

Operazioni con i numeri floating point

Operazioni con i numeri finiti

Dati $x, y \in F(\beta, t, L, U)$, non è detto che il risultato di una operazione tra x e y sia un elemento di F .

Definizione delle operazioni con i numeri finiti

$$x \circ y = fl(x \bullet y) \quad x, y \in F$$

$$\bullet = \begin{cases} + \\ - \\ * \\ / \end{cases}$$

1. eseguire l'operazione tra x e y
2. rappresentare il risultato entro F .

Teorema

$$x, y \in F(\beta, t, L, U)$$



$$\frac{|fl(x \bullet y) - x \bullet y|}{|x \bullet y|} \leq k\beta^{1-t}$$

Che si può scrivere anche come

$$fl(x \bullet y) = (x \bullet y)(1 + \epsilon) \quad |\epsilon| \leq k\beta^{1-t}.$$

- Ogni operazione introduce un errore
- L'errore è maggiorato dalla precisione di macchina

Somma algebrica

- $x, y \in F(\beta, t, L, U) \Leftrightarrow \begin{cases} x = xm \beta^{xe} \\ y = ym \beta^{ye} \end{cases}$
- $z = zm \beta^{ze} = fl(x \pm y)$

- Si scala il numero con l'esponente più basso in modo che gli addendi abbiano lo stesso esponente (quello più alto)
- Si esegue la somma delle mantisse
- Si memorizzano le prime t cifre (per arrotondamento o troncamento) in zm
- Si normalizza il risultato aggiustando l'esponente in modo che la mantissa sia < 1

Esempio con $t=5$, $\beta=10$, arrotondamento

$$x = .64932 \cdot 10^7; y = .53726 \cdot 10^4$$

1. Scalatura di y :
 $y = .00053726 \cdot 10^7$;
2. Somma delle mantisse:
 $.64932 + .00053726 = .64985726$
3. Arrotondamento del risultato alla 5 cifra:
 $zm = .64986$
4. Non necessaria la normalizzazione $ze = 7$

$$\Rightarrow z = .64986 \cdot 10^7$$

$$x = .64937 \cdot 10^7; y = .53726 \cdot 10^7$$

1. Hanno già lo stesso esponente: non c'è bisogno di scalare gli addendi;
2. Somma delle mantisse:
 $.64937 + .53726 = 1.18658$
3. Non è necessario arrotondare
 $zm = 1.18658$
4. Normalizzazione del risultato:
divido la mantissa per 10 e sommo 1 all'esponente
 $zm = .11866$, $ze = 7 + 1$.

$$\Rightarrow z = .11866 \cdot 10^8$$

Cancellazione di cifre

$$x = .75869 \cdot 10^2; y = .75868 \cdot 10^2$$

1. Hanno già lo stesso esponente: non c'è bisogno di scalare gli addendi;
2. Differenza delle mantisse:
 $.75869 - .75868 = .00001$;
3. Non è necessario arrotondare
 $zm = .00001$
4. Normalizzazione del risultato: moltiplico la mantissa per 10^4 e sottraggo 4 all'esponente
 $zm = .1$
5. $ze = 2 - 4$

$$\Rightarrow z = .1 \cdot 10^{-2}$$

Si ha cancellazione quando si sottraggono quantità circa uguali

Cancellazione con dati affetti da errore

$$x = fl(.75868531 \cdot 10^2) = .75869 \cdot 10^2$$

$$E_{ax} = 4.69 \cdot 10^{-4} \quad E_{yx} = .6181 \cdot 10^{-5} \leq \frac{1}{2} \cdot 10^{-4}$$

$$y = fl(.75868100 \cdot 10^2) = .75868 \cdot 10^2$$

$$E_{ay} = 1 \cdot 10^{-4} \quad E_{ry} = .1318 \cdot 10^{-5} \leq \frac{1}{2} \cdot 10^{-4}$$

Il risultato vale $.431 \cdot 10^{-3}$, ma $fl(fl(x) - fl(y)) = .1 \cdot 10^{-2}$

$$E_a = |.10 \cdot 10^{-2} - .431 \cdot 10^{-3}| = .0569 \cdot 10^{-2}$$

$$E_r = \frac{.569 \cdot 10^{-3}}{.431 \cdot 10^{-3}} \approx 1.320186$$

La cancellazione determina un'amplificazione dell'errore sui dati

Errore di incolonnamento: $x+y=x$ anche se $y \neq 0$

$$x = .62379 \cdot 10^7; y = .32881 \cdot 10^1$$

1. Scalatura di y :

$$y = .0000032881 \cdot 10^7;$$

2. Somma delle mantisse

$$.62379 + .0000032881 = .6237932881;$$

3. Arrotondamento della mantissa del risultato

$$zm = .62379$$

4. Non è necessario normalizzare il risultato $ze = 7$

$$z = .62379 \cdot 10^7 \text{ anche se } y \neq 0.$$

Capita ogni volta che $|y| \leq \frac{u}{\beta} |x|$

Questa è la spiegazione del fallimento della formula per le equazioni di secondo grado

- Volevamo calcolare

$$Ms^2 + bs + k = 0$$

$$\Delta = b^2 - 4k \quad \begin{array}{l} b = 10^4 \\ k = 10^{-9} \end{array}$$

$$= 10^8 - 4 \cdot 10^{-9}$$

- La precisione di macchina per i float a 64 bit è circa 10^{-16}

$$4 \cdot 10^{-9} \leq \frac{10^{-16}}{2} \cdot 10^8 = \frac{10^{-8}}{2}$$

- Il fenomeno dell'errore di incolonnamento mostra che non esiste un unico elemento neutro della somma, poiché si ha

$$x + y = x \quad \forall y : |y| \leq \frac{u}{\beta} |x|$$

Prodotto

- $x, y \in F(\beta, t, L, U) \quad \begin{array}{l} x = xm \beta^{xe} \\ y = ym \beta^{ye} \end{array}$
- $z = zm \beta^{ze} = fl(x \cdot y)$
 - Si esegue il prodotto delle mantisse
 - Si esegue arrotondamento o troncamento alle prime t cifre
 - Si sommano gli esponenti, normalizzando il risultato se necessario

Esempio $t=5, \beta=10$, arrotondamento

$$x = .11111 10^7; y = .10202 10^{-2}$$

1. Prodotto delle mantisse $.111111 * .10202 = .0113354422$
2. Arrotondamento $zm = .11335$;
3. Calcolo dell'esponente con normalizzazione $ze = 7 - 2 - 1 = 4$
 $z = .11335 10^4$.

Quoziente

$$\bullet \quad x, y \in F(\beta, t, L, U) \quad \begin{array}{l} x = xm \beta^{xe} \\ y = ym \beta^{ye} \end{array}$$

$$\bullet \quad z = zm \beta^{ze} = fl(x/y)$$

- Si scala x in modo che $xm < ym$
- Si esegue xm/ym
- Si memorizza in zm l'arrotondamento o troncamento alle prime t cifre
- Si calcola l'esponente

Esempio $t=5, \beta=10$, arrotondamento

$$x = .62500 10^0; y = .12500 10^{-2}$$

1. Scalatura di x : $x = .062500 10^1$
2. Divisione delle mantisse $.06250/.12500 = .5$;
3. $ze = 1 + 2 = 3$
 $z = .5 10^3$.

Le operazioni tra numeri finiti si riconducono a:

1. Operazioni tra numeri del tipo $.w_1 w_2 \dots w_\tau \tau \geq t$
2. Moltiplicazioni o divisioni per
3. Somme e sottrazioni di esponenti

- Operazioni fixed point
- Scorrimenti
- Sono riconducibili ad operazioni fixed point:
 - Si ha
 - Quindi si eseguono operazioni fixed point e poi di moltiplica per un opportuno fattore di scala

$$.w_1 w_2 \dots w_\tau = w_1 w_2 \dots w_\tau \beta^\tau$$

ESEMPIO.

$$.312 * .13 = 312 10^{-3} * 13 10^{-2}$$

Si esegue $312 * 13 = 4056$ e poi $4056 10^{-5} = .4056$.

Non validità delle proprietà formali delle operazioni

- F non è chiuso rispetto alle operazioni, ci può essere overflow;
- L'elemento neutro della somma (e del prodotto) non è unico;
- L'opposto di un numero non è unico.

NON VALGONO

- Associativa di somma e prodotto
- Distributiva
- Legge di annullamento del prodotto

NON vale l'associativa della somma

$$fl((x + y) + z) \neq fl(fl(x + y) + z)$$

$\beta = 10, t = 7$, arrotondamento.

$$x = .1234567 \cdot 10^0; y = .6666325 \cdot 10^4; z = -.6666325 \cdot 10^4$$

1. $fl(fl(x + y) + z) = .123 \cdot 10^0$.

$$\begin{aligned} fl(x + y) &= fl(.6666325 + .00001234567) \cdot 10^4 = .6666448 \cdot 10^4 \\ fl(fl(x + y) + z) &= fl(.6666448 - .6666325) \cdot 10^4 = .123 \cdot 10^0 \end{aligned}$$

SI HA CANCELLAZIONE SU DATI PERTURBATI.

2. $fl(x + fl(y + z)) = .1234567 \cdot 10^0$

$$\begin{aligned} fl(y + z) &= 0 \\ fl(x + fl(y + z)) &= .1234567 \cdot 10^0 \end{aligned}$$

IN QUESTO CASO LA CANCELLAZIONE NON DA' PROBLEMI.

Osservazione importante

- L'errore commesso nel calcolo di un'espressione dipende dall'algoritmo usato per calcolarla

ALGORITMO 1	ALGORITMO 2
$s \leftarrow x + y$	$s \leftarrow y + z$
$s \leftarrow s + z$	$s \leftarrow s + x$

NON vale la distributiva

$$fl(x \cdot fl(y + z)) \neq fl(fl(xy) + fl(xz))$$

$\beta = 10, t = 2$, troncamento. $x = .91 \cdot 10^1; y = .92 \cdot 10^1; z = .10 \cdot 10^0$.

1. $fl(y + z) = fl(.92 + .010) \cdot 10^1 = .93 \cdot 10^1$
 $fl(x \cdot fl(y + z)) = fl(.91 \cdot 10^1 \cdot .93 \cdot 10^1) = .84 \cdot 10^2$

2. $fl(xy) = fl(0.8372 \cdot 10^2) = .83 \cdot 10^2$
 $fl(xz) = .91 \cdot 10^0$
 $fl(fl(xy) + fl(xz)) = fl(.83 \cdot 10^2 + .91 \cdot 10^0)$
 $= fl(.83 + .0091) \cdot 10^2$
 $= .83 \cdot 10^2$

NON vale la legge di annullamento del prodotto

$$\beta = 10, t = 7, L = -50, U = 49 \quad x = .2 \cdot 10^{-27}; y = .1 \cdot 10^{-26}.$$

$$fl(xy) = 0 \text{ anche se } x \neq 0, y \neq 0$$

$$fl(xy) = fl(.2 \cdot 10^{-52}) = 0 \text{ UNDERFLOW}$$

Conseguenze

$$\beta = 10, t = 7, L = -50, U = 49 \quad x = .2 \cdot 10^{-27}; y = .1 \cdot 10^{-26}; z = .2 \cdot 10^{-9}$$

$$fl\left(\frac{z}{xy}\right) \neq fl\left(\frac{z}{x} \cdot \frac{1}{y}\right)$$

1. $fl(xy) = fl(.2 \cdot 10^{-52}) = 0 \text{ UNDERFLOW} \implies fl(z/fl(xy))$ non calcolabile.
2. $fl(z/x) = 1.0 \cdot 10^{18} = .1 \cdot 10^{19}$
 $fl(1/y) = .1 \cdot 10^{28}$
 $fl(fl(z/x) * fl(1/y)) = .1 \cdot 10^{46}$

Propagazione degli errori

- Poiché gli errori di arrotondamento capitano potenzialmente ad ogni operazione, ogni risultato intermedio può esserne soggetto e influenzare i risultati di tutte le operazioni successive.
- L'accumulo di questi errori viene chiamato propagazione degli errori.

Esempio di propagazione degli errori

$$\beta = 10, t = 5, \text{ arrotondamento}$$

Si vuole calcolare $(x - y)/z$ dove

$$\begin{array}{rcl} x & = & .554617 \quad y = .554601 \quad z = .1 \cdot 10^{-n} \\ & \downarrow & \downarrow \\ fl(x) & = & .55462 \quad fl(y) = .55460 \end{array}$$

Il risultato esatto dell'espressione è $.16 \cdot 10^{-4+n}$.

$$fl(x - y) = .00002 = .2 \cdot 10^{-4}$$

$$E_a = |.16 \cdot 10^{-4} - .2 \cdot 10^{-4}| = \boxed{.04 \cdot 10^{-4}}$$

$$fl(fl(x - y)/z) = fl(.2 \cdot 10^{-4} / .1 \cdot 10^{-n}) = fl(.02 \cdot 10^{n-4}) = .2 \cdot 10^{n-3}$$

$$E_a = |.02 \cdot 10^{n-4} - .16 \cdot 10^{n-4}| = \boxed{.14 \cdot 10^{n-4}}$$

Amplificazione dell'errore di 10^n volte