# TEORIA DEGLI ERRORI

## Rappresentazione dei numeri

Scelto $\beta \in \mathbb{N}$, $\beta > 1$ (base), ogni numero $x \in \mathbb{R} \setminus \{0\}$ ammette una rappresentazione in base $\beta$

$$x = \text{sign}(x)\, \beta^b \sum_{i=1}^{\infty} \alpha_i \beta^{-i}$$

con

$b \in \mathbb{Z}$ (esponente), $\alpha_i \in \mathbb{N}$: $0 \leq \alpha_i \leq \beta - 1$ (cifra), $\text{sign}(x) = \begin{cases} -1 & x < 0 \\ 1 & x > 0 \end{cases}$

Th. (Unicità della rappresentazione)
Data una base $\beta \in \mathbb{N}$ $\beta > 1$ e un numero $x \in \mathbb{R} \setminus \{0\}$, esiste un'unica rappresentazione in base $\beta$ tale che:

1. $\alpha_1 \neq 0$ (la prima cifra dopo la virgola non può mai essere 0)
2. non esiste $k \in \mathbb{N}$: $a_j = \beta - 1$, $\forall j > k$ (come ultima cifra non può mai esserci 9 – ciò evita che un numero come $1{,}19\overline{9}$, approssimabile con $1{,}2$, abbia appunto la rappresentazione come quest'ultimo → la parte frazionaria è sempre minore di 1)

⇓

Rappresentazione di $x$ in virgola mobile normalizzata.

La serie $\sum_{i=1}^{\infty} \alpha_i \beta^{-i}$ è detta mantissa di $x$ e sappiamo che vale:

$$\beta^{-1} \leq \sum_{i=1}^{\infty} \alpha_i \beta^{-i} < \sum_{i=1}^{\infty} (\beta - 1)\beta^{-i} = (\beta - 1) \sum_{i=1}^{\infty} \beta^{-i} = (\beta - 1) \frac{1}{1 - \frac{1}{\beta}} \beta^{-1} = 1$$

(progressione geometrica di ragione $1/\beta$)

All'interno di un calcolatore si possono rappresentare solo un certo numero $m$ di cifre della mantissa di $x$. La rappresentazione sul calcolatore corrisponde perciò al numero

$$y = \text{tr}(x) = \text{sign}(x)\, \beta^b \sum_{i=1}^{m} \alpha_i \beta^{-i} \qquad \text{troncamento}$$

oppure al numero

$$y = \text{rd}(x) = \begin{cases} \text{tr}(x) & 0 \leq \alpha_{m+1} < \beta/2 \\ \text{sign}(x)\, \beta^b \left[ \sum_{i=1}^{m} \alpha_i \beta^{-i} + \beta^{-m} \right] & \beta/2 \leq \alpha_{m+1} < \beta \end{cases} \qquad \text{arrotondamento}$$

Nel caso in cui usualmente $\alpha_1 = \alpha_2 = \ldots = \alpha_m = \beta - 1$ e $a_{m+1} \geq \beta/2$, in caso di arrotondamento è possibile che si debba aumentare di una unità l'esponente $b$

Si può dimostrare che

$$|\text{tr}(x) - x| < \beta^{b-m} \qquad |\text{rd}(x) - x| \leq \frac{1}{2} \beta^{b-m}$$

per cui, tra le due, la rappresentazione per arrotondamento approssima meglio $x$

Si indica con $M$ l'insieme dei numeri rappresentabili all'interno del calcolatore, i numeri di macchina.

$\mathcal{F}(\beta, m, L, U)$ con $\beta$ base, $m$ cifre della mantissa, $[L, U]$ minimo/massimo esponente, $L, U \in \mathbb{Z}$

È un insieme finito di cardinalità

$$\text{card}(M) = 2 \underbrace{(\beta^m - \beta^{m-1})}_{\substack{\text{la prima cifra della mantissa non può essere 0} \\ (\beta-1)\cdot\underbrace{\beta \cdot \beta \cdot \ldots \beta}_{m-1 \text{ volte}}}} \underbrace{(U - L + 1)}_{\text{possibili esponenti}} + \underbrace{1}_{\substack{\text{rappresentazione} \\ \text{dello zero}}}$$

(2 segni)

MASTER COPY

Tel. 388/9837745

Dato un qualunque numero reale $x \neq 0$ non è assicurata l'esistenza di $rd(x)$ tra i numeri di macchina. Se il numero da rappresentare è "troppo grande" si ha overflow (tutti i calcolatori lo segnalano e poi arrestano l'esecuzione del programma o sostituiscono a $rd(x) = sign(x) \max_{y \in M} |y|$). Se il numero da rappresentare è "troppo piccolo" si ha underflow (non tutti i calcolatori segnalano questa situazione e nel caso in cui proseguono l'esecuzione del programma pongono $rd(x) = 0$)

Si definisce errore assoluto della rappresentazione del numero reale $x$ il valore

$$\delta(x) = rd(x) - x \quad \text{errore assoluto per arrotondamento}$$
$$\delta(x) = tr(x) - x \quad \text{errore assoluto per troncamento}$$

e errore relativo il valore

$$\varepsilon(x) = \frac{rd(x) - x}{x} = \frac{\delta x}{x} \quad \text{errore relativo per arrotondamento}$$
$$\varepsilon(x) = \frac{tr(x) - x}{x} = \frac{\delta x}{x} \quad \text{errore relativo per troncamento}$$

Dalle precedenti considerazioni si ricavano le limitazioni

$$|\delta_{x\,rd}| \leq \frac{1}{2}\beta^{b-m} \qquad |\delta_{x\,tr}| < \beta^{b-m}$$

$$|\varepsilon_{x\,rd}| \leq \frac{1}{2}\frac{\beta^{b-m}}{\beta^{b-1}} = \frac{1}{2}\beta^{1-m} \qquad |\varepsilon_{x\,tr}| < \frac{\beta^{b-m}}{\beta^{b-1}} = \beta^{1-m}$$

## Errore nel calcolo di una funzione

Nel calcolo di una funzione razionale $f(x_1, x_2 \dots x_n)$ in punto assegnato $P_0 = (x_1^{(0)}, x_2^{(0)} \dots x_n^{(0)})$, in generale non si ottiene il valore $f(P_0)$ cercato a causa delle approssimazioni che si introducono. Un primo errore nasce dal fatto che le operazioni aritmetiche che compaiono nella $f(P)$ devono essere sostituite con le corrispondenti operazioni di macchina e organizzate in un certo algoritmo e ciò equivale in definitiva a sostituire $f(P)$ con $f_a(P)$ (errore algoritmico). Un secondo tipo di errore si presenta quando non è possibile rappresentare esattamente le coordinate di $P_0$ e quindi si devono approssimare tali coordinate con i numeri di macchina (errore trasmesso dai dati)

### Errore assoluto

Assegnato $P_0 \in \mathbb{R}^n$ di coordinate $x_i^{(0)}$, $i = 1, 2 \dots n$ in cui si vuole calcolare la funzione $f(P)$, si consideri l'insieme

$$D = \{P \in \mathbb{R}^n : a_i \leq x_i \leq b_i,\ i = 1, 2 \dots n\} \quad a_i, b_i \in \mathbb{R}$$ insieme di indeterminazione

Il valore cercato $f(P_0)$ viene sostituito dal valore calcolato $f_a(P_1)$ dove $P_1$ appartiene a $D$ e quindi l'errore commesso risulta

$$\underbrace{f_a(P_1) - f(P_0)}_{\delta_f} = \underbrace{f_a(P_1) - f(P_1)}_{\delta_a} + \underbrace{f(P_1) - f(P_0)}_{\delta_d}$$

$\delta_f$ = errore assoluto totale; $\delta_a$ = errore assoluto algoritmico; $\delta_d$ = errore assoluto trasmesso dai dati

L'errore algoritmico, una volta fissato l'algoritmo che fornisce $f_a(P)$, risulta definito e stimabile.
All'errore trasmesso dai dati, nell'ipotesi che $f(P) \in C^1(D)$, si può dare la rappresentazione (tramite la formula di Taylor arrestata al primo termine)

$$f(P_1) - f(P_0) = \sum_{i=1}^{n} \frac{df}{dx_i}(x_i^{(1)} - x_i^{(0)})$$

Ponendo poi $\delta x_i = x_i^{(1)} - x_i^{(0)}$, $e_i = \frac{df}{dx_i}$ coefficienti di amplificazione degli errori $\delta x_i$

$$f(P_1) - f(P_0) = \sum_{i=1}^{n} e_i \delta x_i$$

Una limitazione per tale errore è

$$|f(P_1) - f(P_0)| = \left|\sum_{i=1}^{n} e_i \delta x_i\right| \leq \sum_{i=1}^{n} |e_i||\delta x_i|$$

e posto $|e_i| = \left|\frac{df}{dx_i}\right| \leq \sup_{P \in D} \left|\frac{df}{dx_i}\right| = \Delta x_i$

si ottiene $|f(P_1) - f(P_0)| \leq \sum_{i=1}^{n} \Delta x_i |\delta x_i|$

Una limitazione per il modulo dell'errore assoluto $f_a(P_1) - f(P_0)$ è quindi

$$|f_a(P_1) - f(P_0)| \leq E_a + E_d$$

Se si conosce una stima dell'errore di troncamento e degli errori $\delta x_i$ nonchè se $\Delta x_i$ si può stabilire a posteriori un confine superiore per l'errore assoluto con cui si è calcolata la funzione (problema diretto)

(es) $f(x_1, x_2) = x_1 x_2 \quad P_0 = (\sqrt{2}, \pi) \quad D = [1,2] \times [3,4]$

$\frac{df}{dx_1} = x_2 \quad \frac{df}{dx_2} = x_1 \Rightarrow \Delta x_1 = \sup_{P \in D} \left|\frac{df}{dx_1}\right| = 4 \quad \Delta x_2 = \sup_{P \in D} \left|\frac{df}{dx_2}\right| = 2$

approssimo $P_0$ con $P_1(1.41, 3.14)$ sapendo che

$|\delta x_1| = \frac{1}{2} \cdot 10^{-2} \quad |\delta x_2| = \frac{1}{2} \cdot 10^{-2} \Rightarrow$

$|\delta_d| \leq 4 \cdot \frac{1}{2} \cdot 10^{-2} + 2 \cdot \frac{1}{2} \cdot 10^{-2} = 3 \cdot 10^{-2}$

Il problema inverso consiste nel richiedere a priori che $f_a(P_1)$ sia tale che l'errore assoluto $|f_a(P_1) - f(P_0)|$ risulti minore di un valore prefissato per cui va cercato un algoritmo $f_a(P)$ e un punto $P_1$ che soddisfino la richiesta

(es) $f(x_1, x_2) = \frac{x_1}{x_2} \quad P_0 = (\sqrt{2}, \pi) \quad D = [1,2] \times [3,4]$

Voglio un errore assoluto massimo pari a $10^{-3}$

$\frac{df}{dx_1} = \frac{1}{x_2} \quad \frac{df}{dx_2} = -\frac{x_1}{x_2^2} \Rightarrow |\Delta x_1| = \sup_{P \in D} \left|\frac{df}{dx_1}\right| = \frac{1}{3} \quad |\Delta x_2| = \sup_{P \in D} \left|\frac{df}{dx_2}\right| = \frac{2}{9}$

$|\delta_f| \leq |\delta_a| + |\delta_d|$

$10^{-3} \leq |\delta_a| + |\delta_d| \rightarrow$ spezzo l'errore in due contributi

$|\delta_a| \leq \frac{1}{2} \cdot 10^{-3}$ (arrotondo alla terza cifra decimale)

$|\delta_d| \leq \frac{1}{2} \cdot 10^{-3} \quad \sum_{i=1}^{n} \Delta x_i |\delta x_i| \leq \frac{1}{2} \cdot 10^{-3}$

$\Delta x_1 |\delta x_1| + \Delta x_2 |\delta x_2| \leq \frac{1}{2} \cdot 10^{-3} \rightarrow$ spezzo in due contributi

$\Delta x_1 |\delta x_1| \leq \frac{1}{4} \cdot 10^{-3} \quad |\delta x_1| \leq \frac{3}{4} \cdot 10^{-3}$ (arrotondo alla terza cifra decimale)

$\Delta x_2 |\delta x_2| \leq \frac{1}{4} \cdot 10^{-3} \quad |\delta x_2| \leq \frac{9}{8} \cdot 10^{-3}$ (arrotondo alla terza cifra decimale)

Errore relativo

L'errore relativo che si commette nel calcolo di una funzione $f(P)$ in un punto $P_0$ è definito da $\varepsilon_f = \frac{f_a(P_1) - f(P_0)}{f(P_0)} =$

$$= \frac{f(P_1) - f(P_0)}{f(P_0)} + \frac{f_a(P_1) - f(P_1)}{f(P_1)} \left(1 + \frac{f(P_1) - f(P_0)}{f(P_0)}\right)$$

per cui indicando con

$$\varepsilon_a = \frac{fa(P_1) - f(P_1)}{f(P_1)}$$ errore relativo algoritmo

e con

$$\varepsilon_d = \frac{f(P_1) - f(P_0)}{f(P_0)}$$ errore relativo trasmesso dai dati

si ottiene

$$\varepsilon_f = \varepsilon_a + \varepsilon_d + \varepsilon_a \varepsilon_d$$ con $\varepsilon_a \varepsilon_d$ sempre trascurabile perché di ordine superiore

Per l'errore relativo trasmesso dai dati abbiamo una stima

$$\varepsilon_d = \frac{f(P_1) - f(P_0)}{f(P_0)} = \sum_{i=1}^{n} \frac{df}{dx_i} \frac{dx_i}{f(P_0)} = \sum_{i=1}^{n} \frac{df}{dx_i} \frac{x_i^{(0)}}{f(P_0)} \cdot \frac{dx_i}{x_i^{(0)}} = \sum_{i=1}^{n} \frac{df}{dx_i} \frac{x_i^{(0)}}{f(P_0)} \varepsilon_{x_i}$$

con $\frac{df}{dx_i} \frac{x_i^{(0)}}{f(P_0)} = \gamma_i$ coefficiente di amplificazione degli errori relativi

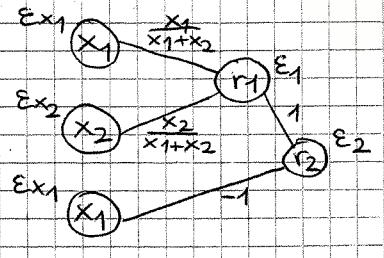
Se i coefficienti $\gamma_i$ sono tali che l'errore $\varepsilon_d$ risulta dello stesso ordine degli errori $\varepsilon_{x_i}$ il problema del calcolo della funzione risulta ben condizionato; si dice invece mal condizionato se a piccoli errori relativi $\varepsilon_{x_i}$ corrispondono un errore $\varepsilon_d$ rilevante.

Se l'algoritmo è tale da produrre errori accettabilmente limitati nella sua applicazione si dice stabile, mentre è detto instabile nel caso contrario.

Analizzando le quattro operazioni fondamentali per quanto riguarda gli errori trasmessi dai dati si ottiene:

| OPERAZIONE | $\delta_d$ | $\varepsilon_d$ | |
|---|---|---|---|
| + | $\delta x + \delta y$ | $\frac{x}{x+y}\varepsilon_x + \frac{y}{x+y}\varepsilon_y$ | → non danno problemi per l'errore assoluto ma possono rendere grande l'errore relativo per $\lvert x \rvert \simeq \lvert y \rvert$ |
| − | $\delta x - \delta y$ | $\frac{x}{x-y}\varepsilon_x - \frac{y}{x-y}\varepsilon_y$ | |
| x | $y\delta x + x\delta y$ | $\varepsilon_x + \varepsilon_y$ | → non danno problemi per l'errore relativo, ma l'errore assoluto può amplificarsi |
| / | $\frac{1}{y}\delta x - \frac{x}{y^2}\delta y$ | $\varepsilon_x - \varepsilon_y$ | |

(es) $f(x_1, x_2) = \frac{x_1 + x_2}{x_1}$




$$\varepsilon_f = \varepsilon_{r_2} = \varepsilon_2 + \varepsilon_{r_1} - \varepsilon_{x_1} =$$

$$= \varepsilon_2 + \varepsilon_1 + \frac{x_1}{x_1+x_2}\varepsilon_{x_1} + \frac{x_2}{x_1+x_2}\varepsilon_{x_2} - \varepsilon_{x_1} =$$

$$= \underbrace{\varepsilon_1 + \varepsilon_2}_{\text{algoritmo}} \underbrace{- \frac{x_2}{x_1+x_2}\varepsilon_{x_1} + \frac{x_2}{x_1+x_2}\varepsilon_{x_2}}_{\text{dati}}$$

MASTER COPY

Tel. 388/9837745

# RICHIAMI DI ALGEBRA LINEARE

## Matrici e vettori

Con $A \in \mathbb{C}^{m\times n}$ si intende una matrice di m righe e n colonne formate da mxn numeri complessi $a_{ij}$ $(i=1,2\dots m,\ j=1,2\dots n)$ detti elementi di A.

$$A=\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

Gli interi m e n si dicono dimensioni di A.
Se $m=n$ la matrice A si dice quadrata di ordine n; in tal caso gli elementi $a_{ii}$ $(i=1,2\dots n)$ si dicono elementi diagonali.

Con $A \in \mathbb{C}^m$ si intende un vettore con m componenti; i vettori possono essere visti come matrici particolari con una sola colonna

Data la matrice $A \in \mathbb{C}^{m\times n}$, la matrice $B \in \mathbb{C}^{n\times m}$ i cui elementi sono $b_{ij}=a_{ji}$ $(i=1,2\dots n,\ j=1,2\dots m)$ si dice matrice trasposta della matrice A e si indica con $A^T$ mentre si dice matrice trasposta coniugata di A la matrice $B \in \mathbb{C}^{n\times m}$ i cui elementi sono $b_{ij}=\bar{a}_{ji}$ e si indica con $A^H$.

La matrice quadrata

$$I=\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix} \qquad a_{ij}=\begin{cases} 1 & i=j \\ 0 & i\neq j \end{cases}$$

si dice matrice identica.
Si dice matrice diagonale una matrice quadrata D con gli elementi non diagonali nulli

$$D=\begin{pmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \dots & \dots & \dots & 0 \\ 0 & 0 & \dots & d_n \end{pmatrix}$$

Le matrici quadrate del tipo

$$L=\begin{pmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \dots & 0 \\ \dots & \dots & \dots & 0 \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{pmatrix} \qquad R=\begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & r_{22} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & r_{nn} \end{pmatrix}$$

si dicono matrici triangolari e precisamente L si dice triangolare inferiore e R triangolare superiore

## Operazioni tra matrici

Date $A, B \in \mathbb{C}^{m\times n}$, la matrice somma $C=A+B$ è definita ponendo

$$c_{ij}=a_{ij}+b_{ij} \qquad (i=1,2\dots m,\ j=1,2\dots n)$$

Valgono la proprietà associativa e commutativa.

se $A \in \mathbb{C}^{m\times k}$ e $B \in \mathbb{C}^{k\times n}$ la matrice prodotto $C=AB$ è definita ponendo

$$c_{ij}=\sum_{e=1}^{k} a_{ie}b_{ej} \qquad (i=1,2\dots m,\ j=1,2\dots n)$$

e risulta quindi $C \in \mathbb{C}^{m\times n}$
Per la moltiplicazione vale la proprietà associativa e distributiva rispetto all'addizione, ma non la proprietà commutativa ($AB \neq BA$ generalmente) e la legge di annullamento del prodotto
Per la trasposta e la trasposta coniugata si ha

$$(AB)^T=B^TA^T \quad e \quad (AB)^H=B^HA^H$$