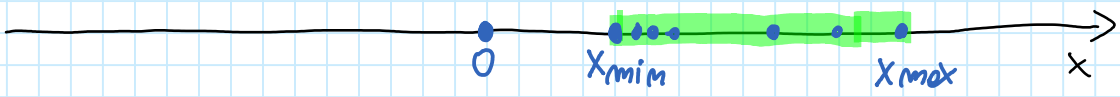
lunedì 10 marzo 2014 09.34

$$\mathbb{F}(\beta, t, L, U) = \{0\} \cup \left\{ x \in \mathbb{R} \mid x = \pm \beta^p \sum_{k=1}^{t} d_k \beta^{-k} \right\}$$

con $L \le p \le U$

$d_k \in \{0, \dots, \beta - 1\}$, $d_1 \ge 1$.

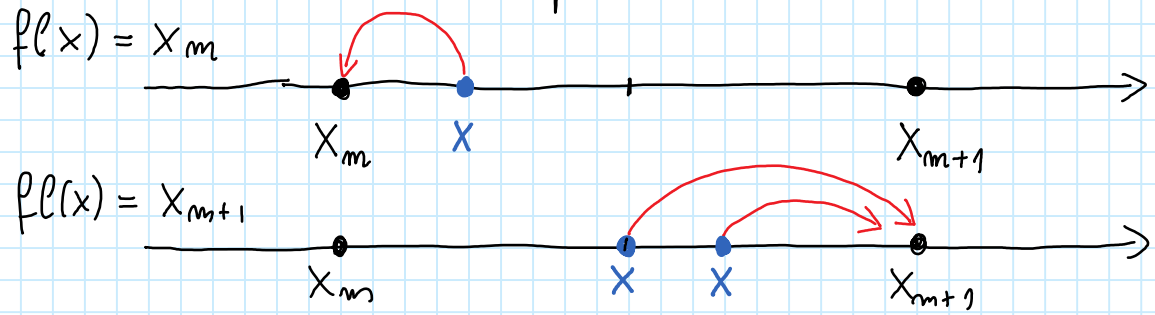


VEDIAMO COME RAPPRESENTARE UN NUMERO $x \in \mathbb{R}$ NELL'INSIEME $\mathbb{F}$

Per semplicità, consideriamo $x \ge 0$.
Si hanno questi casi

1) $x > x_{max}$: viene generato un ERRORE di overflow. Il calcolo si arresta.
2) $0 < x < x_{min}$: viene generata una segnalazione di UNDERFLOW. Il calcolo, di solito, prosegue.
3) $x_{min} \le x \le x_{max}$: in questo caso x ha una rappresentazione, indicata con $fl(x)$, all'interno di $\mathbb{F}$. Ho due possibilità:
   3.1) $x \in \mathbb{R}$ è anche un numero di macchina, cioè di $\mathbb{F}(\beta, t, L, U)$. In questo caso, è rappresentato esattamente (cioè, senza errore) in $\mathbb{F}$:
   $$fl(x) = x$$
   3.2) $x \notin \mathbb{F}(\beta, t, L, U)$. Il rappresentante di x nell'insieme $\mathbb{F}$ viene ottenuto per arrotondamento verso il numero di $\mathbb{F}$

più vicino a x:



dove x_m e x_{m+1} sono due numeri di macchina in $IF(\beta, t, L, U)$ consecutivi.
Nel processo appena descritto viene commesso un errore $|fl(x) - x|$ detto errore di arrotondamento

PROP. Sia $x \in \mathbb{R}$ che non dà né underflow né overflow. Allora, vale la seguente maggiorazione

$$\left| \frac{fl(x) - x}{x} \right| \leq \frac{1}{2} \beta^{1-t}$$

usando $IF(\beta, t, L, U)$.

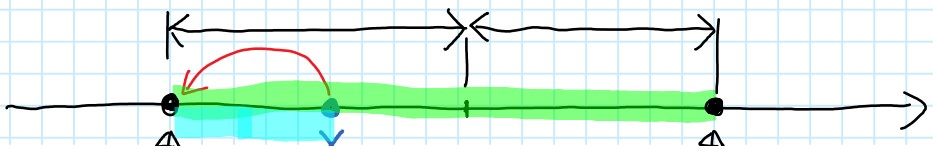
Dim. Sia $x \in \mathbb{R}$ della forma (per semplicità, $x > 0$)

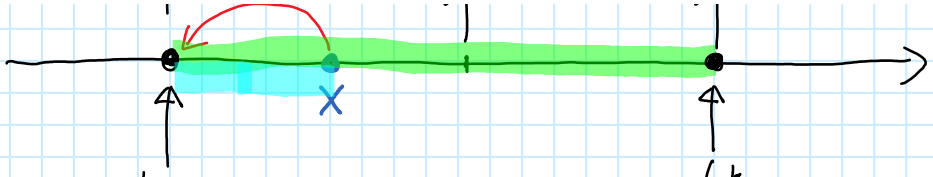
$$x = \beta^p \sum_{k=1}^{+\infty} a_k \beta^{-k}$$

Dato che non ci sono under o over flow, x è nel range dei numeri macchina. Ho due casi.
1) $x \in IF(\beta, t, L, U)$. Allora è $fl(x) = x$ e

$$\left| \frac{fl(x) - x}{x} \right| = \left| \frac{0}{x} \right| = |0| = 0 < \frac{1}{2} \beta^{1-t}$$

2) $x \notin IF(\beta, t, L, U)$. Allora x cade tra due numeri di macchina consecutivi





$\beta^p \sum_{k=1}^{t} d_k \beta^{-k}$ $\qquad$ $\beta^p \left(\sum_{k=1}^{t} d_k \beta^{-k} + \beta^{-t} \right)$

$10^2 (0.345)$ $\qquad$ $10^2 \cdot (0.346) =$

$= 10^2 (0.345 + 0.001)$

Quindi, la distanza tra i due numeri macchina è $\beta^p \cdot \beta^{-t} = \beta^{p-t}$. Ne segue che

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

Perciò ho

$$\left| \frac{fl(x) - x}{x} \right| = \frac{|fl(x) - x|}{|x|} \leq \frac{\frac{1}{2}\beta^{p-t}}{x} \leq$$

$$\leq \frac{\frac{1}{2}\beta^{p-t}}{\beta^{p-1}} = \frac{1}{2}\beta^{1-t}$$

$x = \beta^p \cdot (d_1 \beta^{-1} + d_2 \beta^{-2} + \dots) \geq \beta^p \cdot 1 \cdot \beta^{-1} = \beta^{p-1}$

(è, o potrebbe essere, molto pessimistica).

DEF. Sia $\mathbb{F}(\beta, t, L, U)$. Il numero positivo

$$eps = \frac{1}{2}\beta^{1-t}$$

è detto PRECISIONE DI MACCHINA del sistema floating point $\mathbb{F}$.

La precisione di macchina è il più piccolo numero positivo tale che

$$1 + eps > 1$$

nel sistema floating-point $\mathbb{F}(\beta, t, L, U)$.

Infatti, $1 \in \mathbb{F}(\beta, t, L, U)$ perché

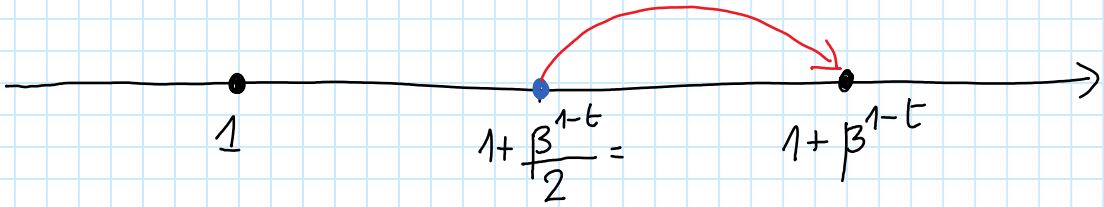
$$1 = \beta^{\textcircled{1}} \left(\textcircled{1} \beta^{-1} + \textcircled{0} \beta^{-2} + \dots + \textcircled{0} \beta^{-t} \right)$$

$$1 = \beta^{\overbrace{1}^{p}} \cdot \left(\overbrace{1}^{d_1} \cdot \beta^{-1} + \overbrace{0}^{d_2} \cdot \beta^{-2} + \cdots + \overbrace{0}^{d_t} \cdot \beta^{-t} \right)$$

Il numero floating point successivo al numero 1 è

$$\beta^{1} \cdot \left(1 \cdot \beta^{-1} + 0 \cdot \beta^{-2} + \cdots + 0 \cdot \beta^{-(t-1)} + 1 \cdot \beta^{-t} \right) =$$

$$= \beta^{1} \left(\beta^{-1} + \beta^{-t} \right) = 1 + \beta^{1-t}$$



$$= 1 + eps$$

Pertanto, se

a) $0 < \varepsilon < eps \Rightarrow 1 + \varepsilon = 1$ in $\mathbb{F}$

b) $eps \leq \varepsilon \Rightarrow 1 + \varepsilon = 1 + \beta^{1-t}$ in $\mathbb{F}$.

ESERCIZIO Cosa succede se consideriamo il numero

$$x = \beta^{p} \sum_{k=1}^{t} d_k \beta^{-k} \quad ?$$

Ossia, quale è il più piccolo numero positivo che devo sommare ad x per ottenere il successivo numero macchine?

OPERAZIONI IN $\mathbb{F}(\beta, t, L, U)$

In modo APPROSSIMATO (la realtà è più complicata!) definiamo le operazioni elementari in $\mathbb{F}(\beta, t, L, U)$ in questo modo:

SOMMA IN $\mathbb{F}$ $\quad x \oplus y = fl(x + y)$ → operazione in $\mathbb{R}$

SOTTRAZIONE IN $\mathbb{F}$ $\quad x \ominus y = fl(x - y)$

SOMMA IN IF $x \oplus y = fl(x+y)$

SOTTRAZIONE IN IF $x \ominus y = fl(x-y)$

MOLTIPLICAZIONE IN IF $x \odot y = fl(x \cdot y)$

DIVISIONE IN IF $x \oslash y = fl(x/y)$

dove $x, y \in IF(\beta, t, L, U)$.

Inoltre, ogni operazione in IF ha due passi:
→ eseguo la corrispondente operazione in $\mathbb{R}$;
→ rappresento il risultato in IF.

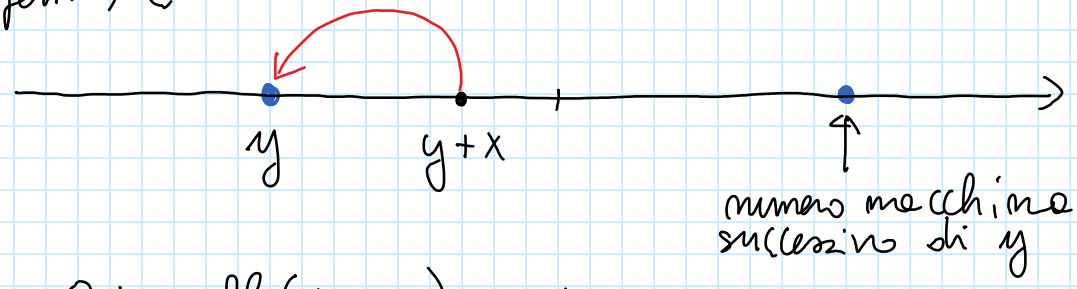
In modo analogo si definiscono le altre operazioni che si possono ritenere elementari, quali, ad esempio, la radice quadrata.

OSSERVAZIONE Le normali regole dell'algebra possono NON essere più valide quando operiamo con la aritmetica di macchina.

Ad esempio, se $x > 0$, $x \in IF(\beta, t, L, U)$ ed $y \in IF(\beta, t, L, U)$ NON è più detto che valga la relazione (valida in $\mathbb{R}$)

$$y \oplus x > y$$

Infatti, è



$$y \oplus x = fl(y+x) = y.$$

ESEMPIO Consideriamo $IF(10, 1, -1, 2)$

Sia $x = 0.1$, $y = 0.08$, numeri di IF.

Allora è

$$x \oplus y = fl(0.1 + 0.08) = fl(0.18) = 0.2$$

IF

$x \oplus y = fl(0.1 + 0.08) = fl(0.18) = 0.2$

perché $0.1 < 0.18 < 0.2$

e 0.2 è più vicino a 0.18.

$x \ominus y = fl(0.1 - 0.08) = fl(0.02) = 0.02$

$x \odot y = fl(0.1 \cdot 0.08) = fl(0.008) = 0$ ← UNDERFLOW

$x \oslash y = fl(0.1/0.08) = fl(1.25) = 1.$

Nelle varie operazioni, tranne le $\pm^{\infty}$, ho degli errori!

ESEMPIO Consideriamo $\mathbb{F}(10, 3, -2, 2)$ e i tre numeri macchina

$x = 0.123$

$y = 45.6$

$z = -45.5$

Calcoliamo $x + y + z$ in due modi:

$(x \oplus y) \oplus z$ o $x \oplus (y \oplus z)$

In $\mathbb{R}$ i due modi sono equivalenti. E in $\mathbb{F}$?

$x \oplus y = fl(x + y) = fl(0.123 + 45.6) =$

$= fl(45.723) = 45.7$ (ho "perso" le cifre 2, 3).

$(x \oplus y) \oplus z = fl(45.7 + (-45.5)) = fl(0.2) = 0.200$

Per l'altra possibilità ho

$y \oplus z = fl(45.6 + (-45.5)) = fl(0.1) = 0.100$

$x \oplus (y \oplus z) = fl(0.123 + 0.100) = fl(0.223) = 0.223$

Pertanto,

→ non vale associativa della somma

→ ho $x+y+z = 0.123 + 45.6 + (-45.5) = 0.223$
per cui la seconda è corretta ma la prima non lo è!

DEF. (STABILITÀ DI UN ALGORITMO) Un algoritmo, che pensiamo come una successione di operazioni macchine elementari, si dice STABILE se e solo se non AMPLIFICA gli errori nelle varie operazioni.
In caso contrario, l'algoritmo si dice INSTABILE.

Chiaramente, ci interessano algoritmi stabili!

ESEMPIO Supponiamo di valutare $f(x) = \sqrt{1+x} - \sqrt{x}$ per $x = 49$ usando IF(10, 1, −1, 2).
In aritmetica esatta è

$$f(49) = \sqrt{50} - \sqrt{49} = 0.0710678\ldots$$

In IF, abbiamo invece

$$\sqrt{50} = fl(\sqrt{50}) = fl(7.071\ldots) = 7$$

$$\sqrt{49} = fl(\sqrt{49}) = fl(7) = 7$$

e quindi, in IF, è

$$\sqrt{50} \ominus \sqrt{49} = 7 - 7 = 0$$

Se cambio la modalità di valutare f, ossia CAMBIO l'ALGORITMO, nel seguente modo

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

$$= \frac{x+1-x}{\sqrt{x+1} + \sqrt{x}} = \frac{1}{\sqrt{x+1} + \sqrt{x}}$$

Ho, in IF,

$$\sqrt{49+1} = \sqrt{50} = 7$$

$\sqrt{49} = 7$

$\sqrt{50} \oplus \sqrt{49} = 7 \oplus 7 = fl(14) = 10$

$\frac{1}{\sqrt{50} \oplus \sqrt{49}} = fl(1/10) = 0.1$

Quindi per lo STESSO PROBLEMA, cambiando algoritmo è possibile contenere meglio gli errori nelle varie operazioni (ossia, avere un algoritmo più stabile).

ESEMPIO Riscrivere un algoritmo stabile per valutare

a) $f(x) = \frac{\sin(x)}{x}$ b) $\frac{1-\cos(x)}{x^2}$

per $x \approx 0$.