ANALISI DEGLI ERRORI

SISTEMI DI NUMERAZIONE

I SISTEMI DI RAPPRESENTAZIONE NUMERICA SONO POSIZIONALI: OGNI CIFRA OCCUPA UNA POSIZIONE CORRISPONDENTE AD UNA POTENZA DELLA BASE DEL SISTEMA ADOTTATO.

SISTEMA DECIMALE (BASE 10)

$$10258.5 = 1\times10^4 + 0\times10^3 + 2\times10^2 + 5\times10^1 + 8\times10^0 + 5\times10^{-1}$$

SISTEMA BINARIO (BASE 2)

$$(10010.1)_2 = 1\times2^4 + 0\times2^3 + 0\times2^2 + 1\times2^1 + 0\times2^0 + 1\times2^{-1} = (18.5)_{10}$$

SISTEMA ESADECIMALE (BASE 16)

$$(2A1)_{16} = 2\times16^2 + 10\times16^1 + 1\times16^0 = (673)_{10}$$

LE CIFRE DI QUESTI SISTEMI SONO:

DECIMALE: $0, 1, \dots, 9$

BINARIO: $0, 1$

ESADECIMALE: $0, 1, \dots, 9, A, B, C, D, E, F$

QUINDI, SCELTA UNA BASE $N>1$ OGNI REALE a PUO' ESSERE SCRITTO:

$$a = \pm(a_m N^m + a_{m-1} N^{m-1} + \dots + a_0 + a_{-1} N^{-1} + \dots)$$

$$0 \le a_i \le N-1$$

TALE RAPPRESENTAZIONE E' UNICA TRANNE SE LA PARTE FRAZIONARIA CONTIENE INFINITE CIFRE

CONSECUTIVE $a_{-k} = N-1$. UNA RAPPRESENTAZIONE EQUIVALENTE E' QUELLA DI CONSIDERARE IL NUOVO NUMERO OTTENUTO SOPPRIMENDO LA SUCCESSIONE E AGGIUNGENDO UNA UNITA' ALL'ULTIMA CIFRA RIMASTA.

PER ES.: NEL SISTEMA DECIMALE:

$$0.72999\ldots9\ldots \quad \text{E} \quad 0.73$$

RAPPRESENTANO LO STESSO NUMERO.

PER LA RAPPRESENTAZIONE DEI NUMERI IN DIVERSE BASI SI HA IL SEGUENTE:

TEOREMA. SIA $b \in \mathbb{N}$, $b \neq 2$, $x \in \mathbb{R}$, $x \neq 0$.

$\Rightarrow \exists_1\ e \in \mathbb{Z}$, $\{a_i\}_{i=1,2,\ldots}$ $a_i \in \mathbb{N}$:

$$\boxed{x = \pm \left(\sum_{i=1}^{\infty} a_i b^{-i} \right) b^e}$$

$0 \leq a_i \leq b-1$, $a_1 \neq 0$, a_i DEFINITIVAMENTE $\neq b-1$ $\square$

RAPPRESENTAZIONE NUMERICA IN UN CALCOLATORE

POICHÉ IN UN CALCOLATORE LO SPAZIO DI MEMORIA È FINITO, LA SOMMATORIA PRECEDENTE PUÒ ESTENDERSI FINO A $t < \infty$:

$$x = \pm \left(\sum_{i=1}^{t} a_i b^{-i} \right) b^{e}$$

$$t < \infty \qquad L \leq e \leq U$$

ABBIAMO DUE TIPI DI RAPPRESENTAZIONE.

RAPPRESENTAZIONE IN VIRGOLA FISSA.

SONO FISSATI IL NUMERO DI CIFRE N CHE RAPPRESENTA IL NUMERO E IL NUMERO DI CIFRE PRIMA E DOPO LA VIRGOLA N_1, N_2: $N = N_1 + N_2$

ES.: $N = 10$, $N_1 = 4$, $N_2 = 6$

$27.325 \rightarrow 0027\ 325000$

$0.024 \rightarrow 0000\ 024000$

QUINDI, SE N SONO LE POSIZIONI DI MEMORIA ED UNA È PER IL SEGNO, $N-K-1$ PER LA PARTE INTERA E K PER QUELLA DECIMALE, SI HA:

$$x = (-1)^{s} b^{-K} \sum_{j=0}^{N-2} a_j b^{j}$$

DOVE b È LA BASE ED s È SCELTO IN BASE AL SEGNO DI x.

RAPPRESENTAZIONE IN VIRGOLA MOBILE

IN QUESTO CASO, LA POSIZIONE DELLA VIRGOLA DI UN NUMERO DECIMALE NON E' FISSA MA E' DATA DALL'ESPONENTE.

$\forall x \in \mathbb{R}$ SI PUO' SCRIVERE COME:

$$x = (-1)^s m \cdot b^e$$

DOVE $m \in \mathbb{R}$. TALE RAPPRESENTAZIONE NON E' UNICA. INFATTI:

$$x = m \cdot b^e = m' b^{e-1} = m'' b^{e+1} = \dots$$

$$m' = m \cdot b, \quad m'' = \frac{m}{b}$$

TALE RAPPRESENTAZIONE SI DICE <u>NORMALIZZATA</u> SE:

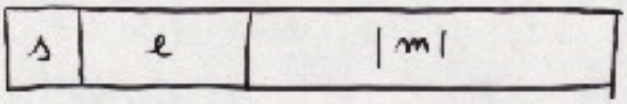
$$b^{-1} \le |m| < 1$$

IN TAL CASO CHIAMIAMO m MANTISSA E e ESPONENTE (CARATTERISTICA) DI x.

$$x = (-1)^s m \cdot b^e, \quad m = .a_1 a_2 \dots a_t$$

$$t \in \mathbb{N}, \quad 0 \le a_i \le b-1, \quad 0 \le m \le b^t - 1, \quad L \le e \le U$$

LO SPAZIO RISERVATO A $\forall x \in \mathbb{R}$ E':



L'INSIEME DEI NUMERI LA CUI MANTISSA E' RAPPRESENTABILE DA t CIFRE E LA CUI CARATTERISTICA E' COMPRESA TRA L ED U, CHE SONO INTERI CHE VARIANO PER $\forall$ CALCOLATORE E' DETTO INSIEME DEI NUMERI MACCHINA.

$$F = F(t, b, L, U) = \{0\} \cup \left\{ x \in \mathbb{R} : x = \pm \left(\sum_{i=1}^{t} a_i b^{-i} \right) b^e \right\}$$

POICHE' 0 HA UNA RAPPRESENTAZIONE PARTICOLARE.

F E' UN INSIEME FINITO E NUMERABILE.

$$\text{card } F = 1 + 2(b-1)b^{t-1}\cdot(U-L+1)$$

PER MEMORIZZARE E SPESSO SI AGISCE NEL SEGUENTE MODO: FISSATO L SI MEMORIZZA: $e^* = e - L$ CHE E' SEMPRE NON NEGATIVA, TENENDO PRESENTE CHE IL NUMERO E' MEMORIZZATO A MENO DI b^L.

IN UN CALCOLATORE A 32 BITS SI HANNO:
1 BIT PER IL SEGNO, 8 BIT PER L'ESPONENTE
$-127 \le e \le 127 \Rightarrow 0 \le e^* \le 255$ $(L = -127, U = 128)$,
23 BIT PER LA MANTISSA.
IL PIU' GRANDE E IL PIU' PICCOLO NUMERO RAPPRESENTABILI SONO:

$$x_{min} = (.1)_2 \cdot 2^{-127} = \frac{1}{2}\cdot 2^{-127} = 2^{-128}$$

$$x_{MAX} = (.1\ldots1)_2 \cdot 2^{128} \sim 2^{128}$$

INFATTI: $(.1\ldots1)_2 = \sum_{i=1}^{23}\left(\frac{1}{2}\right)^i = \frac{1-\left(\frac{1}{2}\right)^{23}}{1-\frac{1}{2}} = 1-\left(\frac{1}{2}\right)^{23} = 1-2^{-23} \sim 1$

CAMBIANDO L'ULTIMO BIT DELLA MANTISSA PER x_{min} SI HA:

$$\left(\frac{1}{2} + 2^{-23}\right)\cdot 2^{-127} = x_{min} + 2^{-150}$$

CHE E' UNA DIFFERENZA MOLTO PICCOLA.

CAMBIANDO L'ULTIMO BIT DELLA MANTISSA PER x_{max} SI HA

$$\left((.1\ldots1)_2 - 2^{-23}\right)2^{128} = x_{MAX} - 2^{105}$$

CHE E' UNA DIFFERENZA MOLTO GRANDE.
E' QUINDI MEGLIO RAPPRESENTARE NUMERI PICCOLI.

UN NUMERO DECIMALE, PER ESSERE RAPPRESENTATO NEL COMPUTER, VIENE CONVERTITO IN BINARIO. TALE CONVERSIONE PUO' COMPORTARE UNA RAPPRESENTAZIONE APPROSSIMATA. PER VEDERE CIO' VEDIAMO COME SI OPERA TALE CONVERSIONE. SI HANNO DUE CASI.

1) NUMERO >1. SI DIVIDE PER 2, SE C'E' RESTO SI METTE 1 ALTRIMENTI 0 E SI ASSEGNANO POTENZE DI 2 CRESCENTI.

ESEMPIO:

37	2	1×2^0
18	2	0×2^1
9	2	1×2^2
4	2	0×2^3
2	2	0×2^4
1	2	1×2^5
0		

(a)

$\rightarrow (100101)_2 = (37)_{10}$

2) NUMERO <1. SI DIVIDE PER $1/2$ OVVERO SI MOLTIPLICA PER 2. SE IL PRODOTTO E' <1 SI HA 0 ALTRIMENTI 1. E SI SOTTRAE 1 PROCEDENDO COME PRIMA.

ESEMPIO:

0.2	2	0×2^{-1}
0.4	2	0×2^{-2}
0.8	2	1×2^{-3}
$1.6 \rightarrow 0.6$	2	1×2^{-4}
$1.2 \rightarrow 0.2$	2	0×2^{-5}
0.4	2	0×2^{-6}
0.8	2	1×2^{-7}
$1.6 \rightarrow 0.6$	2	1×2^{-8}

$\Rightarrow (0.2)_{10} = (.\overline{0011})_2$

RAPPRESENTAZIONE FINITA DECIMALE $\rightarrow$ RAPPR. APPROSS. IN BINARIO : ERRORE DI ARROTONDAMENTO.

CONVERSIONE DA BASE 10 A BASE QUALUNQUE

SIA N UN NUMERO IN BASE 10 E CONVERTIAMOLO IN UN NUMERO IN BASE $B \in \mathbb{N}$.

SIA: $N = (a_j \cdot B^j) + (a_{j-1} \cdot B^{j-1}) + \dots + (a_1 \cdot B^1) + a_0 \cdot B^0 =$

$= (a_j a_{j-1} \dots a_1 a_0)_B$

DIVIDENDO N PER B SI HA:

$$\frac{N}{B} = (a_j \cdot B^{j-1}) + (a_{j-1} B^{j-2}) + \dots + (a_1 B^0) + \frac{a_0}{B} = N_0 + \frac{a_0}{B}$$

DOVE $N_0 < N$. $\Rightarrow N = B N_0 + a_0$

QUINDI a_0, L'ULTIMA CIFRA DELLA RAPPRESENTAZIONE, E' IL RESTO INTERO DI N/B.

PER OTTENERE LA PENULTIMA CIFRA SI DIVIDE N_0/B:

$$\frac{N_0}{B} = (a_j \cdot B^{j-2}) + \dots + (a_2 \cdot B^0) + \frac{a_1}{B} = N_1 + \frac{a_1}{B}$$

IL PROCEDIMENTO SI ARRESTA AD a_j : $N_j = 0$.

ES.: $741 = (a_9 a_8 a_7 \dots a_1 a_0)_2 = (1011100101)_2$

NELL'ARITMETICA IN VIRGOLA MOBILE (FLOATING POINT) SI HA IL PROBLEMA: DATO $x \in \mathbb{R}$ COME SCEGLIERE $fl(x) \in F = \{0\} \cup \{x \in \mathbb{R} : x \neq \left(\sum_{i=1}^{t} a_i b^{-i}\right) b^{\ell}$

SI HANNO I SEGUENTI CASI

1) $\ell < L$ UNDERFLOW $fl(x) = 0$ "WARNING"

2) $\ell > U$ OVERFLOW SEGNALE DI ERRORE E ARRESTO DEL PROGRAMMA

3) $L \leq \ell \leq U$

i) $a_M = 0$, $M > t+1 \Rightarrow x \in F$

ii) $x \notin F$, $fl(x) \neq x$ E SI HA:

a) CHOPPING O TRONCAMENTO: SI ESCLUDE LA PARTE A DESTRA DELLA t-ESIMA CIFRA

$$fl(x) = \pm\left(\sum_{i=1}^{t} a_i b^{-i}\right) b^{\ell}$$

$$fl(x) < x$$

b) ROUNDING O ARROTONDAMENTO:

$$x = \pm\left(\frac{a_1}{b} + \cdots + \frac{a_t}{b^t} + \frac{a_{t+1}}{b^{t+1}} + \cdots\right) b^{\ell}$$

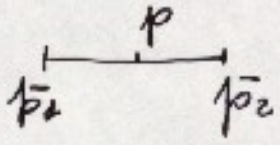
$$fl(x) = \pm\left(\frac{a_1}{b} + \cdots + \frac{a_{t-1}}{b^{t-1}} + \frac{\alpha_t}{b^t}\right) b^{\ell}$$

CON: $\alpha_t = \begin{cases} a_t & 0 \leq a_{t+1} < b/2 \\ a_t + 1 & \frac{b}{2} \leq a_{t+1} \leq b-1 \end{cases}$

$$fl(x) \gtrless x$$

CIOE' SI AGGIUNGE $\frac{1}{2}b^{-t}$ E SI TRONCA ALLA t-ESIMA CIFRA.

POICHE' LE MANTISSE DEI NUMERI MACCHINA $\bar{p}$, CHE SONO $b^{-1} \le |\bar{p}| < 1$, NON HANNO PIU' DI t CIFRE, LA DISTANZA TRA DUE MANTISSE DI MACCHINA CONSECUTIVE E' b^{-t}. QUINDI SE $\bar{p}_1, \bar{p}_2 > 0$, SE $\bar{p}_2$ E' CONSECUTIVO A $\bar{p}_1$ SI HA: $\bar{p}_2 = \bar{p}_1 + b^{-t}$.



QUINDI CON a) TUTTI TALI p VENGONO SOSTITUITI CON $\bar{p}_1$ E QUINDI $|p - \bar{p}_1| < b^{-t}$

INVECE CON b) TUTTE LE MANTISSE IN $(\bar{p}_1, \bar{p}_1 + \frac{1}{2}b^{-t})$ SONO SOSTITUITE CON $\bar{p}_1$ E TUTTE LE MANTISSE IN $(\bar{p}_1 + \frac{1}{2}b^{-t}, \bar{p}_2)$ CON $\bar{p}_2$. PERTANTO SE $\bar{p}$ E' LA MANTISSA ASSOCIATA A p SI HA:

$$|p - \bar{p}| \le \frac{1}{2}b^{-t}.$$

QUINDI, PER L' ERRORE ASSOLUTO SI HA:

$$|x - fl(x)| \le \begin{cases} b^{-t}b^{e} & \text{CHOPPING} \\ \frac{1}{2}b^{-t}b^{e} & \text{ROUNDING} \end{cases}$$

PER L' ERRORE RELATIVO:

$$\left|\frac{x - fl(x)}{x}\right| \le \begin{cases} b^{-t+1} & \text{CHOPPING} \\ \frac{1}{2}b^{-t+1} & \text{ROUNDING} \end{cases}$$

L'ERRORE RELATIVO GENERALMENTE E' PIU' IMPORTANTE:

CALCOLO A: $x = 0.5 \cdot 10^{-4}$, $x_c = 0.4 \cdot 10^{-4}$

$E_A = 0.1 \cdot 10^{-4}$

$E_R = 0.2$ 20%

CALCOLO B: $x = 5000$, $x_c = 4950$

$E_A = 50$

$E_R = 0.01$ 1%

SI DEFINISCE PRECISIONE O EPSILON DI MACHINA IL SUP DELL'ERRORE RELATIVO CIOE':

$$\varepsilon_m = \begin{cases} b^{-t+1} & \text{CHOPPING} \\ \frac{1}{2} b^{-t+1} & \text{ROUNDING} \end{cases}$$

NON HA SENSO CERCARE DELLE APPROSSIMAZIONI CON PRECISIONE INFERIORE AD ε_m.

DISTRIBUZIONE DEI NUMERI FLOATING-POINT

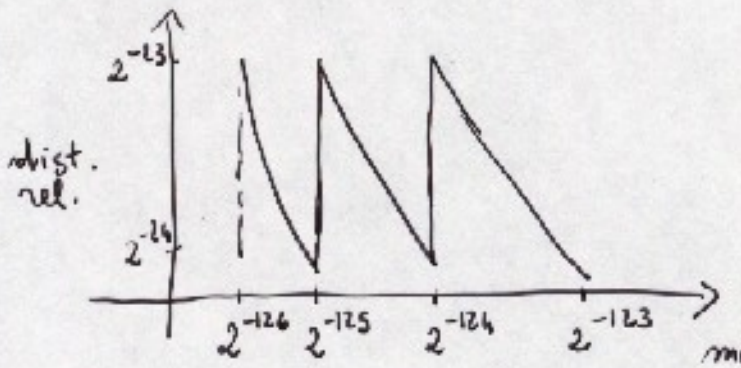
I NUMERI FLOATING-POINT NON SONO EQUISPAZIATI MA SI ADDENSANO IN PROSSIMITA' DEL PIU' PICCOLO NUMERO RAPPRESENTABILE. LA SPAZIATURA TRA DUE $x_1, x_2 \in F$ E' ALMENO $\beta^{-1}\,\varepsilon_M\,|x_2|$ ED AL PIU' VALE $\varepsilon_M\,|x_2|$ SE: $\varepsilon_M = \beta^{1-t}$

EPSILON MACCHINA (ε_M: $1+\varepsilon_M > 1$)

ALL'INTERNO DELL'INTERVALLO $[\beta^e, \beta^{e+1}]$ I PUNTI SONO INVECE EQUISPAZIATI E LA LORO DISTANZA E' β^{e-t}.

QUINDI OGNI VOLTA CHE SI AUMENTA (O DIMINUISCE) E DI UNA UNITA' SI HA UN AUMENTO (O DIMINUZIONE) DI UN FATTORE β DELLA DISTANZA TRA 2 NUMERI CONSECUTIVI. PER QUESTO SI PREDILIGONO BASI PICCOLE

IL FENOMENO (WOBBLING PRECISION) HA QUINDI UN ANDAMENTO OSCILLATORIO.



ALTRA FONTE DI ERRORI: IL RISULTATO DI OPERAZIONI ARITMETICHE PUO' NON ESSERE UN NUMERO DI MACCHINA.

$x = fl(x)$, $y = fl(y)$ SE: $x \,op\, y \notin F$ SI DEVE DEFINIRE $\boxed{OP}$ IL CUI RISULTATO $\in F$:

$$x \,\boxed{OP}\, y = fl(x \,op\, y)$$

VALIDA SE NON C'E' OVERFLOW.

ES.: SOMMA: SI RENDONO UGUALI GLI ESPONENTI SI SOMMANO LE MANTISSE IN ACCUMULATORE CON 2m CIFRE, SI AGGIUSTANO GLI ESPONENTI.

CIO' PUO' PORTARE ALLA NON VALIDITA' DELLE PROPRIETA' COMMUTATIVA, DISTRIBUTIVA ED ASSOCIATIVA DELLA SOMMA E DELLA MOLTIPLICAZIONE.

ESEMPIO: 4 CIFRE PER LA MANTISSA. $\sum_{i=1}^{11} x_i$

$x_1 = 0.5055 \cdot 10^4$, $x_i = 0.4000 \cdot 10^0$ $i = 2, \dots, 11$.

$x_1 + x_2 = 0.5055 \cdot 10^4 + 0.00004 \cdot 10^4 = 0.50554 \cdot 10^4 = 0.5055 \cdot 10^4$

E QUINDI: $\sum_{i=1}^{11} x_i = 0.5055 \cdot 10^4$

SE INVECE SI FA: $\sum_{i=2}^{11} x_i = 0.4 \cdot 10^1$ E:

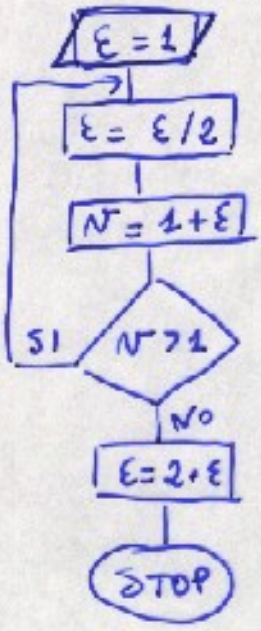
$$\sum_{i=1}^{11} x_i = x_1 + \sum_{i=2}^{11} x_i = 0.5055 \cdot 10^4 + 0.0004 \cdot 10^4 = 0.5059 \cdot 10^4$$

$\Rightarrow$ PER SOMMARE MOLTI NUMERI DELLO STESSO SEGNO SI DEVONO SOMMARE DAL PIU' PICCOLO AL PIU' GRANDE PER MINIMIZZARE LA PROPAGAZIONE DEGLI ERRORI.

DIAGRAMMA DI FLUSSO PER IL CALCOLO DI ε

ε E' IL PIU' PICCOLO NUMERO PER CUI SI HA:

$1 \boxplus \varepsilon > 1$



CODICE FORTRAN

```
      PROGRAM EPSMAC
C     CALCOLO APPROSS. DI E
      EPS = 1.
1     EPS = EPS/2.
      EPS1 = EPS+1.
      IF (EPS1.GT.1.) GOTO 1
      EPS = 2.*EPS
      END
```

MICRO VAX : $\varepsilon = 1.19 \cdot 10^{-7}$ SINGOLA PRECISIONE

$\varepsilon = 2.22 \cdot 10^{-16}$ DOPPIA "

CALCOLO BASE

$m \boxplus \varepsilon > m \qquad m = 1, \ldots, b-1$

$b \boxplus \varepsilon = m$

```
      PROGRAM BASE
      READ(*,*) EPS
      N=1
1     N=N+1
      XN=N+EPS
      IF(XN.GT.N) GOTO 1
      BASE=N
      END
```

CONDIZIONAMENTO E STABILITA'

CONSIDERIAMO IL PROBLEMA: TROVARE x TALE CHE:

$$F(x) = d \qquad (1)$$

DOVE d E' IL DATO O I DATI DA CUI DIPENDE LA SOL. x ED F LA RELAZIONE FUNZIONALE CHE LEGA x E d. DIREMO CHE TALE **PROBLEMA E' BEN POSTO** SE, PER UN CERTO DATO, LA SOLUZIONE ESISTE, E' UNICA E DIPENDE CON CONTINUITA' DAI DATI.
LA DIPENDENZA CONTINUA DAI DATI SIGNIFICA CHE PICCOLE PERTURBAZIONI SUI DATI DANNO LUOGO A A PICCOLE VARIAZIONI DELLA SOLUZIONE, DOVE "PICCOLO" PUO' ESSERE INTESO IN SENSO RELATIVO O ASSOLUTO.

D'ALTRONDE ABBIAMO VISTO CHE SI COMMETTONO ERRORI SIA NEL RAPPRESENTARE I NUMERI REALI, SIA NELL'ESECUZIONE DI OPERAZIONI ARITMETICHE.

NASCONO A QUESTO PUNTO DUE PROBLEMI. CI SI CHIEDE:
1) ALTERANDO I DATI DEL PROBLEMA DI QUANTO SI ALTERA LA SOLUZIONE;
2) COME SI PROPAGANO GLI ERRORI.

IL PRIMO PROBLEMA E' CONNESSO CON LA DIPENDENZA CONTINUA DAI DATI DELLA SOLUZIONE E PUO' ESSERE STIMATO CON IL **NUMERO DI CONDIZIONAMENTO** DEL PROBLEMA, NUMERO CHE NON DIPENDE DALL'USO DELLA ARITMETICA FINITA DEL CALCOLATORE MA DAL TIPO DI PROBLEMA (1).

SUPPONIAMO DI ALTERARE DI δd I DATI DI (d) E CI CHIEDIAMO DI QUANTO SI ALTERA LA SOLUZIONE:

$$d + \delta d \rightarrow x + \delta x$$

DIREMO NUMERO DI CONDIZIONAMENTO RELATIVO:

$$K = \frac{\|\delta x\| / \|x\|}{\|\delta d\| / \|d\|}$$

SE $x=0$, $d=0$ SI CALCOLA IL NUMERO DI COND. ASSOLUTO:

$$K_{ass} = \|\delta x\| / \|\delta d\|$$

SE K E' GRANDE IL PROBLEMA E' MAL CONDIZIONATO
SE UN PROBLEMA E' BEN POSTO MA K E' GRANDE BASTA RIFORMULARE IL PROBLEMA.

IL SECONDO PROBLEMA DIPENDE DALLA STABILITA' DELL'ALGORITMO. RICORDIAMO CHE PER ALGORITMO INTENDIAMO UNA SUCCESSIONE DI PASSI CHE TRASFORMI UN VETTORE DI DATI IN UN CORRISPONDENTE OUTPUT. AD OGNI PROBLEMA NUMERICO SI POSSONO ASSOCIARE PIU' ALGORITMI. UN ALGORITMO E' STABILE SE LA PROPAGAZIONE DEGLI ERRORI, DOVUTI ALLA ARITMETICA DI MACCHINA, E' LIMITATA. UN ALGORITMO E' PIU' STABILE DI UN ALTRO SE IN ESSO L'INFLUENZA DEGLI ERRORI E' MINORE.

QUALE DELLE 4 OPERAZIONI PUO' PROVOCARE UNA PERDITA DI PRECISIONE?

ABBIAMO DETTO CHE I RISULTATI DI OPERAZIONI ARITMETICHE TRA NUMERI DI MACCHINA IN GENERALE NON SONO NUMERI DI MACCHINA. L'OPERAZIONE DI MACCHINA ASSOCIA A DUE NUMERI DI MACCHINA UN TERZO NUMERO DI MACCHINA.

INDICHIAMO CON $\oplus\ \ominus\ \odot\ \oslash$ LE OPERAZIONI DI MACCHINA:

$\bar{a} = fl(a)$
$\bar{b} = fl(b)$

$$\bar{a} \oplus \bar{b} = fl(\bar{a}+\bar{b}) = (\bar{a}+\bar{b})(1+\varepsilon_1)$$
$$\bar{a} \ominus \bar{b} = fl(\bar{a}-\bar{b}) = (\bar{a}-\bar{b})(1+\varepsilon_2)$$
$$\bar{a} \odot \bar{b} = fl(\bar{a}\cdot\bar{b}) = (\bar{a}\cdot\bar{b})(1+\varepsilon_3)$$
$$\bar{a} \oslash \bar{b} = fl(\bar{a}/\bar{b}) = (\bar{a}/\bar{b})(1+\varepsilon_4)$$

$$|\varepsilon_i| \le eps \equiv \varepsilon_m$$

ESAMINIAMO L'ERRORE RELATIVO:

$$\left| \frac{(x_1\ op\ x_2) - (\bar{x}_1\ op\ \bar{x}_2)}{x_1\ op\ x_2} \right| = \left| \frac{(x_1\ op\ x_2) - [x_1(1+\varepsilon_1)]\ op\ [x_2(1+\varepsilon_2)]}{x_1\ op\ x_2} \right|$$

PRODOTTO: $\dfrac{x_1 x_2 - x_1(1+\varepsilon_1)\cdot x_2(1+\varepsilon_2)}{x_1 x_2} = -\varepsilon_1 - \varepsilon_2 - \varepsilon_1\varepsilon_2 \sim -\varepsilon_1 - \varepsilon_2$

DIVISIONE: $\dfrac{\frac{x_1}{x_2} - \frac{x_1(1+\varepsilon_1)}{x_2(1+\varepsilon_2)}}{x_1/x_2} = \dfrac{-\varepsilon_1+\varepsilon_2}{1+\varepsilon_2} \sim -\varepsilon_1+\varepsilon_2$

SOMMA ALGEBRICA: $\dfrac{(x_1+x_2) - [x_1(1+\varepsilon_1)+x_2(1+\varepsilon_2)]}{x_1+x_2} = -\dfrac{x_1}{x_1+x_2}\varepsilon_1 - \dfrac{x_2}{x_1+x_2}\varepsilon_2$

$\left|\dfrac{x_i}{x_1+x_2}\right| \to \infty$ per $x_1+x_2 \to 0$

ERRORE ASSOLUTO ED ERRORE RELATIVO NELLE QUATTRO OPERAZIONI

$$\delta x^* = x - x^*$$

$$\varepsilon_{x^*} = \frac{\delta x^*}{x^*}$$

OPERAZIONE	ERR ASS	ERR REL
$+$	$\delta x_1 + \delta x_2$	$\frac{x_1}{x_1+x_2}\varepsilon_{x_1} + \frac{x_2}{x_1+x_2}\varepsilon_{x_2}$
$-$	$\delta x_1 - \delta x_2$	$\frac{x_1}{x_1-x_2}\varepsilon_{x_1} + \frac{x_2}{x_1-x_2}\varepsilon_{x_2}$
$*$	$x_2\delta x_1 + x_1\delta x_2$	$\varepsilon_{x_1} + \varepsilon_{x_2}$
$:$	$\frac{\delta x_1}{x_2} - \frac{x_1}{x_2^2}\delta x_2$	$\varepsilon_{x_1} - \varepsilon_{x_2}$

$+$ $-$ NESSUN PROBLEMA PER ERR ASS, MENTRE ERR RG GRANDE SE $x_1 \approx x_2$ (FENOMENO DI CANCELLAZIONE)

$*$ OK ERR REL ERR ASS DIPENDE DALL'ORDINE DI GRANDEZZA DEI FATTORI

$:$ OK ERR REL ERR ASS GRANDE SE $x_2 \approx 0$

CANCELLAZIONE

ESEMPIO $f(x) = x \cdot (\sqrt{x+1} - \sqrt{x})$

CON 6 CIFRE DECIMALI PER LA MANTISSA.

x	$\tilde{f}(x)$	$f(x)$
1	0.414210	0.414214
10	1.54340	1.54347
10^2	4.99000	4.98756
10^3	15.8000	15.8074
10^4	50.0000	49.9988
10^5	100.000	158.113

LE CIFRE SOTTOLINEATE SONO AFFETTE DA ERRORE.
PER 10^5 L'INFORMAZIONE E' PERSA.
PER EVITARE L'ERRORE, COMMESSO NEL CALCOLARE LA DIFFERENZA DI NUMERI VICINI SI PUO' PROCEDERE COSI':

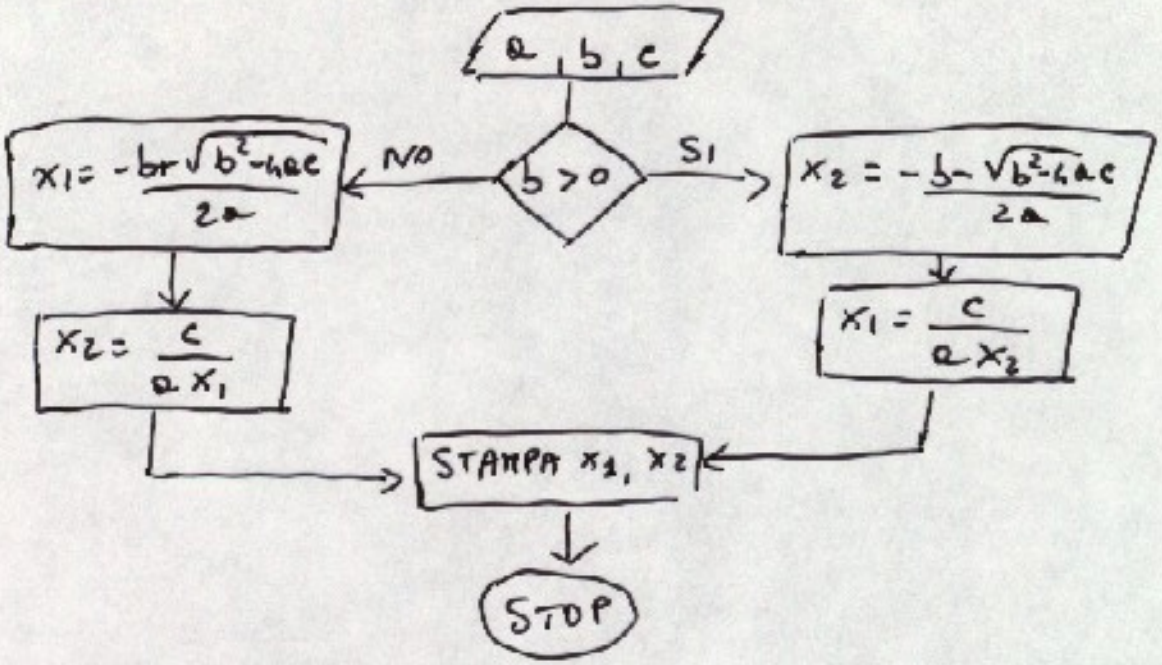
$$f(x) = x(\sqrt{x+1} - \sqrt{x}) = x\frac{(\sqrt{x+1} - \sqrt{x})(\sqrt{x+1} + \sqrt{x})}{\sqrt{x+1} + \sqrt{x}} = x\frac{x+1-x}{\sqrt{x+1} + \sqrt{x}} =$$

$$= \frac{x}{\sqrt{x+1} + \sqrt{x}}$$

ALTRO ESEMPIO: $p(x) = ax^2 + bx + c$

$$x_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}, \quad x_2 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

ALGORITMO STABILE:



SCHEMA DI HÖRNER (NESTED MULTIPLICATION)

SI VUOLE ESEGUIRE IL CALCOLO DI:

$$f(x) = ax^3 + bx^2 + cx + d$$

SONO NECESSARIE 3 SOMME E 6 MOLTIPLICAZIONI PER UN TOTALE DI 9 OPERAZIONI FLOATING POINT (9 flops)

IN GENERALE PER:

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$$

OCCORRONO:

n SOMME

$\frac{n(n+1)}{2}$ MOLTIPLICAZIONI

PER UN TOTALE DI: $\frac{n(n+1)}{2} + n = \frac{n(n+3)}{2} \sim n^2$ flops

SE INVECE SI APPLICA LO SCHEMA DI HÖRNER:

$$f(x) = a_0 + x(a_1 + x(a_2 + \dots + x(a_{n-1} + x a_n) \dots)))$$

n SOMME, n moltiplicazioni. # flops $\sim n$.

- ESEMPIO DI PROBLEMA MAL CONDIZIONATO

$$\begin{cases} x - y = 1 \\ x - 1.00001\, y = 0 \end{cases}$$

LA SOL. E': $x = 100001$, $y = 100000$

MA:
$$\begin{cases} x - y = 1 \\ x - 0.99999\, y = 0 \end{cases}$$

HA SOL.:
$$x = -99999, \quad y = -100000$$

PERTANTO UNA VARIAZIONE DI $2 \cdot 10^{-5}$ NEI DATI HA PORTATO AD UN CAMBIAMENTO DI $2 \cdot 10^{5}$ NELLA SOL. E IL # DI CONDIZ. E' QUINDI DATO DA: $1 \cdot 10^{10}$

- ESEMPIO DI ALGORITMO INSTABILE

CALCOLARE:
$$s_n = 10^6 + \sum_{i=1}^{n} \frac{1}{i} \qquad n = 1, \dots, N$$

OVVERO:
$$s_0 = 10^6, \quad s_n = s_{n-1} + \frac{1}{n} \qquad n = 1, \dots, N.$$

ALGORITMO 1:

```
s = 10^6
Do n = 1, N
s = s + 1/n
END Do
```

PER $n \geq 7,33$ I VALORI NON CAMBIANO PIU'

ALGORITMO 2:

```
s_0 = 0
Do n = 1, N
s_n = s_{n-1} + 1/n
END DO
s_n = s_n + 10^6
```