Analisi Numerica: Aritmetica di macchina

S. Maset

Dipartimento di Matematica e Geoscienze, Università di Trieste

Rappresentazioni in base

- Analizzeremo qui la discretizzazione dei numeri reali con i numeri di macchina. In particolare, studieremo gli errori introdotti nell'approssimare un numero reale con un numero di macchina (detti errori di arrotondamento) e l'effetto di tali errori nelle computazioni.

  Però, prima di introdurre i numeri di macchina, è necessario presentare le (in parte note) rappresentazioni in base dei numeri reali.

Rappresentazione in base dei numeri naturali

- Sia $B > 1$ un numero naturale che chiameremo **base**. Inoltre, chiameremo i naturali $0, 1, 2, \ldots, B-1$ **cifre** nella base B.

Per quanto riguarda i numeri naturali si ha il seguente **teorema di rappresentazione in base**.

Teorema

Per ogni numero $a \in \mathbb{N} \setminus \{0\}$*, esistono unici* $p \in \mathbb{N}$ *e cifre* $b_0, b_1, b_2, \ldots, b_p$ *nella base* B *con* $b_p \neq 0$ *tali che*

$$a = b_p B^p + \cdots + b_2 B^2 + b_1 B + b_0. \tag{1}$$

La (1) è la **rappresentazione in base** B di a. Si scrive

$$a = (b_p \ldots b_2 b_1 b_0)_B .$$

Esempi:

$$\begin{aligned}(100010)_2 &= 2^5 + 2^1 = 34\\ (102)_3 &= 3^2 + 2 = 11\\ (246)_7 &= 2 \cdot 7^2 + 4 \cdot 7 + 6 = 132.\end{aligned}$$

Esercizio. Determinare $(678)_9$ e $(2DB)_{14}$, dove $D = 13$ e $B = 11$.

- La condizione $b_p \neq 0$ è necessaria per avere l'unicità della rappresentazione: non richiedendo più $b_p \neq 0$, data una rappresentazione $(b_p \dots b_2 b_1 b_0)_B$, si potrebbe sempre scrivere

$$(b_p \dots b_2 b_1 b_0)_B = (0 \dots 0 b_p \dots b_2 b_1 b_0)_B$$

aggiungendo un numero arbitrario di zeri in testa.

- Vediamo ora come trovare la rappresentazione in base B del numero naturale non nullo a, assumendo che essa esista.

Sia

$$a_0 := a = b_p B^p + \cdots + b_2 B^2 + b_1 B + b_0.$$

Poichè

$$a_0 = \left(b_p B^{p-1} + \cdots + b_2 B + b_1 \right) B + b_0$$

con $b_0 \in \{0, 1, 2, \ldots, B-1\}$, si ha che b_0 (la prima cifra della rappresentazione) è il resto della divisione intera di a_0 per B e

$$a_1 := b_p B^{p-1} + \cdots + b_2 B + b_1$$

è il quoziente.

Poi si ha

$$a_1 = \left(b_p B^{p-2} + \cdots + b_3 B + b_2 \right) B + b_1$$

con $b_1 \in \{0, 1, 2, \ldots, B-1\}$, e quindi b_1 (la seconda cifra della rappresentazione) è il resto della divisione intera di a_1 per B e

$$a_2 := b_p B^{p-2} + \ldots + b_3 B + b_2$$

è il quoziente.

Così proseguendo si trovano tutte le cifre $b_0, b_1, b_2, \ldots$ della rappresentazione.

La procedura è la seguente:

$a_0 = a$

per $k = 0, 1, 2, \ldots$

$b_k =$ resto della divisione intera di a_k per B

$a_{k+1} =$ quoziente della divisione intera di a_k per B.

Ci si ferma quando si trova un dividendo a_p, per qualche $p \in \{0, 1, 2, \ldots\}$, che è una cifra nella base B, cioè sta in $\{0, 1, \ldots, B-1\}$.

Infatti, se a_p è una cifra nella base B, si ha $b_p = a_p$ e $a_{p+1} = 0$ e dal fatto che $a_{p+1} = 0$ segue $b_n = 0$ e $a_{n+1} = 0$, per ogni $n \geq p+1$. L'ultima cifra della rappresentazione è $b_p = a_p$.
Esercizio. Spiegare perché nella procedura descritta risulta sempre $a_p \neq 0$.

Ad esempio, la rappresentazione in base 2 del numero 100 si trova con

k	Dividendo $= a_k$	Divisore $= B$	Quoziente $= a_{k+1}$	Resto $= b_k$
0	100	2	50	0
1	50	2	25	0
2	25	2	12	1
3	12	2	6	0
4	6	2	3	0
5	3	2	1	1
6	1	2	0	1

Quindi $100 = (1100100)_2 = 2^6 + 2^5 + 2^2$.

Altro esempio. La rappresentazione in base 3 del numero 100 si trova con

k	Dividendo $= a_k$	Divisore $= B$	Quoziente $= a_{k+1}$	Resto $= b_k$
0	100	3	33	1
1	33	3	11	0
2	11	3	3	2
3	3	3	1	0
4	1	3	0	1

Quindi $100 = (10201)_3 = 3^4 + 2 \cdot 3^2 + 1$.

Esercizio. Determinare la rappresentazione di 1000 in base 2, 3, 5, 20 e 30.

Esercizio. Determinare la rappresentazione in base B di $B^p - 1$. Suggerimento: osservare che

$$B^p - 1 = \left(B^{p-1} - 1\right) \cdot B + B - 1.$$

- Esercizio. Data la rappresentazione in base B di $a \in \mathbb{N} \setminus \{0\}$:

$$a = (b_p \dots b_2 b_1 b_0) = b_p B^p + \cdots + b_2 B^2 + b_1 B + b_0,$$

determinare la rappresentazione in base B^2 di a. Suggerimento: scrivere, eventualmente con $b_{2k+1} = 0$ se $p = 2k$ è pari,

$$\begin{aligned} a &= b_{2k+1} B^{2k+1} + b_{2k} B^{2k} + \cdots + b_3 B^3 + b_2 B^2 + b_1 B + b_0 \\ &= (b_{2k+1} B + b_{2k}) B^{2k} + \cdots + (b_3 B + b_2) B^2 + b_1 B + b_0 \end{aligned}$$

e mostrare che, per ogni $i \in \{0, 1, \dots, k\}$, $b_{2i+1} B + b_{2i}$ è una cifra nella base B^2.

Rappresentazione in base dei numeri reali in $[0, 1)$

- Per quanto riguarda i numeri reali nell'intervallo $[0, 1)$ si ha il seguente teorema di rappresentazione in base.

Teorema

Per ogni numero $y \in [0, 1)$*, esiste un'unica successione di cifre* $b_{-1}, b_{-2}, b_{-3}, \ldots$ *nella base* B *non definitivamente uguale a* $B - 1$ *tale che*

$$y = b_{-1}B^{-1} + b_{-2}B^{-2} + b_{-3}B^{-3} + \cdots \tag{2}$$

La (2) è la **rappresentazione in base** B di y. Si scrive

$$y = (0.b_{-1}b_{-2}b_{-3}\ldots)_B\,.$$

Esempi:

$$(0.11)_2 = 2^{-1} + 2^{-2} = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$$

$$\begin{aligned}
\left(0.12\overline{1}\right)_3 &= 3^{-1} + 2 \cdot 3^{-2} + 3^{-3} + 3^{-4} + 3^{-5} + \cdots \\
&= 3^{-1} + 2 \cdot 3^{-2} + 3^{-3}\left(1 + 3^{-1} + 3^{-2} + \cdots\right) \\
&= 3^{-1} + 2 \cdot 3^{-2} + 3^{-3} \cdot \frac{1}{1 - \frac{1}{3}} \\
&= \frac{1}{3} + \frac{2}{9} + \frac{1}{18} = \frac{11}{18}
\end{aligned}$$

$$\begin{aligned}
\left(0.\overline{2}\right)_7 &= 2 \cdot 7^{-1} + 2 \cdot 7^{-2} + 2 \cdot 7^{-3} + \ldots \\
&= 2 \cdot 7^{-1}\left(1 + 7^{-1} + 7^{-2} + \cdots\right) \\
&= 2 \cdot 7^{-1} \cdot \frac{1}{1 - 7^{-1}} = \frac{1}{3}.
\end{aligned}$$

Esercizio. Determinare $\left(0.\overline{1}\right)_2$ e $\left(0.\overline{1}\right)_3$.

Esercizio. Determinare $\left(0.21\overline{12}\right)_3$. Suggerimento: scrivere

$$\begin{aligned}\left(0.21\overline{12}\right)_3 &= 2\cdot 3^{-1}+3^{-2}\\ &\quad +3^{-3}+3^{-5}+3^{-7}+\cdots\\ &\quad +2\cdot 3^{-4}+2\cdot 3^{-6}+2\cdot 3^{-8}+\cdots\end{aligned}$$

- La condizione che la successione di cifre non sia definitivamente $B-1$ è necessaria per avere l'unicità della rappresentazione.

  Consideriamo infatti una rappresentazione

  $$\left(0.b_{-1}\dots b_{-k}\overline{B-1}\right)_B,$$

  dove $b_k \neq B-1$. Si ha

  $$\begin{aligned}
  &\left(0.b_{-1}\cdots b_{-k}\overline{B-1}\right)_B \\
  &= b_{-1}B^{-1}+\cdots+b_{-k}B^{-k} \\
  &\quad +(B-1)B^{-(k+1)}+(B-1)B^{-(k+2)}+(B-1)B^{-(k+3)}+\dots \\
  &= b_{-1}B^{-1}+\cdots+b_{-k}B^{-k}+(B-1)B^{-(k+1)}\left(1+B^{-1}+B^{-2}+\cdots\right) \\
  &= b_{-1}B^{-1}+\cdots+b_{-k}B^{-k}+(B-1)B^{-(k+1)}\cdot\frac{1}{1-B^{-1}} \\
  &= b_{-1}B^{-1}+\cdots+b_{-k}B^{-k}+B^{-k} \\
  &= \left(0.b_{-1}\cdots b_{-k+1}\left(b_{-k}+1\right)\right)_B.
  \end{aligned}$$

  Ad esempio,

  $$(0.44\overline{9})_{10} = (0.45)_{10}.$$

- Vediamo ora come trovare la rappresentazione in base B del numero reale $y \in [0,1)$, assumendo che essa esista.

  Si ponga

$$y_{-1} := y = b_{-1}B^{-1} + b_{-2}B^{-2} + b_{-3}B^{-3} + \cdots .$$

  Poichè

$$By_{-1} = b_{-1} + b_{-2}B^{-1} + b_{-3}B^{-2} + b_{-4}B^{-3} + \cdots ,$$

  con

$$\begin{aligned}
&b_{-2}B^{-1} + b_{-3}B^{-2} + b_{-4}B^{-3} + \cdots = B^{-1}\left(b_{-2} + b_{-3}B^{-1} + b_{-4}B^{-2} + \cdots\right) \\
&< B^{-1}\left((B-1) + (B-1)B^{-1} + (B-1)B^{-2} + \cdots\right) \\
&= B^{-1}(B-1)\left(1 + B^{-1} + B^{-2} + \cdots\right) \\
&= B^{-1}(B-1)\frac{1}{1-B^{-1}} = 1,
\end{aligned}$$

  si ha che b_{-1} (la prima cifra della rappresentazione) è la parte intera di By_{-1} e

$$y_{-2} := b_{-2}B^{-1} + b_{-3}B^{-2} + b_{-4}B^{-3} + \cdots$$

  è la parte frazionaria

Poi

$$By_{-2} = b_{-2} + b_{-3}B^{-1} + b_{-4}B^{-2} + b_{-5}B^{-3} + \cdots$$

con

$$b_{-3}B^{-1} + b_{-4}B^{-2} + b_{-5}B^{-3} + \cdots = B^{-1}\left(b_{-3} + b_{-4}B^{-1} + b_{-5}B^{-2} + \cdots\right)$$

$$< B^{-1}\left((B-1) + (B-1)B^{-1} + (B-1)B^{-2} + \cdots\right) = 1$$

e quindi b_{-2} (la seconda cifra della rappresentazione) è la parte intera di By_{-2} e

$$y_{-3} := b_{-3}B^{-1} + b_{-4}B^{-2} + b_{-5}B^{-3} + \cdots$$

è la parte frazionaria.

Così proseguendo si trovano tutte le cifre $b_{-1}, b_{-2}, b_{-3}, \ldots$ della rappresentazione.

La procedura è la seguente:

$y_{-1} = y$

per $k = 1, 2, 3, \ldots$

$b_{-k} =$ parte intera di By_{-k}

$y_{-k-1} =$ parte frazionaria di By_{-k}.

Ad esempio, la rappresentazione in base 2 di 0.1 è ottenuta con

k	y_{-k}	By_{-k}	Parte intera$= b_{-k}$	Parte frazionaria$= y_{-(k+1)}$
1	0.1	0.2	0	0.2
2	0.2	0.4	0	0.4
3	0.4	0.8	0	0.8
4	0.8	1.6	1	0.6
5	0.6	1.2	1	0.2
6	0.2			

Quindi

$$0.1 = \left(0.0\overline{0011}\right)_2 .$$

Altro esempio, la rappresentazione in base 3 di 0.1 è ottenuta con

k	y_{-k}	By_{-k}	Parte intera $= b\_k$	Parte frazionaria $= y_{-(k+1)}$
1	0.1	0.3	0	0.3
2	0.3	0.9	0	0.9
3	0.9	2.7	2	0.7
4	0.7	2.1	2	0.1
5	0.1			

Quindi

$$0.1 = (0.\overline{0022})_3.$$

Esercizio. Trovare la rappresentazione di 0.5 in base 5, 6 e 7.

Esercizio. Trovare la rappresentazione di 0.1 in base 5, 11 e 20.

Esercizio. Provare che con la procedura descritta si ottiene una successione di cifre $b_{-1}, b_{-2}, b_{-3}, \ldots$ non definitivamente uguale a $B-1$. Suggerimento: se la successione fosse definitivamente uguale a $B-1$, esisterebbe $k \in \{1, 2, 3, \ldots\}$ tale che $b_{-k} = B-1$ e $y_{-k-1} = y_{-k}$. Quindi

$$By_{-k} = b_k + y_{-k-1} = \ldots$$

Esercizio. Sia $y \in [0,1)$ razionale, vale a dire $y = \frac{m}{n}$ con m, n interi positivi, $n \geq 2$ e $m \in \{0, 1, 2, \ldots, n-1\}$. Provare che la rappresentazione in base B di y è periodica con un eventuale antiperiodo. Suggerimento: osservare che la fine del periodo corrisponde al primo $k \in \{1, 2, \ldots\}$ tale che y_{-k} sia un valore già presente in $y_{-1}, \ldots, y_{-(k-1)}$; sia

$$Bm = qn + r$$

la divisione intera di Bm per n; allora

$$By = q + \frac{r}{n};$$

per cui, $b_{-1} = q$ è la parte intera di By e $y_{-2} = \frac{r}{n}$ è la parte frazionaria; ora $y_{-2} = \frac{r}{n}$ è, esattamente come $y_{-1} = y = \frac{m}{n}$, una frazione con denominatore n e numeratore in $\{0, 1, 2, \ldots, n-1\}$.

Rappresentazione in base dei numeri reali

- Si è vista la rappresentazione in base B di un numero in $\mathbb{N} \setminus \{0\}$ e di un numero in $[0, 1)$.

  Vediamo ora la rappresentazione in base B dei numeri reali maggiori o uguali a 1 e dei numeri negativi, così da avere la rappresentazione per ogni numero reale.

- Sia $x \geq 1$. Si ha

$$x = a + y,$$

dove $a \in \mathbb{N} \setminus \{0\}$ e $y \in [0, 1)$ sono, rispettivamente, la parte intera e la parte frazionaria di x.

Quindi, se

$$a = (b_p \dots b_2 b_1 b_0)_B \quad \text{e} \quad y = (0.b_{-1} b_{-2} b_{-3} \dots)_B$$

sono le rappresentazioni in base B di a e y, rispettivamente, si ha

$$x = b_p B^p + \cdots + b_2 B^2 + b_1 B + b_0 + b_{-1} B^{-1} + b_{-2} B^{-2} + b_{-3} B^{-3} + \cdots \quad (3)$$

La (3) è la rappresentazione in base B di x. Si scrive

$$x = (b_p \dots b_2 b_1 b_0 \, . \, b_{-1} b_{-2} b_{-3} \dots)_B.$$

Si è ottenuta così la rappresentazione in base B di ogni numero reale non negativo, sia esso in $[0, 1)$ o in $[1, +\infty)$.

- Se $x < 0$, la rappresentazione in base B di x è

$$x = -\left(b_p \ldots b_2 b_1 b_0 \,.\, b_{-1} b_{-2} b_{-3} \ldots\right)_B,$$

dove

$$|x| = \left(b_p \ldots b_2 b_1 b_0 \,.\, b_{-1} b_{-2} b_{-3} \ldots\right)_B$$

è la rappresentazione in base B di $|x|$.

- Ad esempio, riguardando agli esempi sopra, la rappresentazione in base 2 di

$$100.1 = 100 + 0.1 = (1100100)_2 + \left(0.0\overline{0011}\right)_2$$

è

$$100.1 = \left(1100100 \,.\, 0\overline{0011}\right)_2$$

e la rappresentazione in base 3 di

$$-100.1 = -(100 + 0.1) = -\left((10201)_3 + (0.\overline{0022})_3\right)$$

è

$$-100.1 = -(10201 \,.\, \overline{0022})_3.$$

- Esercizio. Trovare la rappresentazione di -180.25 in base 2, 3, 5 e 12.
- Esercizio. Trovare la rappresentazione in base B di $\frac{B^p+1}{B^k}$, dove p e k sono interi positivi con $p > k$.

Rappresentazione normalizzata

- Sia x un numero reale non nullo. Se $|x| < 1$, allora si ha

$$x = \pm \left(0.0 \ldots 0 b_{-p} b_{-(p+1)} b_{-(p+2)} \cdots\right)_B$$

con $b_{-p} \neq 0$, per qualche $p \in \{1, 2, 3, \ldots\}$. Si può allora scrivere

$$\begin{aligned} x &= \pm \left(b_{-p} B^{-p} + b_{-(p+1)} B^{-(p+1)} + b_{-(p+2)} B^{-(p+2)} + \cdots\right) \\ &= \pm \left(b_{-p} + b_{-(p+1)} B^{-1} + b_{-(p+2)} B^{-2} + \cdots\right) B^{-p} \\ &= \pm \left(b_{-p}.b_{-(p+1)} b_{-(p+2)} \ldots\right)_B B^{-p}. \end{aligned}$$

Invece, per $|x| \geq 1$ si ha

$$x = \pm \left(b_p \ldots b_2 b_1 b_0 . b_{-1} b_{-2} b_{-3} \ldots\right)_B$$

con $b_p \neq 0$ per qualche $p \in \{0, 1, 2, \ldots\}$. Si può allora scrivere

$$\begin{aligned} x &= \pm \left(b_p B^p + \cdots + b_2 B^2 + b_1 B + b_0 + b_{-1} B^{-1} + b_{-2} B^{-2} + b_{-3} B^{-3} + \cdots\right) \\ &= \pm \left(b_p + b_{p-1} B^{-1} + \cdots + b_0 B^{-p} + b_{-1} B^{-(p+1)} + \cdots\right) B^p \\ &= \pm \left(b_p . b_{p-1} \ldots b_0 b_{-1} \ldots\right)_B B^p. \end{aligned}$$

Si ottiene così il seguente teorema.

Teorema

Per ogni numero reale $x \neq 0$ esistono unici $p \in \mathbb{Z}$ e una successione di cifre $b_0, b_{-1}, b_{-2}, \ldots$. nella base B, non definitivamente uguale a $B-1$ e con $b_0 \neq 0$, tali che

$$x = \pm (b_0.b_{-1}b_{-2}b_{-3}\ldots)_B \cdot B^p. \tag{4}$$

La (4) è la **rappresentazione normalizzata in base** B di x. p è detto l'**esponente** della rappresentazione e $(b_0.b_{-1}b_{-2}b_{-3})_B$ la **mantissa** o **caratteristica**.

- Osserviamo che

$$1 \leq (b_0.b_{-1}b_{-2}b_{-3}\ldots)_B < B.$$

Infatti, si ha

$$(b_0.b_{-1}b_{-2}b_{-3}\ldots)_B = b_0 + b_{-1}B^{-1} + b_{-2}B^{-2} + b_{-3}B^{-3} + \cdots \geq b_0 \geq 1$$

e

$$\begin{aligned}
&(b_0.b_{-1}b_{-2}b_{-3}\ldots)_B = b_0 + b_{-1}B^{-1} + b_{-2}B^{-2} + b_{-3}B^{-3} + \cdots \\
&< B - 1 + (B-1)\,B^{-1} + (B-1)\,B^{-2} + (B-1)\,B^{-3} + \cdots \\
&= (B-1)\left(1 + B^{-1} + B^{-2} + B^{-3} + \cdots\right) \\
&= (B-1)\,\frac{1}{1-B^{-1}} = B.
\end{aligned}$$

Per cui,

$$B^p \leq |x| = (b_0.b_{-1}b_{-2}b_{-3}\ldots)_B \cdot B^p < B^{p+1}.$$

Per tale motivo, si dice che x ha **ordine di grandezza** B^p nella base B.

- Esempio. Si consideri il numero $\frac{17}{3}$ e si voglia trovare la sua rappresentazione normalizzata in base 2.

Si ha

$$\frac{17}{3} = \underbrace{5}_{\text{parte intera}} + \underbrace{\frac{2}{3}}_{\text{parte frazionaria}} .$$

Per la parte intera risulta

$$5 = (101)_2$$

e per la parte frazionaria si ha

k	y_{-k}	By_{-k}	Parte intera $= b_{-k}$	Parte frazionaria $= y_{-(k+1)}$
1	$\frac{2}{3}$	$\frac{4}{3}$	1	$\frac{1}{3}$
2	$\frac{1}{3}$	$\frac{2}{3}$	0	$\frac{2}{3}$
3	$\frac{2}{3}$			

e quindi $\frac{2}{3} = (0.\overline{10})_2$.

Per cui,

$$\frac{17}{3} = (101.\overline{10})_2 = (1.01\overline{10})_2 \cdot 2^2.$$

- Esempio. Determiniamo ora la rappresentazione normalizzata in base 3 di 0.01.
  Risulta

k	y_{-k}	By_{-k}	Parte intera $= b_{-k}$	Parte frazionaria $= y_{-(k+1)}$
1	0.01	0.03	0	0.03
2	0.03	0.09	0	0.09
3	0.09	0.27	0	0.27
4	0.27	0.81	0	0.81
5	0.81	2.43	2	0.43
6	0.43	1.29	1	0.29
7	0.29	0.87	0	0.87
8	0.87	2.61	2	0.61
...	...	...	...	...

  Quindi

$$0.01 = (0.00002102\ldots)_3 = (2.012\ldots)_3 \cdot 3^{-5}.$$

  Esercizio. Quanto occorre attendere al massimo per vedere nel precedente esempio la fine del periodo della rappresentazione? Suggerimento: guardare l'esercizio in cui si è provato che ogni numero razionale in $[0, 1)$ ha una rappresentazione periodica.

- Si osservi che lo zero non ha una rappresentazione normalizzata.

  Infatti, per definire la rappresentazione normalizzata è necessario considerare la prima cifra non nulla della rappresentazione e, invece, 0 ha solo cifre nulle: in ogni base B, si ha

  $$0 = (0.000\ldots)_B.$$

- Esercizio. Determinare la rappresentazione normalizzata di 20 e 200 in base 2 e 3.

- Esercizio. Determinare la rappresentazione normalizzata di $\frac{124}{7}$ in base 5.

- Esercizio. Richiamando un precedente esercizio, determinare la rappresentazione normalizzata in base B di $\frac{B^p+1}{B^k}$.

Numeri di macchina

- Nel calcolatore si possono rappresentare solo un numero finito di numeri reali e, per ciascuno di questi, si possono rappresentare solo un numero finito delle cifre di una sua rappresentazione in base.

  I numeri reali che vengono rappresentati nel calcolatore sono detti **numeri di macchina**.

- Vi sono due tipi di rappresentazione dei numeri di macchina nel calcolatore: la **floating-point** e la **fixed-point**.

  La rappresentazione floating point garantisce un piccolo errore relativo quando un numero reale è approssimato con un numero di macchina, mentre la fixed point garantisce un piccolo errore assoluto.

  Per tale motivo, la rappresentazione floating-point è importante in ambito scientifico e ingegneristico, mentre la fixed point è importante in altri ambiti, ad esempio quello bancario.

Qui, noi siamo interessati solo alla rappresentazione floating-point.

- Un insieme dei numeri di macchina in rappresentazione floating point è definito da dei parametri riguardanti la rappresentazione normalizzata dei numeri reali. Questi parametri sono:
  - la base B di rappresentazione;
  - il minimo esponente m ammesso;
  - il massimo esponente M ammesso;
  - il numero t di cifre della mantissa dopo il punto che vengono considerate (dalla $t+1-$ esima in poi le cifre sono uguali a zero).

I corrispondenti numeri di macchina sono i numeri reali non nulli x la cui rappresentazione normalizzata in base B è del tipo

$$x = \pm (b_0.b_{-1}b_{-2}b_{-3}\ldots b_{-t}000\ldots)_B \cdot B^p = \pm (b_0.b_{-1}b_{-2}b_{-3}\ldots b_{-t})_B \cdot B^p,$$

dove $m \leq p \leq M$.

Inoltre, anche lo zero (che non ha una rappresentazione normalizzata) è un numero di macchina.

Osserviamo che i numeri di macchina sono in numero finito di

$$\underbrace{2}_{\text{scelta di } + \text{ o } -} \cdot \underbrace{(B-1)}_{\text{scelte di } b_0 \neq 0} \cdot \underbrace{B}_{\text{scelte di } b_{-1}} \cdot \ldots \cdot \underbrace{B}_{\text{scelte di } b_{-t}} \cdot \underbrace{(M-m+1)}_{\text{scelte di } p} + \underbrace{1}_{\text{lo zero}}$$

$$= 2\,(B-1)\,B^t\,(M-m+1)+1.$$

Inoltre, ogni numero di macchina è rappresentato dalla seguente informazione finita:

- il segno $+$ o $-$;
- le cifre della mantissa $b_0, b_{-1}, \ldots, b_{-t}$ che sono cifre nella base B, cioè numeri in $\{0, 1, \ldots, B-1\}$;
- l'esponente p, che è un numero in $\{m, m+1, \ldots, M-1, M\}$.

- Un esempio didattico di insieme di numeri di macchina è quello definito dai parametri

$$B = 10,\ m = -1,\ M = 1,\ t = 2.$$

I numeri di macchina positivi sono

$$1.00 \cdot 10^{-1}, 1.01 \cdot 10^{-1}, 1.02 \cdot 10^{-1}, \ldots, 9.99 \cdot 10^{-1},\quad p = -1,$$

$$1.00 \cdot 10^{0}, 1.01 \cdot 10^{0}, 1.02 \cdot 10^{-0}, \ldots, 9.99 \cdot 10^{0},\quad p = 0,$$

$$1.00 \cdot 10^{1}, 1.01 \cdot 10^{1}, 1.02 \cdot 10^{1}, \ldots, 9.99 \cdot 10^{1},\quad p = 1,$$

cioè

$$0.100, 0.101, 0.102, \ldots, 0.999,$$

$$1.00, 1.01, 1.02, \ldots, 9.99, \tag{5}$$

$$10.0, 10.1, 10.2 \ldots, 99.9.$$

La (5) spiega perchè il modo di rappresentare i numeri reali che stiamo considerando è detto rappresentazione floating point. La posizione del punto si sposta a seconda dell'ordine di grandezza del numero.

- Esempi concreti di insiemi di numeri di macchina sono quelli definiti dallo **Standard IEEE** (IEEE è l'acronimo di Institute of Electrical and Electronics Engineers).

  Lo standard IEEE prevede che i numeri di macchina siano rappresentati nella base 2 in

  - **semplice precisione** usando una parola di memoria di 32 bit,
  - **doppia precisione** usando una parola di memoria 64 bit,
  - o **quadrupla precisione** usando una parola di memoria 128 bit.

  Nella doppia precisione dello standard IEEE, quello a cui siamo interessati, i parametri che definiscono l'insieme dei numeri di macchina sono

  $$B = 2,\ m = -1022,\ M = 1023,\ t = 52.$$

  I numeri di macchina non zero sono i numeri del tipo

  $$x = \pm (1.b_{-1}b_{-2}b_{-3}\ldots b_{-52})_2 \cdot 2^p$$

  con $-1022 \leq p \leq 1023$ e $b_{-1}, b_{-2}, b_{-3}, \ldots, b_{-52}$ sono cifre binarie. Si noti che $b_0 = 1$, dovendo essere $b_0 \neq 0$.

- Esercizio. Scrivere i numeri di macchina positivi dell'insieme di numeri di macchina definito dai parametri:

$$B = 3,\ m = -2,\ M = 2, t = 1.$$

- Esercizio. Se tutti i numeri dello Standard IEEE in doppia precisione fossero memorizzati in qualche supporto, quale sarebbe l'occupazione complessiva di memoria?

Numero di macchina più piccolo e più grande e precisione di macchina

- In un insieme di numeri di macchina, il più piccolo numero positivo è

$$\left(1.\underbrace{000\ldots0}_{t \text{ cifre}}\right)_B \cdot B^m = B^m.$$

Nell' esempio didattico dove $B = 10$ e $m = -1$, il piu piccolo numero di macchina positivo è $10^{-1} = 0.1$.

Nello standard IEEE in doppia precisione dove $B = 2$ e $m = -1022$, il più piccolo numero di macchina positivo è 2^{-1022}, il cui ordine di grandezza è 10^{-308}.

- Invece, il più grande numero di macchina positivo è

$$\begin{aligned}
&\left((B-1) . \underbrace{(B-1)(B-1)(B-1)\ldots(B-1)}_{t} \right)_B \cdot B^M \\
&= (B-1)\left(1 + B^{-1} + B^{-2} + \cdots + B^{-t}\right) \cdot B^M \\
&= (B-1)\,\frac{1 - B^{-(t+1)}}{1 - B^{-1}} \cdot B^M \\
&= \left(1 - B^{-(t+1)}\right) \cdot B^{M+1} \simeq B^{M+1}.
\end{aligned}$$

Nell'esempio didattico dove $B = 10$ e $M = 1$, il piu grande numero di macchina positivo è circa $10^2 = 100$.

Nello standard IEEE in doppia precisione dove $B = 2$ e $M = 1023$, il piu grande numero di macchina positivo è circa 2^{1024}, il cui ordine di grandezza è 10^{308}.

- Consideriamo l'esempio didattico i cui numeri di macchina positivi sono

$$0.100, 0.101, 0.102, \ldots, 0.999,$$

$$1.00, 1.01, 1.02, \ldots, 9.99,$$

$$10.0, 10.1, 10.2, \ldots, 99.9.$$

Essi sono suddivisi lungo le tre righe a seconda del loro ordine di grandezza.

  - I numeri di macchina della prima riga, quelli con ordine di grandezza 10^{-1}, sono distanziati uno dall'altro di 0.001,
  - i numeri della seconda riga, quelli con ordine di grandezza 10^{0}, sono distanziati l'uno dall'altro di 0.01,
  - e i numeri della terza riga, quelli con ordine di grandezza 10^{1}, sono distanziati l'uno dall'altro di 0.1.

In generale, in un insieme di numeri di macchina in rappresentazione floating point, la distanza tra due numeri di macchina consecutivi varia a seconda dell'ordine di grandezza dei numeri.

I numeri di macchina nell'intervallo $[B^p, B^{p+1})$, $p \in \{m, m+1, \ldots, M\}$, vale a dire quelli di ordine di grandezza B^p che hanno la forma

$$(b_0.b_{-1}b_{-2}b_{-3}\ldots b_{-t})_B \cdot B^p$$

sono distanziati l'uno dall'altro di

$$\left(0.\underbrace{0\ldots01}_{t \text{ cifre}}\right)_B \cdot B^p = B^{-t} \cdot B^p = B^{p-t}.$$

Il numero

$$\text{eps} := B^{-t}$$

(eps è un'abbreviazione di epsilon) è chiamato **precisione di macchina** o **epsilon di macchina**.

La distanza B^{p-t} tra due numeri di macchina consecutivi nell'intervallo $\left[B^p, B^{p+1}\right)$ è quindi $\text{eps} \cdot B^p$.

In particolare, eps è la distanza tra due numeri di macchina consecutivi nell'intervallo $[1, B)$.

Nell' esempio didattico dove $B = 10$ e $t = 2$, si ha $\text{eps} = 10^{-2}$.

Nello standard IEEE in doppia precisione dove $B = 2$ e $t = 52$, si ha $\text{eps} = 2^{-52}$, il cui ordine di grandezza è 10^{-16}.

Esercizio. Quanti sono i numeri macchina nell'intervallo $[B^p, B^{p+1})$, dove $p \in \{m, m+1, \ldots, M\}$ è fissato? Esprimere tale numero in termini di eps e B.

Approssimazione di numeri reali con numeri di macchina

- Assumiamo di avere un insieme di numeri di macchina definito dai parametri B (base di rappresentazione), m (minimo esponente), M (massimo esponente) e t (numero di cifre della mantissa dopo il punto).

  Ogni numero reale viene approssimato nel calcolatore con un numero di macchina.

  Il numero di macchina che approssima un numero reale x viene denotato con $\mathrm{fl}(x)$ (fl sta per "floating").

  Chiaramente, se x è un numero di macchina, si ha $\mathrm{fl}(x) = x$. Così in particolare, $\mathrm{fl}(0) = 0$.

  Nel seguito descriviamo $\mathrm{fl}(x)$ solo per un numero reale positivo x, in quanto, per un numero reale negativo x, si pone

$$\mathrm{fl}(x) = -\mathrm{fl}(|x|).$$

- Sia x un numero reale positivo e sia

$$x = (b_0.b_{-1}b_{-2}b_{-3}\ldots b_{-t}b_{-t-1}b_{-t-2}\ldots.)_B \cdot B^p$$

la sua rappresentazione normalizzata in base B.

Se $p < m$, allora il numero che si deve rappresentare è piu piccolo del più piccolo numero di macchina. Questa è una situazione di **underflow**.

In questo caso, lo standard IEEE prevede che $\mathrm{fl}(x)$ sia zero.

Se $p > M$, allora il numero che si deve rappresentare è piu grande del più grande numero di macchina. Questa è una situazione di **overflow**.

In questo caso, lo standard IEEE prevede che $\mathrm{fl}(x)$ sia un particolare numero di macchina, da aggiungere a quelli precedentemente descritti, che viene interpretato come $+\infty$.

- Supponiamo ora $p \in [m, M]$. Vi sono due modi di approssimare x con un numero di macchina: il **troncamento** e l'**arrotondamento** che ora descriviamo.

  Si considerino i numeri di macchina consecutivi x_1 e x_2 che, rispettivamente, immediamente precedono e immediatamente seguono

  $$x = (b_0.b_{-1}b_{-2}b_{-3}\ldots b_{-t}b_{-t-1}b_{-t-2}\ldots.)_B \cdot B^p,$$

  vale a dire i numeri di macchina consecutivi x_1 e x_2 tali che $x_1 \leq x < x_2$.

  Essi sono

  $$x_1 = (b_0.b_{-1}b_{-2}\ldots b_{-t})_B \cdot B^p$$

  e

  $$\begin{aligned} x_2 &= \text{numero di macchina successivo a } x_1 \\ &= (b_0.b_{-1}\ldots b_{t-1}(b_{-t}+1))_B \cdot B^p \\ &= \left(b_0 + b_{-1}B^{-1} + \cdots + b_{-t}B^{-t} + B^{-t}\right) \cdot B^p. \end{aligned}$$

Infatti si ha

$$
\begin{aligned}
x_1 &= \left(b_0 + b_{-1}B^{-1} + \cdots + b_{-t}B^{-t}\right) \cdot B^p \\
&\leq \left(b_0 + b_{-1}B^{-1} + \cdots + b_{-t}B^{-t} + b_{-t-1}B^{-t-1} + b_{-t-2}B^{-t-2} + \cdots\right) \cdot B^p \\
&= x
\end{aligned}
$$

e

$$
\begin{aligned}
x &= \left(b_0 + b_{-1}B^{-1} + \ldots + b_{-t}B^{-t} + b_{-t-1}B^{-t-1} + b_{-t-2}B^{-t-2} + \cdots\right) \cdot B^p \\
&< \left(b_0 + b_{-1}B^{-1} + \cdots + b_{-t}B^{-t} + (B-1)\,B^{-t-1} + (B-1)\,B^{-t-2} + \cdots\right) \cdot B^p \\
&= \left(b_0 + b_{-1}B^{-1} + \cdots + b_{-t}B^{-t} + (B-1)\,B^{-t-1}\left(1 + B^{-1} + B^{-2} + \cdots\right)\right) \cdot B^p \\
&= \left(b_0 + b_{-1}B^{-1} + \cdots + b_{-t}B^{-t} + (B-1)\,B^{-t-1}\,\frac{1}{1-B^{-1}}\right) \cdot B^p \\
&= \left(b_0 + b_{-1}B^{-1} + \cdots + b_{-t}B^{-t} + B^{-t}\right) \cdot B^p \\
&= x_2.
\end{aligned}
$$

Troncamento e arrotondamento

Nel **troncamento** si definisce

$$\mathrm{fl}(x) = x_1 = (b_0.b_{-1}b_{-2}\ldots b_{-t})_B \cdot B^p,$$

cioè $\mathrm{fl}(x)$ è il numero di macchina che immediatamente precede x.

Nell'esempio didattico, con il troncamento si ha

$$\mathrm{fl}\left(\frac{2}{3}\right) = \mathrm{fl}\left(0.\overline{6}\right) = \mathrm{fl}\left(6.\overline{6} \cdot 10^{-1}\right) = 6.66 \cdot 10^{-1}$$

e

$$\mathrm{fl}(\mathrm{e}) = \mathrm{fl}(2.71828\ldots) = 2.71.$$

- Nell'**arrotondamento** si definisce

$$\mathrm{fl}(x) = \begin{cases} x_1 \text{ se } x < \frac{x_1+x_2}{2} = \text{ punto medio dell'intervallo } [x_1, x_2] \\ \\ x_2 \text{ se } x \geq \frac{x_1+x_2}{2} \end{cases},$$

cioè $\mathrm{fl}(x)$ è il numero di macchina più vicino ad x.

- Nel caso di B pari, per conoscere se valga o non valga la condizione

$$x < \frac{x_1 + x_2}{2},$$

è sufficiente esaminare la cifra b_{-t-1}.

Teorema

Se B è pari, allora

$$x < \frac{x_1 + x_2}{2} \quad \Leftrightarrow \quad b_{-t-1} \leq \frac{B}{2} - 1.$$

Dimostrazione.

Si ha

$$\begin{aligned} x_1 + x_2 &= (b_0 + b_{-1}B^{-1} + \cdots + b_{-t}B^{-t}) \cdot B^p \\ &\quad +(b_0 + b_{-1}B^{-1} + \cdots + b_{-t}B^{-t} + B^{-t}) \cdot B^p \\ &= \left(2(b_0 + b_{-1}B^{-1} + \cdots + b_{-t}B^{-t}) + B^{-t}\right) \cdot B^p \end{aligned}$$

e quindi

$$\frac{x_1 + x_2}{2} = (b_0 + b_{-1}B^{-1} + \cdots + b_{-t}B^{-t} + \frac{B^{-t}}{2}) \cdot B^p.$$

Per cui,

$$\begin{aligned} x &= (b_0 + b_{-1}B^{-1} + \cdots + b_{-t}B^{-t} + b_{-t-1}B^{-t-1} + b_{-t-2}B^{-t-2} + \cdots) \cdot B^p \\ &< \frac{x_1 + x_2}{2} = (b_0 + b_{-1}B^{-1} + \cdots + b_{-t}B^{-t} + \frac{B^{-t}}{2}) \cdot B^p \end{aligned}$$

se e solo se

$$b_{-t-1}B^{-t-1} + b_{-t-2}B^{-t-2} + \cdots < \frac{B^{-t}}{2},$$

vale a dire

$$y := b_{-t-1} + b_{-t-2}B^{-1} + b_{-t-3}B^{-3} + \cdots < \frac{B}{2}.$$

Dimostrazione.

Ora proveremo che

$$y < \frac{B}{2} \quad \Leftarrow \quad b_{-t-1} \leq \frac{B}{2} - 1 \quad \text{e} \quad y < \frac{B}{2} \quad \Rightarrow \quad b_{-t-1} < \frac{B}{2}. \tag{6}$$

Osservare che se B pari e quindi $\frac{B}{2}$ è intero, $b_{-t-1} < \frac{B}{2}$ è equivalente a $b_{-t-1} \leq \frac{B}{2} - 1$. Per cui, se B pari, (6) dimostra il teorema.

Dimostrazione di (6). Osserviamo che si ha

$$b_{-t-1} \leq y < b_{-t-1} + 1.$$

Infatti, avendosi

$$y = b_{-t-1} + b_{-t-2}B^{-1} + b_{-t-3}B^{-3} + \cdots$$

la disuguaglianza di sinistra è banale, mentre quella di destra segue da

$$\begin{aligned} & b_{-t-2}B^{-1} + b_{-t-3}B^{-2} + b_{-t-4}B^{-3} + \cdots \\ & < (B-1)\,B^{-1} + (B-1)\,B^{-2} + (B-1)\,B^{-3} + \cdots \\ & = (B-1)\,B^{-1}\left(1 + B^{-1} + B^{-2} + \cdots\right) = (B-1)\,B^{-1}\,\frac{1}{1 - B^{-1}} = 1. \end{aligned}$$

Dimostrazione.

Per cui, se $b_{-t-1} \leq \frac{B}{2} - 1$, allora

$$y < b_{-t-1} + 1 \leq \frac{B}{2} \quad \text{e quindi} \quad y < \frac{B}{2}.$$

Viceversa, se $y < \frac{B}{2}$, allora

$$b_{-t-1} \leq y < \frac{B}{2} \quad \text{e quindi} \quad b_{-t-1} < \frac{B}{2}.$$

□

Osserviamo che la precedente dimostrazione mostra anche che, indipendentemente dal fatto che B sia dispari o pari, si ha

$$x < \frac{x_1 + x_2}{2} \quad \Leftrightarrow \quad y = b_{-t-1} + b_{-t-2}B^{-1} + b_{-t-3}B^{-3} + \cdots < \frac{B}{2}$$

e

$$y < \frac{B}{2} \quad \Leftarrow \quad b_{-t-1} \leq \frac{B}{2} - 1 \quad \text{e} \quad y < \frac{B}{2} \quad \Rightarrow \quad b_{-t-1} < \frac{B}{2}.$$

Il precedente teorema dice così che nel caso $B = 10$ si ha

$$x < \frac{x_1 + x_2}{2} \quad \Leftrightarrow \quad b_{-t-1} \leq 4$$

e nel caso $B = 2$ si ha

$$x < \frac{x_1 + x_2}{2} \quad \Leftrightarrow \quad b_{-t-1} = 0.$$

- In generale, se B non è pari, non è sufficiente conoscere b_{-t-1} per determinare il numero di macchina più vicino a x.

  Ad esempio, consideriamo l'insieme di numeri di macchina definito da

  $$B = 3,\ m = -1,\ M = 1,\ t = 2.$$

  Sia

  $$x = (1.111 b_{-4})_3\,,$$

  dove $b_{-3} = b_{-t-1} = 1$ è noto.

  Poichè, anche per B dispari, si ha

  $$x < \frac{x_1 + x_2}{2} \quad \Leftrightarrow \quad b_{-t-1} + b_{-t-2}B^{-1} + b_{t-3}B^{-2} + \cdots < \frac{B}{2},$$

  ne viene che

  $$x < \frac{x_1 + x_2}{2} \quad \Leftrightarrow \quad 1 + b_{-4}3^{-1} < \frac{3}{2} \quad \Leftrightarrow \quad b_{-4} < \frac{3}{2} \quad \Leftrightarrow \quad b_{-4} = 0 \text{ o } b_{-4} = 1.$$

  Quindi, per determinare il numero di macchina più vicino a x è necessario conoscere $b_{-4} = b_{-t-2}$ non solo $b_{-3} = b_{-t-1}$.

Esercizio. Nel caso dell'arrotondamento, provare che per una base B dispari si ha

$$\mathrm{fl}(x) = \begin{cases} x_1 \text{ se } b_{-t-1} < \frac{B}{2} - 1 \\ \\ x_2 \text{ se } b_{-t-1} > \frac{B}{2}. \end{cases}$$

Per un solo valore di b_{-t-1} non si è in grado di determinare $\mathrm{fl}(x)$ usando la formula sopra. Quale? Suggerimento: scrivere $B = 2k + 1$ per qualche $k \in \mathbb{N}$. Per questo valore di b_{-t-1} determinare una condizione dipendente dalle cifre $b_{-t-2}, b_{-t-3}, b_{-t-4}, \ldots$ che sia necessaria e sufficiente per avere

$$x < \frac{x_1 + x_2}{2},$$

cioè $\mathrm{fl}(x) = x_1$.

- Dal teorema precedente si ha quindi, nel caso di B pari,

$$\mathrm{fl}(x) = \begin{cases} x_1 = (b_0.b_{-1}b_{-2}\ldots b_{-t})_B \cdot B^p \ \text{ se } \ b_{-t-1} \leq \frac{B}{2} - 1 \\ \\ x_2 = (b_0.b_{-1}b_{-2}\ldots (b_{-t}+1))_B \cdot B^p \ \text{ se } \ b_{-t-1} \geq \frac{B}{2}. \end{cases}$$

Nell'esempio didattico, con l'arrotondamento si ha

$$\mathrm{fl}(x) = \begin{cases} x_1 = b_0.b_{-1}b_{-2} \cdot 10^p \ \text{ se } \ b_{-3} \leq 4 \\ \\ x_2 = b_0.b_{-1}(b_{-2}+1) \cdot 10^p \ \text{ se } \ b_{-3} \geq 5. \end{cases}$$

e così

$$\mathrm{fl}\left(\frac{2}{3}\right) = \mathrm{fl}\left(0.\overline{6}\right) = \mathrm{fl}\left(6.\overline{6} \cdot 10^{-1}\right) = 6.67 \cdot 10^{-1}.$$

e

$$\mathrm{fl}(\mathrm{e}) = \mathrm{fl}(2.71828\ldots) = 2.72.$$

Nello standard IEEE in doppia precisione, con l'arrotondamento si ha

$$\mathrm{fl}(x) = \begin{cases} x_1 = (1.b_{-1}b_{-2}\ldots b_{-52})_2 \cdot 2^p \ \text{ se } \ b_{-53} = 0 \\ \\ x_2 = (1.b_{-1}b_{-2}\ldots (b_{-52}+1))_2 \cdot 2^p \ \text{ se } \ b_{-53} = 1. \end{cases}$$

- Esercizio. Si consideri l'insieme di numeri di macchina definito dai parametri

$$B = 16,\ m = -3,\ M = 3,\ t = 2.$$

Trovare $\mathrm{fl}\left(\frac{1}{13}\right)$ e $\mathrm{fl}(10000)$ sia nel caso del troncamento che in quello dell'arrotondamento.

- Esercizio. Si consideri l'insieme di numeri di macchina definito dai parametri

$$B = 5,\ m = -2,\ M = 2,\ t = 2.$$

Nel caso dell'arrotondamento, trovare $\mathrm{fl}(x)$ per $x = (441.301)_5$. Fare lo stesso ora per $B = 3$ e $x = (11.1b1)_3$ con $b = 0, 1, 2$.

Errori nell'approssimazione con numeri di macchina

- Consideriamo il troncamento.

  Ricordando che numeri di macchina consecutivi nell'intervallo $\left[B^p, B^{p+1}\right)$ distano $\mathrm{eps} \cdot B^p$, per l'errore assoluto $\varepsilon_a = \mathrm{fl}\,(x) - x$ dell'approssimazione $\mathrm{fl}(x)$ di $x \in \left[B^p, B^{p+1}\right)$ risulta

  $$|\varepsilon_a| = |\mathrm{fl}\,(x) - x| = |x_1 - x| < x_2 - x_1 = \mathrm{eps} \cdot B^p.$$

  essendo $|x_1 - x|$ limitato dall'ampiezza dell'intervallo $[x_1, x_2]$ e non uguale a tale ampiezza in quanto $x < x_2$.

  Si osservi che la maggiorazione $\mathrm{eps} \cdot B^p$ per $|\varepsilon_a|$ è la migliore possibile per numeri $x \in \left[B^p, B^{p+1}\right)$, dal momento che x può essere vicino quanto si vuole a x_2.

Per l'errore relativo $\varepsilon_r = \frac{\varepsilon_a}{x}$ dell'approssimazione $\mathrm{fl}(x)$ di x risulta allora

$$|\varepsilon_r| = \left|\frac{\varepsilon_a}{x}\right| = \frac{|\varepsilon_a|}{x} < \frac{\mathrm{eps} \cdot B^p}{B^p} = \mathrm{eps},$$

essendo $x \geq B^p$.

Si ha quindi

$$|\varepsilon_r| < \mathrm{eps}.$$

- Si ha che tale maggiorazione è ottimale: per ogni $c \in (0, 1)$ esiste un x tale che

$$|\varepsilon_r| = (1 - c)\mathrm{eps} + O(\mathrm{eps}^2).$$

Infatti, consideriamo il numero

$$x = B^p + (1 - c)\mathrm{eps}B^p,$$

dove $c \in (0, 1)$.

Risulta

$$\begin{aligned} x_1 &= B^p \\ x_2 &= \text{"numero di macchina successivo a } x_1\text{"} = B^p + \mathrm{eps}B^p \end{aligned}$$

ricordando che i numeri di macchina nell'intervallo $\left[B^p, B^{p+1}\right)$ distano $\mathrm{eps}B^p$.

Per cui $\mathrm{fl}(x) = x_1 = B^p$ e

$$|\varepsilon_r| = \left|\frac{\mathrm{fl}(x) - x}{x}\right| = \left|\frac{B^p - (B^p + (1-c)\,\mathrm{eps}B^p)}{B^p + (1-c)\,\mathrm{eps}B^p}\right|$$

$$= \left|\frac{(1-c)\,\mathrm{eps}}{1 + (1-c)\,\mathrm{eps}}\right| = \frac{(1-c)\,\mathrm{eps}}{1 + (1-c)\,\mathrm{eps}} = (1-c)\,\mathrm{eps} \cdot \underbrace{\frac{1}{1 + (1-c)\,\mathrm{eps}}}_{=1+O(\mathrm{eps})}$$

$$= (1-c)\,\mathrm{eps} \cdot (1 + O(\mathrm{eps})) = (1-c)\,\mathrm{eps} + O\left(\mathrm{eps}^2\right).$$

- Consideriamo ora l'arrotondamento.

  Per l'errore assoluto risulta

$$|\varepsilon_a| = |\mathrm{fl}(x) - x| \leq \frac{x_2 - x_1}{2} = \frac{\mathrm{eps} \cdot B^p}{2}.$$

  essendo $|\mathrm{fl}(x) - x|$ limitato da metà dell'ampiezza dell'intervallo $[x_1, x_2]$.

  Come nel caso del troncamento, la maggiorazione $\frac{\mathrm{eps} \cdot B^p}{2}$ per $|\varepsilon_a|$ risulta essere la migliore possibile per numeri $x \in \left[B^p, B^{p+1}\right)$, dal momento che per x punto medio dell'intervallo $[x_1, x_2]$ si ha $|\mathrm{fl}(x) - x|$ uguale a metà dell'ampiezza dell'intervallo $[x_1, x_2]$.

Per l'errore relativo risulta allora

$$|\varepsilon_r| = \frac{|\varepsilon_a|}{x} \leq \frac{\frac{\text{eps} \cdot B^p}{2}}{B^p} = \frac{\text{eps}}{2}.$$

Si ha quindi

$$|\varepsilon_r| \leq \frac{\text{eps}}{2}$$

- Si ha che tale maggiorazione è ottimale: esiste un x tale che

$$|\varepsilon_r| = \frac{\text{eps}}{2} + O(\text{eps}^2).$$

Infatti, consideriamo il numero

$$x = B^p + \frac{\text{eps}}{2} B^p,$$

per il quale

$$\begin{aligned} x_1 &= B^p \\ x_2 &= \text{"numero di macchina successivo a } x_1\text{"} = B^p + \text{eps} B^p. \end{aligned}$$

Si ha $\text{fl}(x) = x_2 = B^p + \text{eps} B^p$ e

$$|\varepsilon_r| = \left| \frac{\text{fl}(x) - x}{x} \right| = \left| \frac{B^p + \text{eps} B^p - \left(B^p + \frac{\text{eps}}{2} B^p \right)}{B^p + \frac{\text{eps}}{2} B^p} \right|$$

$$= \left| \frac{\frac{\text{eps}}{2}}{1 + \frac{\text{eps}}{2}} \right| = \frac{\frac{\text{eps}}{2}}{1 + \frac{\text{eps}}{2}} = \frac{\text{eps}}{2} \cdot \underbrace{\frac{1}{1 + \frac{\text{eps}}{2}}}_{= 1 + O(\text{eps})}$$

$$= \frac{\text{eps}}{2} \cdot (1 + O(\text{eps})) = \frac{\text{eps}}{2} + O\left(\text{eps}^2\right).$$

- Concludendo, si può dire che nell'approssimare con il troncamento o l'arrotondamento i numero reali con numeri di macchina in rappresentazione floating point,
  - il massimo errore assoluto $\mathrm{eps}B^p$ o $\frac{\mathrm{eps}}{2}B^p$ per numeri di un ordine di grandezza B^p fissato cresce con l'ordine di grandezza,
  - l'errore relativo rimane invece sempre maggiorato dalla precisione di macchina eps (da eps o da $\frac{\mathrm{eps}}{2}$). Questo spiega il nome "precisione di macchina" attribuito a eps.

  Nel caso della rappresentazione fixed point si ha invece che l'errore assoluto rimane sempre maggiorato da una piccola quantità, mentre l'errore relativo cresce con il decrescere dell'ordine di grandezza di x.

- Osserviamo, infine, che nel caso di underflow nello standard IEEE si ha un errore assoluto

$$\varepsilon_a = \mathrm{fl}(x) - x = 0 - x = -x$$

piccolo, ma un errore relativo

$$\varepsilon_r = \frac{\varepsilon_a}{x} = -1.$$

- Esercizio. Sia nel caso del troncamento che in quello dell'arrotondamento, trovare l'errore relativo ε_r dell'approssimazione $\mathrm{fl}(x)$ di

$$x = B^{p+1} - c\mathrm{eps}B^p,$$

dove $c \in (0, 1)$, e scriverlo nella forma

$$|\varepsilon_r| = C\mathrm{eps} + O\left(\mathrm{eps}^2\right)$$

per un'opportuna costante C.

- Esercizio. Qual è l'errore relativo di $\mathrm{fl}(x)$ in caso di overflow nello standard IEEE?

Aritmetica di macchina

- Un calcolatore esegue sui numeri di macchina le operazioni aritmetiche $+$, $-$, $\cdot$ e $/$. Sia $\circ$ una di queste operazioni.

  In generale, non è detto che $a \circ b$, con a e b numeri di macchina, sia un numero di macchina.

  Nell'esempio didattico, si ha che $a = 30$ e $b = 0.11$ sono numeri di macchina ma

  $$a + b = 30.11$$

  non è un numero di macchina. Ancora, $a = 6.1$ è un numero di macchina ma

  $$a \cdot a = 6.1 \cdot 6.1 = 37.21$$

  non è un numero di macchina.

Pertanto, in generale, il risultato dell'operazione $\circ$ dovrà essere approssimato con un numero di macchina e quindi non sarà $a \circ b$, bensì $\mathrm{fl}\,(a \circ b)$.

Così, associata all'operazione $\circ$ vi è una corrispondente **operazione di macchina** $\tilde{\circ}$ definita da

$$a \,\tilde{\circ}\, b = \mathrm{fl}\,(a \circ b)\,,$$

che è quella che viene effettivamente eseguita sul calcolatore.

Esercizio. Nell'esempio didattico, cosa sono $a \,\tilde{+}\, b$ per $a = 30$ e $b = 0.11$ e $a \,\tilde{\cdot}\, a$ per $a = 6.1$?

Ricordando le maggiorazioni per l'errore relativo quando si approssima un numero reale con un numero di macchina, si ha

$$a \,\tilde{\circ}\, b = (a \circ b)\,(1+\varepsilon)\,, \text{ con } |\varepsilon| \leq \begin{cases} \mathrm{eps} \text{ per il troncamento} \\ \\ \frac{\mathrm{eps}}{2} \text{ per l'arrotondamento,} \end{cases}$$

essendo $\varepsilon = \frac{a \,\tilde{\circ}\, b - a \circ b}{a \circ b}$ l'errore relativo dell'approssimazione $a \,\tilde{\circ}\, b$ di $a \circ b$.

- Un calcolatore esegue sui numeri di macchina anche le funzioni matematiche elementari: radici, funzioni trigonometriche e loro inverse, funzione esponenziale e logaritmo.

  Come nel caso delle operazioni aritmetiche, ad ogni funzione matematica elementare g resta associata una corrispondente funzione di macchina $\widetilde{g}$, che è quella che viene effettivamente eseguita sul calcolatore.

  Si garantisce, come nel caso delle operazioni aritmetiche, che

  $$\widetilde{g}(a) = g(a)(1+\varepsilon), \text{ con } |\varepsilon| \text{ al più dell' ordine di grandezza di } \mathrm{eps}.$$

  per ogni numero di macchina a nel dominio di g.

- Gli errori dovuti all'approssimazione di numeri reali con numeri di macchina e gli errori dovuti all'uso di operazioni di macchina o di funzioni matematiche elementari di macchina sono noti come **errori di arrotondamento** (anche nel caso in cui si usi il troncamento per approssimare con numeri di macchina).

- Esercizio. Si consideri l'insieme di numeri di macchina

$$B = 2,\ m = -10,\ M = 10,\ t = 3.$$

Calcolare con tale insieme di numeri di macchina $a \,\widetilde{\circ}\, b$, per $\circ = +, -, \cdot, /$, nel caso $a = 12$ e $b = 15$ come pure nel caso $a = 25$ e $b = 31$. Ricordare che nel caso in cui a e b non siano numeri di macchina, essi vanno approssimati con numeri di macchina prima di effettuare l'operazione $\circ$.

Propagazione degli errori di arrotondamento

- Consideriamo ora un problema matematico caratterizzato da una funzione dato-risultato $f : D \to \mathbb{R}$, dove $D \subseteq \mathbb{R}^n$ è un aperto. Si assume f di classe C^2.

  Sia $x = (x_1, \ldots, x_n) \in D$ un dato e si voglia calcolare $y = f(x)$ utilizzando un calcolatore.

  In generale, ancora prima di inserirlo su un calcolatore, il dato x non sarà noto in maniera esatta, ma si avrà a disposizione solo una sua approssimazione $\widehat{x} = (\widehat{x}_1, \ldots, \widehat{x}_n)$ con errori relativi sulle componenti

$$\widehat{\varepsilon}_i = \frac{\widehat{x}_i - x_i}{x_i}, \ i \in \{1, \ldots, n\}.$$

  Infatti, spesso $x_1, \ldots, x_n$ sono ottenuti attraverso delle misurazioni, le quali sono inevitabilmente affette da errori.

Quando si calcola il risultato $y = f(x)$ con un calcolatore in realtà si calcolerà

$$\widetilde{y} = \widetilde{f}\left(\mathrm{fl}\left(\widehat{x}\right)\right),$$

dove

$$\mathrm{fl}\left(\widehat{x}\right) = \left(\mathrm{fl}\left(\widehat{x}_1\right), \ldots, \mathrm{fl}\left(\widehat{x}_n\right)\right)$$

è l'approssimazione di $\widehat{x} = \left(\widehat{x}_1, \ldots, \widehat{x}_n\right)$ con numeri di macchina del calcolatore e $\widetilde{f}$ è la funzione che viene effettivamente impiegata dal calcolatore al posto di f.

Le funzioni f e $\widetilde{f}$ sono diverse in quanto nel calcolatore le operazioni aritmetiche e le funzioni matematiche elementari sono sostituite dalle corrispondenti operazioni e funzioni di macchina.

- La funzione f viene calcolata utilizzando un algoritmo che a partire dal dato di input x produce l'output $f(x)$. In generale, una stessa funzione può essere calcolata con più di un algoritmo.

  Nel seguito si considereranno i seguenti due esempi di f:

  - Esempio A:
  $$f(x) = \sqrt{x+1} - \sqrt{x},\ x > 0.$$

  - Esempio B:
  $$f(x) = x_1 x_2 + x_1,\ x \in \mathbb{R}^2.$$

Nell'Esempio A vi sono i seguenti due algoritmi per calcolare i valori di f basati sulle due diverse espressioni per $f(x)$:

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

ALGORITMO 1

$a = x + 1$
$b = \sqrt{a}$
$c = \sqrt{x}$
$y = b - c$

ALGORITMO 2

$a = x + 1$
$b = \sqrt{a}$
$c = \sqrt{x}$
$d = b + c$
$y = 1/d$

Nell'Esempio B vi sono i seguenti due algoritmi per calcolare i valori di f basati sulle due diverse espressioni per $f(x)$:

$$f(x) = x_1 x_2 + x_1 = x_1 (x_2 + 1).$$

ALGORITMO 1

$a = x_1 \cdot x_2$

$y = a + x_1$

ALGORITMO 2

$a = x_2 + 1$

$y = x_1 \cdot a$

Come appare dai due esempi, un algoritmo è costituito da una sequenza di istruzioni, ognuna delle quali è l'esecuzione di un'operazione aritmetica o il calcolo di una funzione matematica elementare (realizzate nel calcolatore con operazioni e funzioni di macchina).

Osserviamo anche che il dato iniziale di un algoritmo è indicato con la variabile x, ma il il reale dato iniziale su cui opera l'algoritmo è $\mathrm{fl}(\widehat{x})$, dove x è il dato.

Esercizio. Scrivere due algoritmi per il calcolo di

$$f(x) = x_1^2 - x_2^2, \ x \in \mathbb{R}^2,$$

e tre algoritmi per il calcolo di

$$f(x) = x_1^4 - x_2^4, \ x \in \mathbb{R}^2.$$

- La funzione $\widetilde{f}$ viene a dipendere dal particolare algoritmo usato per calcolare i valori della funzione f.

  Ad esempio, usando l'esempio didattico di numeri di macchina e la funzione dell'Esempio B con $\mathrm{fl}(\widehat{x}) = (-0.75, -0.75)$, l'Algoritmo 1 fornisce

  $$a = (-0.75) \mathbin{\widetilde{\cdot}} (-0.75) = 0.563$$
  $$y = 0.563 \mathbin{\widetilde{+}} (-0.75) = -0.187$$

  mentre l'Algoritmo 2 fornisce

  $$a = (-0.75) \mathbin{\widetilde{+}} 1 = 0.25$$
  $$y = 0.25 \mathbin{\widetilde{\cdot}} (-0.75) = 0.188$$

  Le funzioni $\widetilde{f}$ corrispondenti a due espressioni diverse della f sono diverse, perché le operazioni aritmetiche di macchina e le funzioni matematiche elementari di macchina non soddisfano quelle proprietà delle operazioni aritmetiche e delle funzioni matematiche elementari che permettono di dire che le due espressioni forniscono gli stessi valori.

Errori inerente, di macchina e algoritmico

- Il nostro obiettivo è ora quello di analizzare l'errore sul risultato

$$\varepsilon_y := \frac{\widetilde{y} - y}{y} = \frac{\widetilde{f}\left(\mathrm{fl}\left(\widehat{x}\right)\right) - f(x)}{f(x)}.$$

A tal fine introduciamo le quantità

$$\widehat{y} := f\left(\widehat{x}\right) \quad \text{e} \quad \overline{y} := f\left(\mathrm{fl}\left(\widehat{x}\right)\right)$$

che sono intermedie tra $y = f(x)$ e $\widetilde{y} = \widetilde{f}(\mathrm{fl}(\widehat{x}))$: si ha

$$y = f(x) \;\rightarrow\; \widehat{y} = f\left(\widehat{x}\right) \;\rightarrow\; \overline{y} = f\left(\mathrm{fl}\left(\widehat{x}\right)\right) \;\rightarrow\; \widetilde{y} = \widetilde{f}(\mathrm{fl}(\widehat{x})).$$

e i seguenti tre errori che andiamo ora a definire.

- L' **errore inerente** è

$$\varepsilon_{\mathrm{in}} := \frac{\widehat{y} - y}{y} = \frac{f(\widehat{x}) - f(x)}{f(x)}$$

ed è l'errore che si ottiene sostituendo in $f(x)$ il dato x con la sua approssimazione $\widehat{x}$; questo errore è indipendente dal calcolatore e dall'algoritmo per calcolare i valori di f (non essendo coinvolte né le approssimazioni fl né la funzione $\widetilde{f}$).

- L' **errore di macchina** è

$$\varepsilon_{\text{mac}} = \frac{\overline{y} - \widehat{y}}{\widehat{y}} = \frac{f(\text{fl}(\widehat{x})) - f(\widehat{x})}{f(\widehat{x})}$$

ed è l'errore che si ottiene sostituendo in $f\left(\widehat{x}\right)$ il dato approssimato $\widehat{x}$ con la sua approssimazione di macchina $\text{fl}\left(\widehat{x}\right)$; questo errore dipende dal calcolatore ma è indipendente dall'algoritmo per calcolare i valori di f (essendo coinvolte le approssimazioni fl ma non la funzione $\widetilde{f}$).

- L' **errore algoritmico** è

$$\varepsilon_{\text{alg}} = \frac{\widetilde{y} - \overline{y}}{\overline{y}} = \frac{\widetilde{f}(\text{fl}(\widehat{x})) - f(\text{fl}(\widehat{x}))}{f(\text{fl}(\widehat{x}))}$$

ed è l'errore che si ottiene sostituendo in $f\left(\text{fl}\left(\widehat{x}\right)\right)$ la funzione f con la sua approssimazione di macchina $\widetilde{f}$; questo errore dipende dal calcolatore e dall' algoritmo per calcolare i valori di f (essendo coinvolta $\widetilde{f}$ che dipende dal calcolatore e dall'algoritmo).

- Vogliamo ora mettere in relazione ε_y con ε_{in}, ε_{mac} e ε_{alg}.

Usando le quantità intermedie $\widehat{y}$ e $\overline{y}$, scriviamo

$$\widetilde{y} - y = \widetilde{y} - \overline{y} + \overline{y} - \widehat{y} + \widehat{y} - y$$

e quindi

$$\begin{aligned}
\varepsilon_y &= \frac{\widetilde{y} - y}{y} = \frac{\widetilde{y} - \overline{y} + \overline{y} - \widehat{y} + \widehat{y} - y}{y} \\
&= \frac{\widetilde{y} - \overline{y}}{y} + \frac{\overline{y} - \widehat{y}}{y} + \frac{\widehat{y} - y}{y} \\
&= \frac{\widetilde{y} - \overline{y}}{\overline{y}} \cdot \frac{\overline{y}}{\widehat{y}} \cdot \frac{\widehat{y}}{y} + \frac{\overline{y} - \widehat{y}}{\widehat{y}} \cdot \frac{\widehat{y}}{y} + \frac{\widehat{y} - y}{y} \\
&= \varepsilon_{\text{alg}} \cdot \frac{\overline{y}}{\widehat{y}} \cdot \frac{\widehat{y}}{y} + \varepsilon_{\text{mac}} \cdot \frac{\widehat{y}}{y} + \varepsilon_{\text{in}}.
\end{aligned}$$

Avendosi

$$\frac{\widehat{y}}{y} = 1 + \varepsilon_{\text{in}} \quad \text{e} \quad \frac{\overline{y}}{\widehat{y}} = 1 + \varepsilon_{\text{mac}},$$

si ottiene

$$\begin{aligned}
\varepsilon_y &= \varepsilon_{\text{alg}} (1 + \varepsilon_{\text{mac}}) (1 + \varepsilon_{\text{in}}) + \varepsilon_{\text{mac}} (1 + \varepsilon_{\text{in}}) + \varepsilon_{\text{in}} \\
&= \varepsilon_{\text{alg}} + \varepsilon_{\text{alg}}\varepsilon_{\text{mac}} + \varepsilon_{\text{alg}}\varepsilon_{\text{in}} + \varepsilon_{\text{alg}}\varepsilon_{\text{mac}}\varepsilon_{\text{in}} + \varepsilon_{\text{mac}} + \varepsilon_{\text{mac}}\varepsilon_{\text{in}} + \varepsilon_{\text{in}}.
\end{aligned}$$

Trascurando i monomi di grado ≥ 2 negli errori si ha

$$\varepsilon_y \doteq \varepsilon_{\text{in}} + \varepsilon_{\text{mac}} + \varepsilon_{\text{alg}}. \tag{7}$$

dove $\doteq$ significa uguale a meno di un termine il cui modulo è minore o oguale di una costante volte $\max\{|\varepsilon_{\text{in}}|, |\varepsilon_{\text{mac}}|, |\varepsilon_{\text{alg}}|\}^2$ per $\max\{|\varepsilon_{\text{in}}|, |\varepsilon_{\text{mac}}|, |\varepsilon_{\text{alg}}|\}$ sufficientemente piccolo.

Esercizio. Mostrare che tale termine in (7) ha modulo minore o oguale a $4\max\{|\varepsilon_{\text{in}}|, |\varepsilon_{\text{mac}}|, |\varepsilon_{\text{alg}}|\}^2$ per $\max\{|\varepsilon_{\text{in}}|, |\varepsilon_{\text{mac}}|, |\varepsilon_{\text{alg}}|\} \leq 1$.

Si considerano solo i monomi di grado 1 negli errori in quanto si è interessati solo all'ordine di grandezza di ε_y e questo è determinato dai soli monomi di grado 1 in quanto gli errori ε_{in}, ε_{mac} e ε_{alg} sono quantità piccole.

Lo studio dell'errore ε_y sul risultato è quindi ridotto allo studio dei tre errori inerente, di macchina e algoritmico.

Analisi dell'errore inerente

L'errore inerente è l'errore di $f(x)$ quando il dato x viene perturbato in $\widehat{x}$.

Per quanto visto nella teoria del condizionamento per funzioni dato-risultato $D \subseteq \mathbb{R}^n \to \mathbb{R}$ (come lo è la f) risulta

$$\varepsilon_{\text{in}} \doteq \sum_{i=1}^{n} K_i(x)\, \widehat{\varepsilon}_i,$$

dove i $K_i(x)$, $i \in \{1, \ldots, n\}$, sono gli indici di condizionamento di f sul dato x.

Per cui, se f è ben condizionata sul dato x, allora ε_{in} ha ordine di grandezza non superiore al massimo ordine di grandezza degli $\widehat{\varepsilon}_i$.

Se, invece, f è mal condizionata sul dato x, allora ε_{in} ha "in generale" ordine di grandezza superiore al massimo ordine di grandezza degli $\widehat{\varepsilon}_i$ ("in generale" vuol dire che lo è per qualche n-upla di errori $\widehat{\varepsilon}_i$, ma non è detto che ciò accada per la n-upla di errori $\widehat{\varepsilon}_i$ che effettivamente si hanno).

- Nell' Esempio A, dove

$$f(x) = \sqrt{x+1} - \sqrt{x},\ x > 0,$$

si ha

$$\begin{aligned} K(x) &= \frac{x}{f(x)} \cdot f'(x). = \frac{x}{\sqrt{x+1}-\sqrt{x}} \cdot \left(\frac{1}{2\sqrt{x+1}} - \frac{1}{2\sqrt{x}} \right) \\ &= \frac{x}{\sqrt{x+1}-\sqrt{x}} \cdot \frac{\sqrt{x}-\sqrt{x+1}}{2\sqrt{x+1}\sqrt{x}} \\ &= -\frac{\sqrt{x}}{2\sqrt{x+1}} = -\frac{1}{2}\sqrt{\frac{x}{x+1}} \end{aligned}$$

e quindi

$$|K(x)| \leq \frac{1}{2}.$$

Per cui, f è ben condizionata su ogni dato x e quindi ε_{in} ha ordine di grandezza non superiore a quello dell'errore $\widehat{\varepsilon}$ di $\widehat{x}$.

In particolare, si ha

$$\varepsilon_{\text{in}} \doteq -\frac{1}{2}\sqrt{\frac{x}{x+1}}\widehat{\varepsilon}.$$

- Nell'Esempio B, dove

$$f(x) = x_1 x_2 + x_1, \ x \in \mathbb{R}^2,$$

si ha

$$\begin{aligned} K_1(x) &= \frac{x_1}{f(x)} \cdot \frac{\partial f}{\partial x_1}(x) = \frac{x_1}{x_1(x_2+1)} \cdot (x_2+1) = 1, \\ K_2(x) &= \frac{x_2}{f(x)} \cdot \frac{\partial f}{\partial x_2}(x) = \frac{x_2}{x_1(x_2+1)} \cdot x_1 = \frac{x_2}{x_2+1} \end{aligned}$$

e quindi f è ben condizionata se e solo se x_2 non è vicino a -1.

Si ha

$$\varepsilon_{\text{in}} \doteq \widehat{\varepsilon}_1 + \frac{x_2}{x_2+1}\widehat{\varepsilon}_2.$$

e ε_{in} ha ordine di grandezza non superiore al massimo ordine di grandezza di $\widehat{\varepsilon}_1$ e $\widehat{\varepsilon}_2$ per x_2 non vicino a -1.

Esercizio. Quando x_2 è vicino a -1 si può concludere che ε_{in} ha ordine di grandezza superiore a quello di $\widehat{\varepsilon}_1$ e $\widehat{\varepsilon}_2$, assunti non nulli e dello stesso ordine di grandezza?

Analisi dell'errore di macchina

- L'errore di macchina è l'errore di $f\left(\widehat{x}\right)$ quando $\widehat{x}$ viene perturbato in $\mathrm{fl}\left(\widehat{x}\right)$.

  Avendosi

  $$\mathrm{fl}\left(\widehat{x}_i\right) = \widehat{x}_i\left(1+\delta_i\right),\ i \in \{1,\ldots n\},$$

  dove δ_i è al più dell'ordine di eps (si ha $|\delta_i| \leq \mathrm{eps}$ per il troncamento e $|\delta_i| \leq \frac{\mathrm{eps}}{2}$ per l'arrotondamento) si ottiene, di nuovo dalla teoria del condizionamento per funzioni dato-risultato $D \subseteq \mathbb{R}^n \to \mathbb{R}$,

  $$\varepsilon_{\mathrm{mac}} \doteq \sum_{i=1}^{n} K_i\left(\widehat{x}\right)\delta_i.$$

- Si osservi che la precedente formula contiene gli indici di condizionamento $K_i\left(\widehat{x}\right)$ relativi a $\widehat{x}$, non gli indici di condizionamento $K_i\left(x\right)$ relativi al dato originale x.

Ma, per $i \in \{1, \ldots, n\}$, con una sviluppo di Taylor al grado zero risulta

$$K_i\left(\widehat{x}\right) = K_i\left(x\right) + O\left(\left\|\widehat{x} - x\right\|_\infty\right).$$

Un tale sviluppo esiste in quanto K_i ha derivate prime continue, cosa che segue dall'essere K definito in termini di f e delle sue derivate prime, con f avente derivate prime e seconde continue.

Quindi

$$\begin{aligned} \varepsilon_{\text{mac}} &\doteq \sum_{i=1}^{n} K_i\left(\widehat{x}\right)\delta_i = \sum_{i=1}^{n}\left(K_i\left(x\right) + O\left(\left\|\widehat{x} - x\right\|_\infty\right)\right)\delta_i \\ &= \sum_{i=1}^{n} K_i\left(x\right)\delta_i + \sum_{i=1}^{n} O\left(\left\|\widehat{x} - x\right\|_\infty\right)\cdot\delta_i. \end{aligned} \tag{8}$$

con

$$\begin{aligned} \left\|\widehat{x} - x\right\|_\infty &= \max_{i\in\{1,\ldots,n\}}\left|\widehat{x}_i - x\right| = \max_{i\in\{1,\ldots,n\}}\left|\widehat{\varepsilon}_i x_i\right| \\ &\leq \max_{i\in\{1,\ldots,n\}}\left|\widehat{\varepsilon}_i\right| \max_{i\in\{1,\ldots,n\}}\left|x_i\right| = \max_{i\in\{1,\ldots,n\}}\left|\widehat{\varepsilon}_i\right| \left\|x\right\|_\infty. \end{aligned}$$

Per cui il secondo termine in (8) ha modulo minore o uguale di una costante volte $\max\{\max_{i\in\{1,\ldots,n\}}\left|\widehat{\varepsilon}_i\right|, \max_{i\in\{1,\ldots,n\}}\left|\delta_i\right|\}^2$ per $\max\{\max_{i\in\{1,\ldots,n\}}\left|\widehat{\varepsilon}_i\right|, \max_{i\in\{1,\ldots,n\}}\left|\delta_i\right|\}$ sufficientemente piccolo.

- Si conclude allora che

$$\varepsilon_{\text{mac}} \doteq \sum_{i=1}^{n} K_i(x)\,\delta_i.$$

Per cui, se f è ben condizionata sul dato x, allora l'errore ε_{mac} ha ordine di grandezza non superiore a quello di eps.

Se invece f è mal condizionata sul dato x, allora l'errore ε_{mac} ha "in generale" ordine di grandezza superiore a quello di eps.

Nell'Esempio A, ε_{mac} ha ordine di grandezza non superiore a eps per ogni dato.

Nell'Esempio B, ε_{mac} ha ordine di grandezza non superiore a eps se il dato x ha x_2 non vicino a -1.

Analisi dell'errore algoritmico

- L'errore algoritmico ε_{alg} è dovuto alla sostituzione di f con $\widetilde{f}$ nel calcolo di $f(\text{fl}(\widehat{x}))$ e dipende dal particolare algoritmo usato per calcolare i valori di f.

  Si supponga che l'algoritmo consista di m istruzioni, ognuna delle quali è un'operazione aritmetica o il calcolo di una funzione matematica elementare.

- Si assuma che la j-esima istruzione, $j \in \{1, \ldots, m\}$, consista in un'operazione aritmetica

$$c = a \circ b.$$

  L'operazione non sarà eseguita sui valori a e b ottenibili dai numeri di macchina di input $\text{fl}(\widehat{x}_1), \ldots, \text{fl}(\widehat{x}_n)$ con operazioni aritmetiche esatte e funzioni matematiche elementari esatte eseguite nelle istruzioni precedenti alla j-esima, ma su dalle approssimazioni $\widetilde{a}$ e $\widetilde{b}$ ottenute da $\text{fl}(\widehat{x}_1), \ldots, \text{fl}(\widehat{x}_n)$ con operazioni aritmetiche di macchina e funzioni matematiche elementari di macchina.

Dal momento poi che l'operazione $\circ$ è approssimata dall'operazione di macchina $\widetilde{\circ}$, il risultato dell'operazione sarà allora

$$\widetilde{c} = \widetilde{a} \,\widetilde{\circ}\, \widetilde{b}$$

e non $c = a \circ b$.

Siano

$$\beta_a = \frac{\widetilde{a} - a}{a} \quad \text{e} \quad \beta_b = \frac{\widetilde{b} - b}{b}$$

gli errori delle approssimazioni $\widetilde{a}$ di a e $\widetilde{b}$ di b.

Vogliamo trovare come l'errore

$$\beta_c = \frac{\widetilde{c} - c}{c}$$

dell'approssimazione $\widetilde{c}$ di c è in relazione con β_a e β_b.

Si ha

$$\widetilde{c} = \widetilde{a} \,\widetilde{\circ}\, \widetilde{b} = \left(\widetilde{a} \circ \widetilde{b}\right)(1 + \gamma_j),$$

dove γ_j è al più dell'ordine di eps.

Si consideri ora la funzione

$$h(\alpha, \beta) = \alpha \circ \beta, \ (\alpha, \beta) \in \mathbb{R}^2.$$

L'errore

$$\lambda_j = \frac{\widetilde{a} \circ \widetilde{b} - a \circ b}{a \circ b}$$

di $\widetilde{a} \circ \widetilde{b} = h(\widetilde{a}, \widetilde{b})$ rispetto ad $a \circ b = h(a, b)$ è l'errore di $h(a, b)$ quando il dato (a, b) viene perturbato in $(\widetilde{a}, \widetilde{b})$.

Pertanto, dalla teoria del condizionamento si ottiene

$$\lambda_j \doteq K_1(a, b)\,\beta_a + K_2(a, b)\,\beta_b,$$

dove $K_1(a, b)$ e $K_2(a, b)$ sono gli indici di condizionamento della funzione h sul dato (a, b).

Si ha allora

$$\begin{aligned}\widetilde{c} &= \widetilde{a} \,\widetilde{\circ}\, \widetilde{b} = \left(\widetilde{a} \circ \widetilde{b}\right)(1+\gamma_j) \\ &= (a \circ b)(1+\lambda_j)(1+\gamma_j) \\ &= c\,(1+\lambda_j+\gamma_j+\lambda_j\gamma_j)\end{aligned}$$

e quindi

$$\beta_c = \frac{\widetilde{c}-c}{c} = \lambda_j + \gamma_j + \lambda_j\gamma_j.$$

Trascurando il monomio di grado 2 negli errori si ottiene

$$\beta_c \doteq \lambda_j + \gamma_j \doteq K_1(a,b)\,\beta_a + K_2(a,b)\,\beta_b + \gamma_j. \tag{9}$$

La formula (9) dice come gli errori β_a e β_b sugli operandi a e b dell'operazione $a \circ b$ e l'errore γ_j dell'operazione si propagano sul risultato c.

- Si assuma ora che la $j-$esima istruzione

$$c = g(a).$$

consista nel calcolo di una funzione matematica elementare g.

Il risultato dell'operazione sarà

$$\widetilde{c} = \widetilde{g}(\widetilde{a})$$

e non $c = g(a)$, dove $\widetilde{a}$ è ottenuta da $\mathrm{fl}(\widehat{x}_1), \ldots, \mathrm{fl}(\widehat{x}_n)$ con operazioni aritmetiche di macchina e funzioni matematiche elementari di macchina eseguite nelle istruzioni precedenti la $j-$esima. Sia β_a l'errore dell'approssimazione $\widetilde{a}$.

Si ha

$$\widetilde{c} = \widetilde{g}(\widetilde{a}) = g(\widetilde{a})(1 + \gamma_j),$$

dove γ_j è al più dell'ordine di eps, e

$$g(\widetilde{a}) = g(a)(1 + \lambda_j)$$

con

$$\lambda_j = \frac{g(\widetilde{a}) - g(a)}{g(a)} \doteq K(a)\,\beta_a,$$

dove $K(a)$ è l'indice di condizionamento di g sul dato a.

Per cui

$$\begin{aligned} \widetilde{c} &= g\left(\widetilde{a}\right)\left(1+\gamma_j\right) \\ &= g\left(a\right)\left(1+\lambda_j\right)\left(1+\gamma_j\right) \\ &= c\left(1+\lambda_j+\gamma_j+\lambda_j\gamma_j\right) \end{aligned}$$

e quindi

$$\beta_c \doteq \lambda_j + \gamma_j \doteq K\left(a\right)\beta_a + \gamma_j. \tag{10}$$

La formula (10) dice come l'errore β_a sull'argomento a in $c = g(a)$ e l'errore γ_j nel calcolo di g si propagano sul risultato c.

- Notiamo che l'errore algoritmico ε_{alg} è l'errore β_c quando l'istruzione

$$c = a \circ b \quad \text{oppure} \quad c = g(a)$$

è l'ultima dell'algoritmo.

Infatti, nel caso dell'ultima istruzione si ha

$$c = a \circ b = f\left(\text{fl}\left(\widehat{x}\right)\right) = \overline{y} \quad \text{oppure} \quad c = g(a) = f\left(\text{fl}\left(\widehat{x}\right)\right) = \overline{y}$$

e

$$\widetilde{c} = \widetilde{a} \,\widetilde{\circ}\, \widetilde{b} = \widetilde{f}\left(\text{fl}\left(\widehat{x}\right)\right) = \widetilde{y} \quad \text{oppure} \quad \widetilde{c} = \widetilde{g}\left(\widetilde{a}\right) = \widetilde{f}\left(\text{fl}\left(\widehat{x}\right)\right) = \widetilde{y}.$$

Quindi

$$\varepsilon_{\text{alg}} = \frac{\widetilde{y} - \overline{y}}{\overline{y}} = \frac{\widetilde{c} - c}{c} = \beta_c.$$

- Al fine di ottenere ε_{alg} occorre quindi applicare a tutte le istruzioni dell'algoritmo, in ordine dalla prima all'ultima istruzione, le formule di propagazione. In particolare

  - la formula
    $$\beta_c \doteq K_1(a,b)\,\beta_a + K_2(a,b)\,\beta_b + \gamma_j$$
    se l'istruzione è del tipo $c = a \circ b$;

  - la formula
    $$\beta_c \doteq K(a)\,\beta_a + \gamma_j$$
    se l'istruzione è del tipo $c = g(a)$.

Nell'applicare le formule di propagazione bisogna procedere nell'ordine dalla prima all'ultima istruzione in quanto β_a e β_b sono quantità β_c di precedenti istruzioni, oppure quantità $\beta_{x_1}, \ldots, \beta_{x_n}$, dove $x_1, \ldots, x_n$ sono le variabili dell'algoritmo che contengono i numeri di macchina di input $\mathrm{fl}\left(\widehat{x}_1\right), \ldots, \mathrm{fl}\left(\widehat{x}_n\right)$.

Si osservi che per tali variabili $x_1, \ldots, x_n$, si ha

$$\beta_{x_1} = \cdots = \beta_{x_n} = 0.$$

Infatti, per una variabile c dell'algoritmo, l'errore β_c sorge solo perché c è ottenuta in una istruzione $c = a \circ b$ come risultato di una operazione di macchina o in una istruzione $c = g(a)$ come risultato di una funzione matematica elementare di macchina. Invece, le variabili $x_1, \ldots, x_n$ non sono ottenute in nessuna istruzione come risultato, esse sono le variabili iniziali.

- Per l'Algoritmo 1 dell'Esempio A si ha

1 $a = x + 1$ $\quad \beta_a \doteq \frac{x}{x+1} \cdot \underbrace{\beta_x}_{=0 \text{ essendo } x \text{ un dato iniziale}} + \frac{1}{x+1} \cdot \underbrace{\beta_1}_{=0 \text{ essendo 1 un numero di macchina}} + \gamma_1 = \gamma_1$

2 $b = \sqrt{a}$ $\quad \beta_b \doteq \frac{1}{2}\beta_a + \gamma_2 \doteq \frac{1}{2}\gamma_1 + \gamma_2$

3 $c = \sqrt{x}$ $\quad \beta_c \doteq \frac{1}{2} \cdot \underbrace{\beta_x}_{=0} + \gamma_3 = \gamma_3$

4 $y = b - c$ $\quad \varepsilon_{\text{alg}} = \beta_y \doteq \frac{b}{b-c}\beta_b - \frac{c}{b-c}\beta_c + \gamma_4$

$$\doteq \frac{1}{2} \cdot \frac{b}{b-c}\gamma_1 + \frac{b}{b-c}\gamma_2 - \frac{c}{b-c}\gamma_3 + \gamma_4.$$

Essendo

$$b = \sqrt{\mathrm{fl}\left(\widehat{x}\right) + 1} \text{ e } \; c = \sqrt{\mathrm{fl}\left(\widehat{x}\right)},$$

si ottiene

$$\begin{aligned} \varepsilon_{\mathrm{alg}} \doteq & \frac{1}{2} \cdot \frac{\sqrt{\mathrm{fl}\left(\widehat{x}\right) + 1}}{\sqrt{\mathrm{fl}\left(\widehat{x}\right) + 1} - \sqrt{\mathrm{fl}\left(\widehat{x}\right)}} \gamma_1 + \frac{\sqrt{\mathrm{fl}\left(\widehat{x}\right) + 1}}{\sqrt{\mathrm{fl}\left(\widehat{x}\right) + 1} - \sqrt{\mathrm{fl}\left(\widehat{x}\right)}} \gamma_2 \\ & - \frac{\sqrt{\mathrm{fl}\left(\widehat{x}\right)}}{\sqrt{\mathrm{fl}\left(\widehat{x}\right) + 1} - \sqrt{\mathrm{fl}\left(\widehat{x}\right)}} \gamma_3 + \gamma_4 . \end{aligned}$$

- In generale, si ottiene per $\varepsilon_{\mathrm{alg}}$ un'espressione del tipo

$$\varepsilon_{\mathrm{alg}} \doteq \sum_{j=1}^{m} M_j \left(\mathrm{fl}\left(\widehat{x}\right)\right) \gamma_j,$$

dove, per $j \in \{1, \ldots, n\}$, $M_j : D \to \mathbb{R}$.

Assumendo, per $j \in \{1, \ldots, n\}$, M_j di classe C^1 si ha da uno sviluppo di Taylor di grado zero

$$M_j\left(\mathrm{fl}\left(\widehat{x}\right)\right) = M_j\left(x\right) + O\left(\left\|\mathrm{fl}\left(\widehat{x}\right) - x\right\|_\infty\right)$$

Quindi

$$\begin{aligned}
\varepsilon_{\mathrm{alg}} &\doteq \sum_{j=1}^{m} M_j\left(\mathrm{fl}\left(\widehat{x}\right)\right)\gamma_j = \sum_{j=1}^{m}\left(M_j\left(x\right) + O\left(\left\|\mathrm{fl}\left(\widehat{x}\right) - x\right\|_\infty\right)\right)\gamma_j \\
&= \sum_{j=1}^{m} M_j\left(x\right)\gamma_j + \sum_{j=1}^{m} O\left(\left\|\mathrm{fl}\left(\widehat{x}\right) - x\right\|_\infty\right)\gamma_j. \qquad (11)
\end{aligned}$$

con

$$\begin{aligned}
&\left\|\mathrm{fl}\left(\widehat{x}\right) - x\right\|_\infty = \max_{i\in\{1,\ldots,n\}}\left|\mathrm{fl}\left(\widehat{x}_i\right) - x_i\right| = \max_{i\in\{1,\ldots,n\}}\left|\mathrm{fl}\left(\widehat{x}_i\right) - \widehat{x}_i + \widehat{x}_i - x_i\right| \\
&= \max_{i\in\{1,\ldots,n\}}\left|\delta_i\widehat{x}_i + \widehat{\varepsilon}_i x_i\right| = \max_{i\in\{1,\ldots,n\}}\left|\delta_i\left(1+\widehat{\varepsilon}_i\right)x_i + \widehat{\varepsilon}_i x_i\right| \\
&= \max_{i\in\{1,\ldots,n\}}\left|\delta_i\left(1+\widehat{\varepsilon}_i\right) + \widehat{\varepsilon}_i\right|\left|x_i\right| \leq \max_{i\in\{1,\ldots,n\}}\left|\delta_i\left(1+\widehat{\varepsilon}_i\right) + \widehat{\varepsilon}_i\right| \cdot \left\|x\right\|_\infty.
\end{aligned}$$

Per cui il secondo termine in (11) ha modulo minore o uguale di una costante volte $\max\{\max_{i\in\{1,\ldots,n\}}|\widehat{\varepsilon}_i|, \max_{i\in\{1,\ldots,n\}}|\delta_i|, \max_{j\in\{1,\ldots,m\}}|\gamma_j|\}^2$ per $\max\{\max_{i\in\{1,\ldots,n\}}|\widehat{\varepsilon}_i|, \max_{i\in\{1,\ldots,n\}}|\delta_i|, \max_{j\in\{1,\ldots,m\}}|\gamma_j|\}$ sufficientemente piccolo

- Risulta allora

$$\varepsilon_{\text{alg}} \doteq \sum_{j=1}^{m} M_j(x)\,\gamma_j.$$

Le funzioni M_j, $j \in \{1, \ldots m\}$, sono dette **indici di stabilità** dell'algoritmo. Il valore $M_j(x)$ dice quanto l'errore γ_j introdotto nell'esecuzione della $j-$esima operazione si propaghi sull'errore ε_{alg} quando il dato è x.

L'indice di stabilità M_m dell'ultima istruzione è costantemente uguale a 1:

$$M_m(x) = 1,\ x \in D,$$

dal momento che γ_m compare solo nella formula di propagazione

$\varepsilon_{\text{alg}} = \beta_c \doteq K_1(a,b)\beta_a + K_2(a,b)\beta_b + \gamma_m$ oppure $\varepsilon_{\text{alg}} = \beta_c \doteq K(a)\beta_a + \gamma_m$

dell'ultima istruzione

$$c = a \circ b \text{ oppure } c = g(a).$$

Definizione

L'algoritmo usato per calcolare i valori di f si dice **stabile** sul dato x se tutti gli indici di stabilità $M_j(x)$, $j \in \{1, \ldots, m-1\}$, hanno ordine di grandezza non superiore all'unità. L'algoritmo si dice **instabile** sul dato x se non è stabile, cioè esiste un indice di stabilità $M_j(x)$, $j \in \{1, \ldots, m-1\}$, che ha ordine di grandezza superiore a all'unità.

Pertanto, se l'algoritmo è stabile sul dato x, allora da

$$\varepsilon_{\text{alg}} \doteq \sum_{j=1}^{m} M_j(x)\,\gamma_j$$

segue che l'errore ε_{alg} ha ordine di grandezza non superiore a quello di eps.

Se, invece, l'algoritmo è instabile sul dato x, allora ε_{alg} ha "in generale" ordine di grandezza superiore a quello di eps: come per il condizionamento, "in generale" vuol dire che lo è per qualche $m-$upla di errori γ_j, ma non è detto che ciò accada per la $m-$upla di errori γ_j che effettivamente si hanno.

- Nell'algoritmo esaminato sopra, che è quello dell'Algoritmo 1 dell'Esempio A, si ha, per $x > 0$,

$$\begin{aligned} M_1(x) &= \frac{1}{2} \cdot \frac{\sqrt{x+1}}{\sqrt{x+1}-\sqrt{x}} = \frac{1}{2}\sqrt{x+1}\left(\sqrt{x+1}+\sqrt{x}\right), \\ M_2(x) &= \frac{\sqrt{x+1}}{\sqrt{x+1}-\sqrt{x}} = \sqrt{x+1}\left(\sqrt{x+1}+\sqrt{x}\right), \\ M_3(x) &= -\frac{\sqrt{x}}{\sqrt{x+1}-\sqrt{x}} = -\sqrt{x}\left(\sqrt{x+1}+\sqrt{x}\right). \end{aligned}$$

Poichè

$$x = \frac{1}{2}\sqrt{x}\left(\sqrt{x}+\sqrt{x}\right) \leq M_1(x) \leq \frac{1}{2}\sqrt{x+1}\left(\sqrt{x+1}+\sqrt{x+1}\right) = x+1$$

e, analogamente,

$$\begin{aligned} 2x &\leq M_2(x) = \sqrt{x+1}\left(\sqrt{x+1}+\sqrt{x}\right) \leq 2(x+1) \\ 2x &\leq |M_3(x)| = \sqrt{x}\left(\sqrt{x+1}+\sqrt{x}\right) \leq 2(x+1), \end{aligned}$$

l'algoritmo risulta stabile se e solo se x non è grande.

Per l'Algoritmo 2 dell'Esempio A si ha:

1 $a = x + 1$ $\quad \beta_a \doteq \gamma_1$

2 $b = \sqrt{a}$ $\quad \beta_b \doteq \frac{1}{2}\beta_a + \gamma_2 \doteq \frac{1}{2}\gamma_1 + \gamma_2$

3 $c = \sqrt{x}$ $\quad \beta_c \doteq \gamma_3$

4 $d = b + c$ $\quad \beta_d \doteq \frac{b}{b+c}\beta_b + \frac{c}{b+c}\beta_c + \gamma_4 \doteq \frac{1}{2}\,\frac{b}{b+c}\gamma_1 + \frac{b}{b+c}\gamma_2 + \frac{c}{b+c}\gamma_3 + \gamma_4$

5 $y = 1/d$ $\quad \varepsilon_{\text{alg}} \doteq -\beta_d + \gamma_5 = -\frac{1}{2}\,\frac{b}{b+c}\gamma_1 - \frac{b}{b+c}\gamma_2 - \frac{c}{b+c}\gamma_3 - \gamma_4 + \gamma_5.$

Per cui, per $x > 0$,

$$M_1(x) = -\frac{1}{2} \cdot \frac{\sqrt{x+1}}{\sqrt{x+1} + \sqrt{x}},\ M_2(x) = -\frac{\sqrt{x+1}}{\sqrt{x+1} + \sqrt{x}},$$

$$M_3(x) = -\frac{\sqrt{x}}{\sqrt{x+1} + \sqrt{x}},\ M_4(x) = -1.$$

Poichè $|M_j(x)| \leq 1$, $j \in \{1, 2, 3, 4\}$, l'algoritmo è stabile per ogni dato.

- Consideriamo ora l'Esempio B.

  Per l'Algoritmo 1 si ha

  $$1 \quad a = x_1 \cdot x_2 \qquad \beta_a \doteq \gamma_1$$

  $$2 \quad y = a + x_1 \qquad \varepsilon_{\text{alg}} \doteq \frac{a}{a+x_1}\beta_a + \gamma_2 \doteq \frac{a}{a+x_1}\gamma_1 + \gamma_2.$$

  Quindi, per $x \in \mathbb{R}^2$, si ha

  $$M_1(x) = \frac{x_1 x_2}{x_1 x_2 + x_1} = \frac{x_2}{x_2 + 1}$$

  e l'algoritmo è stabile se e solo se x_2 non è a vicino a -1.

  Per l'Algoritmo 2 si ha

  $$1 \quad a = x_2 + 1 \qquad \beta_a \doteq \gamma_1$$

  $$2 \quad y = x_1 \cdot a \qquad \varepsilon_{\text{alg}} \doteq \beta_a + \gamma_2 \doteq \gamma_1 + \gamma_2.$$

  Quindi, per $x \in \mathbb{R}^2$,

  $$M_1(x) = 1$$

  e l'algoritmo è stabile per ogni dato.

- Esercizio. Si studi il condizionamento di

$$f(x) = \sqrt{x_1 + x_2} - \sqrt{x_1},\ x \in \mathbb{R}^2 \text{ con } x_1, x_2 > 0,$$

e la stabilità dei due algoritmi basati sulle due diverse espressioni

$$f(x) = \sqrt{x_1 + x_2} - \sqrt{x_1} = \frac{x_2}{\sqrt{x_1 + x_2} + \sqrt{x_1}}.$$

- Esercizio. Si studi il condizionamento di

$$f(x) = x_1 x_2 + x_1 x_3,\ x \in \mathbb{R}^3,$$

e la stabilità dei due algoritmi basati sulle due diverse espressioni

$$f(x) = x_1 x_2 + x_1 x_3 = x_1 (x_2 + x_3).$$

- Esercizio. Si studi il condizionamento di

$$f(x) = \log \frac{x+1}{x},\ x > 0,$$

e la stabilità dei tre algoritmi basati sulle tre diverse espressioni

$$f(x) = \log \frac{x+1}{x} = \log \left(1 + \frac{1}{x}\right) = \log(x+1) - \log x.$$

Alcune considerazioni

- Fissato un dato x, nel processo di calcolare $f(x)$ vengono commessi i seguenti errori:
  - errori di misurazione: $\widehat{\varepsilon}_i$, $i \in \{1, \ldots, n\}$;
  - errori di rappresentazione in macchina dei dati misurati: δ_i, $i \in \{1, \ldots, n\}$;
  - errori nell'esecuzione delle operazioni aritmetiche e delle funzioni matematiche elementari: γ_j, $j \in \{1, \ldots, m\}$.

Gli errori δ_i e gli errori γ_j hanno ordine di grandezza non superiore a (quello di) eps.

Assumiamo poi che gli errori $\widehat{\varepsilon}_i$ abbiamo ordine di grandezza non superiore a (quello di) una tolleranza prefissata TOL.

In genere, si ha $\mathrm{TOL} >> \mathrm{eps}$: l'ordine di grandezza di TOL va in genere da 10^{-2} a 10^{-5}, mentre eps ha ordine di grandezza 10^{-16} nello standard IEEE in doppia precisione.

Si possono presentare le seguenti quattro situazioni che ora elenchiamo nel calcolare $f(x)$.

- Situazione 1. f è ben condizionata sul dato x e l'algoritmo usato per calcolare i valori di f è stabile sul dato x.

Nella Situazione 1, ε_{in} ha ordine di grandezza non superiore a TOL, ε_{mac} ha ordine di grandezza non superiore a eps e ε_{alg} ha ordine di grandezza non superiore a eps.

Quindi, $\varepsilon_y \doteq \varepsilon_{\text{in}} + \varepsilon_{\text{mac}} + \varepsilon_{\text{alg}}$ ha ordine di grandezza non superiore a $\max\{\text{TOL}, \text{eps}\}$, che è uguale a TOL nel caso usuale $\text{TOL} >> \text{eps}$.

Nell'Esempio A con l'Algoritmo 1 si è nella Situazione 1 quando x non è grande, con l'Algoritmo 2 si è nella Situazione 1 qualunque sia x.

Nell'Esempio B, sia con l'Algoritmo 1 che con l'Algoritmo 2, si è nella Situazione 1 quando x_2 non è vicino a -1.

- Situazione 2. *f* è ben condizionata sul dato *x* e l'algoritmo usato per calcolare i valori di *f* è instabile sul dato *x*.

  Nella Situazione 2, ε_{in} ha ordine di grandezza non superiore a TOL, ε_{mac} ha ordine di grandezza non superiore a eps e ε_{alg} ha "in generale" ordine di grandezza superiore a eps.

  Quindi, nel caso usuale $\text{TOL} >> \text{eps}$, solo per una forte instabilità dell'algoritmo caratterizzata da indici di stabilità con ordine di grandezza superiore a $\frac{\text{TOL}}{\text{eps}}$, si ha "in generale" che ε_{alg} ha ordine di grandezza superiore a TOL e, quindi, che $\varepsilon_y \doteq \varepsilon_{\text{in}} + \varepsilon_{\text{mac}} + \varepsilon_{\text{alg}} \approx \varepsilon_{\text{alg}}$ ha ordine di grandezza superiore a TOL.

  La Situazione 2 è una non favorevole a causa dell'instabilità dell'algoritmo, il quale va sostituito con un algoritmo stabile per finire nella Situazione 1 che è favorevole.

  Nell'Esempio A con l'Algoritmo 1 si è nella Situazione 2 quando *x* è grande. L'Algoritmo 1 va sostituito con l'Algoritmo 2 che è stabile. Nell'Esempio B non si è mai nella Situazione 2.

- Situazione 3. f è mal condizionata sul dato x e l'algoritmo usato per calcolare i valori di f è stabile sul dato x.

Nella Situazione 3, ε_{in} ha "in generale" ordine di grandezza superiore a TOL e ε_{mac} ha "in generale" ordine di grandezza superiore a eps e ε_{alg} ha ordine di grandezza non superiore a eps.

Quindi, $\varepsilon_y \doteq \varepsilon_{\text{in}} + \varepsilon_{\text{mac}} + \varepsilon_{\text{alg}}$ ha "in generale" ordine di grandezza superiore a $\max\{\text{TOL}, \text{eps}\}$.

La Situazione 3 è una non favorevole a causa del mal condizionamento di f. La stabilità dell'algoritmo non è di alcuna utilità.

Nell'Esempio B con l'Algoritmo 2 si è nella Situazione 3 per x_2 vicino a -1. Nell'Esempio A non si è mai nella Situazione 3.

- Situazione 4. f è mal condizionata sul dato x e l'algoritmo usato per calcolare i valori di f è instabile sul dato x.

  Nella situazione 4, ε_{in} ha "in generale" ordine di grandezza superiore a TOL, ε_{mac} ha "in generale" ordine di grandezza superiore a eps e ε_{alg} ha "in generale" ordine di grandezza superiore a eps.

  Quindi, $\varepsilon_y \doteq \varepsilon_{\text{in}} + \varepsilon_{\text{mac}} + \varepsilon_{\text{alg}}$ ha "in generale" ordine di grandezza superiore a $\max\{\text{TOL}, \text{eps}\}$ come nella situazione 3.

  La Situazione 4 è una non favorevole a causa del mal condizionamento di f e dell'instabilità dell'algoritmo. A differenza della Situazione 2, l'instabilità dell'algoritmo non è il fattore determinante: rimpiazzando l'algoritmo instabile con uno stabile si finisce nella Situazione 3, dove, comunque, ε_y ha "in generale" ordine di grandezza superiore a $\max\{\text{TOL}, \text{eps}\}$.

  Nell'Esempio B con l'Algoritmo 1 si è nella Situazione 4 per x_2 vicino a -1. Nell'Esempio A non si è mai nella Situazione 4.

Esercizio. Si supponga di essere nella Situazione 4 con il dato x tale che $x = \widehat{x} = \mathrm{fl}\left(\widehat{x}\right)$, cioè x è un dato che non si ottiene da una misurazione e le componenti $x_1, \ldots, x_n$ di x sono numeri di macchina. E' utile in questo caso sostituire l'algoritmo instabile per calcolare i valori di f con uno stabile?

Dati con molte componenti e algoritmi con molti passi

- Si assuma di essere nella Situazione 1. Si ha che ε_{in} ha ordine di grandezza non superiore a TOL, ε_{mac} ha ordine di grandezza non superiore a eps e ε_{alg} ha ordine di grandezza non superiore a eps.

  Però questo accade nel caso in cui il numero n di componenti del dato non è grande e il numero m di istruzioni dell'algoritmo non è grande.

  Se questo non è vero, dalle formule

$$\varepsilon_{\text{in}} \doteq \sum_{i=1}^{n} K_i(x)\,\widehat{\varepsilon}_i, \quad \varepsilon_{\text{mac}} \doteq \sum_{i=1}^{n} K_i(x)\,\delta_i, \quad \varepsilon_{\text{alg}} \doteq \sum_{j=1}^{m} M_j(x)\,\gamma_j$$

  si ottiene che ε_{in} ha ordine di grandezza non superiore a $n\,\text{TOL}$, ε_{mac} ha ordine di grandezza non superiore a $n\,\text{eps}$ e ε_{alg} ha ordine di grandezza non superiore a $m\,\text{eps}$., dal momento che i $K_i(x)$ e gli $M_j(x)$ hanno ordine di grandezza non superiore all'unità.

  In realtà, le cose vanno meglio di quanto appena detto, come ora vediamo.

- Assumiamo che gli errori abbiano una struttura probabilistica, vale a dire

  - gli errori $\widehat{\varepsilon}_i$ siano variabili casuali indipendenti di media $\mathbb{E}(\widehat{\varepsilon}_i) = 0$ e deviazione standard $\mathrm{SD}(\widehat{\varepsilon}_i)$ con ordine di grandezza TOL;
  - gli errori δ_i siano variabili casuali indipendenti di media $\mathbb{E}(\delta_i) = 0$ e deviazione standard $\mathrm{SD}(\delta_i)$ con ordine di grandezza eps;
  - gli errori γ_j siano variabili casuali indipendenti di media $\mathbb{E}(\gamma_j) = 0$ e deviazione standard $\mathrm{SD}(\gamma_j)$ con ordine di grandezza eps.

  Quindi

  - i termini $K_i(x)\widehat{\varepsilon}_i$ sono variabili casuali indipendenti di media $K_i(x)\mathbb{E}(\widehat{\varepsilon}_i) = 0$ e deviazione standard $|K_i(x)|\mathrm{SD}(\widehat{\varepsilon}_i)$ con ordine di grandezza TOL;
  - i termini $K_i(x)\delta_i$ sono variabili casuali indipendenti di media $K_i(x)\mathbb{E}(\delta_i) = 0$ e deviazione standard $|K_i(x)|\mathrm{SD}(\delta_i)$ con ordine di grandezza eps;
  - i termini $M_j(x)\gamma_j$ sono variabili casuali indipendenti di media $K_i(x)\mathbb{E}(\gamma_j) = 0$ e deviazione standard $|K_i(x)|\mathrm{SD}(\gamma_j)$ con ordine di grandezza eps.

Poiché la media di una somma di variabili casuali è la somma delle medie e la varianza (il quadrato della deviazione standard) di una somma di variabili casuali indipendenti è la somma delle varianze,

- $\varepsilon_{\text{in}} = \sum_{i=1}^{n} K_i(x)\,\widehat{\varepsilon}_i$ è una variabile casuale di media $\sum_{i=1}^{n} \mathbb{E}(K_i(x)\,\widehat{\varepsilon}_i) = 0$ e deviazione standard $\sqrt{\sum_{i=1}^{n} \mathrm{SD}(K_i(x)\,\widehat{\varepsilon}_i)^2}$ con ordine di grandezza $\sqrt{n\,\mathrm{TOL}^2} = \sqrt{n}\,\mathrm{TOL}$;
- $\varepsilon_{\text{mac}} = \sum_{i=1}^{n} K_i(x)\,\delta_i$ è una variabile casuale di media $\sum_{i=1}^{n} \mathbb{E}(K_i(x)\,\delta_i) = 0$ e deviazione standard $\sqrt{\sum_{i=1}^{n} \mathrm{SD}(K_i(x)\,\delta_i)^2}$ con ordine di grandezza $\sqrt{n\,\mathrm{eps}^2} = \sqrt{n}\,\mathrm{eps}$;
- $\varepsilon_{\text{alg}} = \sum_{j=1}^{m} M_j(x)\,\gamma_j$ è una variabile casuale di media $\sum_{j=1}^{m} \mathbb{E}(M_j(x)\,\gamma_j) = 0$ e deviazione standard $\sqrt{\sum_{i=1}^{m} \mathrm{SD}(M_j(x)\,\gamma_j)^2}$ con ordine di grandezza $\sqrt{m\,\mathrm{eps}^2} = \sqrt{m}\,\mathrm{eps}$.

Ricordiamo ora la disuguaglianza di Cebicev

$$\mathbb{P}(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2},$$

dove $k > 0$ e X è una variabile casuale di media μ e deviazione standard σ.

Da questa risulta, per $k > 0$,

$$\mathbb{P}(|\varepsilon_{\text{in}}| < k\text{SD}(\varepsilon_{\text{in}})) \geq 1 - \frac{1}{k^2}$$

$$\mathbb{P}(|\varepsilon_{\text{mac}}| < k\text{SD}(\varepsilon_{\text{mac}})) \geq 1 - \frac{1}{k^2}$$

$$\mathbb{P}(|\varepsilon_{\text{alg}}| < k\text{SD}(\varepsilon_{\text{alg}})) \geq 1 - \frac{1}{k^2}$$

e quindi con una buona probabilità si ha che ε_{in} ha ordine di grandezza non superiore a $\sqrt{n}\,\text{TOL}$, ε_{mac} ha ordine di grandezza non superiore a $\sqrt{n}\,\text{eps}$ e ε_{alg} ha ordine di grandezza non superiore a $\sqrt{m}\,\text{eps}$.

- Concludendo, essendo $\sqrt{n}$ e $\sqrt{m}$ numeri più piccoli di n e m, rispettivamente, si può dire che ciò che fa esplodere l'errore ε_y rispetto a TOL o eps non è tanto l'elevato numero n di componenti del dato o l'elevato numero m di passi dell'algoritmo, ma piuttosto il mal condizionamento della funzione o l'instabilità dell'algoritmo, che possono portare ad un errore ε_y molto più grande di TOL e eps anche con poche componenti del dato o con pochi passi dell'algoritmo.

  Negli algoritmi bisogna quindi assolutamente evitare di effettuare una sottrazione tra due numeri molto vicini, perché basta anche una sola di queste per far perdere molta accuratezza al risultato.

- L'assunzione probabilistica sugli errori di misurazione $\widehat{\varepsilon}_i$ è ragionevole. In effetti gli errori di misurazione sono variabili casuali con una distribuzione normale (gaussiana) di media 0 e deviazione standard la precisione dello strumento di misura.

  Nonostante gli errori δ_i e γ_j siano quantità deterministiche, non casuali, in quanto non c'è nulla di casuale nel calcolo di $f(\widehat{x})$ utilizzando un calcolatore, è ragionevole assumere che tali errori si comportino, data la complessità del fenomeno della loro generazione, come variabili casuali con le proprietà sopra citate quando n e m sono grandi.

  Tuttavia, il fatto che abbiamo media 0 può essere assunto solo se si usa l'arrotondamento: nel troncamento l'errore di approssimazione di un numero reale con un numero di macchina è sempre negativo e quindi non può avere media 0.

  Questa è la vera ragione per preferire l'arrotondamento al troncamento, non tanto l'avere $\frac{\mathrm{eps}}{2}$ al posto di eps come maggiorazione dell'errore.

- Esercizio. Per calcolare la soluzione x di un sistema lineare

$$Ax = b,$$

dove $A \in \mathbb{R}^{n\times n}$, $x \in \mathbb{R}^n$ e $b \in \mathbb{R}^{n\times n}$, utilizzando l'algoritmo dell'eliminazione di Gauss (il metodo visto nel corso di Geometria) sono richieste circa $\frac{2}{3}n^3$ operazioni aritmetiche per n grande.

Ogni componente x_i, $i \in \{1, \ldots, n\}$, viene quindi calcolata con un algoritmo che richiede circa $m = \frac{2}{3}n^3$ istruzioni per n grande.

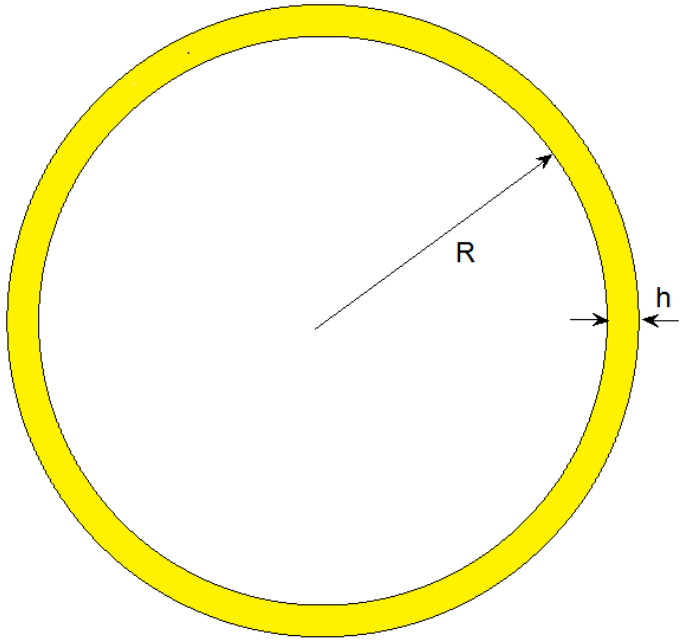
Si assuma che l'algoritmo dell'eliminazione di Gauss sia stabile. Si assuma inoltre la precedente struttura probabilistica degli errori γ_j.

Data una tolleranza tol come deve essere eps , per n grande, se si vuole che l'errore algoritmico nel calcolare ogni componente x_i sia dell'ordine di tol? Si deve ottenere eps come una funzione di tol e n. Se $n \approx 10^6$ e $\mathrm{tol} = 10^{-10}$, va usato lo Standard IEEE in semplice ($\mathrm{eps} \approx 10^{-8}$), doppia ($\mathrm{eps} \approx 10^{-16}$) o quadrupla ($\mathrm{eps} \approx 10^{-32}$) precisione?

- Esercizio. Si assuma di usare il troncamento e che gli errori δ_i come pure gli errori γ_j siano variabili casuali indipendenti di media $-\frac{\mathrm{eps}}{2}$ (nel troncamento l'errore è sempre negativo e varia da 0 a eps) e deviazione standard con ordine di grandezza eps. Determinare la media e la deviazione standard delle variabili casuali $\varepsilon_{\mathrm{mac}}$ e $\varepsilon_{\mathrm{alg}}$.

Un problema particolare

- Consideriamo ora il problema di calcolare l'area A di un anello circolare di raggio interno R e raggio esterno $R + h$, nel caso in cui $h << R$.



Si ha

$$A = f(R, h) = \pi((R + h)^2 - R^2) = \pi(2Rh + h^2).$$

Determiniamo gli indici di condizionamento di $A = \pi(2Rh + h^2)$.

Posto $k = \frac{h}{R}$, si ha

$$\begin{aligned} K_R &= \frac{\frac{\partial A}{\partial R} \cdot R}{A} = \frac{\pi 2h \cdot R}{\pi\left(2Rh + h^2\right)} = \frac{2Rh}{2Rh + h^2} \\ &= \frac{2kR^2}{2kR^2 + k^2R^2} = \frac{2}{2+k} \approx 1 \text{ essendo } k << 1 \end{aligned}$$

e

$$\begin{aligned} K_h &= \frac{\frac{\partial A}{\partial h} \cdot h}{A} = \frac{\pi\left(2R + 2h\right) \cdot h}{\pi\left(2Rh + h^2\right)} = \frac{2R + 2h}{2R + h} \\ &= \frac{2R + 2kR}{2R + kR} = \frac{2 + 2k}{2 + k} \approx 1. \end{aligned}$$

Per cui il problema è ben condizionato e si ha

$$\varepsilon_{\text{in}} \doteq K_R\widehat{\varepsilon}_R + K_h\widehat{\varepsilon}_h \approx \widehat{\varepsilon}_R + \widehat{\varepsilon}_h$$

e

$$\varepsilon_{\text{mac}} \doteq K_R\delta_R + K_h\delta_h \approx \delta_R + \delta_h.$$

- Vi sono ora due algoritmi per calcolare A basati sulle due diverse espressioni

$$A = \pi((R+h)^2 - R^2) = \pi(2Rh + h^2).$$

ALGORITMO 1	ALGORITMO 2
$a = R + h$	$a = R \cdot h$
$b = a \cdot a$	$b = 2 \cdot a$
$c = R \cdot R$	$c = h \cdot h$
$d = b - c$	$d = b + c$
$y = \pi \cdot d$	$y = \pi \cdot d$

L'Algoritmo 1 è quello che la stragrande maggioranza delle persone userebbe.

Per l'Algoritmo 1 si ha

$$a = R + h \quad \beta_a \doteq \gamma_1$$

$$b = a \cdot a \quad \beta_b \doteq 2\beta_a + \gamma_2 = 2\gamma_1 + \gamma_2$$

$$c = R \cdot R \quad \beta_c \doteq \gamma_3$$

$$d = b - c \quad \beta_d \doteq \frac{b}{b-c}\beta_b - \frac{c}{b-c}\beta_c + \gamma_4 = \frac{2b}{b-c}\gamma_1 + \frac{b}{b-c}\gamma_2 - \frac{c}{b-c}\gamma_3 + \gamma_4$$

$$y = \pi \cdot d \quad \varepsilon_{\text{alg}} \doteq \beta_\pi + \beta_d + \gamma_5 = \beta_\pi + \frac{2b}{b-c}\gamma_1 + \frac{b}{b-c}\gamma_2 - \frac{c}{b-c}\gamma_3 + \gamma_4$$

con indici di stabilità

$$\begin{aligned} M_1(R,h) &= \frac{2b}{b-c} = \frac{2(R+h)^2}{(R+h)^2 - R^2} = \frac{2(R+h)^2}{2Rh + h^2} = \frac{2(R+kR)^2}{2kR^2 + k^2R^2} \\ &= \frac{1}{k} \cdot \frac{2(1+k)^2}{2+k} \approx \frac{1}{k} \\ M_2(R,h) &= \frac{1}{2} M_1(R,h) \approx \frac{1}{2} \cdot \frac{1}{k} \\ M_3(R,h) &= -\frac{c}{b-c} = -\frac{R^2}{(R+h)^2 - R^2} = -\frac{R^2}{2Rh + h^2} = -\frac{R^2}{2kR^2 + k^2R^2} \\ &= \frac{1}{k} \cdot \frac{1}{2+k} \approx \frac{1}{2} \cdot \frac{1}{k}. \end{aligned}$$

L'Algoritmo 1 è instabile essendo $\frac{1}{k} >> 1$.

Con l'Algoritmo 1 siamo nella Situazione 2 e l'errore ε_A ha "in generale" ordine di grandezza il massimo tra TOL e $\frac{1}{k} \cdot \mathrm{eps}$ e si può avere un valore per l'area A completamente sbagliato se $\frac{1}{k} \cdot \mathrm{eps} >> \mathrm{TOL}$.

Questo accade ad esempio se si decide di fare il calcolo utilizzando una calcolatrice e si memorizza su un foglio di carta ogni risultato intermedio approssimandolo con due o tre cifre decimali dopo il punto. In questo caso si sta usando una precisione di macchina $\mathrm{eps} = 10^{-2}$ o $\mathrm{eps} = 10^{-3}$.

Per l'Algoritmo 2 si ha

$$a = R \cdot h \;\beta_a \doteq \gamma_1$$
$$b = 2 \cdot a \;\beta_b \doteq \beta_a + \gamma_2 = \gamma_1 + \gamma_2$$
$$c = h \cdot h \;\beta_c \doteq \gamma_3$$
$$d = b + c\beta_d \doteq \frac{b}{b+c}\beta_b + \frac{c}{b+c}\beta_c + \gamma_4 = \frac{b}{b+c}\gamma_1 + \frac{b}{b+c}\gamma_2 + \frac{c}{b+c}\gamma_3 + \gamma_4$$
$$y = \pi \cdot d \;\varepsilon_{\text{a lg}} \doteq \beta_\pi + \beta_d + \gamma_5 = \beta_\pi + \frac{b}{b+c}\gamma_1 + \frac{b}{b+c}\gamma_2 + \frac{c}{b+c}\gamma_3 + \gamma_4 + \gamma_5$$

con indici di stabilità

$$\begin{aligned} M_1(R,h) &= \frac{b}{b+c} = \frac{2Rh}{2Rh+h^2} = \frac{2kR^2}{2kR^2+k^2R^2} = \frac{2}{2+k} \approx 1 \\ M_2(R,h) &= M_1(R,h) \approx 1 \\ M_3(R,h) &= \frac{c}{b+c} = \frac{h^2}{2Rh+h^2} = \frac{k^2R^2}{2kR^2+k^2R^2} = \frac{k}{2+k} << 1. \end{aligned}$$

Si conclude che l'Algoritmo 2 è stabile e quindi dovrebbe essere usato al posto dell'Algoritmo 1 nel problema in esame.

Esercizio. Si consideri $\widehat{R} = 20.5\ \mathrm{m}$ e $\widehat{h} = 0.351\ \mathrm{m}$. Utilizzando una calcolatrice e memorizzando esternamente su un foglio di carta ogni risultato intermedio con due cifre dopo il punto, si confronti il risultato y_1 dell'Algorimo 1 con il risultato y_2 dell'Algoritmo 2. Si utilizzi 3.1416 per π. Si stimi poi l'errore algoritmico come

$$\frac{y_1 - y_2}{y_2}$$

(essendo il risultato dell'Algoritmo 2 più preciso essendo l'Algoritmo stabile) e lo si confronti con $\frac{1}{k} \cdot \mathrm{eps}$ (usare $k \approx \frac{\widehat{h}}{\widehat{R}}$).

Esercizio. Si rifacciamo ora i calcoli dell'esercizio precedente senza memorizzare esternamente i risultati intermedi. Questo vuol dire fare i calcoli scrivendo nella calcolatrice le due espressioni

$$3.1416 \cdot \left((20.5 + 0.351)^2 - 20.5^2 \right)$$

e

$$3.1416 \cdot \left(2 \cdot 20.5 \cdot 0.351 + 0.351^2 \right).$$

Si osservi che le calcolatrici non scientifiche usano la Standard IEEE in semplice precisione con $\mathrm{eps} \approx 10^{-8}$.

- Esercizio. Si analizzi l'analogo problema del calcolo del volume

$$V = \frac{4}{3}\pi \left((R+h)^3 - R^3 \right)$$

di un guscio sferico di raggio interno R e raggio esterno $R+h$, nel caso $h << R$.