OPERAZIONI SUI 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. Può essere un numero maggiore del massimo numero rappresentabile in modulo o avere una mantissa con più di t cifre.

Si devono ridefinire le operazioni di macchina nel seguente modo:

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

ove $\bullet$ è $+, -, *, /$; si tratta di eseguire l'operazione tra x e y e poi rappresentare il risultato entro F. Pertanto, dal Teorema dell'errore di rappresentazione, segue il seguente teorema fondamentale:

TEOREMA. Siano $x, y \in F(\beta, t, L, U)$. Allora

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

ove $k = 1$ o $1/2$ a seconda che la rappresentazione sia per troncamento o per arrotondamento oppure

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

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

$$\begin{aligned} x &= xm\ \beta^{xe} \\ y &= ym\ \beta^{ye} \end{aligned}$$

SOMMA ALGEBRICA. $z = zm\ \beta^{ze} = fl(x \pm y)$.

1. Si confrontano xe e ye; se $xe > ye$, si divide ym per β^{xe-ye};
2. si esegue $xm \pm ym/\beta^{xe-ye}$ e si considerano le prime t cifre più significative (con troncamento o arrotondamento), ponendole in zm. Se il risultato è maggiore o uguale a 1, $h = 1$ altrimenti si pone in h l'opposto del numero degli zeri ottenuti dopo il punto radice;
3. $ze = xe + h$.

ESEMPI. $\beta = 10, t = 5$, arrotondamento.

- $x = .64937\ 10^7; y = .53726\ 10^4$
  1. $xe - ye = 3$;
  2. $.64932 + .00053726 = .64985726$
     $zm = .64986$
  3. $ze = 7$

  $z = .64986\ 10^7$.
- $x = .64937\ 10^7; y = .53726\ 10^7$
  1. $xe - ye = 0$;
  2. $.64932 + .53726 = 1.18658 \quad h = 1$
     $zm = .11866$
  3. $ze = 7 + 1$

  $z = .11866\ 10^8$

- $x = .75869\ 10^2; y = .75868\ 10^2$
  1. $xe - ye = 0;$
  2. $.75869 - .75868 = .00001 \quad h = -4$
     $zm = .1$
  3. $ze = 2 - 4$

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

In questo caso si ha una cancellazione di cifre; vengono introdotte quantità spurie, poichè si esegue una differenza tra due quantità circa uguali, perdendo cifre (effetto smearing). La situazione non è pericolosa se $E_a = \epsilon_r = 0$ eccetto se i dati di partenza sono affetti da errore.
Se invece,

$$\begin{aligned} x &= fl(.75868531\ 10^2) = .75869\ 10^2 \\ & E_{ax} = 4.69\ 10^{-4} \quad \epsilon_{rx} = .6181\ 10^{-5} \leq \frac{1}{2}\ 10^{-4} \\ y &= fl(.75868100\ 10^2) = .75868\ 10^2 \\ & E_{ay} = 1.\ 10^{-4} \quad \epsilon_{ry} = .1318\ 10^{-5} \leq \frac{1}{2}\ 10^{-4} \end{aligned}$$

Il risultato esatto vale $.431\ 10^{-3}$, ma

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

$$\epsilon_r = \frac{.569\ 10^{-3}}{.431\ 10^{-3}} \simeq 1.320186$$

La cancellazione determina una amplificazione dell'errore sui dati.

- $x = .62379\ 10^7; y = .32881\ 10^1$
  1. $xe - ye = 6$;
  2. $.62379 + .00000032881 = .62379032881$
     $zm = .62379$
  3. $ze = 7$

  $z = .62379\ 10^7$ anche se $y \neq 0$.

Quando $x + y = x$ con $y \neq 0$, si verifica un errore di incolonnamento. Questo capita ogni volta che $|y| \leq \frac{u}{\beta}|x|$. Non esiste un solo elemento neutro per la somma.

PRODOTTO.

1. Si esegue il prodotto $xm \cdot ym$, troncando o arrotondando il risultato a t cifre; si memorizza in zm e si pone $h = 1$ se si ottiene uno zero a destra del punto radice, altrimenti $h = 0$;
2. $ze = xe + ye - h$.

ESEMPI. $\beta = 10, t = 5$, arrotondamento.

- $x = .11111\ 10^7; y = .10202\ 10^{-2}$
  1. $.111111 * .10202 = .0113354422 \quad zm = .11335 \quad h = 1$;
  2. $ze = 7 - 2 - 1 = 4$
     $z = .11335\ 10^4$.

QUOZIENTE.

1. Se $xm < ym$ si pone $h = 0$; altrimenti si divide xm per β e si pone $h = 1$;
2. Si esegue $(xm/\beta^h)/ym$ e si pongono le t cifre più significative in zm;
3. $ze = xe - ye + h$.

ESEMPI. $\beta = 10, t = 5$, arrotondamento.

- $x = .62500\ 10^0; y = .12500\ 10^{-2}$
  1. $.062500 \quad h = 1$
  2. $.06250/.12500 = .5;$
  3. $ze = 0 + 2 + 1 = 3$
     $z = .5\ 10^3.$
- sia $x = .554617;\ y = .554601;$ allora $fl(x) = .55462,\ fl(y) = .55460;$ se si esegue $fl(x - y)$, si ottiene $.00002 = .210^{-4}$, invece di $.000016$. Si commette un errore assoluto pari a $.04\ 10^{-4}$. Se si divide il risultato per un numero piccolo (o lo si moltiplica per uno grande), l'errore risulta amplificato ulteriormente. Supponiamo di dividere per $z = .1\ 10^{-n}$. Segue che il risultato $.2\ 10^{-4+n}$, invece di $.16\ 10^{-4+n}$; l'errore assoluto vale $.04\ 10^{-4+n}$, ossia è pari all'errore assoluto precedente amplificato di 10^n.
  Il risultato ottenuto vale $((x - y) \pm E_a)10^n$.

Le operazioni tra numeri finiti si riconducono a:

1. operazioni tra numeri del tipo $(.w_1 w_2 ... w_\tau)$ con $\tau \geq t$, ottenuti troncando o arrotondando a t cifre una mantissa o ottenuti da tale mantissa dividendo per β^k con $k \geq 0$;
2. moltiplicare o dividere per β^k (k intero);
3. sommare e sottrarre esponenti a opportune costanti.

Le operazioni di tipo 3 sono operazioni tra numeri fixed point.

Le operazioni di tipo 2 comportano scorrimenti verso sinistra o destra di k posizioni.

Le operazioni di tipo 1 sono riconducibili a operazioni tra numeri fixed point. Infatti $.w_1 ... w_\tau = w_1 ... w_\tau \ \beta^{-\tau}$. Pertanto si eseguono operazioni tra numeri fixed point e poi si moltiplica per opportuni fattori di scala, con operazioni di tipo 2.

ESEMPIO.
$.312 * .13 = 312 \ 10^{-3} * 13 \ 10^{-2}$.
Si esegue $312 * 13 = 4056$ e poi $4056 \ 10^{-5} = .04056$.

La ridefinizione delle operazioni di macchina comporta la non validità delle proprietà formali.
Dati $x, y \in F$, non è detto che $x \circ y \in F$; infatti può essere che si verifichi OVERFLOW.
F non è chiuso rispetto alle operazioni.

1. Vale la proprietà commutativa per $+$ e per $*$;
2. $\exists 0$ tale che $fl(\alpha + 0) = fl(\alpha)$;
3. $\exists 1$ tale che $fl(\alpha \cdot 1) = fl(\alpha)$;
4. $\forall \alpha, \exists -\alpha$ tale che $fl(\alpha - \alpha) = 0$.

Ma gli elementi neutri rispetto a somma e prodotto e l'opposto di un numero rispetto alla somma non sono unici.

NON VALGONO:

1. associativa per il prodotto e la somma;
2. distributiva;
3. legge di annullamento del prodotto.

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

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

.

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

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

SI HA CANCELLAZIONE SU DATI PERTURBATI.

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

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

IN QUESTO CASO LA CANCELLAZIONE NON DA PROBLEMI.

Non sempre $x + y - y = x$.

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

In 1. c'è cancellazione tra due operandi di cui uno è affetto da errore; si ha perdita di cifre significative (effetto smearing) pericolosa, poichè l'errore viene amplificato.
In 2. l'effetto smearing non crea problemi poichè si esegue su dati ritenuti esatti.

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

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

Infatti:

1. $fl(x + y) = fl((.92 + .010)\ 10^1) = .93\ 10^1$
   $fl(x\ fl(y + z)) = fl(.91\ 10^1 * .93\ 10^1) = .84\ 10^2$
2. $fl(xy) = .83\ 10^2$
   $fl(xz) = .91\ 10^0$
   $fl(fl(xy) + fl(xz)) = fl(.83\ 10^3 + .91\ 10^0) = fl((.83 + .0091)\ 10^2) = .83\ 10^2$

NON VALIDITA' DELLA PROPRIETA' DISTRIBUTIVA.

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

$$fl(z/(xy)) \neq fl((z/x)(1/y))$$

1. $fl(xy) = fl(.2\ 10^{-52}) = 0$ UNDERFLOW
   NON VALIDITA' DELLA LEGGE DI ANNULLAMENTO DEL PRODOTTO
   $fl(z/fl(xy))$ non calcolabile.
2. $fl(z/x) = .1\ 10^{19}$
   $fl(1/y) = .1\ 10^{28}$
   $fl(fl(z/x) * fl(1/y)) = .1\ 10^{46}$

Poichè gli ERRORI DI ARROTONDAMENTO nelle operazioni capitano potenzialmente in ogni operazione, ogni risultato intermedio ne può essere influenzato. L'accumulo degli errori è detto PROPAGAZIONE DEGLI ERRORI DI ARROTONDAMENTO.

CAUSE DI ERRORE

Supponiamo di dover calcolare

$$y = \varphi(x)$$

Se φ non è una funzione razionale, occorre trovare una approssimazione razionale di φ perchè possa essere valutata su un calcolatore.
Se, per esempio, $\varphi(x) = e^x$, si può approssimare e^x con lo sviluppo in serie di Taylor troncato. Si commette in questo modo un errore, detto ERRORE DI TRONCAMENTO, che deriva dall'aver approssimato un procedimento infinito (calcolo di una serie) con un procedimento finito. Tale errore dipende dal metodo che si usa per approssimare e sarà valutato di volta in volta per ogni metodo introdotto.

Restringiamoci al caso in cui $\varphi(x)$ è una funzione razionale. In questo caso ci sono due cause di errore.

ERRORE INERENTE O SUI DATI INIZIALI. Si assume che i dati iniziali x siano perturbati, in modo da calcolare $\tilde{y} = \varphi(\tilde{x})$ invece di $y = \varphi(x)$.

$$E_{dati} = |\varphi(\tilde{x}) - \varphi(x)|$$

Se $\epsilon_x = |\tilde{x} - x|/|x|$ è l'errore relativo sui dati iniziali, si tratta si valutare l'errore relativo sui risultati (CONDIZIONE DI UN PROBLEMA):

$$\epsilon_{dati} = |\varphi(\tilde{x}) - \varphi(x)|/|\varphi(x)| = |\tilde{y} - y|/|y| = \epsilon_y$$

Se ϵ_{dati} è molto grande rispetto a ϵ_x si ha MAL CONDIZIONAMENTO: piccole perturbazioni sui dati iniziali provocano grosse perturbazione sui risultati finali (PROBLEMA MAL CONDIZIONATO) .

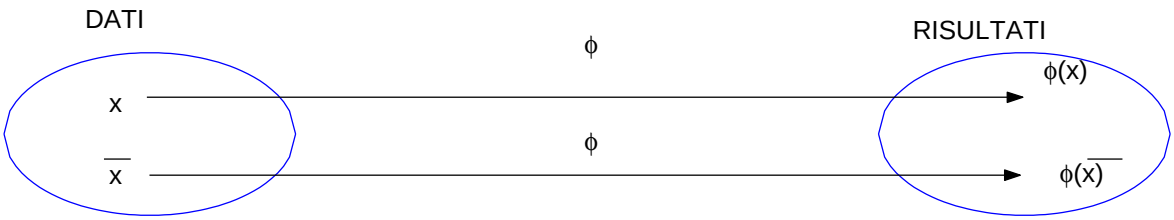
SI ASSUME CHE LE OPERAZIONI SIANO ESATTE (in aritmetica reale).

Il mal condizionamento dipende dal problema, ossia da φ e non dal modo in cui φ è calcolato.

Occorre non solo che φ sia continua, ma anche che sia lipschitziana, con costante di Lipschitz bassa:

$$||\varphi(x) - \varphi(\tilde{x})|| \leq L||x - \tilde{x}||$$

$\forall x, \tilde{x}$ nel dominio di definizione di φ; L è la costante di Lipschitz.



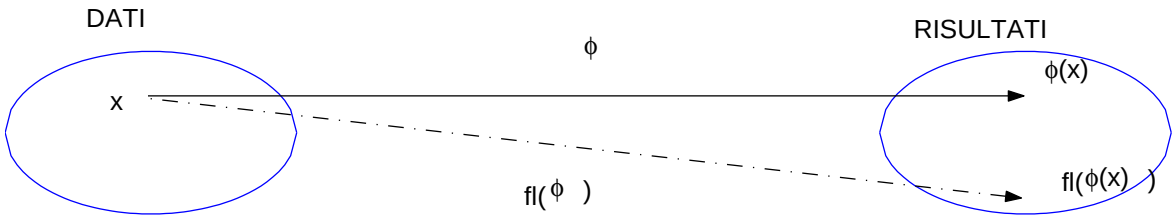
Condizione di un problema.

Si studia $|\varphi(\bar{x}) - \varphi(x)|$ in relazione a $|\bar{x} - x|$.

ERRORE NELLE OPERAZIONI DI MACCHINA O ERRORE DI ARROTONDAMENTO. Si calcola $fl(\varphi(x))$ invece di $\varphi(x)$. Si assume che x sia un elemento dell'insieme dei numeri finiti, ossia sia esatto. Si valuta:

$$E_{alg} = |fl(\varphi(x)) - \varphi(x)|.$$

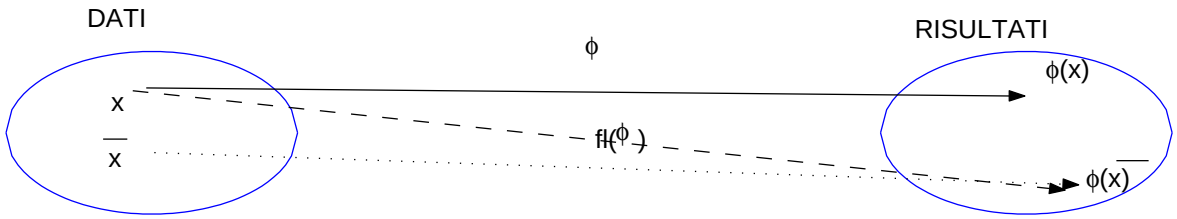
Analisi di stabilità: un algoritmo è stabile se non è troppo sensibile agli effetti degli errori di arrotondamento. La stabilità dipende dall'ordine con cui sono eseguite le operazioni di macchina.



Stabilità di un algoritmo.

Si studia $|fl(\varphi(x)) - \varphi(x)|$ in relazione alla precisione di macchina. Una tecnica per studiare tale errore è quella dell'ANALISI IN AVANTI.

ANALISI ALL'INDIETRO. Si considera il risultato approssimato come risultato esatto di un problema perturbato. Se la perturbazione calcolata è grande, l'algoritmo è instabile.



ANALISI ALL'INDIETRO

$\bar{x} - x$

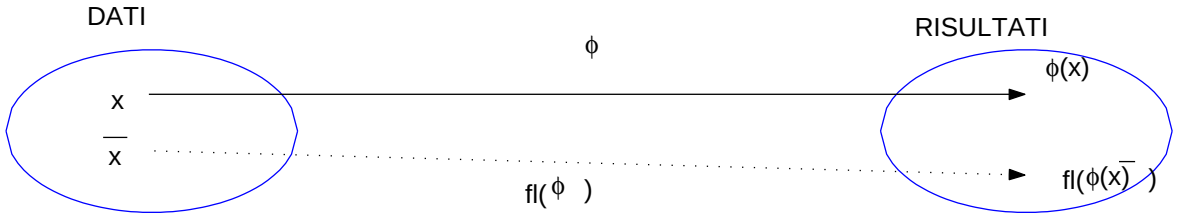
ANALISI IN AVANTI

$$\varphi(\bar{x}) - \varphi(x) =$$
$$= fl(\varphi(x)) - \varphi(x)$$

Infine occorre studiare l'effetto combinato dell'errore sui dati iniziali e dell'aritmetica inesatta.Si studia

$$E_{tot} = |fl(\varphi(\tilde{x})) - \varphi(x)|$$

in relazione a $|x - \tilde{x}|$ e all'aritmetica finita.



In una analisi del I ordine, si prova che l'errore totale è pari alla somma dei contributi dei due tipi di errori (errore inerente+errore di arrotondamento).

$E_{dati} = \varphi(\tilde{x}) - \varphi(x)$ errore assoluto sui dati iniziali

$E_{alg} = fl(\varphi(x)) - \varphi(x)$ errore assoluto algoritmico (dovuto all'uso dell'aritmetica finita)

$E_{tot} = fl(\varphi(\tilde{x})) - \varphi(x) =$ errore assoluto totale

$= fl(\varphi(\tilde{x})) - \varphi(\tilde{x}) + \varphi(\tilde{x}) - \varphi(x) =$
$= E_{alg} + E_{dati}$

$\epsilon_{dati} = \frac{\varphi(\tilde{x}) - \varphi(x)}{\varphi(x)}$ errore relativo sui dati iniziali

$\epsilon_{alg} = \frac{fl(\varphi(x)) - \varphi(x)}{\varphi(x)}$ errore relativo algoritmico (dovuto all'uso dell'aritmetica finita)

$\epsilon_{tot} = \frac{fl(\varphi(\tilde{x})) - \varphi(x)}{\varphi(x)}$ errore relativo totale

$$\begin{aligned}
\epsilon_{tot} &= \frac{fl(\varphi(\tilde{x})) - \varphi(\tilde{x}) + \varphi(\tilde{x}) - \varphi(x)}{\varphi(x)} \\
&= \frac{\varphi(\tilde{x}) - \varphi(x)}{\varphi(x)} + \frac{fl(\varphi(\tilde{x})) - \varphi(\tilde{x})}{\varphi(\tilde{x})}\left(\frac{\varphi(\tilde{x}) - \varphi(x) + \varphi(x)}{\varphi(x)}\right) \\
&= \epsilon_{dati} + \epsilon_{alg}\big(1 + \epsilon_{dati}\big) = \\
&\simeq \epsilon_{dati} + \epsilon_{alg}
\end{aligned}$$

In una analisi del I ordine, $\epsilon_{dati}\epsilon_{alg}$ è trascurato.

Definizione. Supponiamo che ϵ_0 sia l'errore iniziale e ϵ_n l'errore (totale) dopo n passi di un algoritmo. Se

$$\epsilon_n \approx n\epsilon_0$$

la crescita dell'errore è detta lineare. Se

$$\epsilon_n \approx K^n\epsilon_0$$

il comportamento dell'errore è detto esponenziale. In particolare se $K > 1$ l'errore cresce esponenzialmente ($\epsilon_n \to \infty$ per $n \to \infty$), mentre se $0 < K < 1$ l'errore decresce esponenzialmente ($\epsilon_n \to 0$ per $n \to \infty$).