Arrotondamento

$$x \in \mathbb{R} \longmapsto fl(x) \in \mathbb{F} \cup \{0, \pm Inf, \text{"denormali"}\}$$

si legge "float di x" o "arrotondamento di x"

numeri macchina normali

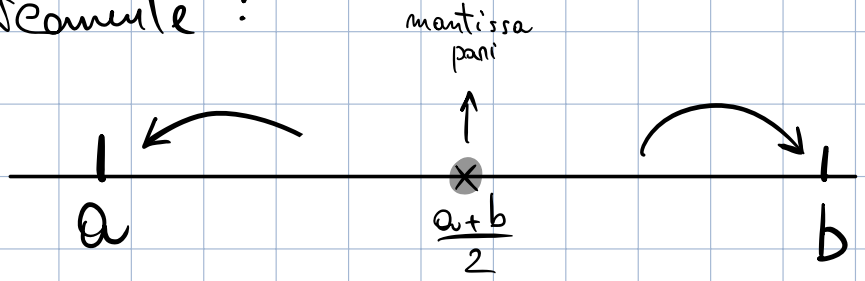
gestione overflow

gestione underflow

Supponiamo che l'arrotondamento non coinvolga underflow o overflow, ovvero $|x| \in [realmin, realmax]$

L'arrotondamento segue la regola "RNTE": "Round To Nearest, Ties to Even"

Graficamente:



dove $a < b \in \mathbb{F}$ consecutivi

Formalmente:

Siano $x \in [a,b]$, $a < b \in \mathbb{F}(\beta, t, m_1, m_2)$ consecutivi, β pari.

- se $x \in [a, \frac{a+b}{2})$, allora $fl(x) = a$
- se $x \in (\frac{a+b}{2}, b]$, allora $fl(x) = b$
- se $x = \frac{a+b}{2}$ e a ha mantissa pari, allora $fl(x) = a$, altrimenti $fl(x) = b$.

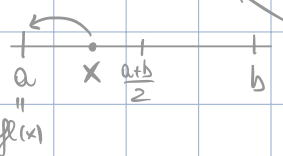
Massimo errore relativo commesso nell'arrotondamento:

$$x \in [\text{realmin}, \text{realmax}] \longmapsto fl(x) \in \mathbb{F}$$

(Per semplicità, supponiamo $x > 0$)

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

errore relativo



$a < b \in [\beta^{p}, \beta^{p+1}]$ numeri macchina consecutivi

Definizione: il numero β^{-t} viene detto "precisione di macchine" e indicato con eps (anche in Matlab).

Dunque il massimo errore relativo commesso nell'arrotondamento RNTE è la metà della precisione di macchina:

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

ESEMPI:

IEEE single prec.: $\frac{1}{2}\beta^{-t} = 2^{-24} \approx 10^{-8}$ (# cifre corrette)

IEEE single prec.: $\frac{1}{2}\beta^{-t} = 2^{-53} \approx 10^{-16}$

Condizionamento dell'aritmetica di macchina

Consideriamo $\tilde{x}$ approx di x $(x, \tilde{x} \in \mathbb{R})$ e definiamo l'errore relativo con segno:

$$\varepsilon_x := \frac{\tilde{x} - x}{x}$$

N.B. Sia $x > 0$. Se $\varepsilon_x > 0$, l'approx è per eccesso $(\tilde{x} > x)$ altrimenti è per difetto.

Si ha:

$$x\,\varepsilon_x = \tilde{x} - x \iff x + x\,\varepsilon_x = \tilde{x} \iff$$

$$\iff \tilde{x} = x(1 + \varepsilon_x)$$

Somma/sottrazione

Siano $x, y \in \mathbb{R}$. Consideriamo

$$(x, y) \longmapsto x + y \in \mathbb{R}$$ ← "aritmetica esatta"

$$(x, y) \longmapsto fl\left(fl(x) + fl(y)\right) \in \mathbb{F}$$

↖ "aritmetica di macchine" o "aritmetica finita"

Abbiamo:

$$fl(x) = x(1 + \varepsilon_x)$$

$$fl(y) = y(1 + \varepsilon_y)$$

$$fl\left(fl(x) + fl(y)\right) = (x + y)(1 + \varepsilon_{x+y})$$

Vogliamo mettere in relazione $\varepsilon_x, \varepsilon_y$ (errori rel. in input) con ε_{x+y} (l'errore rel. in output)

In aritmetica finita:

$$(x + y)(1 + \varepsilon_{x+y}) = x(1 + \varepsilon_x) + y(1 + \varepsilon_y)$$

$$(x+y) + (x+y)\varepsilon_{x+y} =$$

$$= x + x\varepsilon_x + y + y\varepsilon_y$$

$$(x+y)\varepsilon_{x+y} = x\varepsilon_x + y\varepsilon_y \quad \overset{x+y\neq 0}{\Leftrightarrow}$$

$$\varepsilon_{x+y} = \frac{x\varepsilon_x + y\varepsilon_y}{x+y}$$

$$|\varepsilon_{x+y}| = \frac{|x\varepsilon_x + y\varepsilon_y|}{|x+y|} \overset{\text{disug. triang.}}{\leq}$$

$$\leq \frac{|x||\varepsilon_x| + |y||\varepsilon_y|}{|x+y|} \leq$$

maggioro $|\varepsilon_x|$ e $|\varepsilon_y|$ con $\max\{|\varepsilon_x|, |\varepsilon_y|\}$ e metto in evidenza

$$\leq \frac{|x|+|y|}{|x+y|} \max\{|\varepsilon_x|, |\varepsilon_y|\}$$

Riscrivo:

$$|\varepsilon_{x+y}| \leq \underbrace{\frac{|x|+|y|}{|x+y|}}_{K} \max\{|\varepsilon_x|, |\varepsilon_y|\}$$

numero di condizionamento

err. rel. in output

err. rel. in input

Osservazioni

(1) disug. Triang. $\Rightarrow$ $K \geq 1$

(2) K grande se $|X+Y|$ piccolo

Somma / sottrazione:

- ben condizionata se $K \approx 1$
- mal condizionata se $K \gg 1$

Osservazione

X, Y hanno segno concorde $\Rightarrow$

$\Rightarrow |X+Y| = |X| + |Y| \Rightarrow$

$\Rightarrow K = 1$

Dunque:

La somma può essere mal condizionata solo se gli addendi hanno segno opposto e lo è se $y \approx -x$;

È tanto più mal condizionata quanto più y è vicino a $-x$.

Le perdite di precisione dovuta alla somma $x+y$ quando $y \approx -x$ è detta errore di cancellazione (di cifre ...)

"x, y vicini a essere uno l'opposto dell'altro"

Moltiplicazione

Siano $x, y \in \mathbb{R}$. Consideriamo:

$$(x,y) \longmapsto xy \in \mathbb{R} \quad \text{"esatta"}$$

$$(x,y) \longmapsto fl(fl(x)\, fl(y)) \quad \text{"di macchina"}$$

Abbiamo:

$$(xy)(1+\varepsilon_{xy}) = x(1+\varepsilon_x)\,y(1+\varepsilon_y)$$

e dunque

$$1+\varepsilon_{xy} = (1+\varepsilon_x)(1+\varepsilon_y) =$$

$$= 1+\varepsilon_x+\varepsilon_y+\underbrace{\varepsilon_x\varepsilon_y}$$

trascurabile rispetto a ε_x, ε_y, se $|\varepsilon_x|, |\varepsilon_y|$ sono molto piccoli (ipotesi naturale)

Dunque, trascurando l'ultimo addendo:

$$\varepsilon_{xy} \simeq \varepsilon_x + \varepsilon_y$$

Infine

$$|\varepsilon_{xy}| \approx |\varepsilon_x + \varepsilon_y| \overset{\text{disug. } \Delta}{\leq}$$

$$\leq |\varepsilon_x| + |\varepsilon_y| \leq$$

$$\leq 2 \max\{|\varepsilon_x|, |\varepsilon_y|\}$$

Allora

$$|\varepsilon_{xy}| \leq 2 \max\{|\varepsilon_x|, |\varepsilon_y|\}$$

- $|\varepsilon_{xy}|$: err. rel. output
- 2: numero di cond. K
- $\max\{|\varepsilon_x|, |\varepsilon_y|\}$: err. rel. input

L'operazione è sempre ben condizionata

Divisione

Siano $x, y \in \mathbb{R}$, $y \neq 0$. Consideriamo:

$$(x,y) \longmapsto \frac{x}{y} \in \mathbb{R} \quad \text{"esatta"}$$

$$(x,y) \longmapsto fl\left(\frac{fl(x)}{fl(y)}\right) \quad \text{"di macchina"}$$

$$\frac{x}{y}\left(1+\varepsilon_{x/y}\right) = \frac{x(1+\varepsilon_x)}{y(1+\varepsilon_y)}$$

$$1+\varepsilon_{x/y} = \frac{1+\varepsilon_x}{1+\varepsilon_y} \simeq \dots$$

Richiamo (serie geometrica)

$$1+x+x^2+x^3+\dots = \sum_{k=0}^{\infty} x^k = \frac{1}{1-x}, \quad \text{se } |x|<1$$

per x molto piccolo, $x^2, x^3, \dots$ sono trascurabili rispetto a x. Per cui

$$\frac{1}{1\mp x} \simeq 1 \pm x \quad \text{per } x \text{ piccolo}$$

$$\dots \simeq (1+\varepsilon_x)(1-\varepsilon_y) =$$

$$= 1 + \varepsilon_x - \varepsilon_y - \varepsilon_x\varepsilon_y \simeq$$

$$\simeq 1 + \varepsilon_x - \varepsilon_y \quad \left(\text{un } |\varepsilon_x|, |\varepsilon_y| \text{ piccoli}\right);$$

avuto

$$1 + \varepsilon_{x/y} \simeq 1 + \varepsilon_x - \varepsilon_y$$

$$|\varepsilon_{x/y}| \leq |\varepsilon_x| + |\varepsilon_y| \leq$$

$$\leq 2 \max\{|\varepsilon_x|, |\varepsilon_y|\}$$

$$\uparrow$$

$$K$$

Quindi l'operazione è sempre ben condizionata

Ricapitoliamo:

Somma $X+Y$	Mal cond. se $Y \approx -X$
Moltiplicazione	ben cond.
divisione	ben cond.