# Operazioni di macchina tra numeri macchina

L'insieme dei Numeri di Macchina (o Insieme Floating Point) è definito come l'insieme dei numeri reali esattamente rappresentabili con quel calcolatore.

$$F(\beta, t, L, U) = \{0\} \cup \left\{\alpha \in R \setminus \{0\}: \alpha = sign(\alpha) 0.a_1 a_2 \dots a_t \, \beta^p = sign(\alpha) \left(\sum_{i=1}^{t} a_i \beta^{-i}\right) \beta^p\right\}$$

Base di rappresentazione $\beta$

t cifre per la mantissa

Esponenti minimi e massimi L e U (L < 0, U > 0)

$$fl(\alpha) \in F(\beta, t, L, U) = \begin{cases} 0 & se\ \alpha = 0 \\ \pm\ m_t \beta^p & se\ \alpha \neq 0 \end{cases}$$

$$m_t = a_1\beta^{-1} + a_2\beta^{-2} + \cdots \ldots + \tilde{a}_t\beta^{-t} \quad con\ a_1 \neq 0.$$

o equivalentemente:

$$fl(\alpha) = \pm 0.a_1a_2\ldots\tilde{a}_t\beta^p.$$

dove

- nel caso di troncamento $\tilde{a}_t = a_t$;

nel caso di arrotondamento (usato se β è pari) :

$$\tilde{a}_t = \begin{cases} a_t & \text{se} \quad a_{t+1} < \frac{\beta}{2} \\ \\ a_t + 1 & se\ a_{t+1} \geq \frac{\beta}{2} \end{cases} \quad ;$$

$$\frac{|\alpha - fl(\alpha)|}{|\alpha|} \leq \begin{cases} \beta^{1-t} \text{ ( caso di troncamento)} \\ \frac{1}{2}\beta^{1-t} = \boldsymbol{u} \text{ ( caso di arrotondamento)} \end{cases}$$

$|\varepsilon| \leq u$ e $fl(\alpha)=\alpha(1+\varepsilon)$

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

infatti u = $\frac{1}{2}2^{-52} = 2^{-53} \cong 10^{-16}$

Errore relativo della somma finita tra due numeri di macchina e la somma calcolata in aritmetica reale:
Assumiamo $x, y \in R$ , ma $x, y \notin F$ .
Pertanto, x e y devono innanzitutto essere arrotondati in F:

$$x \to fl(x) = x(1+\epsilon_x) \in F, \;\; y \to fl(y) = y(1+\epsilon_y) \in F,$$

$$fl(\, \mathrm{fl}(x) \oplus fl(y)) = [x(1+\epsilon_x) + y(1+\epsilon_y)](1+\epsilon_s) \qquad |\epsilon_x| \le u \,, |\epsilon_y| \le u \,, |\epsilon_s| \le u \,,$$

$$\left| \frac{x + y - fl(fl(x) + \, fl(y))}{x+y} \right| \approx \left| \frac{x}{x+y} \epsilon_x + \frac{y}{x+y} \epsilon_y + \epsilon_s \right| \le \left| \frac{x}{x+y} \right| u + \left| \frac{y}{x+y} \right| u + u$$

**Se x e y hanno lo stesso segno** allora $\left|\frac{x}{x+y}\right| \le 1$ e $\left|\frac{y}{x+y}\right| \le 1$ e di conseguenza

$$\left| \frac{x + y - fl(fl(x) + \, fl(y))}{x+y} \right| \le 3u$$

**Se x e y non hanno lo stesso segno** e sono vicini in modulo allora $\left|\frac{x}{x+y}\right|$ e $\left|\frac{y}{x+y}\right|$ possono crescere in maniera incontrollata e rappresentare dei fattori di amplificazione degli errori di arrotondamento e quindi portano ad avere un errore relativo elevato.

**NB.** Se x e y sono rappresentabili esattamente in F, e quindi $\epsilon_x = \epsilon_y = 0$ , allora la situazione non è pericolosa numericamente.

Errore relativo del prodotto in aritmetica finita tra due numeri di macchina ed il prodotto calcolato in aritmetica reale:

Assumiamo $x, y \in R$ , ma $x, y \notin F$ .

Pertanto, x e y devono innanzitutto essere arrotondati in F:

$$x \rightarrow fl(x) = x(1+\epsilon_x) \in F,\ \ y \rightarrow fl(y) = y(1+\epsilon_y) \in F,$$

$$fl(\,\mathrm{fl}(\mathrm{x}) \otimes fl(y)) = [x(1+\epsilon_x) \cdot y(1+\epsilon_y)](1+\epsilon_p)$$

$$|\epsilon_x| \leq u\,, |\epsilon_y| \leq u\,, |\epsilon_p| \leq u\,,$$

$$\left|\frac{x \cdot y - fl(fl(x) \otimes fl(y))}{x \cdot y}\right| = \frac{|x \cdot y(1+\epsilon_x)(1+\epsilon_y)(1+\epsilon_p) - xy|}{|xy|} =$$

$$|(1+\epsilon_x)(1+\epsilon_y)(1+\epsilon_p) - 1| =$$

$$|\epsilon_x + \epsilon_y + \epsilon_p + \epsilon_x\epsilon_y + \epsilon_x\epsilon_p + \epsilon_y\epsilon_p + \epsilon_x\epsilon_y\epsilon_p|$$

$$\approx |\epsilon_x + \epsilon_y + \epsilon_p| \leq |\epsilon_x| + |\epsilon_y| + |\epsilon_p|$$

$$\left|\frac{x \cdot y - fl(fl(x) \otimes fl(y))}{x + y}\right| \leq 3u$$

Quindi qualunque siano i numeri $x, y \in R$ , da cui si parte, l'errore relativo sul prodotto è sempre minore o uguale a $3u$.

La somma algebrica macchina (addizione e sottrazione) tra due numeri $x, y \in F(\beta, t, L, U)$ richiede le seguenti fasi:

1. Si scala la mantissa del numero con l'esponente minore in modo tale che i due addendi abbiano lo stesso esponente (ovvero quello dell'esponente maggiore);
2. Si esegue la somma tra le mantisse;
3. Si normalizza il risultato aggiustando l'esponente in modo tale che la mantissa sia un numero minore di 1.
4. Si arrotonda (o si tronca) la mantissa alle prime t cifre;

Consideriamo per esempio i numeri $x, y \in F(10, 5, L, U)$

$$x = 0.78546 \cdot 10^2, y = 0.61332 \cdot 10^{-1}$$

e calcoliamo il numero macchina x $\oplus$ y

1. Scaliamo $y$ in maniera tale che abbia la stessa parte esponente di $x$
$$y = 0.61332 \cdot 10^{-1} \cdot (10^{-2}) \cdot (10^2) = 0.00061332 \cdot 10^2$$
2. Sommiamo le mantisse: $0.78546 \cdot 10^2 + 0.00061332 \cdot 10^2 = 0.78607332 \cdot 10^2$
3. Il risultato in questo caso è già normalizzato.
4. $fl(0.78607332 \cdot 10^2) = 0.78607 \cdot 10^2$

Consideriamo per esempio la differenza tra i due numeri

$$x = 0.75869 \cdot 10^2, y = 0.75868 \cdot 10^2$$

esattamente rappresentabili nell'insieme $F(10, 5, L, U)$

$$fl(x-y) = fl(0.75869 \cdot 10^2 - 0.75868 \cdot 10^2) = \mathrm{fl}(0.00001 \cdot 10^2) = 0.1000 \cdot 10^{-2}$$

In aritmetica reale:
$x - y$= $0.75869 \cdot 10^2$- $0.75868 \cdot 10^2$= $0.1000 \cdot 10^{-2}$

Nonostante ci sia la cancellazione di cifre significative, la situazione non è pericolosa, infatti l l'errore relativo è nullo

$$E_{rel} = \frac{|x-y-fl(x-y)|}{|x-y|} = \frac{|0.1000 \cdot 10^{-2} - 0.1000 \cdot 10^{-2}|}{|0.1000 \cdot 10^{-2}|} = 0$$

Questo è giustificato, come abbiamo visto in teoria, dal fatto che i dati x e y sono rappresentabili esattamente in nell'insieme $F(10, 5, L, U)$

Consideriamo adesso la differenza tra i due numeri

$$x = 0.75868531 \cdot 10^2, y = 0.75868100 \cdot 10^2$$

nell'insieme $F(10, 5, L, U)$, in cui $x$ ed $y$ non sono esattamente rappresentabili.

$$fl(x) = 0.75869 \cdot 10^2 \qquad fl(y) = 0.75868 \cdot 10^2$$

$$fl\big(fl(x) - fl(y)\big) = fl(0.75869 \cdot 10^2 - 0.75868 \cdot 10^2) = \mathrm{fl}(0.00001 \cdot 10^2)$$

Normalizzazione della mantissa

$$0.00001 \cdot 10^2 = 0.00001 \cdot 10^4 \cdot (10^{-4}) \cdot 10^2 = 0.10000 \cdot 10^{-2}$$

$$fl\big(fl(x) - fl(y)\big) = 0.10000 \cdot 10^{-2}$$

In aritmetica reale:

$$x - y = 0.430999.. \cdot 10^{-3}$$

Calcoliamo l'errore relativo che commettiamo quando al posto del calcolo della somma in aritmetica reale sostituiamo il calcolo della somma in aritmetica finita:

$$\frac{|x-y-fl(fl(x)-fl(y))|}{|x-y|} = \frac{\left|0.430999.. \cdot 10^{-3} - 0.10000 \cdot 10^{-2}\right|}{|0.430999.. \cdot 10^{-3}|} \approx 1.32016$$

L'errore relativo commesso è del 132.016%

Il prodotto macchina tra due numeri $x, y \in F(\beta, t, L, U)$
richiede le seguenti fasi:
1. Si esegue il prodotto tra le mantisse;
2. Si esegue l'arrotondamento (o il troncamento) alle prime t cifre;
3. Si sommano gli esponenti, normalizzando, se necessario, la mantissa ad un numero minore di 1.

**Esempio:**
Consideriamo per esempio il prodotto tra i due numeri
$$x = 0.11111 \cdot 10^3 , y = 0.52521 \cdot 10^2$$
nell'insieme $F(10, 5, L, U)$.
1. Il prodotto delle mantisse produce 0.05835608;
2. L'arrotondamento a 5 cifre produce $0.58356 \cdot 10^{-1}$
3. Somma degli esponenti x ∗ y = $0.58356 \cdot 10^{-1} \cdot 10^3 \cdot 10^2 = 0.58356 \cdot 10^{-1} \cdot 10^5 = 0.58356 \cdot 10^4$

La divisione macchina tra due numeri numeri $x, y \in F(\beta, t, L, U)$ richiede le seguenti fasi:

1. Si scala il dividendo x finchè la sua mantissa non risulti minore di quella del divisore y;
2. Si esegue la divisione tra le mantisse;
3. Si esegue l'arrotondamento (o il troncamento) alle prime t cifre;
4. Si sottraggono gli esponenti

Consideriamo la divisione tra i due numeri

$$x = 0.12100 \cdot 10^5, y = 0.11000 \cdot 10^2$$

nell'insieme $F(10, 5, L, U)$

1. Scaliamo il dividendo di una cifra decimale $\mathbf{0.012100 \cdot 10^6}$ (in questo modo la sua mantissa risulta minore di quella del divisore);
2. Dividiamo le mantisse 0.01210/0.11000 = 0.11000;
3. Il troncamento fornisce lo stesso numero 0.11000;
4. Si sottraggono gli esponenti ottenendo il risultato

$$fl\left(\frac{x}{y}\right) = 0.11000 \cdot 10^4$$

In $F(\beta, t, L, U)$

Non vale la proprietà  associativa di somma e prodotto;
Non vale la proprietà distributiva della somma rispetto al prodotto.

**Esempio 1: (La somma tra numeri finiti non gode della proprietà associativa)**

Sia $\beta$=10, e t=8. Dati i numeri reali a=0.23371258 $\cdot$ $10^{-4}$ ,b=0.33678429 $\cdot$ $10^{2}$ ,c= -0.33677811 $\cdot$ $10^{2}$

a

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| + | 0. | 2 | 3 | 3 | 7 | 1 | 2 | 5 | 8 | $10^{-4}$ |

b

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| + | 0. | 3 | 3 | 6 | 7 | 8 | 4 | 2 | 9 | $10^{2}$ |

c

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| - | 0. | 3 | 3 | 6 | 7 | 7 | 8 | 1 | 1 | $10^{2}$ |

I numeri $a, b, c$ appartengono ad F. Calcoliamo $fl(fl(a+b)+c)$ $e$ $fl(a+fl(b+c))$

1) $fl(fl(a+b)+c)$

Si trasforma il numero $a$ con esponente minore rispetto $b$ in modo che i numeri abbiano lo stesso esponente

$$0.23371258 \cdot 10^{-4} = (0.23371258 \cdot 10^{-4})\mathbf{10^{-2}} \cdot \mathbf{10^{2}} = 0.0000023371258 \cdot 10^{2}$$

a

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| + | 0. | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 3 | 7 | 1 | 2 | 5 | 8 | $10^2$ |

b +

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| + | 0. | 3 | 3 | 6 | 7 | 8 | 4 | 2 | 9 | $10^2$ |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| + | 0. | 3 | 3 | 6 | 7 | 6 | 4 | 5 | 2 | 3 | 7 | 1 | 2 | 5 | 8 | $10^2$ |

$\mathrm{fl}(a+b) = \mathrm{fl}(0.33678452371258 \cdot 10^2) = 0.33678452 \cdot 10^2$

fl(a + b)

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| + | 0. | 3 | 3 | 6 | 7 | 8 | 4 | 5 | 2 | $10^2$ |

c

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| - | 0. | 3 | 3 | 6 | 7 | 7 | 8 | 1 | 1 | $10^2$ |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| + | 0. | 0 | 0 | 0 | 0 | 0 | 6 | 4 | 1 | $10^2$ |

**Normalizzazione della mantissa e conseguente aggiustamento dell'esponente:**

$$\underbrace{0.00000641\ (10^{+5})}(10^{-5})(10^{+2}) = \underbrace{0.6410000}\ (10^{-3})$$

$fl(fl(a+b)+c) =$

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| + | 0. | 6 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | $10^{-3}$ |

2) Calcoliamo adesso $fl(a + fl(b+c))$

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| b | + | 0. | 3 | 3 | 6 | 7 | 8 | 4 | 2 | 9 | $10^2$ |
| | | | | | | | | | | | + |
| c | - | 0. | 3 | 3 | 6 | 7 | 7 | 8 | 1 | 1 | $10^2$ |
| | + | 0. | 0 | 0 | 0 | 0 | 0 | 6 | 1 | 8 | $10^2$ |

**Normalizzazione della mantissa e conseguente aggiustamento dell'esponente:**

$$fl(b+c) = 0.00000618\,(10^{+5})(10^{-5})(10^{+2}) = 0.6180000\,(10^{-3})$$

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| a | + | 0. | 0 | 2 | 3 | 3 | 7 | 1 | 2 | 5 | 8 | $10^{-3}$ |
| fl(b + c) | + | 0. | 6 | 1 | 8 | 0 | 0 | 0 | 0 | 0 | $10^{-3}$ |
| | + | 0. | 6 | 4 | 1 | 3 | 7 | 1 | 2 | 5 | 8 | $10^{-3}$ |

$fl(a+fl(b+c))$=

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| + | 0. | 6 | 4 | 1 | 3 | 7 | 1 | 2 | 6 | $10^{-3}$ |

Il risultato in aritmetica reale è a+b+c = $0.641371258 \cdot 10^{-3}$

**Esercizio 2:** Si consideri l'equazione

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

e sia $\beta = 10$ e t=4 cifre.

$a = 1, \qquad b = -6.433, c = 0.009474$

Calcoliamo le radici dell'equazione in aritmetica finita.

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

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| a= | + | 0. | 1 | 0 | 0 | 0 | $10^1$ |
| b= | - | 0. | 6 | 4 | 3 | 3 | $10^1$ |
| c= | + | 0. | 9 | 4 | 7 | 4 | $10^{-2}$ |

Calcoliamo $\mathrm{fl}(\sqrt{fl(b^2) - \mathrm{fl}(4ac)}$

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $b^2 =$ | - | 0. | 6 | 4 | 3 | 3 | $10^1$ | x | - | 0. | 6 | 4 | 3 | 3 | $10^1$ |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| = | + | 0 | 4 | 1 | 3 | 8 | 3 | 4 | 8 | 9 | $10^2$ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $fl(b^2) =$ | + | 0. | 4 | 1 | 3 | 8 | $10^2$ |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $4ac$ | + | 0. | 9 | 4 | 7 | 4 | $10^{-2}$ | x | + | 0. | 4 | 0 | 0 | 0 | $10^1$ | = |

$$(0.9474 \cdot 0.4)10^{-2} \cdot 10^1 = 0.37896\, 10^{-2} \cdot 10^1 = 0.37896\, 10^{-1}$$

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $fl(4ac\quad) =$ | + | 0. | 3 | 7 | 8 | 9 | $10^{-1}$ |

**Approssimazione per troncamento**

| $fl(b^2)$ | + | 0. | 4 | 1 | 3 | 8 | $10^{2}$ |
|---|---|---|---|---|---|---|---|

\-

| $fl(4ac)$ | + | 0. | 3 | 7 | 8 | 9 | $10^{-1}$ |
|---|---|---|---|---|---|---|---|

Si trasforma il numero con esponente minore in modo che i numeri abbiano lo stesso esponente

$$0.3789\,(10^{-1})(10^{-2})(10^{+2}) = 0.0003789\,(10^{2})$$

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $fl(b^2)$ | + | 0. | 4 | 1 | 3 | 8 | | | | | $10^{2}$ |
| $fl(4ac)$ | + | 0. | 0 | 0 | 0 | 3 | 7 | 8 | 9 | | $10^{2}$ |

\-

| $fl(b^2) - fl(4ac)$ | + | 0 | 4 | 1 | 3 | 4 | 2 | 1 | 1 | $10^{2}$ |
|---|---|---|---|---|---|---|---|---|---|---|

| $\eta = fl(fl(b^2) - fl(4ac)) =$ | + | 0. | 4 | 1 | 3 | 4 | $10^{2}$ |
|---|---|---|---|---|---|---|---|

$\sqrt{\eta} =$

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| + | 0. | 6 | 4 | 2 | 9 | 6 | 2 | 2 | 7 | 1 | 4 | 0 | 4 | . | $10^1$ |

$fl(\sqrt{\eta}) =$

| | | | | | |
|---|---|---|---|---|---|
| + | 0. | 6 | 4 | 2 | 9 | $10^1$ |

**Approssimazione per troncamento**

$-b$

| | | | | | | |
|---|---|---|---|---|---|---|
| + | 0. | 6 | 4 | 3 | 3 | $10^1$ |

-

$fl(\sqrt{\eta})$

| | | | | | | |
|---|---|---|---|---|---|---|
| + | 0. | 6 | 4 | 2 | 9 | $10^1$ |

$-\mathrm{b} - fl(\sqrt{\eta}) =$

| | | | | | | |
|---|---|---|---|---|---|---|
| + | 0. | 0 | 0 | 0 | 4 | $10^1$ |

**Normalizzazione della mantissa e conseguente aggiustamento dell'esponente:**

$$\gamma = \mathrm{fl}\left(-\mathrm{b} - fl(\sqrt{\eta})\right) = 0.0004 \cdot (10^3)\,(10^{-3})\mathbf{10^1} = \mathbf{0.4000} \cdot 10^{-2}$$

$fl(x_1) = fl\left(\frac{\gamma}{2}\right)$

| | | | | | | |
|---|---|---|---|---|---|---|
| + | 0. | 2 | 0 | 0 | 0 | $10^{-2}$ |

Calcoliamo adesso la radice $x_2$

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $-b$ | + | 0. | 6 | 4 | 3 | 3 | $10^1$ |
| $fl(\sqrt{\eta})$ | + | 0. | 6 | 4 | 2 | 9 | $10^1$ |

+

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $-\mathrm{b} + fl(\sqrt{\eta}) =$ | + | 1. | 2 | 8 | 6 | 2 | $10^1$ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| + | 0. | 1 | 2 | 8 | 6 | 2 | $10^2$ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $\gamma = \mathrm{fl}(-\mathrm{b} + fl(\sqrt{\eta})) =$ | + | 0. | 1 | 2 | 8 | 6 | $10^2$ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $fl(x_2) \; = \; fl\left(\frac{\gamma}{2}\right)$ | + | 0. | 6 | 4 | 3 | 0 | $10^1$ |

In conclusione, le due soluzioni dell'equazione di partenza calcolare in aritmetica finita con $\beta = 2$ e $t = 4$

$fl(x_1) = fl\left(\frac{\gamma}{2}\right)$

| | | | | | | |
|---|---|---|---|---|---|---|
| + | 0. | 2 | 0 | 0 | 0 | $10^{-2}$ |

$fl(x_2) = fl\left(\frac{\gamma}{2}\right)$

| | | | | | | |
|---|---|---|---|---|---|---|
| + | 0. | 6 | 4 | 3 | 0 | $10^{1}$ |

In aritmetica reale:

$$x_1 = \frac{6.433 - \sqrt{41.383489 - 0.037896}}{2} = \frac{6.433 - 6.4300538}{2} = 0.001473056 \ldots$$
$$= 0.1473056 \ldots 10^{-2}$$

$$x_2 = \frac{6.433 + \sqrt{41.383489 - 0.037896}}{2} = \frac{6.433 + 6.4300538}{2} = 6.431526943899537$$
$$= 0.6431526943899537 \ldots \cdot 10^{1}$$

**Errore relativo sul calcolo delle due soluzioni in aritmetica finita rispetto alle soluzioni calcolare in aritmetica reale:**

$$\frac{|x_1 - fl(x_1)|}{|x_1|} = \frac{|0.1473056\ldots\cdot 10^{-2} - 0.2000\cdot 10^{-2}|}{|0.1473056\ldots\cdot 10^{-2}|} \approx 0.36$$

L'errore relativo sul calcolo della soluzione $x_1$ è del 36%

$$\frac{|x_2 - fl(x_2)|}{|x_1|} = \frac{|0.6431526943899537\ldots\cdot 10^1 - 0.6430\cdot 10^1|}{|0.6431526943899537\ldots\cdot 10^1|} \approx 0.2374\cdot 10^{-3}$$

L'errore relativo sul calcolo della soluzione $x_2$ è dello 0.02374%

Un algoritmo più efficiente per il calcolo delle radici di un'equazione di II° grado è il seguente:

calcolo la radice che, in base al segno di b, non porta problemi, cioè evita la differenza tra numeri molto vicini in modulo, nel nostro caso si suggerisce di calcolare $x_2$ e poi calcolare $x_1$, tenendo conto della relazione che lega le due radici di un'equazione di secondo grado: $\boldsymbol{x_1 x_2 = \frac{c}{a}}$ e quindi $\boldsymbol{x_1 = \frac{c}{a x_2}}$

$$fl(x_1) = \mathrm{fl}\left(\frac{c}{ax_2}\right) = \mathrm{fl}\left(\frac{0.9474 \cdot 10^{-2}}{0.6430 \cdot 10^{1}}\right) = fl(0.1473405909797823.. \cdot\ 10^{-2}\ ) = 0.1473 \cdot 10^{-2}$$

Rivalutiamo adesso l'errore relativo sul calcolo della soluzione $x_1$

$$\frac{|x_1 - fl(x_1)|}{|x_1|} = \frac{|0.1473056 \ldots \cdot 10^{-2} - 0.1473 \cdot 10^{-2}|}{|0.1473056 \ldots \cdot 10^{-2}|} \approx\ 0.3808440341125482 \cdot\ \ 10^{-6}$$

L'errore relativo sul calcolo della  soluzione $x_2$ è dello 0.000038%