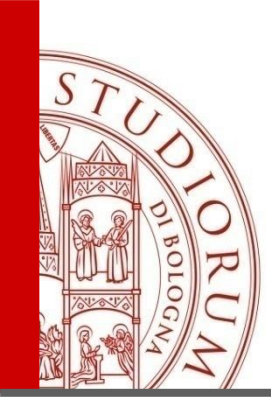


Numeri Finiti

…un’approssimazione inevitabile…

Sistemi di numerazione

Sistema di numerazione posizionale

$(0.1)_{10}=(0.0011\ 0011\ 0011\ \ldots)_2$

Base 10, numeri interi

$37294 = 4+90+200+7000+30000$

$= 4 \times 10^0+9 \times 10^1+2 \times 10^2+7 \times 10^3+3 \times 10^4$

In generale:

$$a_n \ldots a_0 = \sum_{k=0}^{n} a_k 10^k$$

Sistemi di numerazione

Base 10, numeri reali

$$0.7217 = 7 \times 10^{-1} + 2 \times 10^{-2} + 1 \times 10^{-3} + 7 \times 10^{-4}$$

In generale:

$$\pm a_n \ldots a_0 \, . \, b_1 b_2 \ldots = \pm \sum_{k=0}^{n} a_k \, 10^k + \sum_{k=1}^{\infty} b_k \, 10^{-k}$$

Nota: esistono numeri reali, es. irrazionali, che hanno un numero infinito di cifre decimali in ogni base razionale, es. $\pi, \sqrt{2}$

Se in una base sono necessarie infinite cifre ⇏ Numero irrazionale

⇐

$(0.1)_{10}$ non ha una rappresentazione finita in base 2

Altre basi

- Base 8, cifre 0, 1, 2, 3, 4, 5, 6, 7

$$(0.36207)_8 = 3 \times 8^{-1} + 6 \times 8^{-2} + \ldots = (0.47286\ldots)_{10}$$

- Base 2, cifre 0, 1, o sui calcolatori "*off*" e "on",
  "**bit**" = binary digit

$$(0.111)_2 = 1 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3} = (0.875)_{10}$$

- Base B

$$\pm(a_n \ldots a_0 . b_1 b_2 \ldots)_B = \pm \sum_{k=0}^{n} a_k B^k + \sum_{k=1}^{\infty} b_k B^{-k}$$

Numeri reali: Notazione scientifica normalizzata

32.213 → 0.003221×10^4 non normalizzata

0.3221×10^2 normalizzata

Sia x un numero reale non nullo e B (base) un numero intero, allora esiste una rappresentazione univoca di x in base B:

$$x = \pm 0.d_1 d_2 .... \times B^p = \pm \left(\sum_{i=1}^{\infty} d_i B^{-i} \right) B^p$$

$$x = \pm m \times B^p ,$$

dove:

$$0 \le d_i \le B-1, \qquad d_1 \neq 0$$

$$\frac{1}{B} \le m < 1$$

m Mantissa B base p esponente

Insieme dei numeri finiti

NON TUTTI I NUMERI SONO RAPPRESENTABILI DAL CALCOLATORE

Numeri di macchina:

- un insieme finito di numeri interi:
  SIGNED/UNSIGNED INTEGER
- un insieme finito di numeri reali:
  NUMERI A PUNTO MOBILE
  (**floating point**)

Numeri interi

Supponiamo:
base B=10, spazio di memoria riservato di lunghezza t+1=8

Esempi:

- X=+1320 è rappresentato come 00001320
- X=+9999999 è rappresentato come 09999999

X>+9999999 (B^t -1) provocano errore di OVERFLOW

Numeri interi **negativi**

1)Rappresentazione complemento alla base:

Un numero negativo **–x** viene rappresentato in base B in t+1 posizioni come:

$$B^{t+1} - x$$

Esempi:

- X=-31152 è rappresentato come 99968848

100000000- 31152= 99968848

- X=-10000000 è rappresentato come 90000000

X<-10000000 (-B^t) provocano errore di OVERFLOW

Numeri reali: floating point

- In generale la rappresentazione normalizzata in base B di un numero reale al calcolatore è una rappresentazione univoca data da:

$$x = \pm 0.d_1 d_2 \ldots . \times B^p = \pm \left(\sum_{i=1}^{t} d_i B^{-i} \right) B^p$$

in cui la mantissa è rappresentata da un numero finito di cifre **t**

$$0 \le d_i \le B-1, \qquad d_1 \neq 0$$

$$x = \pm m \times B^p, \qquad \frac{1}{B} \le m < 1$$

Segno	*esponente* p	*mantissa* m

Numeri reali: floating point

ESPONENTE: rappresentazione per traslazione

Si aggiunge il fattore costante $\frac{B^i}{2}$ dove *i* sono le cifre riservate all'esponente

ESEMPIO : B=10 $\frac{10^2}{2} = 50$

x=.1039 x 10^{-6} è rappresentato come 04401039

Nell'esempio gli esponenti da –50 a 49
vengono memorizzati con i valori da 00 a 99

NUMERI FINITI

- Insieme finito di numeri normalizzati a virgola mobile:

$$F(B, t, L, U) =$$

$$\left\{ x \in \Re : x = \pm\left(\sum_{i=1}^{t} d_i B^{-i}\right) B^{p} \right\} \cup \{0\}$$

$$0 \le d_i \le B-1, \qquad d_1 \neq 0$$

$$\boldsymbol{L \le p \le U}, \qquad \boldsymbol{L < 0, U > 0}$$ Generalmente L=-U

Non tutti i numeri reali sono rappresentabili in F, vediamo le motivazioni:

NUMERI FINITI

Poiché un numero finito ha a disposizione solo un numero limitato di bits, nel rappresentare un numero reale x si possono verificare i seguenti casi

I. Esponente $L \le p \le U, \quad L < 0, U > 0$

Se $p \notin [L, U]$

p > U errore di overflow

p < L errore di underflow

II. Mantissa, t cifre disponibili

Se il numero di cifre nella mantissa è superiore a t:
I numeri NON sono esattamente rappresentabili, occorre approssimarli mediante i criteri:
troncamento o arrotondamento

Approssimazione per TRONCAMENTO

$$\mathbf{f}l_T(x) = \pm\left(d_1 B^{-1} + d_2 B^{-2} + \ldots + d_t B^{-t}\right) B^{p}$$

- Esempio base B=10 t=4

  X=0.372145

  troncamento

$$fl_T(x) = \mathbf{0.3721}$$

Approssimazione per ARROTONDAMENTO

$$\mathrm{fl}_A(x) = \pm\left(d_1 B^{-1} + d_2 B^{-2} + \ldots + \tilde{d}_t B^{-t}\right) B^p$$

$$\tilde{d}_t = \begin{cases} d_t & se\ d_{t+1} < \dfrac{B}{2} \\ d_t + 1 & se\ d_{t+1} \geq \dfrac{B}{2} \end{cases}$$

approssimazione al più vicino

Approssimazione per ARROTONDAMENTO

- Esempio B=10 t=4 $\frac{1}{2}B^{-t} = 0.00005$

x=0.3798165
$+0.00005=$
0.3798665

$fl_A(x) = \mathbf{0.3798}$

y=0.1265873
$+0.00005=$
0.1266337

$fl_A(y) = \mathbf{0.1266}$

Numeri finiti: esempi di approssimazione

Sia B=10, t=5,

$$\frac{1}{2}B^{-t} = 0.5\times10^{-5} \qquad \begin{array}{l} .314159226+ \\ .000005 \quad = \\ .\underbrace{31416}4226\overline{6} \end{array}$$

$\pi = 3.14159226$ $\quad fl_T(\pi) = 0.31415\times10^{1}$ troncamento

$\quad fl_A(\pi) = 0.31416\times10^{1}$ arrotondamento

Insieme dei numeri finiti

$F(B, t, L, U)$ contiene $2(U-L+1)(B-1)B^{t-1}+1$

F non è una perfetta simulazione di R

- Non sono uniformemente distribuiti sull'asse reale
- La loro densità decresce con l'aumentare del valore assoluto del numero
- Tutti i numeri reali compresi tra due consecutivi numeri finiti vengono approssimati da uno dei due valori

Esempio

Sia F(2,3,-1,2) allora 0.100, 0.101, 0.110, 0.111 sono tutte le possibili mantisse, p in [-1,2]

$$0.100 \cdot 2^{-1} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \quad 0.101 \cdot 2^{-1} = (\frac{1}{2} + \frac{1}{8}) \cdot \frac{1}{2} = \frac{5}{16} \quad 0.110 \cdot 2^{-1} = (\frac{1}{2} + \frac{1}{4}) \cdot \frac{1}{2} = \frac{3}{8} \quad 0.111 \cdot 2^{-1} = \frac{7}{16}$$

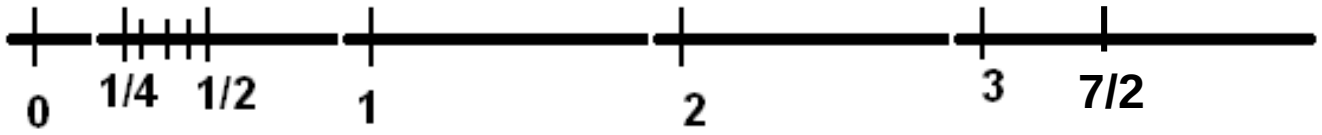
$$0.100 \cdot 2^{0} = \frac{1}{2} \cdot 1 = \frac{1}{2} \quad 0.101 \cdot 2^{0} = (\frac{1}{2} + \frac{1}{8}) \cdot 1 = \frac{5}{8} \quad 0.110 \cdot 2^{0} = (\frac{1}{2} + \frac{1}{4}) \cdot 1 = \frac{3}{4} \quad 0.111 \cdot 2^{0} = \frac{7}{8}$$

$$0.100 \cdot 2^{1} = \frac{1}{2} \cdot 2 = 1 \quad 0.101 \cdot 2^{1} = (\frac{1}{2} + \frac{1}{8}) \cdot 2 = \frac{5}{4} \quad 0.110 \cdot 2^{1} = (\frac{1}{2} + \frac{1}{4}) \cdot 2 = \frac{3}{2} \quad 0.111 \cdot 2^{1} = \frac{7}{4}$$

$$0.100 \cdot 2^{2} = \frac{1}{2} \cdot 4 = 2 \quad 0.101 \cdot 2^{2} = (\frac{1}{2} + \frac{1}{8}) \cdot 4 = \frac{5}{2} \quad 0.110 \cdot 2^{2} = (\frac{1}{2} + \frac{1}{4}) \cdot 4 = 3 \quad 0.111 \cdot 2^{2} = \frac{7}{2}$$

F(2, 3, -1, 2)

Elementi positivi



33 elementi

Errori di rappresentazione

$$Sia \quad x = mB^p, \quad \bar{x} = fl(x) = \bar{m}B^p$$

Errore assoluto

$$\left| x - fl(x) \right|$$

Errore relativo

$$\frac{\left| x - fl(x) \right|}{\left| x \right|} \qquad x \neq 0$$

L'errore più significativo è l'errore relativo

Teorema

$$Sia \quad x = mB^p, \quad \bar{x} = fl(x) = \bar{m}B^p$$

Errore assoluto

$$|x - fl_T(x)| < B^{p-t}$$

$$|x - fl_A(x)| \leq \frac{1}{2} B^{p-t}$$

Errore relativo

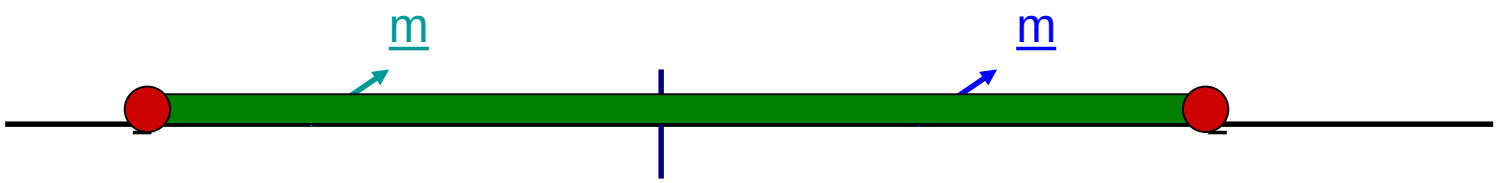
$$\frac{|x - fl_T(x)|}{|x|} < B^{1-t}$$

$$\frac{|x - fl_A(x)|}{|x|} \leq \frac{1}{2} B^{1-t}$$

non dipende dal numero rappresentato

Troncamento: distanza tra due mantisse

Sia **x=0.2352**, $fl_T(x)$? Base 10 , t =3



$\overline{m}_1$

0.235

$\overline{m}_2 = \overline{m}_1 + B^{-t}$

0.236 =0.235+10^{-3}

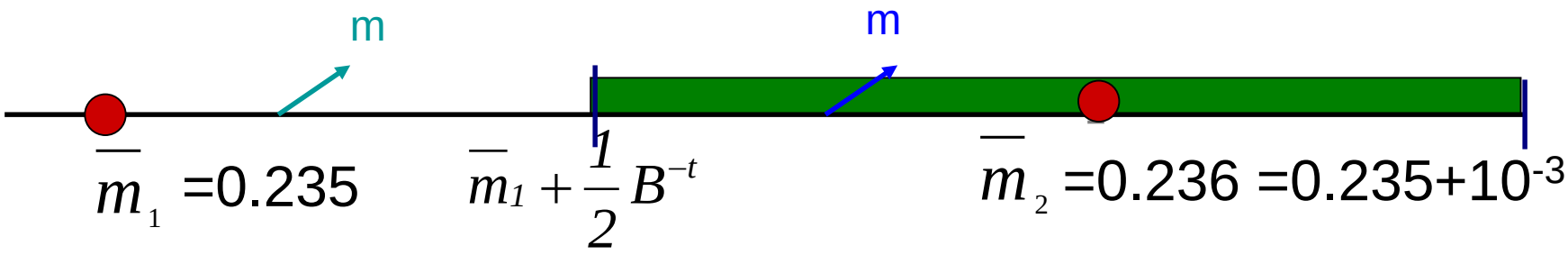
$$Se \quad m \in [\overline{m}_1, \overline{m}_2) \quad allora \quad \overline{m} = \overline{m}_1$$

$$\left| m - \overline{m}_1 \right| < B^{-t}$$

Arrotondamento: distanza tra due mantisse

Sia **x=0.2352**, $fl_A(x)$?

Base 10 , t =3



$$Se \quad m \in [\overline{m}_1, \overline{m}_1 + \frac{1}{2}B^{-t}) \quad allora \quad \overline{m} = \overline{m}_1$$

$$Se \quad m \in [\overline{m}_1 + \frac{1}{2}B^{-t}, \overline{m}_2) \quad allora \quad \overline{m} = \overline{m}_2$$

$$\left| m - \overline{m}_1 \right| \leq \frac{1}{2}B^{-t}$$

Per concludere la dim. del teorema..

$$\frac{\left|x - fl_T(x)\right|}{\left|x\right|} = \frac{\left|mB^p - \overline{m}B^p\right|}{\left|mB^p\right|} < \frac{B^{p-t}}{0.d_1 \ldots d_t B^p} <$$

$$(poichè \quad B^{-1} \le \left|m\right| < 1) \qquad < \frac{B^{-t}}{B^{-1}} = B^{1-t}$$

Analogamente per l'arrotondamento #c.v.d.

Precisione di macchina ***eps*** o unità di arrotondamento

$$\frac{|x - fl(x)|}{|x|} \leq eps \qquad eps = \begin{cases} B^{1-t} & troncamento \\ \frac{1}{2} B^{1-t} & arrotondamento \end{cases}$$

È il più piccolo numero finito positivo tale che
fl(1+eps)>1

La formula fornisce una misura di quanto accuratamente i numeri reali possono essere "*approssimati*" da numeri finiti F(B,t,L,U) e fornisce quindi una misura della "*precisione del calcolatore*".

Massimo errore relativo che si commette nel rappresentare un reale al calcolatore.

Precisione di macchina

Esempio: B=10, t=5

x=1 $y=0.5\ 10^{-4}$ ⟵ eps

$fl_A(x+y)= fl_A(1.00005)=0.10001 \times 10^1 > 1$

x=1 $y=0.4\ 10^{-4}$

$fl_A(x+y)= fl_A(1.00004)=0.10000 \times 10^1 = 1$

Errore di rappresentazione *err*

L'errore relativo che si commette nel rappresentare in F un numero reale x con fl(x) è:

$$err = \frac{fl(x) - x}{x}$$

Dal teorema discende che:

$$Sia\ x \in R,\ rappresentabile\ in\ F$$

$$allora\ \ fl(x) = x(1+err),\quad |err| \le eps$$

Ogni numero reale rappresentabile in F può essere approssimato da un elemento di F con un errore relativo *err* non più grande di *eps*.

Lo Standard IEEE

Lo standard IEEE 754 pubblicato nel 1985 (versione attuale IEEE 754-2008) definisce un sistema aritmetico floating point binario ed è adottato dalla maggior parte delle case costruttrici di elaboratori.

Round to nearest (EVEN)
Esso adotta la tecnica di arrotondamento al pari (in binario, zero nel bit meno significativo) per x esattamente equidistante da due numeri finiti x_1 e x_2 consecutivi,

e l' "hidden bit": poiché il primo bit di mantissa è sempre 1 possiamo fare a meno di memorizzarlo e guadagnare così un bit. HIDDEN BIT

Formati dello Standard IEEE per numeri floating point

Binary32: Precisione semplice 32 bit base 2

Segno (0/1) Esponente p* Mantissa m

t=23, m ha 24 bit , incluso l'hidden bit $p^* = p + 127, \quad 0 \le p^* \le 255$

Binary64: Precisione doppia 64 bit base 2

Segno (0/1) Esponente p* Mantissa m

t=52, m ha 53 bit , incluso l'hidden bit $p^* = p + 1023, \quad 0 \le p^* \le 2047$

Binary128: Precisione quadrupla 128 bit base 2

Segno (0/1) Esponente p* Mantissa m

t=112, m ha 113 bit , incluso l'hidden bit $p^* = p + 16383, \quad 0 \le p^* \le 32767$

Esempio

Utilizzando la rappresentazione standard IEEE per numeri floating point su 32 bit, si determini il valore decimale N della sequenza di bit

0	01111110	100 0000 0000 0000 0000 0000

- segno: +
- esponente: $p^* = (01111110)_2 = 126$

  $p = 126 - 127 = -1$

Il valore del numero rappresentato è calcolabile come:

$$(-1)^s * m * 2^p ,$$

$\Rightarrow N = 1.1 * 2^{-1} = (0.11)_2 = 1*2^{-1} + 1*2^{-2} = 0.5 + 0.25 = 0.75$

Esempio

Utilizzando la rappresentazione standard IEEE per numeri floating point su 32 bit, si determini la sequenza di bit per rappresentare il valore decimale N=1

$\Rightarrow \quad (1)_{10} = (1.00)_2 * 2^0$

- segno: +
- esponente: p=0, p* = p + 127 =127 = $(0111\ 1111)_2$

Il valore del numero rappresentato è calcolabile come:

$$(-1)^s * m * 2^p ,$$

$N = 1.0 * 2^0 \Rightarrow (1.00)_2$ $\qquad$ $m = (0.00)_2$

0	01111111	000 0000 0000 0000 0000 0000

Esempio

Utilizzando la rappresentazione standard IEEE per numeri floating point su 32 bit, si determini la sequenza di bit per rappresentare il valore decimale N=0.375

$\Rightarrow \quad (0.375)_{10} = (1.10)_2 * 2^{-2}$

- segno: +
- esponente: p = -2, p* = -2 + 127 =125 = $(0111\ 1101)_2$

Il valore del numero rappresentato è calcolabile come:

$$(-1)^s * m * 2^p ,$$

$N = 1.10 * 2^{-2} \quad \Rightarrow \quad m = (0.10)_2$

0	0111 1101	100 0000 0000 0000 0000 0000

Cifre significative

- In singola precisione t=24 corrisponde ad avere circa 8 cifre decimali significative, infatti

$$2^{-24} \approx 10^{-8}$$

- In doppia precisione t=53 corrisponde ad avere circa 16 cifre decimali significative, infatti

$$2^{-53} \approx 10^{-16}$$

Formati dello Standard IEEE per floating point

Precisione semplice 32 bit base 2

$$eps = \frac{1}{2} 2^{1-t} = \frac{2^{1-24}}{2} \approx 5.96 \times 10^{-8}$$

Precisione doppia 64 bit base 2

$$eps = \frac{1}{2} 2^{1-t} = \frac{2^{1-53}}{2} \approx 1.11 \times 10^{-16}$$

Matlab usa aritmetica IEEE doppia precisione

Formati dello Standard IEEE per floating point

zero

0/1	0	0

NaN=not a number

0/1	111.....1111	Qualsiasi configurazione di bit diversa da zero

Infinito (overflow)

0/1	111.....1111	0

MATLAB

- eps
- X > Realmax (1.7977e+308)

  ---->overflow
- X < Realmin (2.2251e-308)

  --->underflow (posti a 0)

  --->numeri subnormali nell'intervallo [eps*realmin,realmin]

  Per colmare il gap tra 0 e realmin

Aritmetica finita (o floating point)

I risultati di operazioni aritmetiche tra numeri finiti F generalmente non sono numeri finiti; pertanto in un calcolatore risulterà impossibile implementare correttamente le operazioni aritmetiche.

Operazioni in aritmetica floating point associano a due numeri finiti un terzo numero finito, ottenuto approssimando il risultato esatto dell'operazione aritmetica.

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

$$\bar{a} \quad op \quad \bar{b} = fl(\bar{a}\ op\ \bar{b}) = (\bar{a}\ op\ \bar{b})(1+\varepsilon)$$

$$op: \quad +,-,*,/, \qquad con \quad |\varepsilon| \le eps$$

L’aritmetica finita presenta due fondamentali differenze dall'aritmetica reale:

- l'aritmetica in virgola mobile non è associativa: in generale, per i numeri in virgola mobile,

$$x + (y + z) \neq (\mathrm{x} + \mathrm{y}) + \mathrm{z}$$

$$(x \cdot y) \cdot z \neq x \cdot (y \cdot z)$$

- l'aritmetica in virgola mobile non è distributiva: in generale,

$$x \cdot (y + z) \neq (\mathrm{x} \cdot \mathrm{y}) + (x \cdot z)$$

- esistono l'elemento neutro della moltiplicazione, l'elemento neutro dell'addizione e l'opposto, ma non sono unici.

Es: $1.0 + (10^{100} - 10^{100})$ dà come risultato 1.0, mentre

$(1.0 + 10^{100}) - 10^{100}$ dà 0.0

Propagazione degli errori

Oltre agli errori introdotti nelle singole operazioni si deve considerare l'esecuzione di una sequenza di queste nella quale si verifica una propagazione degli errori la cui entità è tutt'altro che trascurabile.

Il controllo e la gestione di tale fenomeno risulta fondamentale quando si approssima un modello matematico teorico al fine di determinare l'attendibilità dei risultati ottenuti.

Applichiamo l'analisi degli errori 'in avanti' sulle quattro operazioni aritmetiche.

Propagazione degli errori nella moltiplicazione

$$\frac{fl(\,fl(x)\cdot fl(y))-(x\cdot y)}{x\cdot y}=$$

$$=\frac{x(1+\varepsilon_1)\cdot y(1+\varepsilon_2)(1+\varepsilon_3)-x\cdot y}{x\cdot y}=$$

$$=(1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3)-1\approx$$

$$\approx \varepsilon_1+\varepsilon_2+\varepsilon_3$$

Propagazione degli errori nell'addizione

$$\frac{fl(fl(x)+fl(y))-(x+y)}{x+y}=$$

$$=\frac{(x(1+\varepsilon_1)+y(1+\varepsilon_2))(1+\varepsilon_3)-(x+y)}{x+y}=$$

$$=\frac{x\varepsilon_1+y\varepsilon_2+x\varepsilon_3+y\varepsilon_3+x\varepsilon_1\varepsilon_3+y\varepsilon_1\varepsilon_3}{x+y}\approx$$

$$\approx\frac{x}{x+y}\varepsilon_1+\frac{y}{x+y}\varepsilon_2+\varepsilon_3$$

Propagazione degli errori nella sottrazione

$$\frac{fl(\,fl(x)-\,fl(y))-(x-y)}{x-y}=$$

$$=\frac{(x(1+\varepsilon_1)-y(1+\varepsilon_2))(1+\varepsilon_3)-(x-y)}{x-y}=$$

$$\approx \frac{x}{x-y}\varepsilon_1-\frac{y}{x-y}\varepsilon_2+\varepsilon_3$$

Quando x è quasi uguale ad y, e_1 ed e_2 vengono enormemente amplificati e così l'errore.

Cancellazione numerica

Nelle operazioni di sottrazione quando i due operandi sono "quasi uguali" si ha una *perdita di cifre significative*

Es: X1=0.14754326 X2=0.14751742, t=6, B=10

fl(X1)=0.147554 fl(X2)=0.147252

$fl(fl(X1)-fl(X2))=0.302000x10^{-3}$

mentre la vera differenza è

$X1-X2=0.302584x10^{-3}$

l'errore relativo commesso sarà ~ $0.2x10^{-2}$, le ultime cifre della mantissa sono alterate dovuto al fatto che dopo aver eseguito fl(x1)-fl(x2)=0.000302 la rappresentazione normalizzata ha introdotto 3 zeri alla fine della mantissa.

Cancellazione numerica

Se prendiamo i numeri ancora più 'vicini':

X1=0.147554326 X2=0.147551742, t=6, B=10

fl(X1)=0.147554 fl(X2)=0.147552

fl(fl(X1)-fl(X2))=$0.2x10^{-5}$

mentre la vera differenza è

X1-X2=$0.2584x10^{-5}$

l'errore relativo commesso sarà ~ $0.2x10^{-0}$.

L'operazione di sottrazione in se' non introduce alcuna perdita di precisione, ma può amplificare gli errori presenti negli operandi

Cancellazione numerica: ESEMPIO

$x^2 - 6.433\,x + 0.009474 = 0$

B=10 t=4

$$x_2 = \frac{6.433 - \sqrt{(6.433)^2 - 4(0.009474)}}{2}$$

fl(x)=0.2000 x 10^{-2} x=0.0014731 esatto

$$\left|\frac{x - fl(x)}{x}\right| \cong 0.357$$

36%

Cancellazione numerica: ESEMPIO

x^2-6.433 x + 0.009474 = 0 B=10 t=4

$$b - \sqrt{\Delta} = b - \sqrt{(6.433)^2 - 4(0.009474)} = 0.6433 \times 10^1 - 0.6429 \times 10^1$$

differenza fra numeri molto simili in modulo.

Pur partendo da dati affetti da un piccolo errore relativo di arrotondamento,

Errore in b

$$\left| \frac{6.433 - 6.433}{6.433} \right| = 0.0$$

Errore in $\sqrt{\Delta}$

$$\left| \frac{6.4300538 - 6.429}{6.4300538} \right| \approx 0.16 \times 10^{-3}$$

si ottiene un errore significativo sul risultato.

Cancellazione numerica: ESEMPIO

- La cancellazione numerica si può evitare:

$$ax^2 + bx + c = 0$$

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Errori di canc. num. in x_1 se $\sqrt{\Delta} \approx b$ o in x_2 se $\sqrt{\Delta} \approx -b$

Per eliminare il fenomeno, si cambia procedimento numerico:

$$x_1 \cdot x_2 = \frac{c}{a} \rightarrow x_2 = \frac{c}{x_1}$$

Se $b^2 \approx 4ac$ allora è bene usare maggior precisione

Cancellazione numerica: ESEMPIO

- Calcolare e^x in x=5.5 con lo sviluppo in serie:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ... = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

- Consideriamo la somma dei primi 25 termini si ottiene:

$$e^{-5.5} = 1.0000 - 5.5000 + 15.1250 - 27.730 + 38.1290 - 41.9420 +$$
$$+ 38.4460 - 30.208....$$

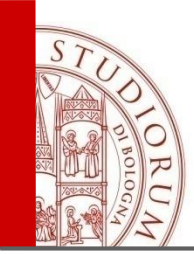
$$e^{-5.5} \approx \quad 0.26363 \cdot 10^{-2}$$

- Il valore corretto è $e^{-5.5} = \quad 0.408677 \cdot 10^{-2}$

1) Poiché $e^{-5.5} = \frac{1}{e^{5.5}} = 0.40865 \cdot 10^{-2}$

si può applicare la formula usando tutti termini positivi

2) Si possono sommare tutti i negativi e poi tutti i positivi e si fa un'unica differenza finale



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Serena Morigi

Dipartimento di Matematica

serena.morigi@unibo.it

http://www.dm.unibo.it/~morigi