1}{4} \right) = \frac{3}{32} = 0,09375$

$\rightarrow \frac{2^{-3} \cdot 0,10 + 2^{-3} \cdot (0,10 + 0,01)}{2} = \frac{2^{-3} \cdot 0,10 + 2^{-3} \cdot 0,11}{2} = 2^{-3} \cdot (0,10 + 0,001) = 2^{-3} \cdot 0,101$

Proprietà di rd:

- non è invertibile
- è dispari [$rd(-x) = -rd(x)$]
- è non decrescente
- $x < y \Rightarrow rd(x) \leq rd(y)$
- $rd(x) = x \Leftrightarrow x \in M$
- Se $M = F(\beta, m)$ allora $rd(x) = 0 \Leftrightarrow x = 0$



E1) Calcolare l'arrotondato di $1/4$ in $F(3, 2)$

$\frac{1}{9} \leq \frac{1}{4} < \frac{1}{3}$

$x = 3^{-1} \cdot \frac{1}{4} = 3^{-1} \cdot \frac{3}{4} = 3^{-1} \cdot 0,20$

$\xi_- = 0,20 \cdot 3^{-1}$ $\xi_+ = 0,21 \cdot 3^{-1}$
 $3 \cdot m = [3^{-1} \cdot (0,20 + 0,21)] \cdot 3^{-1}$

$\Rightarrow rd(x) = \xi_+ = 3^{-1} \cdot 0,21 = \frac{1}{3} \left(\frac{2}{3} + \frac{1}{9} \right) = \frac{7}{27} \approx 0,259$

$3 \cdot m = \frac{1}{3^2} \left(\frac{2}{3} + \frac{2}{3} + \frac{1}{9} \right) = \frac{13}{54} \approx 0,24$
 $0,25 > 0,24$
 Ho parlato in $\beta = 10$

E2) Calcolare l'arrotondato di $2^2 \cdot 0,1011$ in $F(10, 2)$

$4 \left(\frac{1}{2} + \frac{1}{8} + \frac{1}{16} \right) = \left(2 + \frac{1}{2} + \frac{1}{4} \right) = \left(\frac{11}{4} \right) = 2,75$

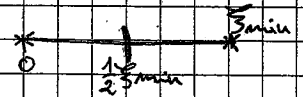
$rd(10^1 \cdot 0,275) = 10^1 \cdot 0,28$ (RTTA)

(anche RTTE \rightarrow arrotondamento \rightarrow RTA)

E3) Calcolare l'arrotondato di $\frac{1}{2} \xi_{min}$ in $F(2, 5, -9, 9)$

$\xi_{min} = \min \{ \xi \in M \mid \xi > 0 \} = 2^{-9} \cdot 0,10000$

$\frac{1}{2} \xi_{min} = 0,1 \cdot 2^{-10}$



$rd\left(\frac{1}{2} \xi_{min}\right) = \xi_{min}$ (RTTA, RTTE \rightarrow RTA)

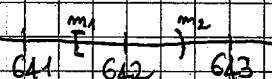
E4) Calcolare l'arrotondato di $\frac{1}{2} \xi_{min}$ in $F_d(2, 5, -9, 9)$

$\xi_{min} = 2^{-9} \cdot 0,00001 = 2^{-14}$

$\xi_+ = \xi_{min}$ $\xi_- = 0$

$rd(x) = 0$ (RTTE)

E5) Determinare $\{ x \in \mathbb{R} \mid rd(x) = 642 \}$. $M = F(10, 3)$, rd con RTTE. ($642 = 10^3 \cdot 0,642 \in M$)



$641,5 < x < 642,5$ (RTTE) oppure $641,5 \leq x \leq 642,5$ (RTA)

E6) Determinare $\max \{ y \in \mathbb{R} \mid rd(314 + y) = 314 \}$. $M = F(10, 3)$ rd con RTTE. ($314 = 10^3 \cdot 0,314 \in M$)
 $313,5 < x < 314,5$ $\forall_{max} = 0,5$

FUNZIONI ERRORE: $\delta(x) = rd(x) - x$ Funzione errore assoluto

$\epsilon(x) = \frac{rd(x) - x}{x} = \frac{\delta(x)}{x}$ Funzioni err. relative
 • $\delta(x)$ dispari $\delta(x) = -\delta(-x)$ con $x \neq 0$
 • $\epsilon(x)$ e $\eta(x)$ pari

$$\eta(x) = \frac{\delta(x)}{rd(x)} = \frac{\delta(x)}{x} \cdot \frac{x}{rd(x)} = \epsilon(x) \cdot \left[\frac{x}{rd(x)} - 1 \right] + 1 = \epsilon(x) [1 - \eta(x)] \Rightarrow \epsilon(x) = \frac{\eta(x)}{1 - \eta(x)}$$

è $\neq 1$ per hp
o x sarebbe ∞

Calcolare gli errori commessi approssimando $x = \frac{1}{3}$ in $F(10, 3) = M$

$x = 10^0 \cdot 0,3$ $rd(x) = \frac{x}{3} = 0,333$ $\delta(\frac{1}{3}) = 0,333 - \frac{1}{3} = -\frac{1}{3000}$

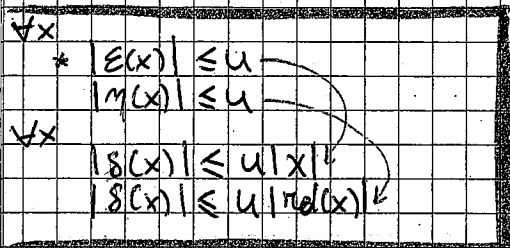
$\eta(\frac{1}{3}) = \frac{\delta(\frac{1}{3})}{rd(\frac{1}{3})} = \frac{-\frac{1}{3000}}{\frac{1}{333}} = -\frac{1}{999}$ $\epsilon(\frac{1}{3}) = \frac{\delta(\frac{1}{3})}{\frac{1}{3}} = -\frac{1}{1000}$

TEOREMA (stima delle funzioni errore)

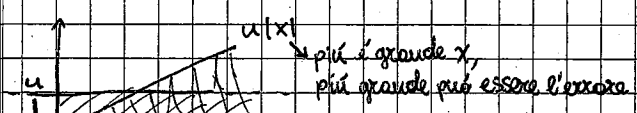
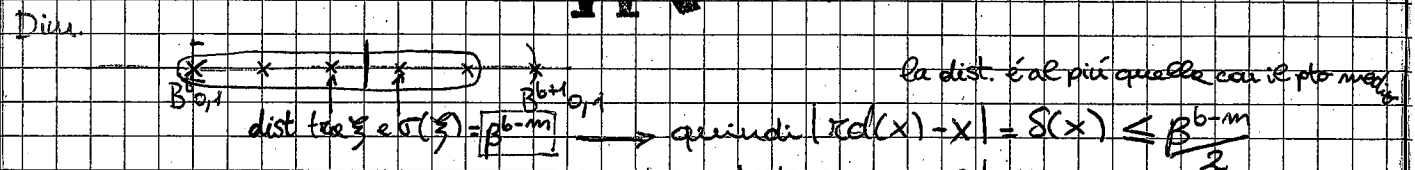
$M = F(\beta, m)$; $x = \beta^b \cdot q$ con $0 < q < 1$

$$\begin{cases} |\delta(x)| \leq \frac{1}{2} \beta^{b-m} \\ |\epsilon(x)| \leq \frac{1}{2} \beta^{b-m} \\ |\eta(x)| \leq \frac{1}{2} \beta^{b-m} \end{cases}$$

la LIMITAZIONE non dipende da x



$$\frac{1}{2} \beta^{b-m} = u \text{ Precisione di macchina}$$



$$|\epsilon(x)| = \frac{|\delta(x)|}{x} \leq \frac{\beta^{b-m}}{q \beta^b \cdot 2} \leq \frac{1}{2} \beta^{b-m}$$

qualità + piccola possibile

limitazione uniforme (M è un insieme di numeri in VIRGOLA MOBILE)

Quale limitazione è più stringente sull'errore dell'approssimato in $F(2, 53)$ e $F(10, 12)$

$u_2 \approx \frac{1}{2} \cdot 2^{1-53} = 2^{-53} \approx 10^{-16}$ $u_{10} \approx \frac{1}{2} \cdot 10^{1-12} = 5 \cdot 10^{-12}$

limitaz. più stringente

ciò non implica che l'approssimazione che si ha approssimando in $F(2, 53)$ sia migliore di quella in $F(10, 12)$, "a parità di sfortuna l'errore peggiore sarà + piccolo".
 Il teorema fornisce una LIMITAZIONE SUPERIORE dell'ERRORE.

$M = F(\beta, m)$
 $\forall x \exists d$ (dipendente da x) t.c. $\begin{cases} |rd(x) - x| = |d| \\ |d| \leq u|x| \end{cases}$ d può interpretare come PERTURBAZIONE ADDITIVA di x
 ($d = rd(x) - x = \delta(x)$)

$\forall x \exists e$ (dipendente da x) t.c. $\begin{cases} |rd(x) - x| = |(1+e)x| \\ |e| \leq u \end{cases}$ $rd(x)$ è una PERTURBAZIONE MOLTIPLICATIVA di x

$\forall x \exists t$ (dipendente da x) t.c. $\begin{cases} x = (1+t)rd(x) \\ |t| \leq u \end{cases}$

E28) Siano $x = \frac{5}{4}$ e rd la funzione arrotondamento in $F(2,2)$ con RTE.
 Determinare $rd(x)$ e gli errori assoluto e relativo commessi approssimando x con il suo arrotondato. Verificare le condizioni sugli errori.

$x = \frac{5}{4} \quad 2^0 < \frac{5}{4} < 2^1 \quad x = 2^1 \cdot \frac{5}{4} = 2^1 \cdot \frac{5}{8} = 2^1 \cdot 0,101 \quad (\frac{5}{8}) = 2^1 \cdot 0,10 \quad \sigma(\frac{5}{8}) = 2^1 \cdot 0,11$
 $x = 1,25 \quad rd(x) = 2^1 \cdot 0,10 = 1 \quad S(x) = \frac{0,25 \cdot \frac{1}{2}}{2} = \frac{0,125}{2} = \frac{1}{16} \quad rd(x) = \frac{5}{8} \quad m = \frac{\frac{5}{8} + \frac{5}{8}}{2} = \frac{1 + \frac{3}{2}}{2} = \frac{5}{4} = x$
 $E(x) = \frac{0,25}{1,25} = \frac{1}{5} \leq \frac{1}{2} = \frac{1}{4}$

Il tipo in virgola mobile e precisione finita è definito dall'insieme M e da funzioni PREDEFINITE, date dall'unione di 3 sottoinsiemi:

- Funzioni predefinite corrispondenti a 2) FUNZIONI ARITMETICHE: $\oplus, \ominus, \otimes : M \times M \rightarrow M$ t.c.
 - ⊕ Simmetrica, non associativa (arrotondamenti diversi)
 - ⊗ Simmetrica, non associativa
 anche $\odot : M \times M \rightarrow M \quad \frac{5}{8} \odot \frac{5}{8} = rd(\frac{5}{8} \cdot \frac{5}{8})$
- Funzioni predefinite corrispondenti a 2) FUNZIONI ELEMENTARI: $f : D \rightarrow R$
 - SEN ha solo uno zero $\rightarrow rd(\text{sen } \frac{\pi}{2}) = 0$
 - (non ne ha infiniti) $\Rightarrow \text{sen } \frac{\pi}{2} = 0 + (\pi)$ se $\pi \neq 0 \in M$

$F : D \subset M \rightarrow M$ t.c.
 $F(\frac{\pi}{2}) = rd(f(\frac{\pi}{2}))$

- Funzioni predefinite corrispondenti ai CONFRONTI: $\langle, \leq, =, \geq, > : M \times M \rightarrow \{V, F\}$
 Il confronto avviene esattamente come nei Reali.
 al peggio $e = u$ (precisione di macchina)

Le funzioni PREDEFINITE sono definite nel MODO MIGLIORE POSSIBILE, il valore di una f predefinita è l'elemento di M che dista meno dal RISULTATO ESATTO.

E31) Sia $M = F(10,2)$. Dimostrare, utilizzando le proprietà della funzione rd , che:
 1) $\forall \xi$ si ha $\xi \oplus (-\xi) = 0 \rightarrow rd(\xi + (-\xi)) = rd(0) = 0$
 2) $\forall \xi \exists$ un solo α t.c. $\xi \oplus \alpha = 0 \rightarrow rd(\xi + \alpha) = 0 \rightarrow$ poiché 0 è punto di accumulazione dovrà essere $\xi + \alpha = 0 \Rightarrow \alpha = -\xi$.

PASSAGGIO da P a P^*

- 1) Sostituire a ciascuna costante a valore in R il suo arrotondato in M .
- 2) Sostituire a ciascuna operazione aritmetica o f elementare la corrispondente funzione PREDEFINITA aggiungendo, se è il caso, opportuna precedenza tra operatori.

STUDIO dell'ERRORE $y = f(x)$ in $R \quad y = \phi(x)$ in M
 $\forall x \in D$ t.c. $f(x) \neq 0$, si vuole determinare informazioni sull'errore commesso approssimando $f(x)$ con $\phi(x)$, errore sulla quantità: $e_t = \frac{\phi(x) - f(x)}{f(x)}$

Algoritmo accurato: l'algoritmo ϕ è accurato quando utilizzato per approssimare f in x se $\phi(x)$ è una piccola perturbazione multiplicativa di $f(x)$ $\rightarrow \phi(x) = (1 + e_t) f(x) \Rightarrow e_t = \frac{\phi(x) - f(x)}{f(x)}$
 $|e_t| \leq K u$
 K non troppo grande

Quantitativamente e_t è un multiplo non troppo grande di u .
 $e_t = \phi(x) - f(x)$

Nel caso migliore possibile si ha che $\phi(x) = rd(f(x)) \Rightarrow e_t = \frac{rd(f(x)) - f(x)}{f(x)} \leq u$

Quando passiamo da P a P^* ogni operazione porta un "contributo" (una perturbazione moltiplicativa) del tipo $(1 + e_t)$
 Raccolgendo questi termini ci accorgiamo che i termini significativi saranno i singoli errori in quanto i prodotti tra gli stessi saranno di ordine trascurabile. Pertanto n operazioni provocheranno un errore $\leq n \cdot u$ - precisione di macchina. Se non volessimo trascurare: $|e_t| \approx nu + u^n$
 n approssimazioni.