

Esempio 1.2 Definito l'insieme dei numeri finiti $\mathbb{F}(10, 5, -50, 49)$ resta definito il numero di posizioni (bit nel caso di base 2) necessarie per rappresentare in memoria un numero finito dell'insieme. Nel caso specifico saranno 1 per il segno (0 se positivo e 1 se negativo), 2 per l'esponente (si usa la tecnica di memorizzazione per traslazione, cioè nei due campi per l'esponente si memorizzano i valori da 00 a 99 intendendo gli esponenti da -50 a 49) e 5 posizioni per la mantissa (si memorizza a partire da sinistra). Vediamo qualche esempio numerico:

$$\alpha = 0.1039 \times 10^{-6} \quad 04410390$$

dove 0 indica che il numero è positivo, 44 rappresenta l'esponente -6, quindi la mantissa 10390 arrotondata alla quinta cifra.

| | |
|------------------------|----------|
| $\alpha = 0.05302$ | 04953020 |
| $\alpha = -237.141$ | 95323714 |
| $\alpha = -0.00321665$ | 94832167 |

RISULTA EVIDENTE CHE L'INSIEME DEI NUMERI FINITI RAPPRESENTA SOLO UN RISTRETTO SOTTOINSIEME DI QUELLO DEI NUMERI REALI. LA MAGGIORPARTE DEI VALORI $d \in \mathbb{R}$ RISULTA $\notin \mathbb{F}(B, t, \lambda, w)$, QUINDI TALI VALORI POSSANO ESSERE SOLAMENTE APPROSSIMATI MEDIANTE UN $\tilde{d} \in \mathbb{F}(B, t, \lambda, w)$, COMMETTENDO UN CERTO ERRORE DI RAPPRESENTAZIONE.

Per valutare l'entità si definiscono le seguenti quantità:

$$E_{\text{ABS}} = |\tilde{d} - d| \quad \text{ERRORE ASSOLUTO} \quad \text{e} \quad E_{\text{REL}} = \left| \frac{\tilde{d} - d}{d} \right| \quad \text{SE } d \neq 0 \quad \text{ERRORE RELATIVO}$$

La rappresentazione discreta della retta reale descritta in precedenza:



È TALE CHE FORNISCE UN ERRORE RELATIVO DI RAPPRESENTAZIONE MASSIMO COSTANTE PER OGNI d , MENTRE QUELLO ASSOLUTO, DI CONSEGUENZA, AUMENTA PROPORTIONALMENTE AL VALORE DI d .

SULLA RETTA REALE (CONTINUA), I NUMERI RAPPRESENTABILI SONO PUNTI DISCRETI (ISOLATI). Lo spazio tra questi punti (il buco sulla retta) cresce proporzionalmente al valore di d :

- INTORNO A 1: |---|---| (BUCHI PICCOLI = 0.1)
 - INTORNO A 100: |---|---|---| (BUCHI GRANDI = 10)
 - INTORNO A 10000: |-----|-----| (BUCHI ENORMI = 10000)
- $\left. \begin{array}{l} \text{1) ERRORE MASSIMO POSSIBILE = METÀ' DELLA DISTANZA TRA DUE N° CONSECUTIVI} \\ \text{2) DISTANZA CRESCHE = ERRORE ASSOLUTO CRESCE} \\ \text{3) HA LA DISTANZA È SEMPRE UNA FRAZIONE COSTANTE DI } d = \text{ERRORE RELATIVO COSTANTE} \end{array} \right\}$

ESEMPIO: Sia $d = 1.234567$, i numeri rappresentabili vicini sono $\frac{1.2}{1.234567} - \frac{1.3}{1.234567}$

↳ Approssima a $\tilde{d} = 1.2$

$$\hookrightarrow E_{\text{ABS}} = |1.234567 - 1.2| = 0.034567$$

$$|1.2 - 1.234567|$$

$$E_{\text{REL}} = \frac{0.034567}{1.234567} = 2.8\%$$

ESEMPIO: Sia $d = 12345.67$, i numeri rappresentabili vicini sono $\frac{12000}{12345.67} - \frac{12345.67}{12345.67}$

↳ Approssima a $\tilde{d} = 1200$

$$\hookrightarrow E_{\text{ABS}} = |12345.67 - 12000| = 345.67$$

$$|1200 - 12345.67|$$

$$E_{\text{REL}} = \frac{345.67}{12345.67} = 2.8\%$$

NOTA CHE t È FONDAMENTALE, ESSO DETERMINA LA DENSITÀ DEI PUNTI SULLA RETTA, SEGUO LA RELAZIONE: " t PIÙ GRANDE = PIÙ NUMERI RAPPRESENTABILI = BUCHI PIÙ PICCOLI = MASSIORE PRECISIONE"; t DETERMINA LA RISOLUZIONE / PRECISIONE DELLA TUA RAPPRESENTAZIONE SULLA RETTA.

NEL CALCOLO SCIENTIFICO, DOVE LA RISPOSTA AI PROBLEMI POSSONO VARIARE GRANDEMENTE IN VALORE, SOLITAMENTE SI USA L'ERRORE RELATIVO IN QUANTO È SCAFFING VARIANT (L'ERRORE RELATIVO NON CAMBIA SE MULPIPLICHI IL NUMERO PER UNA COSTANTE s), INFATTI PER $d \rightarrow s \cdot d$ E $\tilde{d} \rightarrow s \cdot \tilde{d}$, E_{REL} RESTA UGUALE.

QUESTO RENDE L'ERRORE RELATIVO LA MISURA GIUSTA PER IL CALCOLO SCIENTIFICO, PERCHE' È INDEPENDENTE DALL'ORDINE DI GRANDEZZA DEI NUMERI (AL CONTRARIO DELL'ERRORE ASSOLUTO).

Th. LIMITI DEGLI ERROTI ASSOLUTI

Per ogni $d \in \mathbb{R}$ risulta: $|fl_T(d) - d| < \beta^{p-t}$, $|fl_A(d) - d| \leq \frac{1}{2}\beta^{p-t}$ dove il segno di ugualanza vale solo se $d_{t+2} = \beta_{1/2}$ e $d_{t+i} = 0$ per $i \geq 2$.

Sulla retta β^{p-t} indica la distanza tra due numeri finiti consecutivi.

Troncamento: errore $< \beta^{p-t}$

↳ Tronchi sempre per difetto

↳ L'errore può essere quasi quanto la distanza tra due numeri consecutivi (ma mai uguale)

↳ Nel caso peggiore però quasi una cifra intera

Arrotondamento: errore $\leq \frac{1}{2}\beta^{p-t}$

↳ Arrotondi al più vicino

↳ L'errore è massimo metà della distanza tra due numeri consecutivi

↳ Due volte più preciso del troncamento

ESEMPIO NUMERICO ($p=10, t=2$): Supponiamo di avere $d = 1.299 \rightarrow d = 0.1299 \cdot 10^3$ con $p=1$; distanza tra

numeri consecutivi: $\beta^{p-t} = 10^{-2} = 0.1$

↳ Con Troncamento: $x = 1.2, y = 1.3 \quad fl_T(1.299) = 1.2$

↳ Errore $= |1.299 - 1.2| = 0.099$, limite teorema: < 0.1 (rispettato)

↳ Con Arrotondamento: $fl_A(1.299) = 1.3$ (più vicino)

↳ Errore $= |1.299 - 1.3| = 0.001$, limite teorema: ≤ 0.05 (rispettato, molto meglio)

$|fl_A(d) - d| \leq \frac{1}{2}\beta^{p-t}$ { l'ugualanza vale solo se $d_{t+2} = \beta_{1/2}$ (esattamente sul punto medio) e $d_{t+i} = 0$ per $i \geq 2$ (tutte le altre cifre sono zero)}

ESEMPIO: $d = \frac{x+y}{2}$ esattamente, $x = 1.2, y = 1.3, d = 1.25$ (punto medio perfetto)

$fl_A(d)$ può essere 1.2 o 1.3; Errore $= 0.05 = \frac{1}{2}\beta^{p-t}$ (ugualanza). Se $d = 1.251$ (non più sul punto medio, errore $= 0.049 < 0.05$ (disugualanza stretta))

Dim. Siano x ed y i due numeri consecutivi tali che $x \leq d < y$ riprendendo l'osservazione:

OSSERVAZIONE (IL SANDWICH): Siano x, y due numeri finiti consecutivi tali che $x \leq d < y$ allora

$$x = \left(\sum_{i=1}^t d_i \beta^{p-i} \right) \beta^p, \quad y = \left(\sum_{i=1}^t d_i \beta^{p-i} + \beta^{-t} \right) \beta^p \quad \text{e risulta:}$$

$$fl_T(d) = x; \quad fl_A(d) = \begin{cases} x & \text{se } d < \frac{x+y}{2} \\ y & \text{se } d \geq \frac{x+y}{2} \end{cases}$$

Troncamento è sempre x

$$\text{per } fl_T(d) = x; \quad fl_A(d) = \begin{cases} x & \text{se } d < \frac{x+y}{2} \\ y & \text{se } d \geq \frac{x+y}{2} \end{cases}$$



"Distanza tra due numeri consecutivi"

$$\text{sarà } d - fl_T(d) < y - x = \beta^{p-t}; \\ d - x < y - x = \beta^{p-t};$$

$d < y$ (per ipotesi: d è tra x e y), quindi $d - x < y - x = \beta^{p-t}$, conclusione:

$$|fl_T(d) - d| = d - fl_T(d) = d - x < \beta^{p-t}.$$

ancora sarà $|fl_A(d) - d| \leq \frac{y-x}{2} = \frac{1}{2}\beta^{p-t}$ e l'ugualanza vale solo se $d = \frac{x+y}{2}$, cioè $d_{t+2} = \beta_{1/2}$ e $d_{t+i} = 0$ per $i \geq 2$.

DEFINIZIONE: Dato l'insieme di n° finiti $F(\beta, t, \lambda, w)$, si dice UNITÀ DI ARROTONDAMENTO E LA SI INDICA CON u , la quantità:

$$u = \begin{cases} \beta^{t-t} \text{ PER TRONCAMENTO} \\ \frac{1}{2}\beta^{t-t} \text{ PER ARROTONDAMENTO} \end{cases}$$

L'UNITÀ DI ARROTONDAMENTO È UNA MISURA DELLA PRECISIONE MASSIMA DEL SISTEMA FLOATING-POINT.
È L'ERRORE RELATIVO MASSIMO CHE PUOI COMMETTERE RAPPRESENTANNO UN NUMERO.

PENCHE' β^{t-t} ? LA DEFINIZIONE DI u CERCA IL CASO PEGGIORE POSSIBILE. L'ERRORE RELATIVO PER UN NUMERO A SPECIFICO È: ERRORE RELATIVO = $\frac{\text{ERRORE ASSOLUTO}}{|\delta(\alpha) - \alpha|}$, SAPPIAMO CHE $|\delta(\alpha) - \alpha| \leq \beta^p$ (PER ARROTONDAMENTO CON $\frac{1}{2}\beta^{t-t}$) DOVE p È L'ESPOLENTE DI α . VOLIAMO TROVARE IL MASSIMO POSSIBILE DI E_{REL} :

$E_{\text{REL}} = \frac{\beta^{p-t}}{|\alpha|}$, PER MASSIMIZZARE QUESTA FRAZIONE: NUMERATORE β^{p-t} È FISSO (DATA DA p), MENTRE IL DENOMINATORE $|\alpha|$ DEVE ESSERE MINIMO. QUALE' IL MINIMO $|\alpha|$ CON ESPOLENTE p : PER UN NUMERO NORMALIZZATO CON ESPOLENTE p : $\alpha = (0.d_1 d_2 d_3 \dots) \cdot \beta^p$ DOVE $d_1 \geq 1$. IL MINIMO È QUANDO $d_1 = 1$ E TUTTE LE ALTRE CIFRE SONO 0: $|\alpha|_{\text{MIN}} = 0.1000 \dots \cdot \beta^p = \beta^{-1} \cdot \beta^p = \beta^{p-1}$

Quindi il caso peggiore: $E_{\text{REL MAX}} = \frac{\beta^{p-t}}{\beta^{p-1}} = \beta^{((p-t)-(p-1))} = \beta^{(p-t-p+1)} = \beta^{(t-1)} = \beta^{t-t}$

Th. $\forall \alpha \in \mathbb{R}$ E $\alpha \neq 0$ VALE: $\left| \frac{\delta(\alpha) - \alpha}{\alpha} \right| < u$ | $u = \begin{cases} \beta^{t-t} \text{ PER TRONCAMENTO} \\ \frac{1}{2}\beta^{t-t} \text{ PER ARROTONDAMENTO} \end{cases}$
(QUANTO SBAGLIAMO RISPETTO AL VALORE ESATTO)

Dim. Caso di troncamento:
"LIMITE INFERIORE SEMPLICE"
 $\text{FORMA PIU' SEMPLICE}$ \rightarrow
 $|\alpha| = (d_1 \beta^{-1} + d_2 \beta^{-2} + \dots) \beta^p \geq d_1 \beta^{-1} \beta^p \geq \beta^{p-1}$ E QUINDI

$$\frac{\text{ERRORE ASSOLUTO}}{|\alpha|} < \frac{\beta^{p-t}}{\beta^{p-1}} = \beta^{t-t}$$

SAPPIAMO CHE $|\alpha| \geq \beta^{p-1}$, QUINDI $\frac{1}{|\alpha|} \leq \frac{1}{\beta^{p-1}}$ È IL LIMITE SUPERIORE. PASSIAMO DAL LIMITE INFERIORE AL LIMITE SUPERIORE.

NON POSSO CALCOLARE L'ERRORE ESATTO OGNI VOLTA, SAREBBERE TROPPO COSTOSO. VOLGO SAPERE "QUAL'È IL MASSIMO ERRORE CHE POSSO ASPETTARMI?"

IL TEOREMA DICE: "QUALUNQUE SIA IL NUMERO CHE RAPPRESENTA IN FLOATING-POINT, L'ERRORE RELATIVO SARÀ SEMPRE MINORE DI $u = \beta^{t-t}$ ".

IN SINTESI, MASSIMIZZARE L'ERRORE MUOGLIO DIRE TROVARE UNA GARANZIA SUL MASSIMO ERRORE POSSIBILE. È COME DIRE "NEL PEGGIOR CASO, SBASLI AL MASSIMO DI QUESTA QUANTITÀ". QUESTO CI DA' SICUREZZA NEI CALCOLI NUMERICI.

$$\left| \frac{\alpha - \delta(\alpha)}{\alpha} \right| < \frac{\beta^{t-t}}{\beta^{p-1}} = \beta^{t-t}$$

BASE DEL SISTEMA NUMERICO

$t = \text{NUMERO DI CIFRE DI MANTISSA}$

$p = \text{POSIZIONE DELL'ESPOLENTE}$

$\left. \begin{array}{l} \text{l'ERRORE RELATIVO È LIMITATO SUPERIORMENTE DA } \beta^{t-t}, \text{ O} \\ \text{ANCORA PIU' SEMPLICEMENTE, È MINORE DI } \beta^{t-t}. \end{array} \right\}$

1) PRENDIAMO UN N° REALE α LO RAPPRESENTAMO IN $\delta(\alpha)$ (CALCOLIAMO E_{REL}) 2) Th. DICE CHE E_{REL} È LIMITATO A β^{t-t}

DIM. CASO DI ARROTONDAMENTO:

$$\left| \frac{d - f_l(d)}{d} \right| < \frac{1}{2} \frac{\beta^{p-t}}{\beta^{p-1}} = \frac{1}{2} \beta^{t-p}$$

"QUALUNQUE SIA IL NUMERO CHE RAPPRESENTA IN FLOATING-POINT,
L'ERRORE RELATIVO SARÀ SEMPRE MINORE DI $u = \frac{1}{2} \beta^{t-p}$ ".

NELLA CATENA DI MAGGIORANZA: $|d| = (d_1 \beta^{-1} + d_2 \beta^{-2} + \dots) \beta^p \geq d_1 \beta^{-1} \beta^p \geq \beta^{p-1}$, L'UQUALIANZA SI POTREBBE AVERE S.S.S. $d_{t+1} = \beta/2$ E $d_{t+i} = 0 \quad \forall i \geq 2$, MA IN TAL CASO CORRISPONDEREBBE:

$d \geq (d_1 \beta^{-1} + d_2 \beta^{-2} + \dots) \beta^p > d_1 \beta^{-1} \beta^p \geq \beta^{p-1}$, QUESTO DICE CHE IL LIMITE INFERIORE È "ROBUSTO" ANCHE SENZA IL VALORE ASSOLUTO.

L'ERRORE RELATIVO CHE SI COMMETTE NEL RAPPRESENTARE IL NUMERO REALE d CON UN NUMERO FINITO $f_l(d)$ LO INDICHIAMO CON ϵ E LO CHIAMEREMO "ERRORE RELATIVO DI RAPPRESENTAZIONE":

$$\epsilon = \frac{|f_l(d) - d|}{d}$$

IL TH. DI PRIMA DICE CHE L'UNITÀ DI ARROTONDAMENTO LIMITA SUPERIORMENTE IL MODULO DELL'ERRORE RELATIVO DI RAPPRESENTAZIONE.

* HOGO ALTERNATIVO E ELEGANTE DI ESPRIMERE L'ERRORE DI RAPPRESENTAZIONE IN FLOATING-POINT!

* (COROLARIO): $\forall d \in \mathbb{R}, d \neq 0$ VALE $f_l(d) = d(1+\epsilon)$, CON $|\epsilon| < u$ ($f_l(d) \approx d$ CON PERTURBAZIONE RELATIVA PICCOLA)

DIM.

1) $\left| \frac{f_l(d) - d}{d} \right| < u \Rightarrow \left| \frac{f_l(d) - \frac{d}{1+\epsilon} d}{d} \right| < u \Rightarrow \left| \frac{f_l(d)}{d} - 1 \right| < u$

$\epsilon = \frac{f_l(d) - d}{d}$ } ERRORE RELATIVO
RISPETTO AD d

2) $\epsilon = \frac{f_l(d)}{d} - 1 \quad |\epsilon| < u \Rightarrow \frac{f_l(d)}{d} = 1 + \epsilon \Rightarrow \frac{f_l(d)}{d} = 1 + \epsilon \Rightarrow f_l(d) = d(1 + \epsilon)$

ANALOGAMENTE AL TH. DI PRIMA, SE $f_l(d) \neq 0$ VALE
DEFINENDO $\epsilon = \frac{d - f_l(d)}{f_l(d)}$ VALE $f_l(d) = \frac{d}{1 + \epsilon}$ CON $|\epsilon| < u$.

ESEMPIO: PER $f_l(d) = d(1 + \epsilon)$, SE VOGLIO SOMMARE

DUE NUMERI x E y CON IL COMPUTER:

1) PREndo x ESATTO E LO RAPPRESENTO: $f_l(x) = x(1 + \epsilon_x)$

2) PREndo y ESATTO E LO RAPPRESENTO: $f_l(y) = y(1 + \epsilon_y)$

SI FA LA SOMMA IN FLOATING-POINT: $f_l(x+y) = (x+y)(1 + \epsilon_z)$

ESEMPIO: PER $f_l(d) = \frac{d}{1 + \epsilon}$, SUPPONIAMO DI AVERE GIA' CALCOLATO DAL COMPUTER E VUOI CAPIRE QUAL'ERA IL VALORE ESATTO:

1) IL COMPUTER TI DA $f_l(d) = 3.14159$

2) TU SAI CHE $f_l(d) = \frac{d}{1 + \epsilon}$ CON $|\epsilon| < 10^{-6}$

3) QUINDI $d = f_l(d) \cdot (1 + \epsilon) \approx 3.14159 \cdot (\text{QUALSOA DI VICINO A } 1)$

$\left| \frac{f_l(d) - d}{f_l(d)} \right| < u$, DA QUESTO SI HA POI CHE

$\epsilon = \frac{f_l(d) - d}{f_l(d)}$ } ERRORE RELATIVO
RISPETTO AD $f_l(d)$

PARTO DAI VALORI ESATTI E MOLTIPLICO PER $(1 + \epsilon)$
PER OTTENERE LA RAPPRESENTAZIONE.

} PARTO DAL VALORE RAPPRESENTATO $f_l(d)$ E LO USO PER STIMARE L'ORIGINALE d .

• $d(1 + \epsilon) =$ VAI DA ESATTO \rightarrow RAPPRESENTATO (ANALISI SIRETTA)

• $\frac{d}{(1 + \epsilon)} =$ VAI DA RAPPRESENTATO \rightarrow ESATTO (ANALISI INVERSA)

ARITMETICA FLOATING-POINT: OLTRE ALL'ERRORE INTRODOTTO NELLA RAPPRESENTAZIONE DI UN NUMERO REALE, ANCHE LE SINGOLE OPERAZIONI RISULTANO APPROSSIMATE. AD ESEMPIO, LA SOMMA DI DUE NUMERI FINITI x ED y CHE E $\tilde{F}(p,t,\lambda,w)$ E' UN NUMERO CHE PUO' NON APPARTENERE ALL'INSIEME $F(p,t,\lambda,w)$.

ES: SIANO $0.1 \cdot 10^{-3}$ E $0.1 \cdot 10^3$ E $\tilde{F}(10,3,2,w)$. ESEGUENDO LA SOMMA SI HA:

$$\frac{100.0000 + }{0.0001} = \left\{ \begin{array}{l} \text{MA, } 100.0001 = 0.1000001 \cdot 10^3 \in F \\ 7 \text{ CIFRE PER LA MANTISSA. QUESTO PRESENTA IL PROBLEMA DI APPROSSIMARE IL RISULTATO} \end{array} \right.$$

DI UN'OPERAZIONE ARITMETICA FRA DUE NUMERI FINITI CON UN NUMERO FINITO.

OCCORRE DEFINIRE QUINDI UN'ARITMETICA DI MACCHINA. PER CONVENIENZA SI ASSUME CHE: SE $\tilde{\text{OP}}$ E' L'OPERAZIONE DI MACCHINA CHE APPROSSIMA L'OPERAZIONE ESATTA OP , PER TUTTI I NUMERI FINITI x, y PER CUI L'OPERAZIONE NON DIA LUOGO A CONDIZIONI DI OVERFLOW OPPURE UNDERFLOW, SIA: $x \tilde{\text{OP}} y = f(x \text{OP} y)$.

QUESTA CONVENTIONE E' IL **Th.** DI PRIMA, COMPORTANO CHE, SE $x \tilde{\text{OP}} y \neq 0$, VALE: (ERRORE RELATIVO OPERAZIONE MACCHINA)

$$\left| \frac{x \tilde{\text{OP}} y - (x \text{OP} y)}{x \text{OP} y} \right| < u$$

| #

NOTA: IL COMPUTER QUANDO ESEGUE UN'OPERAZIONE, PRIMA CALCOLA IL RISULTATO ESATTO $x \text{OP} y$, Poi LO RAPPRESENTA IN $f(x \text{OP} y)$

Ogni singola operazione introduce un errore relativo minore di u . Questo e' il "MATTONCINO FONDAMENTALE" PER CAPOIRE COME GLI ERRORI SI PROPAGANO IN UN ALGORITMO COMPLESSO (ERRORE MASSIMO DI u PER OGNI OPERAZIONE).

ERRORE RELATIVO DELL'OPERAZIONE MACCHINA: $E = \frac{x \tilde{\text{OP}} y - (x \text{OP} y)}{x \text{OP} y}$

Per il corollario possiamo scrivere: $x \tilde{\text{OP}} y = (x \text{OP} y)(1+E)$, $|E| < u$ con E l'errore relativo commesso nell'operazione. Il risultato di un'operazione macchina deve essere uguale all'approssimazione con un numero finito del risultato dell'operazione esatta. \rightarrow Piccolo fattore introdotto per ogni operazione, e questi fattori si moltiplicano man mano che esegui piu' operazioni.

L'UNITA DI ARROTONDAMENTO u E' DETTA PRECISIONE DI MACCHINA NEL SISTEMA FLOATING-POINT.

Questo valore rappresenta un limite fondamentale sulla precisione del sistema: indica il massimo errore relativo che puo' verificarsi quando rappresentiamo un numero reale in formato floating-point. Con il termine "PRECISIONE MACCHINA" si indicano in realta' due aspetti distinti ma collegati:

• **PRECISIONE DI RAPPRESENTAZIONE**: DESCRITA DA Th. DI PRIMA, CHE QUANTIFICA L'ERRORE COMMESO NELL'RAPPRESENTARE UN SINGOLO NUMERO IN MEMORIA; QUALUNQUE NUMERO x VIENE APPROSSIMATO CON UN ERRORE RELATIVO MINORE DI u .

• **PRECISIONE DI CALCOLO**: RIGUARDA L'ERRORE INTRODOTTO DALLE OPERAZIONI ARITMETICHE; OGNI OPERAZIONE DI MACCHINA INTRODUCE UN ULTERIORE ERRORE RELATIVO LIMITATO DA u . L'UNITA DI ARROTONDAMENTO u E' LA "RISOLUZIONE MINIMA" DEL SISTEMA: E' IMPOSSIBILE DISTINGUERE DUE NUMERI CHE DIFFERISCONO PER MEZO DI u IN TERMINI RELATIVI. Es: $u = 10^{-6}$, il sistema non puo' distinguere 1.000000 da 1.000001 (differenza relativa uguale ad u)

L'UNITA DI ARROTONDAMENTO u E' IL PIU' PICCOLO NUMERO FINITO POSITIVO t.c.: $u + 1 = f(u + 1) > 1$ QUESTO IMPLICA CHE OGNI NUMERO FINITO $v < u$ SARÀ $v + 1 = f(v + 1) = v$.

ANALISI DEGLI ERROTI : ANALISI DEGLI ERROTI CHE SI POSSONO AVERE NEL RISOLVERE UN PROBLEMA BEN POSTO, PRINCIPALMENTE ERROTI CHE NASCONO NELL'ESECUZIONE DI UNA SEQUENZA DI OPERAZIONI (PER RISOLVERE UN PROBLEMA BEN POSTO) NELLA QUALE AL OGNI PASSO VENGONO UTILIZZATI COME INPUT VALORI CALCOLATI PRECEDENTEMENTE. IN QUESTO CASO SI VERIFICA UNA PROPAGAZIONE DEGLI ERROTI. IL VERO PROBLEMA E' QUINDI LA SEQUENZA DI OPERAZIONI. BANCAMENTE, IN UN ALGORITMO TIPICO, OGNI PASSO USA COME INPUT I RISULTATI DEI PASSI PRECEDENTI CHE PERO' CONTENGONO GIÀ ERROTI, E QUINDI GLI ERROTI SI PROPAGANO E SI ACCUMULANO ATTRAVERSO LA SEQUENZA DI CALCOLI. NELLA PROPAGAZIONE DEGLI ERROTI E' MOLTO IMPORTANTE CONSIDERARE QUESTO TIPO DI ERRORE PERCHE' PUO' RISULTARE PIÙ CHE SIGNIFICATIVO / IMPORTANTE.

UN PROBLEMA E' BEN POSTO SE SODDISFA: (MATEMATICAMENTE CORRETTO E STABILE)

- **ESISTENZA DELLA SOLUZIONE**: IL PROBLEMA HA ALMENO UNA SOLUZIONE

- **UNICITA' DELLA SOLUZIONE**: IL PROBLEMA HA AL MASSIMO UNA SOLUZIONE (QUINDI ESATTAMENTE UNA)

- **STABILITA'** (O DIPENDENZA CONTINUA DAI DATI): PICCOLE PERTURBAZIONI NEI DATI DI INPUT PRODUcono PICCOLE PERTURBAZIONI NEI DATI DI OUTPUT. PERTURBARE VOGLIO DIRRE MODIFICARE LEGGERMENTE UN DATO, INTRODUCENDO UNA PICCOLA VARIAZIONE RISPETTO AL VALORE ORIGINALE.

ES: DATO ORIGINALE $\rightarrow x = 5.0$, PERTURBATO $\rightarrow \tilde{x} = 5.0001$, PERTURBAZIONE $\rightarrow \delta x = 0.0001$

LA GESTIONE DELLA PROPAGAZIONE DELL'ERRORE E' FONDAMENTALE NELLE APPLICATION INFORMATICHE CHE IMPLEMENTANO UN METODO DI CALCOLO, AL FINE DI DETERMINARE L'ATTENSIBILITA' (O AFFIDABILITA') DEI RISULTATI OTTENUTI.

PRECISIONE E ACCURATEZZA (STRUMENTI BASE DELLA TEORIA DELLA PROPAGAZIONE DELL'ERRORE):

IL TERMINE "ACCURATEZZA" SI RIFERISCE ALL'ERRORE ASSOLUTO O RELATIVO DI UNA QUANTITA' APPROSSIMATA (QUANTO SIAMO VICINI AL VALORE VERO), MENTRE INVECE "PRECISIONE" E' L'ACCURATEZZA CON LA QUALE LE OPERAZIONI ARITMETICHE $+, -, \cdot, /$, VENGONO EFFETTUATE (QUANTO SONO ACCURATE LE SINGOLE OPERAZIONI), E NELL'ARITMETICA FLOATING-POINT QUESTA VIENE MISURATA DALL'UNITA' DI ARROTONDAMENTO " u " (SEMPRE LIMITATA DA u).

Usando le formule

(\rightarrow STIMA DELL'ERRORE PRIMA DEL CALCOLO, "Se uso questo algoritmo, quanto posso aspettarmi di sbagliare?" ANALISI IN AVANTI DELL'ERRORE: PER VALUTARE L'ENTITA' (IMPORTANZA) DELLA PROPAGAZIONE DEGLI ERROTI IN UNA SEQUENZA DI OPERAZIONI, SI CONSIDERA L'ERRORE TOTALE DATO DALLA SOMMA DEI SINGOLI ERROTI. UN PROBLEMA PUO' ESSERE VISTO COME UNA FUNZIONE f CHE A PARTIRE DA UN DATO x PRODUCE UN RISULTATO $\hat{f}(x)$ ($f: x \rightarrow \hat{f}(x)$).

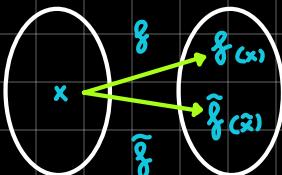
Per ogni dato ed operazione si introduce un errore che rappresenta l'approssimazione introdotta nella rappresentazione e dall'aritmetica finita (\mathbb{Q} , \mathbb{R}) e si ricava un corrispondente risultato approssimato $\tilde{f}(x)$. Quindi:

$$\begin{array}{l} x \rightarrow f \rightarrow \hat{f}(x) \\ \downarrow \quad \downarrow \quad \downarrow \\ \tilde{x} \rightarrow \tilde{f} \rightarrow \tilde{\hat{f}}(\tilde{x}) \\ \downarrow \quad \downarrow \\ \text{DATO APPROSS. OP. APPROSS.} \quad \text{RIS. APPROSS.} \end{array}$$

1) Errore di rappresentazione: $\hat{f}(x) = x(1+\varepsilon)$, $|\varepsilon| < u$
 2) Errore nelle operazioni: $x \hat{o} y = (x \hat{o} y)(1+\varepsilon)$, $|\varepsilon| < u$

La quantita': $\left| \frac{\hat{f}(x) - \tilde{\hat{f}}(\tilde{x})}{\hat{f}(x)} \right|$ fornisce un'indicazione sull'entita' dell'errore relativo totale che affligge la soluzione del problema f con dato x .

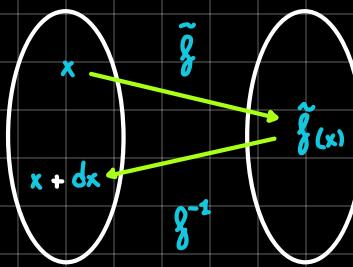
ANALISI IN AVANTI DELL'ERRORE:



Analisi all'indietro degli errori: approccio opposto al precedente, considero il risultato finale $\tilde{f}(x)$ come risultato esatto derivato da dati iniziali perturbati rispetto a quelli reali. La valutazione dell'entità dell'errore è data quindi da un fattore dx sul dato iniziale x t.c. $\tilde{f}(x+dx) = \tilde{f}(x)$. "Ho un dato x , faccio calcoli approssimati, quanto sbaglio nel risultato?"

Analisi Avanti: esatto
 $x \rightarrow \tilde{f} \rightarrow \tilde{f}(x)$ approssimato
 errore = $|\tilde{f}(x) - f(x)|$

Analisi Indietro: perturbato
 $x + dx \rightarrow \tilde{f} \rightarrow f(x+dx) = \tilde{f}(x)$
 Quanto devo perturbare x ?



Ese: Si consideri l'addizione di due numeri finiti x ed y , avremo:

$$\underbrace{\tilde{f}(x,y)}_{\text{PROBLEMA POSTO: } \tilde{f} \text{ FA LA SOMMA TRA } x, y} = f(x+y) = (x+y)(1+\varepsilon) = x(1+\varepsilon) + y(1+\varepsilon) \quad \text{QUINDI: } \tilde{f}(x,y) = f(x+dx, y+dy)$$

Dove: $dx = x\varepsilon$, $dy = y\varepsilon$ che indica $\tilde{f}(x,y)$ come il risultato esatto a partire da due dati $x+dx$ e $y+dy$ che rappresentano le perturbazioni dei dati iniziali.

ESEMPIO: Sia $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ (sono numeri reali): (**MOLTIPLICAZIONE**)

$$x, y \rightarrow x \cdot y$$

$$x \rightarrow \tilde{x} \equiv f(x) = x(1+\varepsilon_1) \text{ con } |\varepsilon_1| < u$$

$$y \rightarrow \tilde{y} \equiv f(y) = y(1+\varepsilon_2) \text{ con } |\varepsilon_2| < u$$

$$\tilde{x} \cdot \tilde{y} = f(\tilde{x} \cdot \tilde{y}) = (\tilde{x} \cdot \tilde{y})(1+\varepsilon_3) = (x \cdot (1+\varepsilon_1) \cdot y \cdot (1+\varepsilon_2))(1+\varepsilon_3) \text{ con } |\varepsilon_3| < u$$

* (**ERRORE NELLE OPERAZIONI MACCINA - ANALISI AVANTI**)

QUESTI PASSACCI MOSTRANO QUELLO CHE VUOLE FARE IL NOSTRO CALCOLATORE IN ARITMETICA FINITA.

$$\begin{aligned} \varepsilon_{\text{rel}} &= \left| \frac{x \cdot y - (x \cdot (1+\varepsilon_1) \cdot y \cdot (1+\varepsilon_2))(1+\varepsilon_3)}{x \cdot y} \right| = \left| \frac{x \cdot y - (x \cdot (1+\varepsilon_1) \cdot y \cdot (1+\varepsilon_2))(1+\varepsilon_3)}{x \cdot y} \right| \\ &= \left| 1 - (1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3) \right| = \left| 1 - (1 + \underbrace{\varepsilon_1}_{\text{PIU' PICCOLO DI}}, \underbrace{\varepsilon_2}_{\text{PIU' PICCOLO DI}}, \underbrace{\varepsilon_3}_{\text{PIU' PICCOLO DI}} + \underbrace{\varepsilon_1 \varepsilon_2}_{\text{PIU' PICCOLO DI}}, \underbrace{\varepsilon_2 \varepsilon_3}_{\text{PIU' PICCOLO DI}}, \underbrace{\varepsilon_1 \varepsilon_3}_{\text{PIU' PICCOLO DI}}, \underbrace{\varepsilon_1 \varepsilon_2 \varepsilon_3}_{\text{PIU' PICCOLO DI}}) \right| \\ &\approx \left| \varepsilon_1 + \varepsilon_2 + \varepsilon_3 \right| \leq \left| \varepsilon_1 \right| + \left| \varepsilon_2 \right| + \left| \varepsilon_3 \right| < 3u \end{aligned}$$

* Probabilità di piccole quantità sono

quantità ancora più piccole

DISEGUALANZA TRIANGOLARE: la somma in valore assoluto è al

massimo la somma dei valori assoluti.

MASSIMIZZAZIONE DI OGNI TERMINE: $|\varepsilon_i| < u$, $|\varepsilon_1| + |\varepsilon_2| + |\varepsilon_3| < 3u$

CONCLUSIONE: $|\varepsilon_1 + \varepsilon_2 + \varepsilon_3| < 3u$, l'errore cresce al massimo linearmente con il numero di operazioni.

STIMA PESSIMISTICA

GENERALIZZAZIONE: Per n errori $\rightarrow |\varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_n| \leq |\varepsilon_1| + |\varepsilon_2| + \dots + |\varepsilon_n| < \overbrace{n \cdot u}^{\uparrow}$

ANALISI DEGLI ERRORE NELLA MOLTIPLICAZIONE: VOGLIAMO CALCOLARE $c = a \cdot b$, MA OTTEMIAMO $\tilde{c} = a \cdot \tilde{b} = (a \cdot b) \cdot (1 + \varepsilon)$, $|\varepsilon| < u$.

ANALISI INSIEME: IL RISULTATO APPROSSIMATO \tilde{c} E' IL RISULTATO ESATTO DI QUALE PROBLEMA PERTURBATO?

$$\tilde{c} = (a \cdot b) \cdot (1 + \varepsilon) = a(1 + \varepsilon) \cdot b = \tilde{a} \cdot b \text{ OPPURE } \tilde{c} = a \cdot b(1 + \varepsilon) = a \cdot \tilde{b}$$

INTERPRETAZIONE: \tilde{c} E' IL PRODOTTO ESATTO DI a E b PERTURBATI DI QUANTITA' RELATIVA S U

PERCHE' E' STABILE?: LA PERTURBAZIONE RELATIVA SUI DATI E' DELLO STESSO ORDINE DELLA PRECISIONE

MACHINA: $\frac{|\tilde{a} - a|}{a} \approx u$
CIRCA

ESEMPIO: Sia $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ (Sono numeri reali): (DIVISIONE)

$$x, y \rightarrow x / y$$

$$\cdot x \rightarrow \tilde{x} \equiv f(x) = x(1 + \varepsilon_1) \text{ con } |\varepsilon_1| < u$$

$$\cdot y \rightarrow \tilde{y} \equiv f(y) = y(1 + \varepsilon_2) \text{ con } |\varepsilon_2| < u$$

$$\cdot \tilde{x} / \tilde{y} = f(\tilde{x} / \tilde{y}) = (\tilde{x} / \tilde{y})(1 + \varepsilon_3) \text{ con } |\varepsilon_3| < u$$

ANALISI IN AVANTI PER CALCOLARE L'ERRORE RELATIVO TRA QUESTO APPROSSIMATO E QUELLO ESATTO:

$$E_{\text{REL}} = \left| \frac{\text{RISULTATO APPROSSIMATO} - \text{RISULTATO ESATTO}}{\text{RISULTATO ESATTO}} \right| = \left| \frac{f\left(\frac{f(x)}{f(y)}\right) - \frac{x}{y}}{\frac{x}{y}} \right|$$

ANALOGAMENTE ALLA MOLTIPLICAZIONE, ANCHE LA DIVISIONE E' STABILE.

$$= \left| \frac{x(1 + \varepsilon_1) \cdot \frac{(1 + \varepsilon_2)}{y} \cdot (1 + \varepsilon_3) - \frac{x}{y}}{\frac{x}{y}} \right|$$

$$= \left| (1 + \varepsilon_1)(1 + \varepsilon_2)(1 + \varepsilon_3) - 1 \right| \simeq |\varepsilon_1 + \varepsilon_2 + \varepsilon_3| < 3u$$

ESEMPIO: Sia $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ (Sono numeri reali): (ADDITIONE E SOTTRAZIONE)

$$x, y \rightarrow x \pm y$$

$$\cdot x \rightarrow \tilde{x} \equiv f(x) = x(1 + \varepsilon_1) \text{ con } |\varepsilon_1| < u$$

$$\cdot y \rightarrow \tilde{y} \equiv f(y) = y(1 + \varepsilon_2) \text{ con } |\varepsilon_2| < u$$

$$\cdot \tilde{x} \pm \tilde{y} = f(\tilde{x} \pm \tilde{y}) = (\tilde{x} \pm \tilde{y})(1 + \varepsilon_3) = (x \cdot (1 + \varepsilon_1) \pm y \cdot (1 + \varepsilon_2)) \cdot (1 + \varepsilon_3) \text{ con } |\varepsilon_3| < u$$

$$E_{\text{REL}} = \left| \frac{f(f(x) \pm f(y)) - (x \pm y)}{x \pm y} \right| = \left| \frac{(x \pm y) - (x(1 + \varepsilon_1) \pm y(1 + \varepsilon_2))(1 + \varepsilon_3)}{x \pm y} \right|$$

→ IPOTESI DEL +

$$= \left| \frac{(x+y) - (x + x\varepsilon_1 + y + y\varepsilon_2)(1 + \varepsilon_3)}{x+y} \right|$$

$$= \left| \frac{(x+y) - (x+y) - (x+\gamma)\varepsilon_3 - (x\varepsilon_1 + y\varepsilon_2)(1 + \varepsilon_3)}{(x+y)} \right|$$

TRASCURABILI

$$= \left| \frac{-(x+y)\varepsilon_3 - x\varepsilon_1 - y\varepsilon_2 - \underbrace{x\varepsilon_1\varepsilon_2}_{\text{TRASCURABILI}} - \underbrace{y\varepsilon_2\varepsilon_3}_{\text{TRASCURABILI}}}{x+y} \right|$$

$$\left| \frac{x}{|x+y|} \varepsilon_1 + \frac{y}{|x+y|} \varepsilon_2 + \varepsilon_3 \right| \leftarrow = \frac{x}{|x+y|} \varepsilon_1 + \frac{y}{|x+y|} \varepsilon_2 + \varepsilon_3$$

(**caso stabile**: addizione di numeri con stesso segno (es: $5+3$) o sottrazione di numeri con segno opposto (es: $5 - (-3)$). L'errore relativo è < 3u. L'operazione è stabile e ben composta.

(**caso instabile**: sottrazione di numeri simili con stesso segno (es: $5.00001 - 5.00000$). Se $x \approx y \rightarrow 0$ l'errore relativo è molto grande. L'errore dell'operazione (ϵ_3) è sempre < u. Il problema sono gli errori già presenti su x e y (ϵ_1, ϵ_2). Gli errori relativi ϵ_1, ϵ_2 sui dati iniziali vengono amplificati drammaticamente quando fai $x - y$ con $x \approx y$. Si verifica il fenomeno di **cancellazione numerica**, che è la conseguenza più grave della rappresentazione finita. Causa errori anche con precisione alta.

ESEMPIO NUMERICO:

$$x = 1.23456789 \text{ (con } \epsilon_1 \approx 10^{-8})$$

$$y = 1.23456780 \text{ (con } \epsilon_2 \approx 10^{-8})$$

$$x - y = 0.00000009$$

• E_{ass} su $x, y \approx 10^{-8}$, ma $x - y \approx 10^{-8}$, quindi errore relativo $\approx 100\%$

Erori numerici più comuni ($IF(10, 5, \lambda, w)$, $u = \frac{1}{2} 10^{i-5} = 0.00005$): i computer usano rappresentazioni binarie approssimate dei numeri, generando inevitabilmente errori. Nei calcoli bisogna sempre valutare quanto questi errori impattino l'attendibilità del risultato.

1) **ERRORE DI CONVERSIONE**: poiché l'input dei dati avviene attraverso il formato decimale, mentre la base interna dei computer è binaria, non sempre è possibile rappresentare in modo esatto i valori numerici introdotti. Ad esempio, il numero 0.6 ha una rappresentazione binaria periodica (0.10011001...), e quindi verrà arrotondato con un numero finito t di cifre introducendo un certo errore. Numeri periodici in base 10 o numeri irrazionali non possono essere rappresentati in maniera esatta.

2) **ERRORE DI ARROTONDAMENTO**: poiché $a \cdot b$ in generale richiede una precisione maggiore dei rispettivi operandi a e b per essere rappresentato esattamente, il risultato di un prodotto può essere arrotondato risultando inesatto nell'ultima cifra. Ad esempio, $0.24665 \cdot 0.63994 = 0.1578412010$ viene arrotondato a 0.15784 con un errore relativo di $1.2 \cdot 10^{-6}$.

3) **ERRORE DI ASSORBIMENTO**: sommando due numeri a e b di due ordini di grandezza diversi, quello più piccolo può essere assorbito e non influenzare il risultato.

Per esempio $0.1 \cdot 10^3 + 0.3 \cdot 10^{-3}$ è:

| | | |
|--------|----------|---|
| 100.0 | + | RISULTATO FINALE APPROSSIMATO E' $0.1 \cdot 10^3$ |
| 0.0003 | 100.0003 | |

E quindi $a+b=a$ anche se $b \neq 0$.

4) **ERRORE DI CANCELLAZIONE NUMERICA**: differenza di due numeri a e b quasi uguali ha meno cifre significative rispetto sia ad a che a b . Ad esempio: $0.90905 \cdot 10^2 - 0.90903 \cdot 10^2 = 0.2 \cdot 10^2$. La mantissa risultante 0.20000, a meno che a e b siano rappresentabili in modo esatto, contiene quattro zeri non significativi che sono dovuti alla perdita di precisione. Il numero di cifre significative affidabili si riduce drasticamente. Perdere precisione è come ridurre l'affidabilità, ha meno cifre sulle quali porre fiducia.