1

RAPPRESENTAZIONE FLOATING-POINT DI UN NUMERO REALE a

$a = (-1)^s \, p \, N^q$ $\quad s \in \{0, 1\}$, $p \geq 0$ REALE, q INTERO

$N \rightarrow$ BASE DEL SISTEMA DI NUMER.

$s = 0 \Rightarrow a > 0$
$s = 1 \Rightarrow a < 0$

$(N, s, p, q) \Rightarrow a$ ⚠ NON VALE IL CONTRARIO

$a = 0.015 \cdot 10^{-1} = 0.15 \cdot 10^{-2} = 0.0015 \cdot 10^{0}$

- RAPP. F.P. NORMALIZZATA

$N^{-1} \leq p < 1$ SE VIENE SODDISFATTA LA SEGUENTE CONDIZIONE p e q SONO UNIVOCAMENTE DETERMINATI
E p (MANTISSA DI A)
q (ESPONENTE / CARATTERISTICA DI A)

ES. SE $N = 10$ $\quad 0.1 \leq p < 1$

$a = 0.15 \cdot 10^{-2}$ $\quad p = 0.15 \quad q = -2$
$a = 12.4 \cdot 10^{0} \longrightarrow a = 0.124 \cdot 10^{2}$ $\quad p = 0.124 \quad q = 2$
$a = -0.0013 \cdot 10^{4} \longrightarrow a = -0.13 \cdot 10^{2}$ $\quad p = 0.13 \quad q = 2 \quad s = 1$

SE F.P. NOR
$(s, p, q) + N \Leftrightarrow a \in R$

NUMERI DI MACCHINA

CALCOLATORE $\Rightarrow$ SPAZIO FINITO DI MEMORIA

$p \rightarrow$ PUÒ AVERE AL MAX t CIFRE NEL SISTEMA DI NUMERAZIONE SCELTO

$q \rightarrow L \leq q \leq U \quad L < 0$ e $U > 0$ INTERI

N.M. $\Rightarrow$ NUMERI CON MANTISSA ED ESPONENTE ESATTAMENTE RAPPRESENTABILI NEGLI SPAZI A LORO RISERVATI DAL CALCOLATORE

INSIEME DEI NUMERI MACCHINA $\quad F = \{0\} \cup \{(-1)^s \, 0.\underbrace{a_1 a_2 \ldots a_t}_{\text{CIFRE DOPO LA VIRGOLA}} \cdot N^q, \; 0 \leq a_i < N, \; L \leq q \leq U\}$
$a_1 \neq 0$

ES. $N = 10 \quad t = 5 \quad L = -127 \quad U = 128$
È UN NUMERO DI MACCHINA?

t MAX NUMERO DI CIFRE DELLA MANTISSA RAPPRESENTABILI

$a = 1.58291 \cdot 10^{0} = 0.\underbrace{158291}_{6} \cdot 10^{1}$ NO

$a = 0.0038245 \cdot 10^{0} = 0.\underbrace{38245}_{5} \cdot 10^{-2}$ SÌ

$a = 12.29 \cdot 10^{128} = 0.1229 \cdot 10^{\underbrace{130}_{>128}}$ NO

PIÙ PICCOLO NUMERO POSITIVO, RAPP. COME NUM. DI MAC

$$m = 0.\underbrace{1000\ldots0}_{t \text{ cifre}} \cdot N^{L}$$

ES. $N = 10 \quad t = 8 \quad L = -125$

$$m = 0.1\underbrace{0000000}_{t-1 \text{ cifre}} \cdot 10^{-125}$$

$8 - 1 \rightarrow 7$ zeri

REGIONE DI UNDERFLOW
↓
INSIEME DI NUMERI REALI $\neq 0$ $(-m, m)$ NON È UN UNICO INT. $(-m, 0) \cup (0, m)$

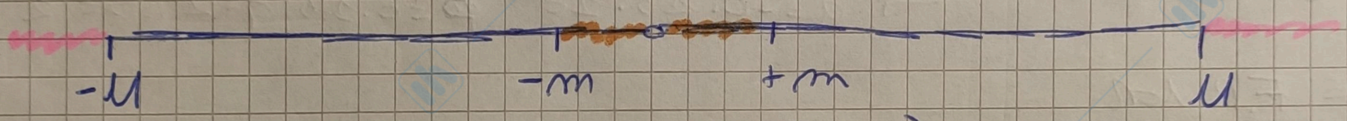
PIÙ GRANDE NUMERO POS., RAP. COME NUM. DI MAC.

$$M = 0.\underbrace{N-1\; N-1\; N-1}_{t \text{ cifre}} \cdot N^{U}$$

ES. $N = 10, \; t = 8, \; U = 128$

$$M = 0.\underbrace{99999999}_{8 \text{ volte } N-1} \cdot 10^{128}$$

REGIONE DI OVERFLOW
↓
INSIEME DI NUMERI APPARTENENTI A $(-\infty, -M) \cup (M, +\infty)$



→ UNDERFLOW $(-m, 0) \cup (0, m)$ ⟹ APPROSSIMATI A 0

→ OVERFLOW $(-\infty, -M) \cup (M, +\infty)$ ⟹ APPROSSIMATI A INF

PER OTTENERE UN RISULTATO BISOGNA EVITARE L'OVERFLOW

TECNICA DI ARROTONDAMENTO "ROUND TO EVEN"

PERMETTE DI RAPPRESENTARE I NUMERI REALI, LA CUI MANTISSA HA PIÙ DI t CIFRE, ARROTONDANDO p (MANTISSA) A t CIFRE

I NUMERI DI MAC LI INDICHEREMO CON $\bar{a}$

$$a = (-1)^{s} p N^{q}$$
$$\bar{a} = (-1)^{s} \bar{p} N^{q}$$

T. R. TO EVEN → APPROSSIMA AL N. M. PIÙ VICINO
→ SE EQUIDISTANTE, APPROSSIMA AL N. DI M. CON L'ULTIMA CIFRA PARI

ES. $N = 10 \quad t = 3$

$p = 0.15814, \; \bar{p} = 0.158$

$p = 0.158\underline{5}432, \; \bar{p} = 0.159$

$p = 0.158\underline{8}12, \; \bar{p} = 0.159$

$p = 0.158\underline{5}, \; \bar{p} = 0.158$

$p = 0.159\underline{5}, \; \bar{p} = 0.160$

SE a È UN N. DI MAC $a = \bar{a}$ SE NO $a \approx \bar{a}$

SE MEMORIZZIAMO $\bar{a}$ AL POSTO DI a STIAMO COMMETTENDO UN ERRORE

ERRORI CHE COMMETTIAMO CON L'ARROTONDAMENTO

ERRORE ASSOLUTO $e_A = |x - \tilde{x}|$ ERRORE RELATIVO $e_r = \frac{|x - \tilde{x}|}{|x|}$, $x \neq 0$

SE X NON È UN NUMERO REALE, AD ES. UN VETTORE O UNA MATRICE, $\tilde{x}$ DENOTA ANCORA LA SUA APPROSSIMAZIONE.

⚠ DEVO SOSTITUIRE, PERÒ, IL VALORE ASSOLUTO CON UNA NORMA OPPORTUNA

E. RELATIVO → INDICAZIONE PIÙ REALISTICA

ES
$x = 10^3$ $\tilde{x} = 999$ $e_A = 1$ $e_r = 10^{-3}$
$x = 10^6$ $\tilde{x} = 999999$ $e_A = 1$ $e_r = 10^{-6}$

ERR. DI ARROTONDAMENTO QUANDO SOSTITUISCO UN NUM. REALE $a \neq 0$ CON IL CORRISPONDENTE NUM. DI MAC $\bar{a}$

APPROSSIMAZIONE ERRORI

$|a - \bar{a}| \le \frac{1}{2} N^{q-t}$ (ER. ASSOLUTO)

$\frac{|a - \bar{a}|}{|a|} \le \frac{1}{2} N^{1-t}$ (ER. RELATIVO)

L'ERRORE REL È SEMPRE ≤ ALLA PRECISIONE DI MAC

- EPSILON DI MACCHINA
  $EPS = N^{1-t}$
- PRECISIONE DI MACCHINA
  $\varepsilon_m = \frac{1}{2} N^{1-t}$ → MAX ERRORE REL. CHE SI PUÒ AVERE

$\varepsilon = \frac{\bar{a} - a}{a}$ → $\bar{a} = a(1 + \varepsilon)$

$|\varepsilon| \le \varepsilon_m$

MATLAB → BASE 2, 64 BIT
→ 1 BIT SEGNO MANTISSA
→ 52 BIT CIFRE MANT.
→ 11 BIT ESPONENTE

COMANDI

REALMIN $m \approx 2.2 \cdot 10^{-308}$ EPS $2.2 \cdot 10^{-16}$

REALMAX $M \approx 1.8 \cdot 10^{308}$ ε = EPS/2 $1.1 \cdot 10^{-16}$

REALMAX → PIÙ GRANDE NUM POSITIVO DI MAC

REALMIN → PIÙ PICCOLO NUM DI MAC (POS)

ε → MAX ERRORE RELATIVO DOVUTO ALL'ARROTONDAMENTO

OPERAZIONI DI MACCHINA

NON È POSSIBILE ESEGUIRE ESATTAMENTE LE OPER. ARITMETICHE
↓
OPERAZIONI DI MACCHINA $\oplus, \ominus, \otimes$ e $\oslash$
↓
ASSOCIA A DUE N. DI M. UN TERZO NUM. DI MAC., OTTENUTO ARROTONDANDO L'ESATTO RISULTATO DELL'OPERAZIONE IN QUESTIONE

$\bar{a}_1 \oplus \bar{a}_2 = \overline{\bar{a}_1 + \bar{a}_2}$

$N = 10$ $t = 4$ $\bar{a}_1 = 0.5823$ $\bar{a}_2 = 0.6214$ $\bar{a}_1 + \bar{a}_2 = 1.2037 = 0.1204 \cdot 10^1$

- VALE SOLO LA PR. COMMUTATIVA $\bar{a}_1 \oplus \bar{a}_2 = \bar{a}_2 \oplus \bar{a}_1$
$\bar{a}_1 \otimes \bar{a}_2 = \bar{a}_2 \otimes \bar{a}_1$

- RELAZIONE ANOMALA

$\bar{a}_1 \oplus \bar{a}_2 = \bar{a}_1$, QUANDO $|\bar{a}_2| << |\bar{a}_1|$

IN PARTICOLARE $|\bar{a}_2| < |\bar{a}_1|$ EPS

IL VALORE DI $|\bar{a}_2|$ È TROPPO PICCOLO RISPETTO A $|\bar{a}_1|$ E NON CONTRIBUISCE ALLA SOMMA

EQUIVALENZA

DUE QUANTITÀ e_1 ED e_2 SI DICONO EQUIVALENTI SE

$$\frac{|\bar{e}_1 - \bar{e}_2|}{|\bar{e}_1|} \lesssim \varepsilon_m \qquad \frac{|\bar{e}_1 - \bar{e}_2|}{|\bar{e}_2|} \lesssim \varepsilon_m$$

N STESSO ORDINE DI GRANDEZZA

ε_m → DIPENDE DALLO CON, IN MAT $1.1 \cdot 10^{-16}$

ES $e_1 = 0.999999$ $e_2 = 1$ $N = 10$ $t = 5$

$$\frac{|\bar{e}_1 - \bar{e}_2|}{|\bar{e}_2|} = \frac{|0.99999 - 1|}{|1|} = 0.1 \cdot 10^{-4} \leq \frac{1}{2} N^{1-t} = 0.5 \cdot 10^{-4}$$

ES MATLAB

$x = 1.0e-11$

CANCELLAZIONE NUMERICA

⚠ NON SI VERIFICA MAI CON ESPONENTI DIVERSI

CONSEGUENZA DELLA RAPPRESENTAZIONE CON PRECISIONE FINITA DEI NUMERI REALI

$N = 10$ $t = 5$ $\varepsilon_m = \frac{1}{2} N^{1-t} = 0.5 \cdot 10^{-4}$

CASO 1

$a_1 = 0.15782$ $a_2 = 0.15735$ $a_1 = \bar{a}_1$ $a_2 = \bar{a}_2$

SONO ENTRAMBI N. DI M

ARIT. ESATTA

$a_1 - a_2 = 0.47 \cdot 10^{-3}$
$\bar{a}_1 \ominus \bar{a}_2 = 0.47 \cdot 10^{-3}$

ARIT. FINITA

STESSO RISULTATO

CASO 2

$a_1 = 0.157824831$ $\bar{a}_1 = 0.15782$
$a_2 = 0.157348212$ $\bar{a}_2 = 0.15735$

$a_1 - a_2 = 0.476619 \cdot 10^{-3}$
$\bar{a}_1 \ominus \bar{a}_2 = 0.47000 \cdot 10^{-3}$

PERDITA DI 3 CIFRE SIGN.!

- LA C.N. CONSISTE IN UNA PERDITA DELLE CIFRE SIGNIFICATIVE DELLA MANTISSA, SI VERIFICA QUANDO SI ESEGUE L'OP DI SOTTRAZIONE FRA DUE RAPP. DI MACCHINA $\bar{a}_1$ e $\bar{a}_2$

$s_1 = s_2$, $q_1 = q_2$, $\bar{p}_1 \simeq \bar{p}_2$, $\bar{p}_1 \neq p_1$ e/o $\bar{p}_2 \neq p_2$

MATLAB

CONDIZIONAMENTO PROBLEMA NUMERICO

Bisogna distinguere il ruolo del problema dal ruolo dell'algoritmo utilizzato per risolvere il problema

- Un problema è ben condizionato se l'errore relativo associato a $\bar{y}$ (risultati ottenuti a partire dai dati $\bar{x}$) è dello stesso ordine di grandezza dell'errore relativo associato a $\bar{x}$ (perturbazione dei dati in input) o minore, altrimenti è mal condizionato

$$\boxed{\frac{|y-\bar{y}|}{|y|} \leq K(f,x)\,\frac{|x-\bar{x}|}{|x|}}$$

Più piccolo valore per cui vale la sup. dis (è sempre positivo)

Se i dati sono un vettore o una matrice, sostituisco al valore assoluto l'opportuna norma

BEN CONDIZIONATO $K(f,x) < 10$

MAL CONDIZIONATO $K(f,x) > 10$

$K(f,x) \geq 0$

ES

$y = x_1 + x_2$ $\quad$ $\bar{x}_1$ e $\bar{x}_2$ → dati perturbati, l'err. relativo, aumenterà

$\bar{y} = \bar{x}_1 + \bar{x}_2$

$\bar{x}_1 = x_1(1+\varepsilon_1)$ $\quad$ $\bar{x}_2 = x_2(1+\varepsilon_2)$

$$\frac{y-\bar{y}}{y} = \frac{x_1+x_2-(\bar{x}_1+\bar{x}_2)}{x_1+x_2} = \frac{x_1+x_2-x_1(1+\varepsilon_1)-x_2(1+\varepsilon_2)}{x_1+x_2}$$

$$= -\frac{x_1}{x_1+x_2}\varepsilon_1 - \frac{x_2}{x_1+x_2}\varepsilon_2 \qquad K_i = \left|\frac{x_i}{x_1+x_2}\right|$$

$$\frac{|y-\bar{y}|}{|y|} \leq K_1|\varepsilon_1| + K_2|\varepsilon_2| \leq K\left(\frac{|x_1-\bar{x}_1|}{|x_1|} + \frac{|x_2-\bar{x}_2|}{|x_2|}\right)$$

Se $x_1 + x_2 \to 0$ allora $K \to \infty$

Il problema sarà mal condizionato quando x_1 e x_2 sono opposti e simili in modulo

STABILITÀ DI UN ALGORITMO

$\bar{x}$ arrotondamento dei dati x di input

$\bar{y}$ risultati algoritmo ottenuti a partire dai dati $\bar{x}$ in precisione infinita di calcolo

$\tilde{y}$ ris. alg. ottenuti a partire dai dati $\bar{x}$ in precisione finita di calcolo

Un algoritmo è numericamente stabile se l'errore relativo associato al risultato $\tilde{y}$ ha lo stesso ordine di grandezza della precisione di macchina o minore, altrimenti si dice instabile.