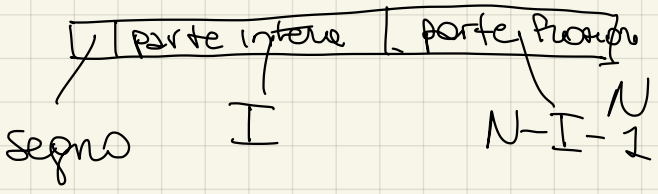
Lezione del 27/1/2023

Venerdì a lezione con il portatile con installato matlab / octave

Rappresentazione dei numeri

→ virgola fissa




es: quanti sono i numeri rappresentabili con questo tipo di rappresentazione? (in generale per una base $\beta$)

Teorema di rappresentazione in base

$x \in \mathbb{R}$, $x \neq 0$ allora scelta una base di rappresentazione $\beta$ esistono e sono unici

① $p \in \mathbb{Z}$ detto esponente

② una successione $\{d_i\}_{i \geq 1}$ i numeri sono normalizzati

a. $d_1 \neq 0$

b. $0 \leq d_i \leq \beta - 1$ interi

c. $d_i$ non definitivamente uguali a $\beta - 1$

per $i > M$

tali che

$$x = \text{segno}(x) \cdot \beta^{p} \underbrace{\sum_{i=1}^{\infty} d_i \beta^{-i}}_{\text{mantissa}}$$

a e c servono per garantire l'unicità della rappresentazione

$$0.123 = 10\,(0.0123) = 10^{2}(0.00123)$$

($d_1 \neq 0$)

$\beta = 10$

$$0.345 = 10^{0}\left(3 \cdot 10^{-1} + 4 \cdot 10^{-2} + 5 \cdot 10^{-3}\right)$$

$$0.\overline{9} = 1$$

$$0.\overline{9} = \sum_{i=1}^{\infty} 9 \cdot 10^{-i} = 9 \cdot \sum_{i=1}^{\infty} 10^{-i} = 1$$

$$\sum_{i=0}^{\infty} q^{i} = \frac{1}{1-q}$$

converge solo se $|q| < 1$

$$\sum_{i=0}^{n} q^{i} = \frac{1-q^{n+1}}{1-q}$$

# Floating point (virgola mobile)

$$\mathcal{F}(\beta, t, m, M) = \{0\} \cup \{x \in \mathbb{R},$$

$\beta$: base; $t$: numero cifre delle rappresentazione; $m, M$: range per esponente

$$x = \text{segno}(x)\ \beta^p \sum_{i=1}^{t} d_i \beta^{-i},$$

$\beta^p (0.d_1 d_2 \dots d_t)_\beta$

(a) $d_1 \neq 0$

(b) $0 \leq d_i \leq \beta - 1$ intere

$-m \leq p \leq M \}$ INSIEME DEI NUMERI DI MACCHINA

es: $\mathcal{F}(2, 3, 1, 1)$

$x \in \mathcal{F} \Rightarrow x = \text{segno}(x) \cdot \beta^p \cdot f$ ← mantissa, $0 \leq f < 1$

- numero più piccolo positivo rappresentabile?

$$\omega = \beta^{-m} (0.10\dots0)_\beta = \beta^{-m-1}$$

- numero più grande positivo rappresentabile

$$\Omega = \beta^M (0.(\beta-1)\dots(\beta-1))_\beta = \beta^M \sum_{i=1}^{t} (\beta-1)\beta^{-i}$$

$t$ cifre

$$= \beta^M \cdot (\beta - 1) \sum_{i=1}^{t} \beta^{-i} = \boxed{\beta^M \left(1 - \beta^{-t}\right)}$$

$\mathcal{F}(2, 3, 1, 1)$

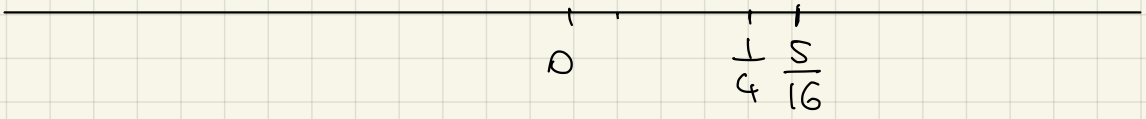
$\omega = 2^{-2} = 1/4$

$2^{-1}(.100)$

$.101$

$.110$

$.111$




numero successivo a $\frac{1}{4}$ in $\mathcal{F}$ è

$$\frac{1}{4} + 2^{-1} \cdot 2^{-3} = \frac{1}{4} + \frac{1}{16} = \frac{5}{16}$$

$$\Omega = \beta^M \left(1 - \beta^{-t}\right) = 2\left(1 - 2^{-3}\right) = 2 - 2^{-2} = \frac{7}{4}$$

cardinalità dell'insieme $\mathcal{F}$

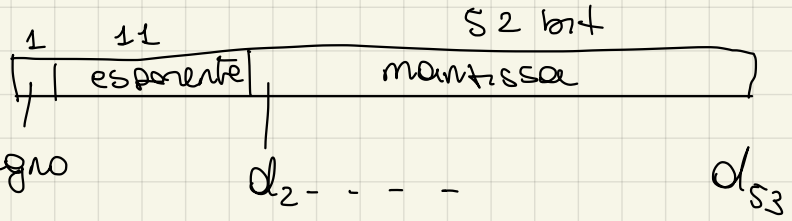
$$\# \mathcal{F} = 1 + 2 \left(m + M + 1\right)\left(\beta - 1\right) \cdot \beta^{t-1}$$

- $1$: zero
- $2$: $\pm$
- $m + M + 1$: possibili esponenti
- $(\beta - 1) \cdot \beta^{t-1}$: possibili mantisse

Standard IEEE

base 2

double 64 bit




$$\mathcal{F}(2, 53, 1021, 1024)$$

↑ perché $d_1 \neq 0 \Rightarrow d_1 = 1$

esponente 1 - - - 1 → numeri speciali

$f = 0$ (mantissa) → $+Inf$ segno $= 0$

$-Inf$ segno $= 1$

$x + Inf = Inf$

$Inf / 0 = Inf$

$f \neq 0 \rightarrow NaN$ not a number

$x/0 \rightarrow NaN$

esponente = 00000 (11 bit)

$f = 0.0\, d_2 - - - d_{53}$ ($d_1 = 0$)

numeri denormalizzati

$$(0.0d_2 - - - d_{53}) \cdot 2^{-1021}$$

# Limitazioni dell'aritmetica floating point

$\hat{x} = segno(x) f \cdot \beta^p$ numero di macchina

1. $p \notin [-m, M]$

$x \in \mathbb{R}$

$x = segno(x)\ \beta^p \cdot \sum_{i=1}^{\infty} d_i \beta^{-i}$

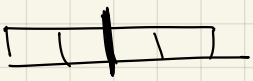
| | | |
|---|---|---|
| $p > M$ | "overflow" | $\hat{x} = \pm Inf$ |
| $p < -m$ | "underflow" | $\hat{x} = 0$ |

esempio di scaling per evitare overflow

$$c = \sqrt{a^2 + b^2}$$

$\beta = 10 \quad t = 2$

$p \in [-99, 99]$



$a = 10^{60}$

$b = 1$

$a^2 = 10^{120}$ → errore di overflow → $c = Inf$

$$c = s\sqrt{\left(\frac{a}{s}\right)^2 + \left(\frac{b}{s}\right)^2} \qquad s \neq 0$$

$s = \max\{|a|, |b|\} = 10^{60}$

$$c = 10^{60} \cdot \sqrt{1 + \left(\frac{1}{10^{60}}\right)^2} \approx \underline{10^{60}}$$

↖ underflow ($10^{-120} \notin \mathbb{F}$)

$$c = \sqrt{10^{120}+1} \approx$$

$\boxed{2}$ $x \in \mathbb{R}$, $x = \text{segno}(x)\beta^{p} \cdot \underbrace{\sum_{i=1}^{\infty} d_i \beta^{-i}}$

$\exists k : d_k \neq 0$ per $k > t$

$$\tilde{x} \begin{cases} \text{trunc}(x) = \text{segno}(x)\beta^{p} \sum_{i=1}^{t} d_i \beta^{-i} \\ \text{arr}(x) = \text{segno} \end{cases}$$

$$\pi = 3.\underbrace{1415}\,9\,2 \ldots$$

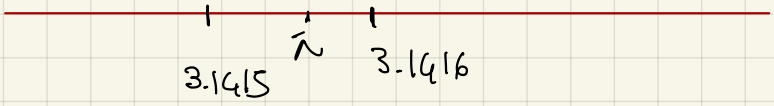
$\boxed{t=4}$

$$\tilde{\pi} = \begin{cases} \text{trun}(\pi) = 3.1415 \\ \text{arr}(\pi) = 3.1416 \end{cases}$$

$$3.999 \to 4.00$$

$x \in \mathbb{R} \to \tilde{x} \in \mathcal{F}$ $\quad \varepsilon_x = x - \tilde{x}$ errore assoluto

$$\boxed{\varepsilon_x = \frac{x - \tilde{x}}{x} \quad \text{errore relativo}}$$




$$a = \beta^p \cdot \sum_{i=1}^{t} d_i \beta^{-i}$$

il successivo di $a$ $\quad b = a + \beta^{p-t}$

$$\beta^p (.d_1 \ldots d_t)_t + \beta^p (.00\ldots1)$$

**Teorema**: se non si verifica underflow o overflow

$x \in \mathbb{R}, \; x \neq 0$

$$|trunc(x) - x| < \beta^{p-t}$$

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

Lezione del 1/2/2023

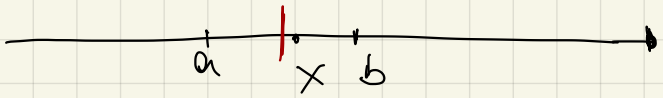
**Teorema:** se non si verifica underflow o overflow

$x \in \mathbb{R}, \quad x \neq 0$

$$|trunc(x) - x| < \beta^{p-t}$$

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

**Dim** $\boxed{x > 0}$




$x = \Omega \qquad trunc(x) = arr(x) = \Omega \Rightarrow |trunc(x) - x| = 0 < \beta^{p-t}$

$|arr(x) - x| = 0 < \frac{1}{2} \beta^{p-t}$

$$x = \beta^p \cdot \sum_{i=1}^{\infty} d_i \beta^{-i}$$

$$a = \beta^p \sum_{i=1}^{t} d_i \beta^{-i} \in \mathbb{F}$$

$$b = \beta^p \left( \sum_{i=1}^{t} d_i \beta^{-i} + \beta^{-t} \right)$$

se opero per troncamento

$\Rightarrow \quad trunc(x) = a$

$$|a - x| = x - a < \beta^{p-t}$$

$b - a$

se opero per arrotondamento

$$arr(x) = \begin{cases} a & \text{se } x < \frac{a+b}{2} \\ b & \text{se } x \geq \frac{a+b}{2} \end{cases}$$

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

**Teorema:** Se non abbiamo overflow o underflow

$$\left|\frac{fl(x) - x}{x}\right| < u$$

$u$ = precisione di macchina

$$u = \begin{cases} \beta^{1-t} & \text{se } fl(x) = trunc(x) \\ \frac{1}{2}\beta^{1-t} & \text{se } fl(x) = arr(x) \end{cases}$$

$$\tilde{x} = fl(x) = \begin{cases} arr(x) \\ trunc(x) \end{cases}$$

Notazione

eps $\leftarrow$ in Matlab

in IEEE $t = 53$, $\beta = 2$

$$u = \frac{1}{2} \cdot 2^{1-53} = 2^{-53} \approx 10^{-16}$$

**Dim:** Dal teo precedente

$$|fl(x) - x| \leq \begin{cases} \beta^{p-t} & fl(x) = trunc(x) \\ \frac{1}{2}\beta^{p-t} & fl(x) = arr(x) \end{cases}$$

$$x = \beta^p \cdot \sum_{i=1}^{t} d_i \beta^{-i}$$

$$|x| \overset{?}{\geq} \beta^p (0.1)_\beta = \beta^{p-1}$$

$$\frac{|fl(x) - x|}{|x|} \leq \begin{cases} \dfrac{\beta^{p-t}}{\beta^{p-1}} = \beta^{1-t} & \text{tronc.} \\ \dfrac{1}{2}\dfrac{\beta^{p-t}}{\beta^{p-1}} = \dfrac{1}{2}\beta^{1-t} & \text{arr.} \end{cases}$$

$$\varepsilon_x = \frac{fl(x) - x}{x}$$ ← errore di rappresentazione

Il modulo dell'errore di rappresentazione è limitato dalla precisione di macchina.

$$fl(x) = x(1 + \varepsilon_x) \quad \text{con} \quad |\varepsilon_x| \leq u$$

Operazioni di macchina

$\mathcal{F}(10, 3, m, M)$

$fl(a) = (0.123)\,10^0$ $\qquad$ $fl(b) = 10^{-2}(0.456)$

$fl(a) + fl(b)$

$0.12300$
$0.00456$
$0.12756$

$0.128$ arr.

$0.127$ tronc

$$fl\Big(fl(a)+fl(b)\Big) = fl\Big(a(1+\varepsilon_a)+b(1+\varepsilon_b)\Big) =$$

$$= \Big(a(1+\varepsilon_a)+b(1+\varepsilon_b)\Big)(1+\varepsilon_+)$$

$\varepsilon_+$: errore locale dell'operazione

$$|\varepsilon_a| < u \quad , \quad |\varepsilon_b| < u$$

Anche $|\varepsilon_+| < u$ : posso applicare il teorema al numero reale $\big(a(1+\varepsilon_a)+b(1+\varepsilon_b)\big)$

$$a + a\varepsilon_a + b + b\varepsilon_b + a\varepsilon_+ + a\varepsilon_a\varepsilon_+ + b\varepsilon_+ + b\varepsilon_b\varepsilon_+$$

$$\doteq a+b + a\varepsilon_a + b\varepsilon_b + (a+b)\varepsilon_+$$

$\doteq$ approssimo al 1° ordine

$$\varepsilon_{TOT} = \frac{(a+b) + a\varepsilon_a + b\varepsilon_b + (a+b)\varepsilon_+ - (a+b)}{a+b} =$$

$$= \boxed{\frac{a}{a+b}\varepsilon_a + \frac{b}{a+b}\varepsilon_b} + \boxed{\varepsilon_+}$$

coefficienti di amplificazione — errore inerente o inevitabile — errore algoritmico

## Coefficienti di amplificazione

| | | | |
|---|---|---|---|
| somma | $a+b$ | $\frac{a}{a+b}$ | $\frac{b}{a+b}$ |
| sottrazione | $a-b$ | $\frac{a}{a-b}$ | $\frac{-b}{a-b}$ |
| moltiplicazione | $a*b$ | $1$ | $1$ |
| divisione | $a/b$ | $1$ | $-1$ |

$$f(x_1, x_2 \dots x_n) \longrightarrow f\left(fl(x_1), fl(x_2) \dots, fl(x_n)\right)$$

$$\downarrow$$

$$g\left(fl(x_1) \dots fl(x_n)\right)$$

$$f(x_1, x_2, x_3) = (x_1 + x_2)\, x_3$$

$$f\left(\underbrace{fl(x_1)}_{\tilde{x}_1}, \underbrace{fl(x_2)}_{\tilde{x}_2}, \underbrace{fl(x_3)}_{\tilde{x}_3}\right) = \left(fl(x_1) + fl(x_2)\right) \cdot fl(x_3)$$

$$g(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) = fl\left(fl(\tilde{x}_1 + \tilde{x}_2) \cdot \tilde{x}_3\right)$$

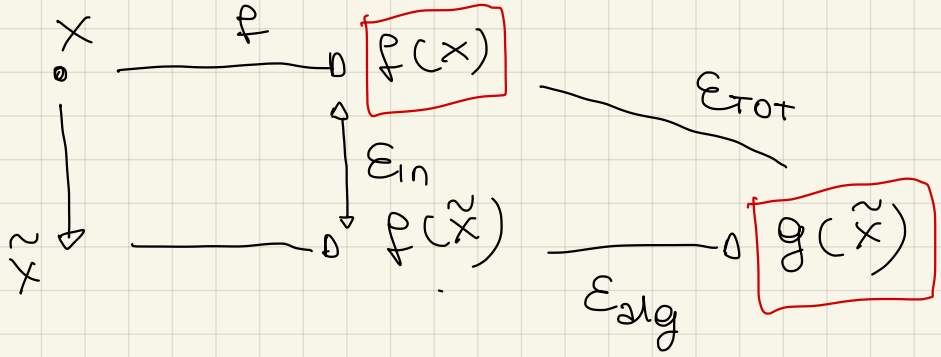
$$fl\left(fl(\tilde{x}_1 \cdot \tilde{x}_3) + fl(\tilde{x}_2 \cdot \tilde{x}_3)\right)$$

**Def:** Sia $f$ una funzione razionale, $f(x_1 \dots x_n) \to \mathbb{R}$

sia $x \neq 0$, $x = (x_1 \dots x_n)$ : $f(x) \neq 0$

$$\varepsilon_{in} = \frac{f(\tilde{x}_1, \tilde{x}_2, \dots \tilde{x}_n) - f(x_1 \dots x_n)}{f(x_1 \dots x_n)}$$

errore inerente




$$\varepsilon_{alg} = \frac{g(\tilde{x}_1, \dots \tilde{x}_n) - f(\tilde{x}_1 \dots \tilde{x}_n)}{f(\tilde{x}_1 \dots \tilde{x}_n)}$$

Errore algoritmico

**Teorema:** $\varepsilon_{TOT} \doteq \varepsilon_{IN} + \varepsilon_{ALG}$

$f(x) \neq 0$
$f(\tilde{x}) \neq 0$

$$\varepsilon_{TOT} = \frac{g(\tilde{x}) - f(x)}{f(x)} = \frac{g(\tilde{x}) - f(\tilde{x})}{f(x)} + \frac{f(\tilde{x}) - f(x)}{f(x)}$$

$\varepsilon_{in}$

$$= \underbrace{\frac{g(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})}}_{\varepsilon_{alg}} \cdot \underbrace{\frac{f(\tilde{x})}{f(x)}}_{1 + \varepsilon_{IN}} + \varepsilon_{IN}$$

$$= \varepsilon_{ALG}(1 + \varepsilon_{IN}) + \varepsilon_{IN} \leq \varepsilon_{ALG} + \cancel{\varepsilon_{ALG}\varepsilon_{IN}} + \varepsilon_{IN}$$

$$\doteq \varepsilon_{IN} + \varepsilon_{ALG}$$

Se l'errore inerente è alto (rispetto alla precisione di macchina) allora si dice che il PROBLEMA è MALCONDIZIONATO

Se l'errore algoritmico è alto allora si dice che L'ALGORITMO è INSTABILE

Ben condizionato se $|\varepsilon_{IN}| < K \cdot u$ (K costante bassa)

stabile se $|\varepsilon_{ALG}^{(g)}| < h \cdot u$

Formula per il calcolo dell'errore inerente:

$f(x)$

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

e $f \in C^1([a,b])$

$$f(\tilde{x}) = f(x) + xf'(x)\underbrace{\frac{(\tilde{x}-x)}{x}}_{\varepsilon_x} + \cancel{O((\tilde{x}-x)^2)}$$

$$\varepsilon_{IN} = \frac{f(\tilde{x})-f(x)}{f(x)} \doteq \boxed{\frac{x}{f(x)} f'(x)} \cdot \varepsilon_x = c_x \cdot \varepsilon_x$$

$c_x$ coefficiente di amplificazione

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + f''(x_0)\frac{(x-x_0)^2}{2} + \dots + f^{n+1}(\xi)\frac{(x-x_0)^{n+1}}{(n+1)!}$$

$|\xi - x_0| \leq |x - x_0|$

$f(x_1 - - x_n)$

$$\varepsilon_{IN} = \sum_{i=1}^{n} c_{x_i} \cdot \varepsilon_{x_i}$$

$$c_{x_i} = \frac{x_i}{f(x_1 - - x_n)} \frac{\partial f}{\partial x_i}$$

$$f(a, b) = a + b$$

$$c_a = \frac{a}{a+b} \qquad c_b = \frac{b}{a+b}$$

$$f(a, b) = a^2 - 3ab$$

$$\varepsilon_{IN} = c_a \varepsilon_a + c_b \varepsilon_b$$

$$c_a = \frac{a}{a^2 - 3ab} \cdot (2a - 3b) = \frac{2a - 3b}{a - 3b} =$$

$$= \frac{a}{a - 3b} + 1$$

$$c_b = \frac{b}{a^2 - 3ab} (-3a) = \frac{-3b}{a - 3b}$$

$$\varepsilon_{IN} = \left(\frac{a}{a-3b} + 1\right) \varepsilon_a + \left(\frac{-3b}{a-3b}\right) \varepsilon_b$$

Il problema è ben condizionato o esistono valori di

a e b per cui è malcondizionato

$a \approx 3b$ il problema è malcondizionato

$|\varepsilon_a| < u$, $|\varepsilon_b| < u$

$$|\varepsilon_{in}| = \left| \left( \frac{a}{a-3b} + 1 \right) \varepsilon_a + \left( \frac{-3b}{a-3b} \right) \varepsilon_b \right| \leq$$

$$\left| \frac{a}{a-3b} + 1 \right| |\varepsilon_a| + \left| \frac{-3b}{a-3b} \right| |\varepsilon_b|$$

$|\varepsilon_a| \leq u$, $|\varepsilon_b| \leq u$

Esercizio

$$f(x_1, \dots x_n) = \sum_{i=1}^{n} x_i$$

Errore inerente con la formula

Errore algoritmico con 1 o più algoritmi:

Esercizio

$$f(x) = x^2 - 1 \quad \begin{cases} x \cdot x - 1 = g_1(x) \\ (x+1)(x-1) = g_2(x) \end{cases}$$

$$\tilde{x} = fl(x)$$

$$\varepsilon_{TOT}^{(1)} = \frac{g_1(\tilde{x}) - f(x)}{f(x)}$$

$$\varepsilon_{TOT}^{(2)} = \frac{g_2(\tilde{x}) - f(x)}{f(x)}$$

Vorremo trovare che

$$\varepsilon_{TOT}^{(1)} = \boxed{\varepsilon_{IN}} + \varepsilon_{ACC}^{(1)}$$

$$\varepsilon_{TOT}^{(2)} = \boxed{\varepsilon_{IN}} + \varepsilon_{ACC}^{(2)}$$

L'errore inerente non dipende dall'algoritmo utilizzato

$$\tilde{x} = \hat{\rho}_1(x) = x(1+\varepsilon_x)$$

$$g_1(\tilde{x}) = \hat{\rho}_1\left(\hat{\rho}_1(\tilde{x}\cdot\tilde{x}) - 1\right)$$

$$|\varepsilon_x|<u = \hat{\rho}_1\left(\hat{\rho}_1\left(x(1+\varepsilon_x)\cdot x(1+\varepsilon_x)\right) - 1\right)$$

$$|\varepsilon_\cdot|<u \quad |\varepsilon_-|<u = \hat{\rho}_1\left(x^2(1+\varepsilon_x)^2(1+\varepsilon_\cdot) - 1\right)$$

$$= \left[x^2(1+\varepsilon_x)^2(1+\varepsilon_\cdot) - 1\right](1+\varepsilon_-)$$

$$= \left[x^2(1+\cancel{\varepsilon_x^2} + 2\varepsilon_x)(1+\varepsilon_\cdot) - 1\right](1+\varepsilon_-)$$

$$= \left[x^2(1+2\varepsilon_x+\varepsilon_\cdot+\cancel{2\varepsilon_x\varepsilon_\cdot}) - 1\right](1+\varepsilon_-)$$

$$= x^2(1+2\varepsilon_x+\varepsilon_\cdot+\varepsilon_-) - (1+\varepsilon_-)$$

$$= x^2 - 1 + 2x^2\varepsilon_x + x^2\varepsilon_\cdot + (x^2-1)\varepsilon_-$$

$$\varepsilon_{tot}^{(a)} = \frac{\cancel{x^2-1} + 2x^2\varepsilon_x + x^2\varepsilon_\cdot + (x^2-1)\varepsilon_- - \cancel{(x^2-1)}}{(x^2-1)}$$

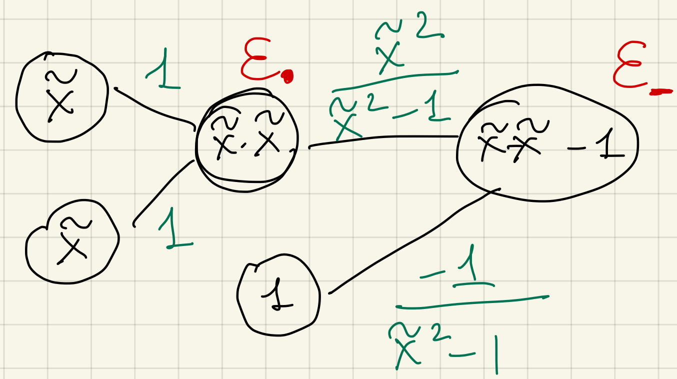
$$= \underbrace{\frac{2x^2}{x^2-1}\varepsilon_x}_{\text{erreur relative}} + \frac{x^2}{x^2-1}\varepsilon_\cdot + \varepsilon_-$$

$$\varepsilon_{IN} = c_x \, \varepsilon_x \qquad c_x = \frac{x}{f(x)} f'(x)$$

Infatti: $c_x = \frac{x}{x^2-1} \cdot 2x = \frac{2x^2}{x^2-1}$

Uso del grafo per lo studio dell'errore algoritmico

$$g_1(\tilde{x}) = \tilde{x} \cdot \tilde{x} - 1$$

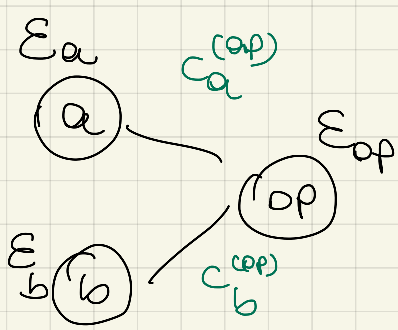



$a - b$

$c_a = \frac{a}{a-b}$

$c_b = \frac{-b}{a-b}$

**Regole**



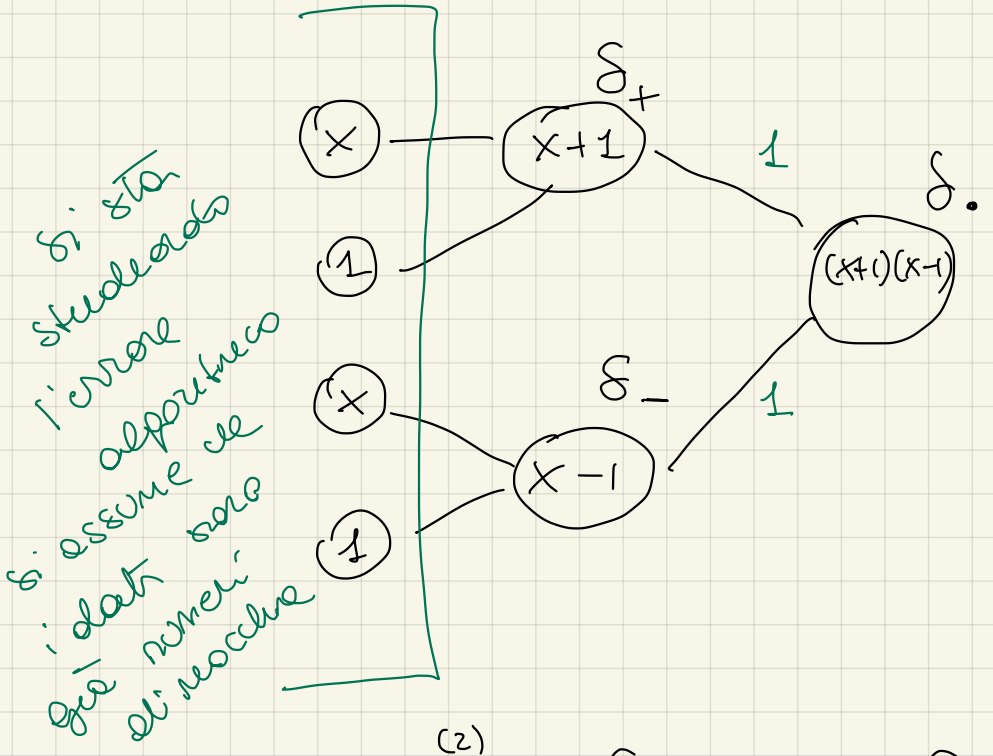

$$\varepsilon_{op} + c_a^{(op)} \varepsilon_a + c_b^{(op)} \varepsilon_b$$

$$\varepsilon_{ALG}^{(1)} = \varepsilon_- + \frac{-1}{x^2-1} \varepsilon_1 + \frac{x^2}{x^2-1} \left( \varepsilon_\circ + 1 \cdot \varepsilon_{\tilde{x}} + 1 \varepsilon_{\tilde{x}} \right)$$

($\varepsilon_1 = 0$ ← errore di rappr. di 1; $\varepsilon_{\tilde{x}} = 0$, $\varepsilon_{\tilde{x}} = 0$)

$$= \boxed{\varepsilon_- + \frac{x^2}{x^2-1} \varepsilon_\circ}$$

$$g_2(x) = (x+1)(x-1)$$




si sta studiando l'errore algoritmico si assume che i dati siano già numeri di macchina

$$|\delta_+| < u$$

$$|\delta_-| < u$$

$$\delta_+ = \frac{fl(x+1) - (x+1)}{x+1}$$

$$\varepsilon_{ACG}^{(2)} = \delta_\cdot + 1 \cdot \delta_+ + 1\,\delta_-$$

$$= \delta_\cdot + \delta_+ + \delta_-$$

$$|\varepsilon_{ACG}^{(2)}| = |\delta_\cdot + \delta_+ + \delta_-| \leq |\delta_\cdot| + |\delta_+| + |\delta_-| \leq 3u$$

disuguaglianza triangolare

l'algoritmo 2 è sempre stabile!

$$|\varepsilon_{ACG}^{(a)}| = \left|\varepsilon_\cdot + \frac{x^2}{x^2-1}\,\varepsilon_\cdot\right| \leq |\varepsilon_-| + \left|\frac{x^2}{x^2-1}\right| |\varepsilon_\cdot|$$

$$< u + \frac{x^2}{|x^2-1|} \cdot u = u\left(1 + \frac{x^2}{|x^2-1|}\right)$$

l'algoritmo è instabile per $x \to \pm 1$

$$|\varepsilon_{IN}| = \left| \frac{2x^2}{x^2-1} \right| |\varepsilon_x|$$

$\gg 1$

il problema è malcondizionato per $x \to \pm 1$

$x = 1 + 2^{-50}$

IEEE $t = 53$

$n_1 = 1021$

$n = 1024$

$x \in \mathbb{F}$ ?

$$x = 2^1 \left( 2^{-1} + 2^{-51} \right) =$$

$$= 2^1 \left( 0.\underset{\uparrow}{1}0\cdots\underset{\uparrow}{1}00 \right)_2$$

1 51

$\varepsilon_x = 0$ $\varepsilon_{IN} = 0$

$g_1(x) = x \cdot x - 1$

$$fl(x \cdot x) = fl\left(\left(1+2^{-50}\right)\left(1+2^{-50}\right)\right) =$$

$$= fl\left(1 + 2^{-49} + 2^{-100}\right)$$

$$fl\left(2^{1}\left(0.\overset{1}{1}\,0 - - - - 0\,\overset{50}{1}\,0 - \underset{q}{\cdots} - \overset{101}{1}\right)\right) =$$

$$2^{1}\left(1 + 2^{-50}\right)$$

$$\varepsilon_{\bullet} = \frac{\left(1+2^{-49}\right) - \left(1+2^{-49}+2^{-100}\right)}{1+2^{-49}+2^{-100}} = \frac{-2^{-100}}{\cdots}$$

$$fl(x \cdot x) = 1 + 2^{-49}$$

$$fl\left(1+2^{-49} - 1\right) = 2^{-49} = g_1(x)$$

$$f\left(1+2^{-50}\right) = 2^{-49} + 2^{-100} \in \mathcal{F}$$

$$2^{-48}\left(2^{-1} + 2^{-52}\right)$$

$$2^{-48}\left(0.1\,0 - - - - \cdot 1\,0\right)_2$$

$g_2(x) = (x+1)(x-1)$

$fl(x+1) = fl(1 + 2^{-50} + 1) = fl(2 + 2^{-50}) =$

$= fl\left(2^2\left(2^{-1} + 2^{-52}\right)\right) = 2^2\left(2^{-1} + 2^{-52}\right)$

$fl(x-1) = fl(1 + 2^{-50} - 1) = fl(2^{-50}) = 2^{-49}(0.1)_2$

$2^{-1}$

$fl\left(2^2(2^{-1} + 2^{-52}) \cdot 2^{-49}\left(2^{-1}\right)\right) =$

$= fl\left(2^{-50}\left(2 + 2^{-50}\right)\right) = fl\left(2^{-49} + 2^{-100}\right) =$

$2^{-49} + 2^{-100} \in \mathbb{F}$

In questo caso $\varepsilon_{ACC}^{(2)} = 0$

$1,100$

$c = 10^{20} + 2 \cdot 10^{14} + 10^8$

$c - a^2 =$

Lezione del 8/2/2023

$$f(x) = x_1 + \dots + x_n \qquad x_i \in \mathbb{R}^+ \qquad x_i > 0$$

$\varepsilon_{in}$ è indipendente dall'algoritmo.

$$\varepsilon_{in} = \frac{f(\tilde{x}) - f(x)}{f(x)}$$

$$\tilde{x}_i = x_i(1+\varepsilon_{x_i}) \qquad |\varepsilon_{x_i}| = \left|\frac{\tilde{x}_i - x_i}{x_i}\right| \text{ errore di rappr.} \qquad |\varepsilon_{x_i}| < u$$

$$f(\tilde{x}) = x_1(1+\varepsilon_{x_1}) + x_2(1+\varepsilon_{x_2}) + \dots + x_n(1+\varepsilon_{x_n})$$

$$\varepsilon_{in} \doteq \sum_{i=1}^{n} c_{x_i} \cdot \varepsilon_{x_i} \qquad c_{x_i} = \frac{x_i}{f(x)} \frac{\partial f}{\partial x_i}$$

$$\frac{\partial f}{\partial x_i} = 1 \qquad c_{x_i} = \frac{x_i}{f(x)}$$

$$\varepsilon_{in} \doteq \sum_{i=1}^{n} \frac{x_i}{f(x)} \varepsilon_{x_i} = \frac{1}{f(x)} \sum_{i=1}^{n} x_i \varepsilon_{x_i}$$

discuss. teorema

$$|\varepsilon_{in}| = \left|\frac{1}{f(x)} \sum_{i=1}^{n} x_i \varepsilon_{x_i}\right| = \frac{1}{f(x)} \left|\sum_{i=1}^{n} x_i \varepsilon_{x_i}\right| \leq$$

$$\frac{1}{f(x)} \sum_{i=1}^{n} |x_i| |\varepsilon_{x_i}| < \frac{1}{f(x)} \sum_{i=1}^{n} |x_i| u = u \frac{1}{f(x)} \boxed{\sum_{i=1}^{n} |x_i|} = f(x)$$

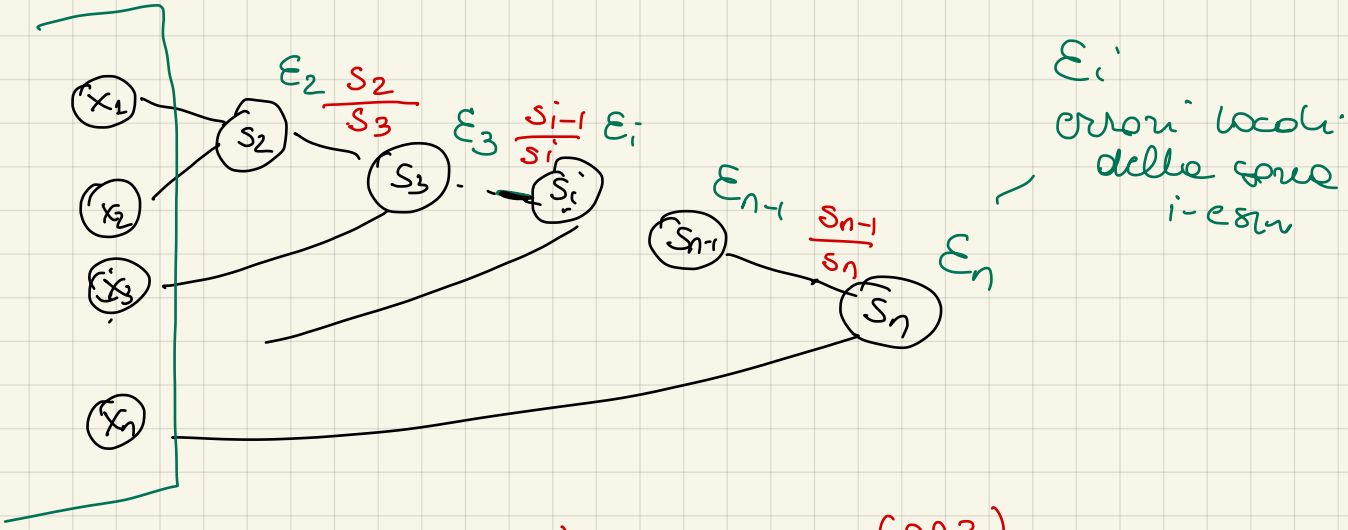
($|\varepsilon_{x_i}| < u$)

$\leq u$ ha senso di $n$ numeri dello stesso segno è ben condizionato

Algoritmo 1

$$S_1 = x_1$$
$$S_2 = S_1 + x_2$$
$$\vdots$$
$$S_i = S_{i-1} + x_i$$
$$\vdots$$
$$S_n = S_{n-1} + x_n$$

$$s_i = \sum_{k=1}^{i} x_k$$

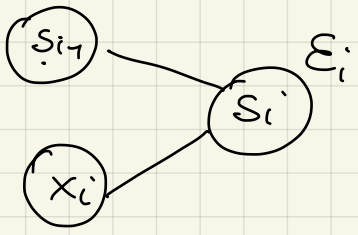



$$\varepsilon_i + c_{op1}\,\delta^{(op1)} + c_{op2}\,\delta^{(op2)}$$

errore accumulato per il calcolo del primo operando

$$\delta^{(i)} = \varepsilon_i + \frac{S_{i-1}}{S_i}\,\delta^{(i-1)}$$

errore accumulato per il calcolo di $S_i$




$$\varepsilon_{ALG} = \delta^{(n)} = \varepsilon_n + \frac{S_{n-1}}{S_n}\,\delta^{(n-1)} =$$

$$\varepsilon_n + \frac{S_{n-1}}{S_n}\left(\varepsilon_{n-1} + \frac{S_{n-2}}{S_{n-1}}\,\delta^{(n-2)}\right) =$$

$$= \varepsilon_n + \frac{s_{n-1}}{s_n}\left(\varepsilon_{n-1} + \frac{s_{n-2}}{s_{n-1}}\left(\varepsilon_{n-2} + \frac{s_{n-3}}{s_{n-2}}\left(\varepsilon_{n-3} + \dots + 1\right)\right)\right)$$

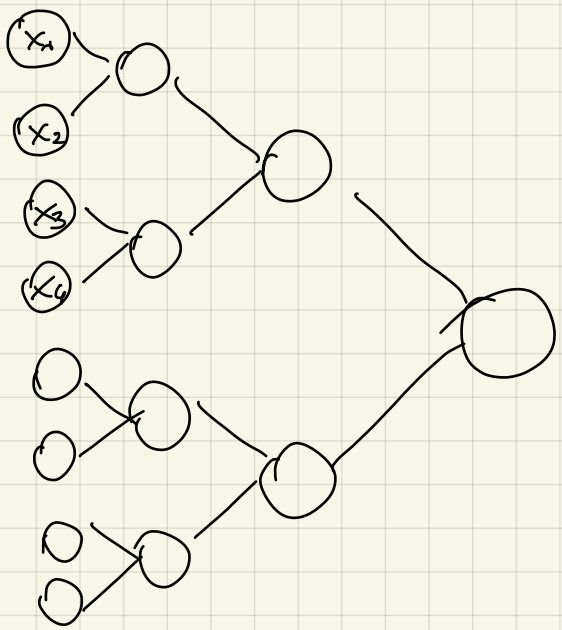
$$= \frac{s_n}{s_n}\varepsilon_n + \frac{s_{n-1}}{s_n}\varepsilon_{n-1} + \frac{s_{n-2}}{s_n}\varepsilon_{n-2} + \frac{s_{n-3}}{s_n}\varepsilon_{n-3} + \dots + \frac{s_2}{s_n}\varepsilon_2$$

$$= \frac{1}{s_n}\sum_{i=2}^{n} s_i \varepsilon_i$$

$$|\varepsilon_{ACC}| = \frac{1}{s_n}\cdot\left|\sum_{i=2}^{n} s_i \cdot \varepsilon_i\right| \le \frac{1}{s_n}\sum_{i=2}^{n} |s_i|\,|\varepsilon_i| < u$$

$$\hookleftarrow \frac{1}{s_n}\, u \cdot \sum_{i=2}^{n} s_i = u \cdot \frac{1}{s_n}\sum_{i=2}^{n} \underbrace{\sum_{k=1}^{i} x_k}_{s_n}$$

$$\hookleftarrow \frac{u}{s_n}\sum_{i=2}^{n} s_n = (n-1)\,u$$




$$|\varepsilon^{(2)}_{ACC}| < u \cdot \log_2 n$$

## Errore analitico

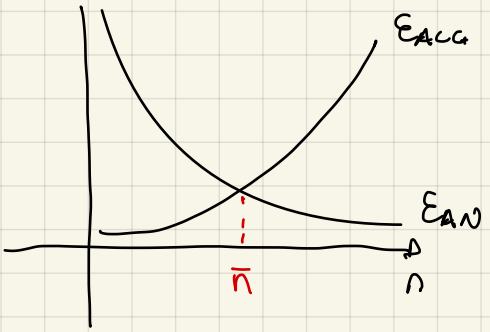
$f(x)$ → $h(x)$ razionale

es $\Rightarrow$ $f(x) = \exp(x) = e^x$

$$h(x) = \sum_{i=0}^{n} \frac{x^i}{i!}$$

$f(x) \xrightarrow{\varepsilon_{AN}} h(x) \xrightarrow{\varepsilon_{IN}} h(\tilde{x}) \xrightarrow{\varepsilon_{ALG}} g(\tilde{x})$ sostituisco le operazioni di macchina

$$\varepsilon_{TOT} = \frac{g(\tilde{x}) - f(x)}{f(x)} \doteq \varepsilon_{IN} + \varepsilon_{ALG} + \varepsilon_{AN}$$




$$\varepsilon_{AN} = \frac{h(x) - f(x)}{f(x)}$$ valutato $$\frac{h(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})}$$

## Esercizio

Si vuole approssimare $g(x) = \log(1+x)$ con la funzione $f(x) = \frac{2x}{x+2}$. Si studi l'errore analitico, algoritmico ed inerente che si commette approssimando $g(x)$ con $f(x)$ nell'intervallo $x \in [-\frac{1}{2}, 0]$

$$E_{an} = \frac{f(x) - g(x)}{g(x)}$$

Analizziamo indipendentemente numeratore e denominatore.

Si osserva che $g(x)$ e $f(x)$ condividono vari termini del proprio sviluppo di Taylor. Infatti:

$$g'(x) = \frac{1}{1+x} \qquad g''(x) = -(1+x)^{-2} \qquad g''' = 2(1+x)^{-3}$$

$$f'(x) = \frac{4}{(x+2)^2} \qquad f''(x) = -\frac{8}{(x+2)^3} \qquad f'''(x) = \frac{24}{(x+2)^4}$$

e $g(x) = g(0) + g'(0)\cdot x + g''(0)\frac{x^2}{2} + g'''(\xi)\frac{x^3}{3!}$

con $|\xi| < |x|$.

Abbiamo quindi che

$$g(x) = x - \frac{x^2}{2} + \frac{1}{(1+\xi)^3}\frac{x^3}{3}$$

$$f(x) = f(0) + f'(0)x + f''(0)\frac{x^2}{2} + f'''(\eta)\frac{x^3}{3!} \quad \text{con } |\eta| < |x|$$

Abbiamo quindi de

$$f(x) = x - \frac{x^2}{2} + \frac{24}{(\eta+2)^4}\cdot\frac{x^3}{3}$$

27+8 35

Da questa si ottiene de

$$|f(x) - g(x)| = \left|\frac{x^3}{3}\left(\frac{24}{(\eta+2)^4} - \frac{1}{(1+5)^3}\right)\right| \le \frac{|x|^3}{3}\left(\frac{24}{\left(-\frac{1}{2}+2\right)^4} + \frac{1}{\left(1-\frac{1}{2}\right)^3}\right)$$

sono decrescenti tra $-\frac{1}{2} \le x \le 2$

$$= \frac{|x|^3}{3}\left(\frac{24}{3^4}\cdot 2^4 + 2^3\right) = \frac{|x|^3}{3}\left(\frac{2^7}{3^3} + 2^3\right) =$$

$$= \frac{2^3}{3}\cdot|x|^3\left(\frac{8}{27} + 1\right) < \frac{8}{3}|x|^3\left(\frac{9}{27} + 1\right)$$

$$= \frac{8}{3}|x|^3\left(\frac{4}{3}\right) = \frac{32}{9}|x|^3$$

$$\frac{|f(x) - g(x)|}{|g(x)|} \le \frac{\frac{32}{9}|x|^3}{|\log(1+x)|} = \frac{32}{9}\frac{x^2}{\left|\frac{\log(1+x)}{x}\right|} \le \frac{32}{9}\cdot x^2 \le \frac{8}{9}\cdot\frac{1}{4} < 1$$

Infatti $\left|\frac{\log(1+x)}{x}\right| < 1$ nell'intervallo considerato

Quindi $|\varepsilon_{AV}| < \frac{8}{9} < 1$

Esaminiamo adesso il condizionamento del calcolo di $f(x)$

$$\varepsilon_{in} = c_x \cdot \varepsilon_x \quad \text{con} \quad c_x = \frac{x}{f(x)}f'(x) \quad \text{e } |\varepsilon_x| < u$$

Nel nostro caso abbiamo:

$$|C_x| = \frac{\cancel{x}}{\cancel{2x}}(\cancel{x+2}) \cdot \frac{\cancel{4}^{2}}{(x+2)^{\cancel{2}}} = \left|\frac{2}{x+2}\right|$$

poiché sull'intervallo $\left[-\frac{1}{2}, 0\right]$ $\left|\frac{2}{x+2}\right|$ è decrescente e limitata

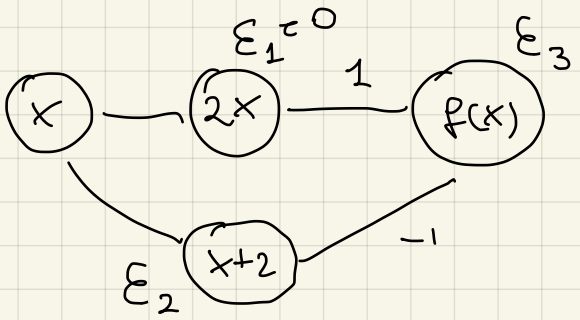
$$|C_x| < \frac{2}{-\frac{1}{2}+2} = \frac{2}{3/2} = \frac{4}{3}$$

quindi la formula è sempre ben condizionata dal momento che

$$|\varepsilon_{in}| \leq \frac{4}{3} \cdot |\varepsilon_x| \leq \frac{4}{3} u$$

Errore algoritmico:

Analizziamo l'errore algoritmico con il grafo:




Si osserva che $\varepsilon_1 = 0$ poiché la moltiplicazione e la divisione per 2 o multipli di 2 non introduce errore comportando solo una variazione nell'esponente della rappresentazione del numero in macchina.

$$\varepsilon_{alg} = \varepsilon_3 + 1 \cdot \underbrace{\varepsilon_1}_{=0} - \varepsilon_2 \Rightarrow |\varepsilon_{alg}| \leq 2u$$

**Esercizio** Si studi condizionamento e stabilità del calcolo di

$$f(x) = \sqrt{1-\sqrt{x}}, \quad x \geq 0$$

assumendo che $\sqrt{x}$ sia realizzata da una funzione di libreria che introduce un errore (analitico + algoritmico) limitato dalla precisione di macchina.

La funzione è definita per $x \geq 0$ e $1-\sqrt{x} \geq 0$ cioè $x \in [0,1]$

$$\varepsilon_{in} = c_x \varepsilon_x$$

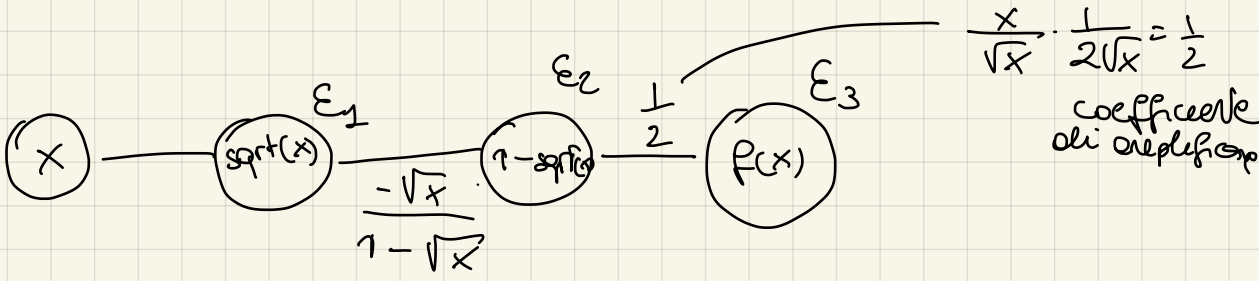
$$c_x = \frac{x}{\sqrt{1-\sqrt{x}}} \cdot \frac{1}{2\sqrt{1-\sqrt{x}}} \cdot \left(-\frac{1}{2}\frac{1}{\sqrt{x}}\right)$$

$$= \frac{1}{4}\frac{\sqrt{x}}{(1-\sqrt{x})} \qquad x \neq 1$$

$$|\varepsilon_{in}| = |c_x| \cdot |\varepsilon_x| = \frac{1}{4}\frac{\sqrt{x}}{(1-\sqrt{x})} \cdot |\varepsilon_x|$$

si nota che $|c_x| \xrightarrow[x \to 1^-]{} +\infty$ quindi per $x \to 1^-$ il problema è malcondizionato.

Analizziamo la stabilità: il fatto che $\sqrt{x}$ è realizzata da una funzione di libreria ci permette di utilizzarla come una qualsiasi altra operazione che genera un errore dell'ordine di $u$




$$\varepsilon_{oeg} = \varepsilon_3 + \frac{1}{2}\left(\varepsilon_2 - \frac{\sqrt{x}}{1-\sqrt{x}}\,\varepsilon_1\right)$$

$$|\varepsilon_{oeg}| = \left|\varepsilon_3 + \frac{1}{2}\varepsilon_2 - \frac{1}{2}\frac{\sqrt{x}}{1-\sqrt{x}}\,\varepsilon_1\right| \le |\varepsilon_3| + \frac{1}{2}|\varepsilon_2| + \frac{1}{2}\left|\frac{\sqrt{x}}{1-\sqrt{x}}\right||\varepsilon_1|$$

$$\le u\left(\frac{3}{2} + \frac{1}{2}\left|\frac{\sqrt{x}}{1-\sqrt{x}}\right|\right)$$

si osserva che l'algoritmo è instabile per $x \to 1^-$

perché $\lim_{x\to 1^-} \frac{\sqrt{x}}{1-\sqrt{x}} = +\infty$