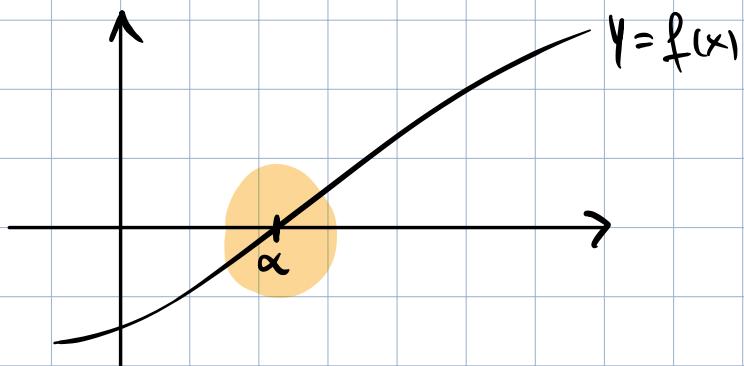


# SOLUZIONE NUMERICA DEUE EQUAZIONI NON LINEARI

Problema : data  $f: [a, b] \rightarrow \mathbb{R}$ ,  
trovare  $\alpha \in [a, b]$  t.c.

$$f(\alpha) = 0.$$

Graficamente cerchiamo l'intersezione del grafico di  $f$  con l'asse delle ascisse.



Esempio : trovare  $x$  t.c.  $3s x - x = 0$ .

TEOREMA (di Bolzano).

Sia  $f: [a, b] \rightarrow \mathbb{R}$  continua e t.c.  
 $f(a)f(b) < 0$ . Allora  $\exists \alpha \in (a, b)$  t.c.

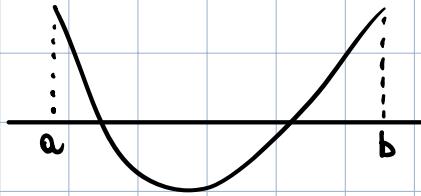
$$f(\alpha) = 0.$$

DEFINIZIONE : se  $f(x) = 0$ , allora diremo  
che  $x$  è "zero per  $f$ " oppure "zero di  $f$ ".

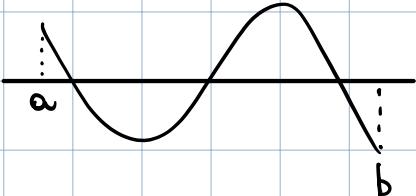
Osservazioni :

1)  $f(a)f(b) < 0$  è condizione sufficiente ma  
non necessaria. Ad esempio, si pensi a

a ~



2) Le ipotesi non garantiscono l'unicità  
dello zero di funzione. Ad esempio,  
si pensi a



L'unicità può seguire da ipotesi  
supplementari. Ad esempio:

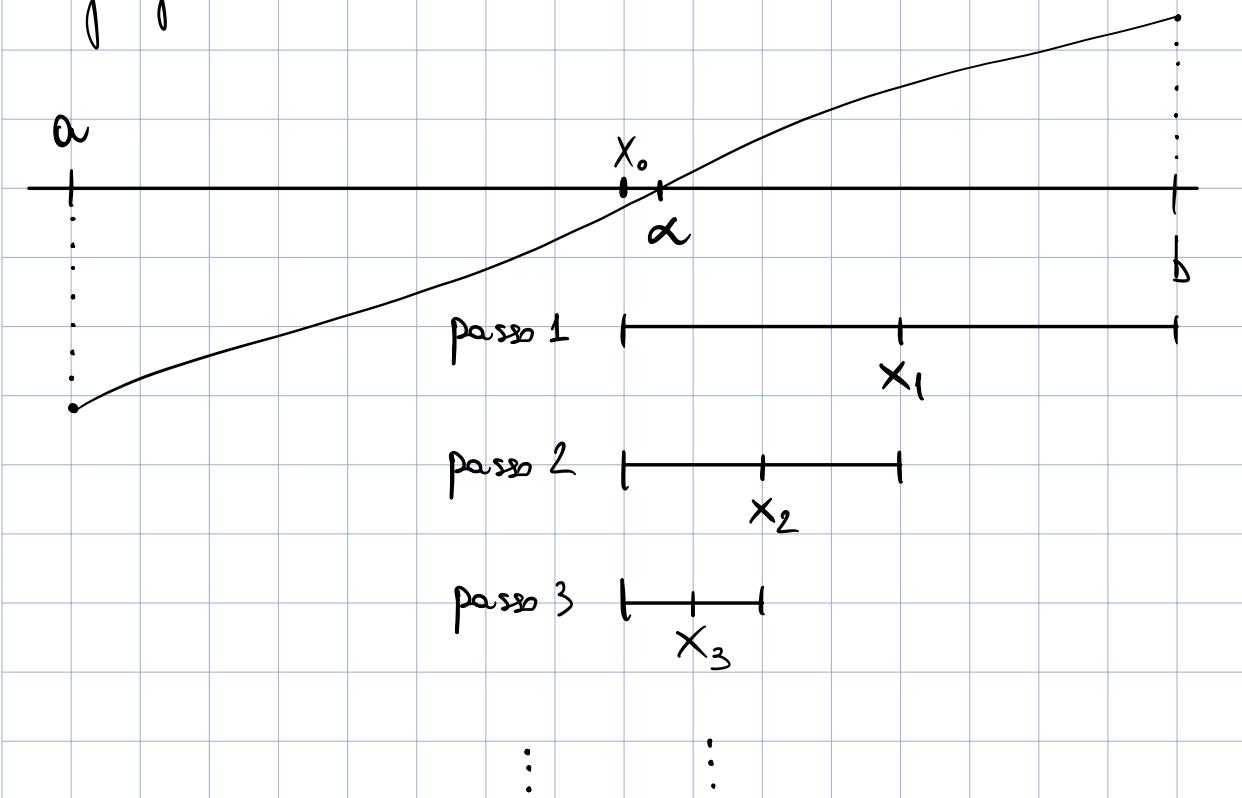
- $f$  monotona in  $[a, b]$
- $f(a)f(b) < 0$  e  $f$  di segno costante in  $[a, b]$

## METODO DELLE SUCCESSIVE BISEZIONI

Ipotesi:  $f$  continua,  $f(a)f(b) < 0$ .

Idee: ripetutamente si suddivide l'intervalllo  
in 2 semi-intervalli e si conserva  
quello che verifica le ipotesi.

Graficamente:



$x_0, x_1, x_2, x_3, \dots$  sono le successive approssimazioni  
di  $\alpha$  generate dal metodo.

Formalmente:

dati  $f$ ,  $a$ ,  $b$ , poniamo:  
 $a^{(0)} := a$ ,  $b^{(0)} := b$ ,  $x^{(0)} := \frac{a^{(0)} + b^{(0)}}{2}$  e iteriamo

per ogni  $k = 0, 1, 2, \dots$

① se  $f(x^{(k)}) = 0$ , poniamo  $\alpha = x^{(k)}$  e  
usciamo dal ciclo

② se  $f(a^{(k)}) f(x^{(k)}) < 0$ ,  
poniamo  $a^{(k+1)} := a^{(k)}$ ,  $b^{(k+1)} := x^{(k)}$

altrimenti

poniamo  $a^{(k+1)} := x^{(k)}$ ,  $b^{(k+1)} := b^{(k)}$

③ poniamo  $x^{(k+1)} := \frac{a^{(k+1)} + b^{(k+1)}}{2}$

fine

Osservazioni

L'algoritmo si arresta in un

numero finito di passi solo se  
per qualche  $K$  si ha  $f(x^{(K)}) = 0$ .

Altrimenti si generano tre succ. m.

$$\{a^{(k)}\}_{k \in \mathbb{N}}, \{b^{(k)}\}_{k \in \mathbb{N}}, \{x^{(k)}\}_{k \in \mathbb{N}}$$

tali che

- $a^{(k)} \leq x^{(k)} \leq b^{(k)} \quad \forall k \in \mathbb{N}$
- $x \in [a^{(k)}, b^{(k)}]$
- $a^{(k)} \xrightarrow[k \rightarrow +\infty]{} x$  in modo monotono crescente
- $b^{(k)} \xrightarrow[k \rightarrow +\infty]{} x$  in modo monotono decrescente
- $x^{(k)} \xrightarrow[k \rightarrow +\infty]{} x$
- $f(x) = 0$

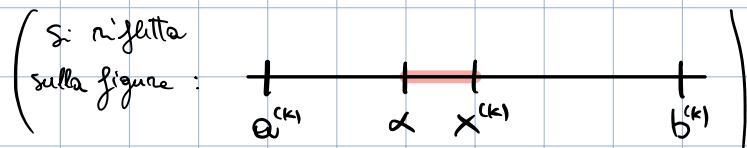
La probabilità che si realizzi l'evento  
di punto ① è trascurabile e rende  
il controllo superfluo.

È possibile prendere a parola il numero di  
passi sufficiente a garantire una  
predeterminata qualità dell'approssimazione.

Vediamo come:

Siano  $I^{(k)} := [a^{(k)}, b^{(k)}]$  e  $|I^{(k)}| := |b^{(k)} - a^{(k)}|$ ,  
per ogni  $k \geq 0$ . Allora si ha: lunghezza dell'int.  $[a^{(k)}, b^{(k)}]$

$$1) |x^{(k)} - \alpha| \leq \frac{1}{2} |b^{(k)} - a^{(k)}|, \forall k \geq 0$$



$$\begin{aligned} 2) |I^{(k)}| &= \frac{1}{2} |I^{(k-1)}| = \frac{1}{2} \frac{1}{2} |I^{(k-2)}| = \\ &= \frac{1}{2^2} |I^{(k-2)}| = \dots = \frac{1}{2^k} |I^{(0)}| = \frac{|b - a|}{2^k} \end{aligned}$$

Unendo 1) e 2) si ottiene

$$|x^{(k)} - \alpha| \leq \frac{1}{2} |I^{(k)}| = \dots = \frac{1}{2^{k+1}} |b - a|$$

Sufficiente di voler calcolare una approssimazione  $x^{(k)}$  di  $\alpha$  tale che:

$$|x^{(k)} - \alpha| < \varepsilon$$

toleranza

(vogliamo che  $x^{(k)}$  disti da  $\alpha$  meno di  $\varepsilon$ )

$|x^{(k)} - \alpha|$  non è noto, ma possiamo scegliere  $k$

Tale che :

$$\frac{1}{2^{k+1}} |b-a| < \varepsilon \iff K > \log_2 \frac{|b-a|}{\varepsilon} - 1$$

n'solvendo  
rispetto a K

ESEMPIO: Siano  $a=1$ ,  $b=2$ ,  $\varepsilon = 10^{-10}$ .

Allora si ha

$$K > \log_2 10^{10} - 1 = 10 \overline{\log_2 10} - 1 \approx \\ \approx 32.2$$

Ovvio sono sufficienti 33 passi a generare che  $x^{(k)}$  disti da  $a$  meno di  $10^{-10}$ .

Come quantificidiamo l'errore associato ad una approssimazione?

### DEFINIZIONI

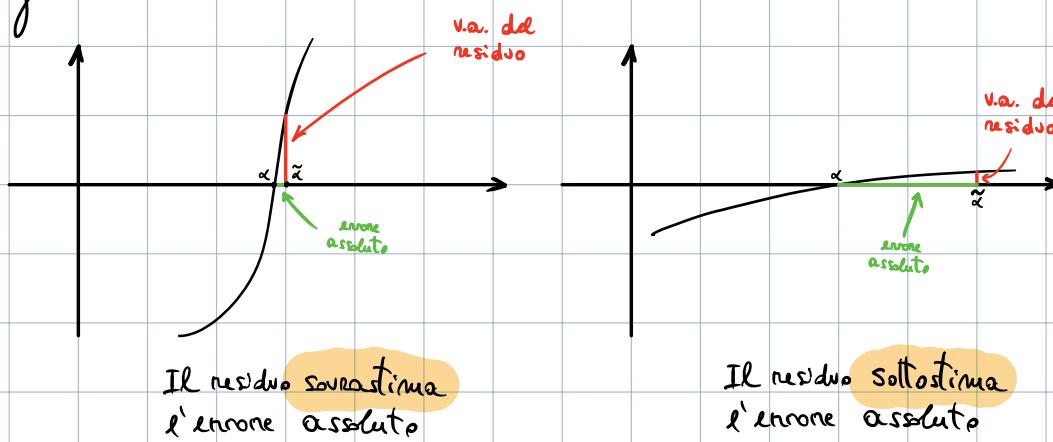
Siano  $\alpha, \tilde{\alpha} \in \mathbb{R}$ , con  $\tilde{\alpha}$  approssimazione del "dato esatto"  $\alpha$ . Definiamo

Errore assoluto :  $E_A = |\alpha - \tilde{\alpha}| \leftarrow$  distanza di  $\tilde{\alpha}$   
da  $\alpha$  sulla  
retta reale

$$\text{Errore relativo : } E_R = \frac{|\alpha - \tilde{\alpha}|}{|\alpha|}$$

→ Residuo :  $f(\tilde{\alpha})$   
 (è legato al problema)

Osservazione : il valore assoluto del residuo può sovrestimare o sottostimare l'errore assoluto. Si rifletta sulle figure :



## CRITERI DI ARRESTO PER IL METODO DELLE SUCC.VE BISEZIONI

Ricordiamo che  $|x^{(k)} - \alpha| < \frac{1}{2} |b^{(k)} - a^{(k)}|$ ,  $k \geq 0$ ,  
 dove  $x^{(k)} = \frac{a^{(k)} + b^{(k)}}{2}$ .

1) stima errore assoluto:

$$\frac{1}{2} |b^{(k)} - a^{(k)}| < \varepsilon$$

toleranza

2) stima errore relativo:

$$\frac{\frac{1}{2} |b^{(k)} - a^{(k)}|}{|x^{(k)}|} < \varepsilon ,$$

o equivalentemente

$$\frac{|b^{(k)} - a^{(k)}|}{|a^{(k)} + b^{(k)}|} < \varepsilon$$

3) Residuo:

$$|f(x^{(k)})| < \varepsilon$$

4) Errore misto assoluto/relativo:

$$\frac{|b^{(k)} - a^{(k)}|}{|a^{(k)} + b^{(k)}| + 2} < \epsilon$$

Vediamo se  $\frac{1}{2}|b^{(k)} - a^{(k)}|$  può essere molto piccolo  
e se  $\frac{|b^{(k)} - a^{(k)}|}{|a^{(k)} + b^{(k)}|}$  può essere molto grande; è una  
soluzione contro possibili problemi  
dovuti all'uso dell'errore relativo  
(risp. assoluto) quando  $a$  è molto  
piccolo (risp. grande).

# VANTAGGI E SVANTAGGI DEL METODO DELLE SUCCESSIONI

## VANTAGGI :

- Semplicità
- Convergenza "globale" (se  $f(a)f(b) < 0$ , non occorre che  $a$  e  $b$  siano vicini allo zero di  $f$  affinché il metodo sia convergente)
- Stima esatta dell'errore ad ogni passo :

$$a^{(k)} \leq x \leq b^{(k)}, \forall k$$

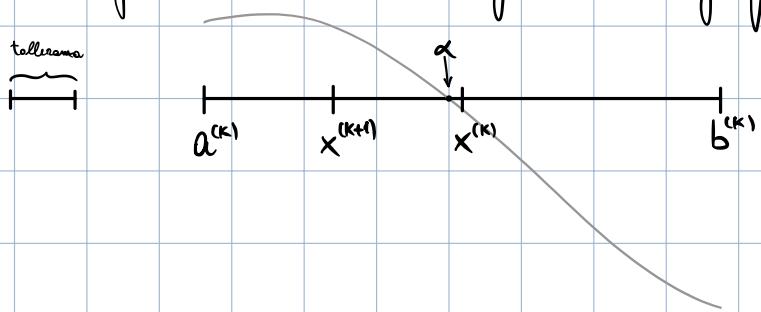
↑ stima per  
dritto                      ↑ stima per  
                                escesso

## SVANTAGGI

- È lento
- Di frequente scatta buona stima dell'errore :

$$|x^{(k+1)} - x| > |x^{(k)} - x|$$

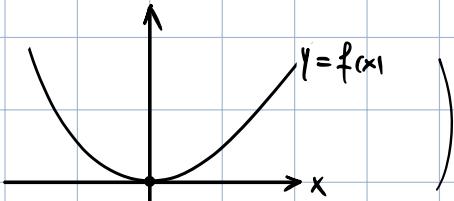
Si riflette sulla seguente figura:



- Per implementarla è necessario che

$f(a)f(b) < 0$  (Ad esempio non si

può applicare se



## PSEUDOCODE DEL ALGORITMO

dati:  $f, a, b, \varepsilon$

( $f, a, b$  verificano le ipotesi del Teorema di Bolzano;  $\varepsilon > 0$  tolleranza)

1.  $x := \frac{1}{2}(a+b)$

stima della  
soluzione

2. se  $b-x < 0$ , pongo  $\alpha := x$  e esco

3. se  $\text{sign}(f(a)) \neq \text{sign}(f(x_1))$ , allora

$b := x$ , altrimenti  $a := x$

4. torna al punto 1.

L'algoritmo "leggi" solo il segno di  $f$ ,

non il suo effettivo valore. La seguente

variante mi sfrutta anche il valore,

ottenendo generalmente una soluzione nel

numero di passi necessari per approssimare

la soluzione a portata di tolleranza.

## METODO REGOLA FALSE

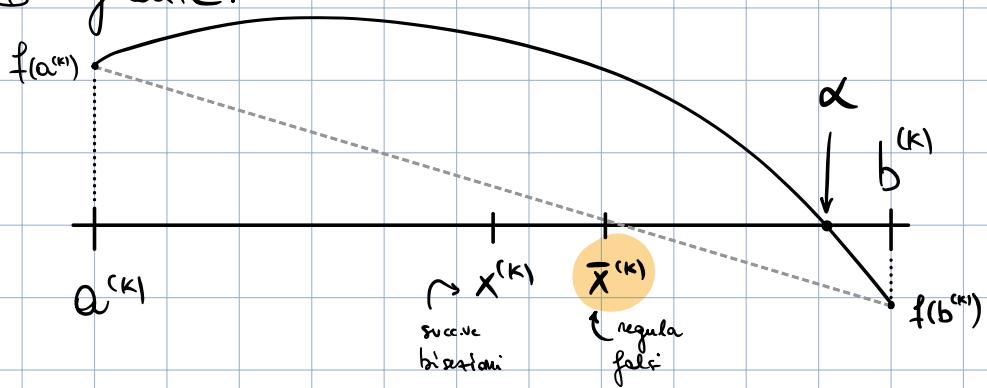
S'ragione come nel metodo delle successive

bisezioni, ma al generico passo  $k$

s'rimposta il punto medio  $x^{(k)} = \frac{a^{(k)} + b^{(k)}}{2}$

con l'ascissa  $\bar{x}^{(k)}$  del punto di intersezione della

Secante al grafico di  $f$  per i punti d'ascisse  $a^{(k)}, b^{(k)}$  con l'asse delle ascisse. Si vede il grafico seguente.



### ESERCIZIO

1. determinare l'espressione per  $\bar{x}^{(k)}$  in funzione di  $a^{(k)}, f(a^{(k)}), b^{(k)}, f(b^{(k)})$ .

2. osservare che generalmente non si

$b^{(k)} - a^{(k)} \xrightarrow[k \rightarrow +\infty]{} 0$ .

(si consideri  $f$  con  $f(a) > 0, f(b) < 0$  e concavità rivolta verso il basso in  $[a, b]$ )

Si rende necessario utilizzare nuovi criteri di arresto. Si considerino i seguenti:

- $|x^{(k+1)} - x^{(k)}| < \varepsilon$  Stima errore assoluto
- $\frac{|x^{(k+1)} - x^{(k)}|}{|x^{(k+1)}|} < \varepsilon$  Stima errore relativo
- $\frac{|x^{(k+1)} - x^{(k)}|}{|x^{(k+1)}| + 1} < \varepsilon$  Stima errore "misto" ass/rel

## CONDIZIONAMENTO DI UN PROBLEMA

Quando dobbiamo risolvere al computer un problema che viene dalle applicazioni, dobbiamo tener

erono di diverse origini d'  
 errore : semplificazioni nel modello,  
 errori di misurazione nei dati,  
 discretizzazione (esempio :  $f'(x)$   
 sostituita con  $\frac{f(x+h) - f(x)}{h}$ ,  $h$  "passo"),  
 arrotondamento di un numero reale  
 ad un "numero macchina".

Ci chiediamo : Qual è l'effetto d'  
 questi errori sulla soluzione del  
 problema? Ci concentreremo maggiormente  
 sugli errori derivati dall'utilizzo del  
 computer.

Definizione : Sia  $P$  un problema.

Abbiamo :

$$\begin{array}{c} \text{dato} \qquad \text{soluzione} \\ \hline x \longmapsto y \\ x + \delta x \longmapsto y + \delta y \end{array}$$

Il problema  $P$  si dice ben condizionato se e solo se piccole perturbazioni  $\delta x$  sul dato corrispondono altrettanto piccole perturbazioni  $\delta y$  sulla soluzione. Altrimenti,  $P$  è detto mal condizionato.

È possibile quantificare il condizionamento mediante un numero.

Per semplicità, supponiamo  $x, y$  numeri reali. Idealmente escludiamo le più piccole esatte  $K > 0$  t.c.

$$\frac{|\delta y|}{|y|} \leq K \frac{|\delta x|}{|x|}$$

,  $|\delta x|$  piccolo

$\nearrow$  errore relativo sulla soluzione

$\nwarrow$  errore relativo sul dato

In generale,  $K$  dipende anche da  $x$ :

$$K = K(x).$$

Se  $K \approx 1$ ,  $P$  è ben condizionato.

Se  $K \gg 1$ ,  $P$  è mal condizionato.

↑  
"di ordini di grandezza  
superiore a"

$K$  è detto fattore di condizionamento.

$K(x)$  è il maggior fattore d' amplificazione  
dell' errore relativo del positivo osservato  
per piccole perturbazioni di  $x$ .

Esempio:  $\frac{\|Sx\|}{\|x\|} = 10^{-6}$ ,  $K(x) = 1000$ .

Allora  $\frac{\|Sy\|}{\|y\|} \leq 10^{-3}$ . Il dato ha

6 cifre corrette, la soluzione almeno 3 :

"possiamo perdere 3 cifre"!

Note: il condizionamento può essere  
studiato anche rispetto all'errore

assoluto :

più piccolo  $K(x) > 0$  t.c.  $|f_y| \leq K(x)|f_x|$ ,  
 $|f_x|$  piccolo.

In questo caso  $K(x)$  è essenzialmente  
la derivata delle funzione

$$x \mapsto y(x)$$

voluta in  $x$  (supponendo che questa  
sia derivabile)

Ricchiamo : Polinomio d' Taylor

Teorema : Se  $f: [a, b] \rightarrow \mathbb{R}$

derivabile  $m+1$  volte in  $[a, b]$

e continua in  $[a, b]$  assieme a  
tutte le sue  $m+1$  derivate. Se

$x \in (a, b)$ . Allora,  $\forall x \in (a, b)$

$\exists c \in (a, b)$  t.c.

$$f(x) = T_m(x) + R_m(x), \forall x \in (a, b)$$

dove

$$\begin{aligned} T_m(x) &= f(x_0) + f'(x_0)(x-x_0) + \frac{1}{2} f''(x_0)(x-x_0)^2 + \dots \\ &\dots + \frac{f^{(m)}(x_0)}{m!} (x-x_0)^m \end{aligned}$$

$$R_m(x) = \frac{f^{(m+1)}(c)}{(m+1)!} (x-x_0)^{m+1}$$

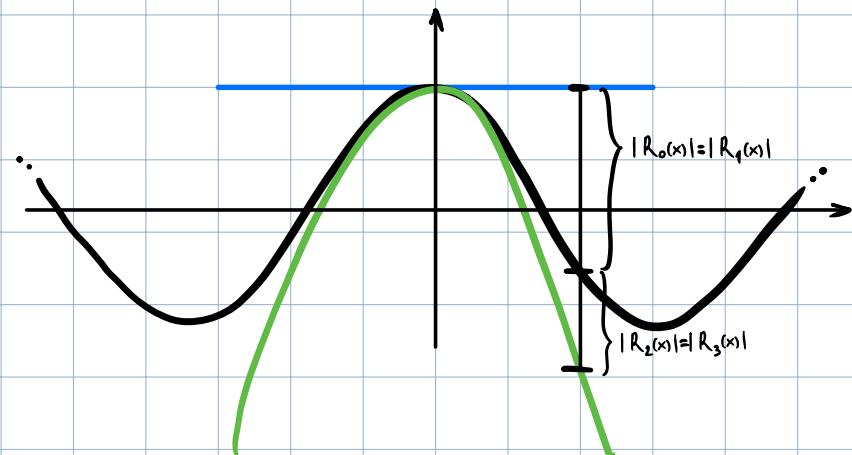
$T_m$  è detto polinomio di Taylor di grado  $m$  centrato in  $x_0$ ;  $R_m$  è detto resto di Lagrange di grado  $m+1$ .

Notare :  $|R_m(x)| = |f(x) - T_m(x)|$

errore assoluto che commettiamo  
nell'approssimare  $f(x)$  con  $T_m(x)$

Esempio:  $f(x) = \cos x$ ,  $x_0 = 0$ .

Allora  $T_0(x) = T_1(x) = 1$ ,  $T_2(x) = T_3(x) = 1 - \frac{x^2}{2}$



Per un ulteriore approfondimento, si vede la raccolta di esercizi sul calcolo degli zeri di funzione.

## STUDIO DEL CONDIZIONAMENTO DELLO ZERO DI UNA FUNZIONE

Problema: data  $f: [a, b] \rightarrow \mathbb{R}$ , determinare  $x$  t.c.  $f(x) = 0$ .

Studiamo l'effetto delle perturbazioni  
su  $f$ .

dato	soluzione
$f$	$\alpha$ t.c. $f(\alpha) = 0$
$\tilde{f} = f + e$ <small>↑ errore</small>	$\tilde{\alpha}$ t.c. $\tilde{f}(\tilde{\alpha}) = 0$

Ipotesi : 1)  $|e(x)| < \varepsilon \leftarrow$  stima sull'errore

2)  $f, \tilde{f}$  continue,  $f(a) f(b) < 0$

Per fissare le idee, supponiamo  $f(a) > 0$

Osserviamo che :

- se  $f(a) > \varepsilon$ , allora  $\tilde{f}(a) > 0$
- se  $f(b) < -\varepsilon$ , allora  $\tilde{f}(b) < 0$

Quindi, se  $f$  è sufficientemente grande

in  $a < b$ , anche  $\tilde{f}(a)\tilde{f}(b) < 0$  e  
dunque  $\tilde{f}$  ha uno zero  $\tilde{x}$  in  $[a, b]$ .

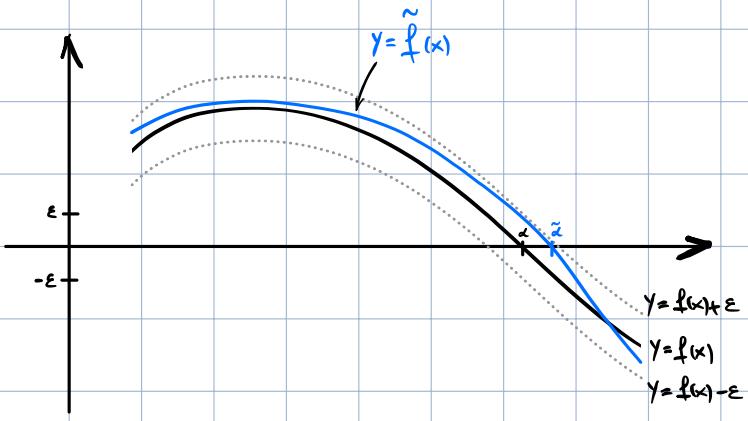
Se invece  $|f(x)| < \varepsilon$  per  $x = a \circ x = b$ ,  
allora il segno di  $f$  può non cambiare  
con quello di  $\tilde{f}$  in  $a \circ b$ , e  
 $\tilde{f}$  potrebbe non avere zero in  $[a, b]$ .

I ruoli di  $f$  e  $\tilde{f}$  possono essere intercambiati.

Conclusione: leggendo  $|f(x)| < \varepsilon$ ,  $\tilde{f}$

potrebbe essere "dominata dall'errore" e dunque  
informazioni errate sugli zeri di  $f$ .

Graficamente:

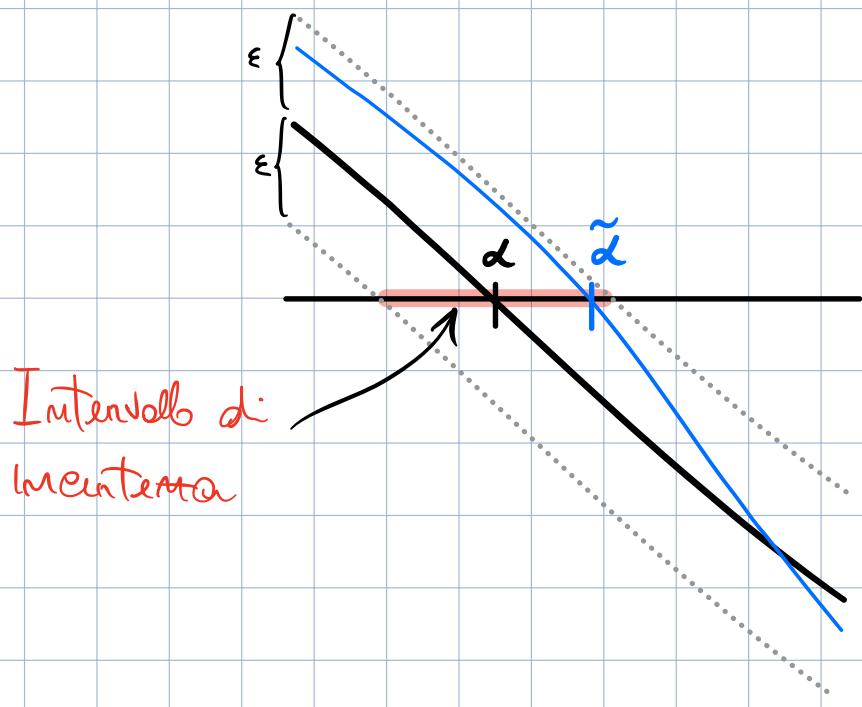


Definizione: il più grande intervallo  $I \subset [a, b]$  contenente  $\alpha$  e t.c.  $|f(x)| < \varepsilon$ ,  $\forall x \in I$ , è detto intervallo di incertezza per  $\alpha$ .

Ottivio di un metodo: fornire una

approssimazione di  $\alpha$  che appartiene all'intervallo di incertezza per  $\alpha$ .

Non possono richiedere maggiorazione precorsore.



Adesso stimeremo la lunghezza dell'int. llo di incertezza. Vorremmo risolvere la d' segugl'ona  $|f(x)| \leq \varepsilon$

Polinomio di Taylor di grado 1 di  $f$  centrato in  $\alpha$ :

$$f(x) = f(\alpha) + f'(\alpha)(x-\alpha) + \underbrace{R_1(x)}_{\text{trascurabile per } x \approx \alpha}$$

$$= 0$$

Da cui

$$f(x) \approx f'(\alpha)(x-\alpha), \text{ per } x \approx \alpha.$$

Dunque per  $x \approx \alpha$  rimpiccioliamo  $|f(x)| \leq \varepsilon$

con  $|f'(\alpha)| |x-\alpha| \leq \varepsilon$ .

Per risolverla bisogna supporre  $f'(\alpha) \neq 0$ .

Introduciamo una

Definizione: Sia  $f : [a, b] \rightarrow \mathbb{R}$ ,  $f$  derivabile in  $(a, b)$  e  $\alpha$  t.c.  $f(\alpha) = 0$ ,  $f'(\alpha) \neq 0$ .

Allora  $\alpha$  è detto per  $f$  simile.

Continuiamo il ragionamento. Se  $\alpha$  perno simile per  $f$ . Allora

$$|f'(\alpha)| |x - \alpha| \leq \varepsilon \iff |x - \alpha| \leq \frac{\varepsilon}{|f'(\alpha)|} \iff$$

$$\iff x \in \left[ \alpha - \frac{\varepsilon}{|f'(\alpha)|}, \alpha + \frac{\varepsilon}{|f'(\alpha)|} \right]$$

stima dell'intervolo  
di incertezza

Sintetizziamo:

incertezza su  $f$ :  $\varepsilon$

incertezza su  $\alpha$ :  $\frac{\varepsilon}{|f'(\alpha)|}$

Ne segue che il fattore di condizionamento (rispetto all'errore assoluto) è

$$K(\alpha) = \frac{1}{|f'(\alpha)|}$$

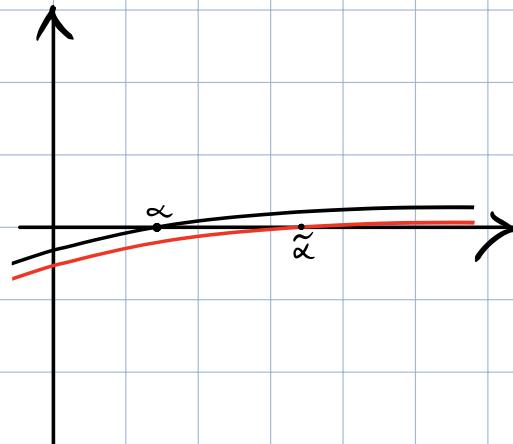
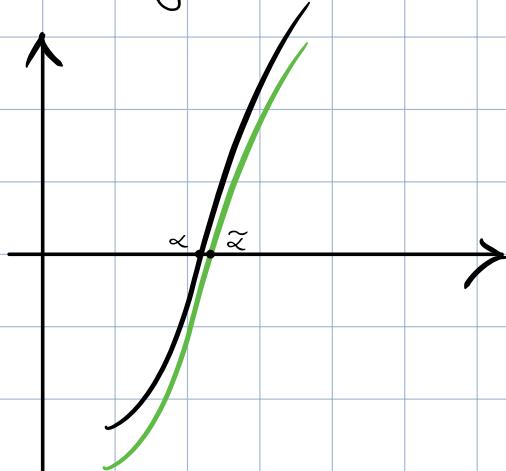
Conclusione:

$$|f'(\alpha)| \geq 1 : \text{zen ben condizionato}$$

"molto più  
presto di"

$$|f'(\alpha)| \ll 1 : \text{Aero Mid condizionato}$$

S' rifletta sulle figure:

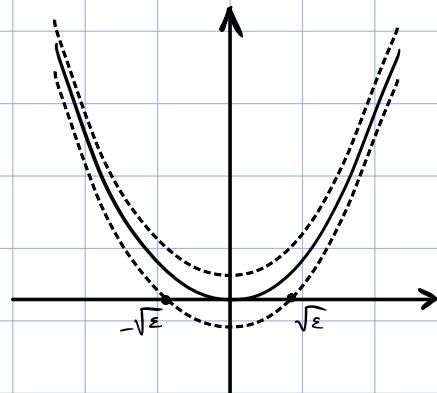


Cose succederebbe se  $\alpha$  non fosse Aero semplice?

ESEMPIO:

$$f(x) = x^2$$

$$\tilde{f}(x) = x - \varepsilon$$



Per  $\varepsilon < 0$  non c'è più soluzione!

Per  $\varepsilon > 0$  abbiamo:  $\lambda = 0$ ,  $\tilde{\lambda} = \pm\sqrt{\varepsilon}$ .

Errore assoluto su  $f$ :  $\varepsilon$

Errore assoluto su  $\lambda$ :  $\sqrt{\varepsilon}$

Se  $0 < \varepsilon < 1$ , si ha  $\sqrt{\varepsilon} > \varepsilon$ .

Se, ad esempio, poniamo  $\varepsilon = 10^{-6}$ ,

allora avremo  $\sqrt{\varepsilon} = 10^{-3}$ , cioè

"errore assoluto sulle soluzioni" = 1000 "errore assoluto sul dato"

Ovvero problema molto condizionato!

Adesso studiamo un modo per quantificare la "Velocità" di un metodo iterativo per il calcolo degli zeri di funzione.

## ORDINE DI CONVERGENZA

Sia  $\{x^{(k)}\}_{k \in \mathbb{N}}$  successione t.c.

$\lim_{k \rightarrow +\infty} x^{(k)} = \alpha \in \mathbb{R}$ . Dico che

$\underbrace{x^{(k)}}_{\substack{\text{"leggi } "x^{(k)}\text{"} \\ \text{converge a } \alpha}} \text{ con ordine di convergenza}$

"leggi "x<sup>(k)</sup>"  
converge a α"

(Odc)  $p \geq 1$  su  $\exists 0 < C < +\infty$  t.c.

$\lim_{k \rightarrow +\infty} \frac{|x^{(k+1)} - \alpha|}{|x^{(k)} - \alpha|^p} = C$ . In tal caso,

C è detto fattore anzitutto di convergenza.

Si utilizza la seguente nomenclatura:

P	C	Convergenza
1	$0 < C < 1$	lineare
1	$C = 1$	sublineare
$> 1$	qualsiasi	superlineare
2	qualsiasi	quadratica
3	qualsiasi	cubica

Osservazione: per "K suff. grande" si ha

$$|x^{(k+1)} - \alpha| \approx C|x^{(k)} - \alpha|^p.$$

Per noi  $|x^{(k)} - \alpha|$  è sempre un numero

più o meno minore di 1, per cui

$|x^{(k)} - \alpha|^p$  è tanto più piccolo quanto più grande è  $p$ .

ESEMPIO: supponiamo  $|x^{(0)} - \alpha| = 10^{-1}$  e  $C = 10^{-1}$ . Allora:

Passo	$p=1$	$p=2$	$p=3$
0	$10^{-1}$	$10^{-1}$	$10^{-1}$
1	$10^{-2}$	$10^{-3}$	$10^{-9}$
2	$10^{-3}$	$10^{-7}$	$10^{-13}$
3	$10^{-4}$	$10^{-15}$	
:	:		
12	$10^{-13}$		

Se ad esempio impostato una tolleranza  $\varepsilon = 10^{-13}$ , i tre metodi avrebbero raggiunto l'affine.  
 Ricorda che in 12, 3 e 2 passi, rispettivamente.  
 In sostanza (tenendo l'effetto di  $C$ ),

Maggior Odc  $\Rightarrow$  minor numero  
 di passi

Note: ovviamente conta anche il costo computazionale per passo.

Domanda: Qual è l'ordine di convergenza aspetto alle successive generate dal metodo delle successive bisezioni?

Sì può dimostrare che generalmente accade

che  $\frac{|x^{(k+1)} - \alpha|}{|x^{(k)} - \alpha|} \rightarrow c$  arbitrariamente

grande / piccolo per infiniti valori di  $k$ , ma  
 cui non esistono  $P \geq 1$  e  $c > 0$  t.c.

$$\lim_{k \rightarrow +\infty} \frac{|x^{(k+1)} - \alpha|}{|x^{(k)} - \alpha|^P} = c.$$

Per un metodo lineare ( $P=1$ ) si può facilmente dimostrare che  $\exists$   
 $c > 0$  t.c., per  $k$  sufficientemente grande,

$$\begin{aligned} |x^{(k)} - \alpha| &\leq c |x^{(k-1)} - \alpha| \leq \\ &\leq c^2 |x^{(k-2)} - \alpha| \leq \dots \\ &\dots \leq c^k |x^{(0)} - \alpha|, \end{aligned}$$

una relazione simile a quella dimostrata per il metodo delle successive bisezioni:

$$|x^{(k)} - \alpha| \leq \left(\frac{1}{2}\right)^k \frac{(b-a)}{2}.$$

Per questo motivo diciamo che "il metodo delle successive bisezioni" è un metodo lineare".

Attenzione: quando attribuiremo (impostiamo)

un OdC ad un metodo, lo faremo  
intendendo da quelli se l'OdC esiste  
generalmente dalle successive generazioni da  
quel metodo.

# Il metodo di Newton

Problema: Risolvere  $f(x) = 0$ .

Idea: "linearizziamo il problema"

Vicino alla soluzione.

Consideriamo  $x^{(0)}$  vicino a  $x$  falso per  $f$ , e scriviamo il polinomio di Taylor:

$$f(x) = f(x^{(0)}) + \frac{f'(x^{(0)})}{1!}(x - x^{(0)}) + \dots + \Theta((x - x^{(0)}))$$

resto

Proseguiamo il resto, e  
risolviamo

$$f(x^{(0)}) + f'(x^{(0)})(x - x^{(0)}) = 0$$

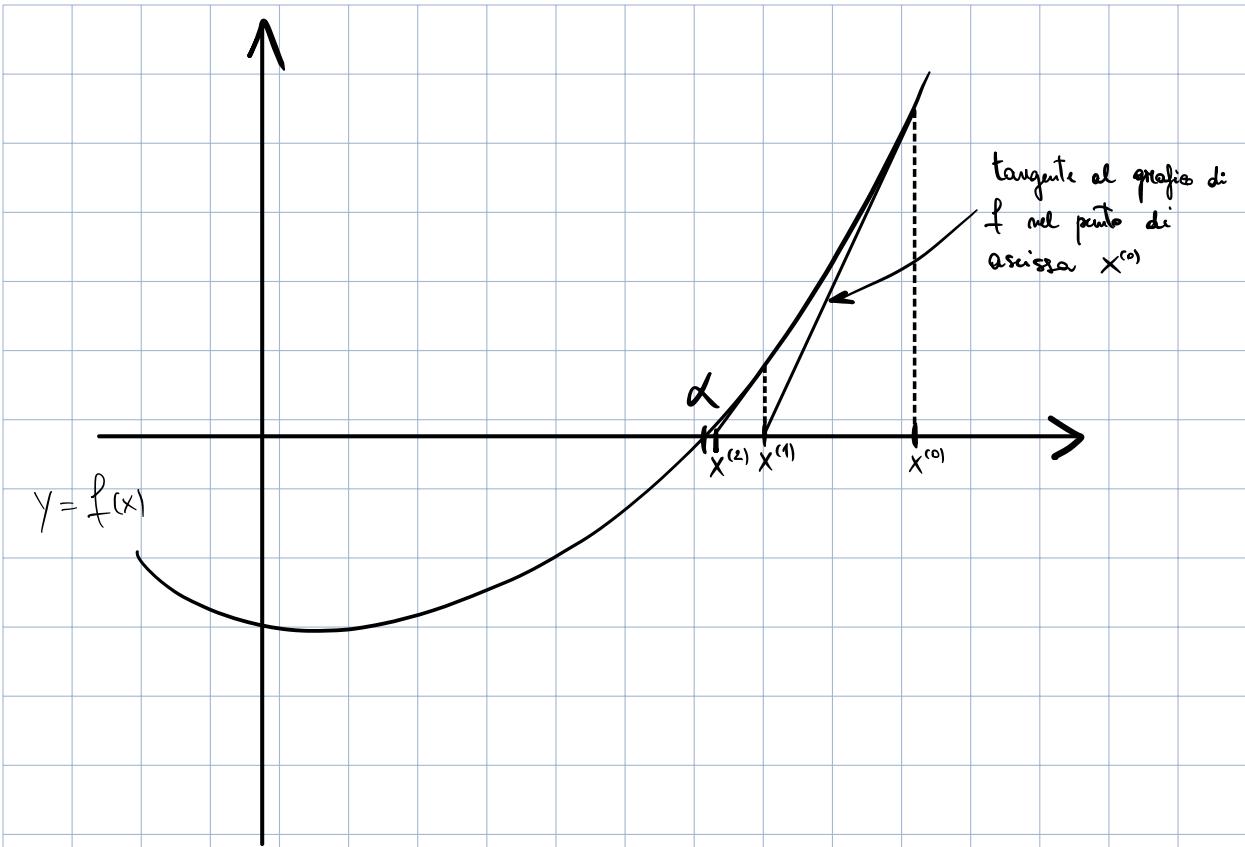
rispetto a  $x$ , se  $f'(x^{(0)}) \neq 0$ .

Denotiamo la soluzione con

$$x^{(1)} : x^{(1)} = x^{(0)} - \frac{f(x^{(0)})}{f'(x^{(0)})}.$$

Procediamo analogamente per  
ottenere  $x^{(2)}, x^{(3)}, \dots$

Questo procedimento ha la seguente  
interpretazione geometrica.



Per questo il metodo è anche detto  
"metodo delle Tangenti".

La formula iterata è

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}, \quad k \geq 0,$$

$f'(x^{(k)}) \neq f'(x^{(k)}) \neq 0.$

Consideriamo il polinomio di Taylor di grado 1 di  $f$  con resto

di Lagrange, centrato in  $x^{(k)}$  e

valutato in  $\alpha$  zero per  $f$ :

$$f(\alpha) = f(x^{(k)}) + f'(x^{(k)})(\alpha - x^{(k)}) +$$

$$+ \frac{f''(c^{(k)})}{2} (\alpha - x^{(k)})^2,$$

con  $c^{(k)}$  nell'intervallo  
di estremi  $\alpha \in x^{(k)}$

Essendo  $f(\alpha) = 0$ , dividendo per

$f'(x^{(k)})$  e moltiplicando i termini:

$$\alpha - \frac{f(x^{(k)})}{f'(x^{(k)})} - \alpha = \frac{f''(c^{(k)})}{2 f'(x^{(k)})} (\alpha - x^{(k)})^2$$

$x^{(k+1)} \rightarrow$

Da cui segue

$$x^{(k+1)} - \alpha = \frac{f''(\alpha^{(k)})}{2f'(\alpha^{(k)})} (\alpha - \alpha^{(k)})^2 \quad [\ast]$$

Cioè suggerisce che, se il fattore

$$\frac{f''(\alpha^{(k)})}{2f'(\alpha^{(k)})}$$
 si mantenga limitato per

$\leftarrow +\infty$ , le succ.  $x^{(k)}$  convergono a  $\alpha$ ,

e la convergenza sarà almeno quadratica.

Formalizziamo l'idea:

Totemma (di Convergenza del metodo di Newton). Se  $f: I \rightarrow \mathbb{R}$  derivabile 2 volte in  $I$  intorno di  $\alpha$ , e siano  $f, f', f''$  continue in  $I$ . Se  $\alpha$  è zero semplice per  $f$  ( $f(\alpha) = 0, f'(\alpha) \neq 0$ ).

Allora, se  $x^{(0)}$  è sufficientemente

vicino a  $\alpha$ , le successive

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}, \quad k \geq 0$$

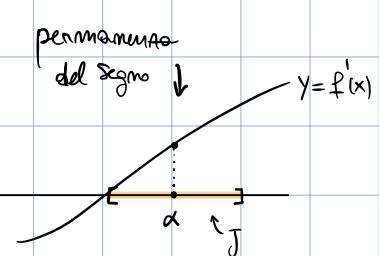
converge a  $\alpha$  con O.d.c. almeno 2.

Dimostrazione:

Dato che  $f'$  è continua e  $f'(\alpha) \neq 0$ ,

$$\exists J = [\alpha - \varepsilon, \alpha + \varepsilon] \text{ t.c.}$$

$$f'(x) \neq 0 \quad \forall x \in J$$



Consideriamo

$$M := \frac{1}{2} \frac{\max_{x \in J} |f''(x)|}{\min_{x \in J} |f'(x)|}.$$

$\exists \epsilon > 0 \in J$  t.c.  $M|x^{(0)} - \alpha| < 1$ .

Essendo  $x^{(0)} \in J$ , dall'eq. me  $[*]$  segue che

$$|x^{(1)} - \alpha| \leq M |x^{(0)} - \alpha|^2,$$

da cui :

$$1) |x^{(1)} - \alpha| \leq \underbrace{M |x^{(0)} - \alpha|}_{< 1} |x^{(0)} - \alpha| < \epsilon \leq \epsilon$$

$$2) M |x^{(1)} - \alpha| \leq (M |x^{(0)} - \alpha|)^2 < 1$$

Ne segue che  $x^{(1)} \in J$  e  $M|x^{(1)} - \alpha| < 1$ .

Per induzione possiamo provare che

$$x^{(k)} \in J \text{ e } M|x^{(k)} - \alpha| < 1, \forall k \geq 0.$$

Dunque, ammetta delle  $[*]$  :

$$|x^{(k)} - \alpha| \leq M |x^{(k-1)} - \alpha|^2 \Rightarrow$$

$$\begin{aligned} \Rightarrow M |x^{(k)} - \alpha| &\leq (M |x^{(k-1)} - \alpha|)^2 \leq \\ &\leq (M |x^{(k-2)} - \alpha|)^4 \leq \dots \leq \underbrace{(M |x^{(0)} - \alpha|)^{2^k}}_{< 1}, \end{aligned}$$

ovvero

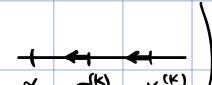
$$0 \leq M |x^{(k)} - \alpha| \leq \underbrace{(M |x^{(0)} - \alpha|)^{2^k}}_{\downarrow \text{ per } k \rightarrow +\infty}.$$

Posiamo dunque concludere che

$$\lim_{k \rightarrow +\infty} x^{(k)} = \alpha.$$

Ancora dalle  $[*]$  abbiamo:

$$|x^{(k+1)} - \alpha| = \underbrace{\frac{1}{2} \frac{f''(c^{(k)})}{f'(x^{(k)})}}_{\downarrow \text{ per } k \rightarrow +\infty} |x - x^{(k)}|^2$$

( poiché  $f', f''$  continue e  
 $f'(\alpha) \neq 0$ , e 
)

Né segue da :

$$\lim_{k \rightarrow +\infty} \frac{|x^{(k+1)} - \alpha|}{|x^{(k)} - \alpha|^2} = \frac{1}{2} \left| \frac{f''(\alpha)}{f'(\alpha)} \right|,$$

e quindi :

- $f''(\alpha) \neq 0 \Rightarrow$  O.d.c. 2

- $f''(\alpha) = 0 \Rightarrow$  O.d.c. > 2



Criteri di arresto: come stimare l'errore assoluto?

Abbiamo visto che, scegliendo  $x^{(0)}$  suff. vicino a  $\alpha$ , si ha:

$$|x^{(K+1)} - \alpha| \leq M |x^{(K)} - \alpha|^2, K \geq 0$$

dove  $M = \frac{1}{2} \frac{\max_{x \in J} |f''(x)|}{\min_{x \in J} |f'(x)|}$ ,  $J$  opportuno intorno di  $\alpha$  (vedere Teo. di convergenza del metodo di Newton).

Umetti le due diseguaglianze:

$$|x^{(K)} - \alpha| = |x^{(K)} - x^{(K+1)} + x^{(K+1)} - \alpha| \leq$$

$$\leq |x^{(K)} - x^{(K+1)}| + |x^{(K+1)} - \alpha| \leq$$

$$\leq |x^{(k+1)} - x^{(k)}| + M |x^{(k)} - \alpha|^2$$

Ricondimando :

$$(1 - M|x^{(k)} - \alpha|) |x^{(k)} - \alpha| \leq |x^{(k+1)} - x^{(k)}|$$

scrivere strettamente  
positivo se  $x^{(0)}$   
è scelto opportunamente

Dunque :

$$|x^{(k)} - \alpha| \leq \frac{1}{1 - M|x^{(k)} - \alpha|} |x^{(k+1)} - x^{(k)}|$$

$$\underbrace{1 - M|x^{(k)} - \alpha|}_{\text{Tende a } 1}$$

per  $k \rightarrow +\infty$

Quindi asintoticamente <sup>(\*)</sup> :

$$|x^{(k)} - \alpha| \approx |x^{(k+1)} - x^{(k)}|$$

↓  
 errore al  
 passo k      ↑  
 passo corrente

"la stima ha un passo di ritardo"

(\*) Formalmente:

$$\forall \varepsilon > 0 \exists N \geq 0 \text{ t.c.}$$

$$|x^{(k)} - \alpha| \leq (1 + \varepsilon) |x^{(k+1)} - x^{(k)}|, \\ \text{per } k \geq N$$

Criteri di arresto

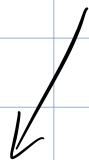
$$1) |x^{(k+1)} - x^{(k)}| \leq \text{Tol} \quad (\text{errore assoluto})$$

$$2) \frac{|x^{(k+1)} - x^{(k)}|}{|x^{(k+1)}|} \leq \text{Tol} \begin{pmatrix} \text{errore} \\ \text{relativo} \end{pmatrix}$$

$$3) \frac{|x^{(k+1)} - x^{(k)}|}{|x^{(k+1)}| + 1} \leq \text{Tol} \begin{pmatrix} \text{errore} \\ \text{"misto"} \end{pmatrix}$$

Si "polarità" su errore  $\rightarrow$  <sup>relativo</sup>  
 $\rightarrow$  <sup>assoluto</sup> per  $\alpha$   $\rightarrow$  <sup>grande</sup>  
 $\rightarrow$  <sup>piccolo</sup>

$$4) |f(x^{(k+1)})| \leq \text{Tol} \begin{pmatrix} \text{errore} \\ \text{relativo} \end{pmatrix}$$



Ma se siamo di  
 puo' sotto stimare o  
 soverstimare l'errore  
 assoluto

## Osservazioni pratiche

(1) Verifichiamo che  $f(x^{(k)}) \neq 0$

(oppure  $f(x^{(k)})$  "non troppo piccolo")

$$\forall k \geq 0$$

(2) Se "Spettiamo" il passo in due:

$$dx^{(k)} = -\frac{f(x^{(k)})}{f'(x^{(k)})},$$

↑ incremento

$$x^{(k+1)} = x^{(k)} + dx^{(k)}$$

Motiamo:  $|dx^{(k)}| = |x^{(k+1)} - x^{(k)}|$

↑ stima dell'errore assoluto

Ricchiamiamo il metodo di Newton:

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}, k \geq 0$$

PSEUDOCODICE

Input:  $f, f', x^{(0)}, tol, mitmax$

1.  $mit = 1$

2.  $dfx \leftarrow f'(x^{(0)})$

3. se  $dfx = 0$ , uscisci con errore

4.  $\Delta x = -f(x^{(0)})/dfx$

$x^{(1)} = x^{(0)} + \Delta x$

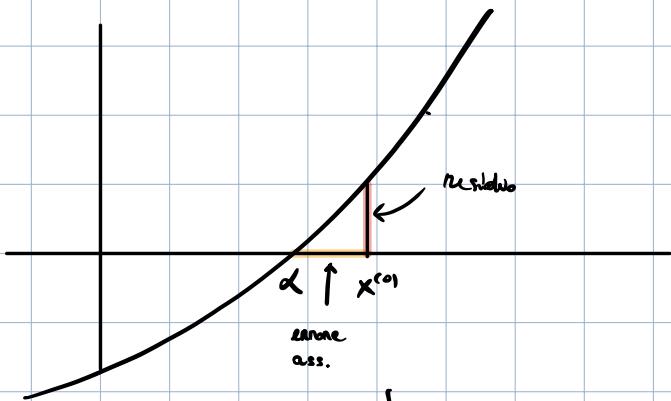
5. Se  $|\Delta x| \leq tol$ , uscisci con  $\alpha = x^{(1)}$

6. se  $mit = mitmax$ , uscisci con errore

7.  $mit = mit + 1, x^{(0)} \leftarrow x^{(1)}$

8. torna al punto 2.

Output:  $\alpha$  approssimazione dello zero di  $f$



"residuo"  $\approx f(\alpha) \cdot$  "errore ass."

$$\text{Se } f'(x) = 1$$

"residuo"  $\approx$  "errore ass."

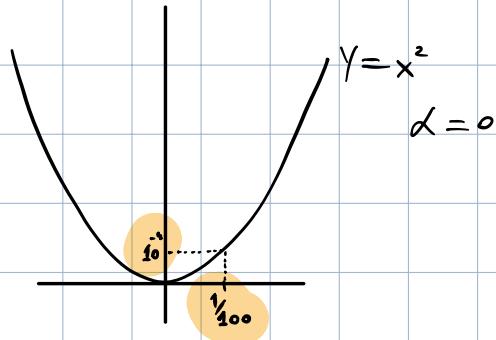
Ricchiamo:

stimiamo  $|x^{(k)} - \alpha|$  con  $|x^{(k+1)} - x^{(k)}|$

$$\text{Se } \frac{|x^{(k+1)} - \alpha|}{|x^{(k)} - \alpha|^2} \rightarrow C \quad , \quad \text{allora}$$

$k \rightarrow +\infty$

$$|x^{(k+1)} - \alpha| \approx C |x^{(k)} - \alpha|^2 \text{ per } k \text{ suff. grande!}$$



Verificare sperimentalmente:

1) Ateno doppio (non semplice): O.d.C. 1

2) Ateno p.t. di flussi: O.d.C. 3

OSSERVAZIONE:

Se  $|x^{(k+1)} - \alpha| \approx C |x^{(k)} - \alpha|^p$ ,  $k$  grande,

Passando ai logaritmi:

$$\log(|x^{(k+1)} - \alpha|) \approx \log(C |x^{(k)} - \alpha|^p) =$$

$$= \log C + p \log |x^{(k)} - \alpha|, \text{ da cui}$$

divido per  $\log |x^{(k)} - \alpha|$  e riorganizzo:

$$p \approx \frac{\log |x^{(k+1)} - \alpha|}{\log |x^{(k)} - \alpha|} - \frac{\log C}{\log |x^{(k)} - \alpha|} \leftarrow \begin{matrix} \text{infinitesima} \\ \text{per } k \rightarrow +\infty \end{matrix}$$

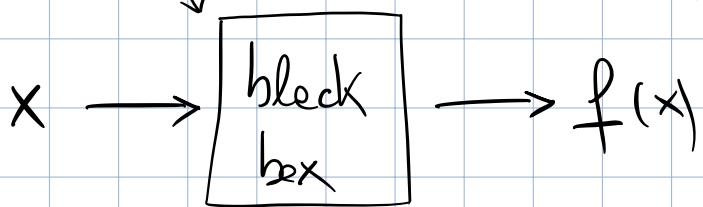
quindi per K grande:

$$P \approx \frac{\log |X^{(k+1)} - \alpha|}{\log |X^{(k)} - \alpha|}$$

Eperimenti in Matlab

Il metodo di Newton richiede la valutazione di  $f'$  ad ogni passo. Ciò potrebbe essere impossibile oppure troppo oneroso.

Ese.

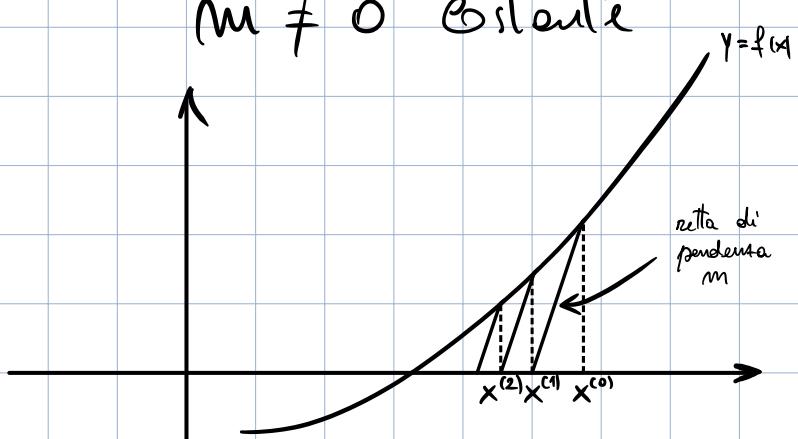


Varianti "derivative free" del metodo di Newton :

(1) metodo delle 3onde :

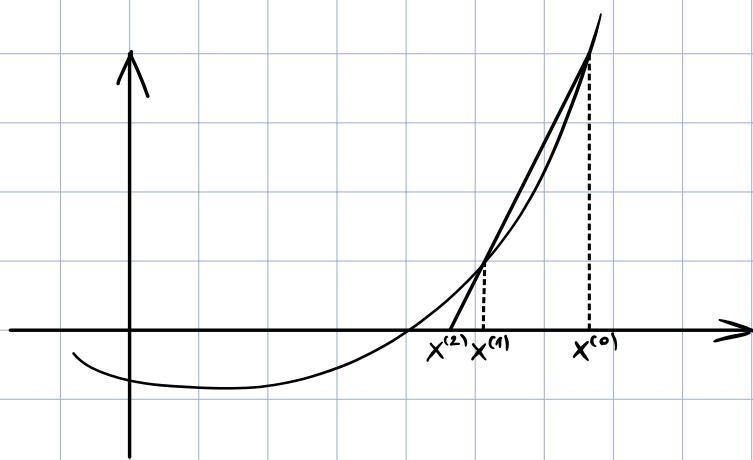
$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{m}, \quad k \geq 0,$$

$m \neq 0$  costante



- è generalmente lineare
- scelta ricorrente:  $\mu = f'(x^{(0)})$ , tanto efficace quanto più  $x^{(0)}$  è vicino a  $\alpha$

## (2) metodo delle secanti:



Iterata generale:

$$x^{(k+1)} = x^{(k)} - f(x^{(k)}) \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})}, \quad k \geq 1.$$

- osserviamo che occorrono due stime iniziali:  $x^{(0)}$  e  $x^{(1)}$ .
- si può dimostrare che, per  $x^{(0)}, x^{(1)}$  suff. vicini a  $\alpha$  e no simile per  $f$ , il metodo genera una successione  $x^{(k)}$  che converge a  $\alpha$  con o.d.c. (dimessi)  $\frac{1+\sqrt{5}}{2} \approx 1.6$

(3) ques) - Newton :

$$\begin{cases} d^{(k)} = \frac{f(x^{(k)} + h) - f(x^{(k)})}{h}, & \begin{matrix} h \neq 0 \text{ scelto} \\ \text{"opportunamente"} \\ \text{perob} \end{matrix} \\ x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{d^{(k)}} \end{cases}$$

$$k \geq 0$$

- è "Tecnicamente" lineare ma in pratica si computa come il metodo di Newton se  $h$  è scelto opportunamente

## Confronto fra i metodi

Consideriamo solo il costo  
Computazionale delle valutazioni di  
funzione:

metodo	# valutazioni di funzione per passo	O.d.c. attesa
bisezione	1f	1
regula falsi	1f	1
Newton	1f + 1f'	2
Conde secanti	1f	1
quasi-Newton	1f	$\approx 1.6$
	2f	1

↙

S' osserva ordine 2 per h scelto opportunamente

## Newton vs Secanti

Ipotesi: Valutare  $f'$  costi Tanto quanto  
Valutare  $f$ . Allora

1 passo di Newton costa quanto 2  
passi di secanti.

Osserviamo che, per secanti, si ha

$$\begin{aligned} |x^{(k+2)} - \alpha| &\approx C |x^{(k+1)} - \alpha|^{1.6} \approx \\ &\stackrel{\text{delle def.}}{\approx} C \left( C |x^{(k)} - \alpha|^{1.6} \right)^{1.6} \\ &= C^{2.6} |x^{(k)} - \alpha|^{2.56} \quad \text{quindi} \end{aligned}$$

Secanti "2 passi alla Volta" ha  
ordine di Convergenza 2.56, e  
quindi è "più veloce" di Newton.

Un ragionamento più sofistico porta a  
concludere che se senti è più "efficiente"  
di Newton se: " $\text{Costo } f'$ " > 44% " $\text{Costo } f$ ".

Gambi sui metodi iterativi ed  
un passo (one-step).

Il metodo di Newton

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} \quad , \quad k \geq 0$$

è un caso particolare di "metodo  
iterativo ad un passo":

$$x^{(k+1)} = g(x^{(k)}), \quad k \geq 0$$

in cui si "itera" sempre la stessa funzione  $g$ . Per Newton:

$$g(x) = x - \frac{f(x)}{f'(x)}.$$

Più precisamente, a partire da  $x^{(0)}$ , detta una funzione  $g$ , definiamo per ricorrenza le successive

$$x^{(k+1)} = g(x^{(k)}), \quad k \geq 0.$$

Diciamo che  $\{x^{(k)}\}_{k \geq 0}$  è definita

"Tenendo" la funzione  $g$ , da è detta "funzione iterativa".

## Osservazioni :

(1) Supponieren wir  $X^{(k)} \rightarrow x$ ,  
 d.h. sie continua im  $x$ .

Passiamo al limite per  $K \rightarrow +\infty$ :

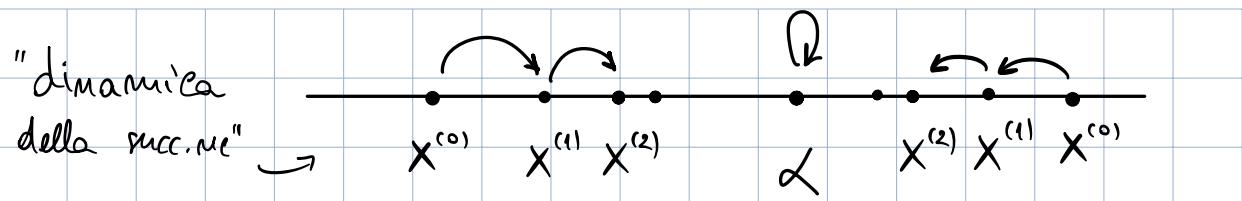
$$x^{(k+1)} = g(x^{(k)})$$

continuité

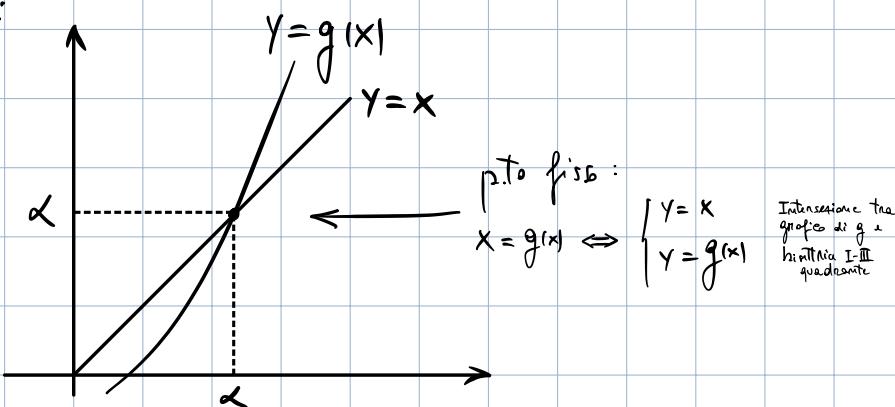
Ne deduciamo da

$$\alpha = g(\zeta)$$

Mel quid esse & è detto: "punto fiso per  $g$ ". Perdù? Interpretare le seguenti figure.



Graficamente :



Ci sono molti modi di tradurre il problema  $f(x) = 0$  in un problema equivalente  $g(x) = x$ .

Esempi :  $f(x) = x^2 - 2$ .

$$(1) \quad f(x) = 0 \Leftrightarrow x^2 - 2 = 0 \Leftrightarrow$$

$$\Leftrightarrow \underbrace{x^2 + x - 2}_{g(x)} = x \Leftrightarrow g(x) = x$$

$$g(x)$$

$$(2) \quad x^2 - 2 = 0 \iff x^2 = 2 \iff x = \sqrt{2} \quad \text{s.t. } x \neq 0$$

$$\iff x = \frac{2}{x} \quad \Leftrightarrow g(x) = x$$

$\underbrace{\phantom{x = \frac{2}{x}}}_{g(x)}$

$$(3) \quad g(x) = \frac{1}{2} \left( x + \frac{2}{x} \right); \quad \text{dimostrare}$$

per esercizio che

$$g(x) = x \iff x^2 - 2 = 0$$

$\text{s.t. } x \neq 0$

OSSERV. sulla funzione Teorema del  
mетод di Newton:

$$x = g(x) = x - \frac{f(x)}{f'(x)} \iff f(x) = 0$$

$\underbrace{\phantom{x = g(x) = x - \frac{f(x)}{f'(x)}}_{x \text{ p.t. fisso per } g}}$

$\uparrow$   
 $\text{s.t. } f'(x) \neq 0$

Mettiamo i 3 metodi alla prova

$K$	(1)	(2)	(3)
0	2	2	2
1	4	1	$\frac{3}{2} = 1.5$
2	18	$\begin{pmatrix} 2 \\ 1 \end{pmatrix}$	$\frac{17}{12} = 1.\overline{416}$
3	340		$\frac{577}{408} = 1.414215\dots$
:	$\downarrow$	"ciclo periodico"	$\downarrow$
	$+\infty$		$\sqrt{2} = 1.414213\dots$

La scelta di  $g$  è fondamentale. Notare che il metodo (3) esibisce convergenza quadratica... perché è il metodo di Newton!

Considerazioni Teoriche: affinché

$x^{(k+1)} = g(x^{(k)})$  se è ben definita, occorre

che  $x^{(k)}$  si trovi nel dominio di  $g$   $\forall k \geq 0$ .

Formalizziamo. Se  $g: [a, b] \rightarrow \mathbb{R}$ .

Sia  $g(x) \in [a, b] \quad \forall x \in [a, b]$ , diremo che

" $g$  manda  $[a, b]$  in se stesso", e

se ne veniamo  $g([a, b]) \subset [a, b]$ .

Sia  $g$  manda  $[a, b]$  in se stesso e

$x^{(0)} \in [a, b]$ , la succ.  $\left\{ x^{(k+1)} = g(x^{(k)}) \right\}_{k \geq 0}$

è ben definita.

TEOREMA (esistenza del pto fiss)

Sia  $g: [a, b] \rightarrow \mathbb{R}$  continua che

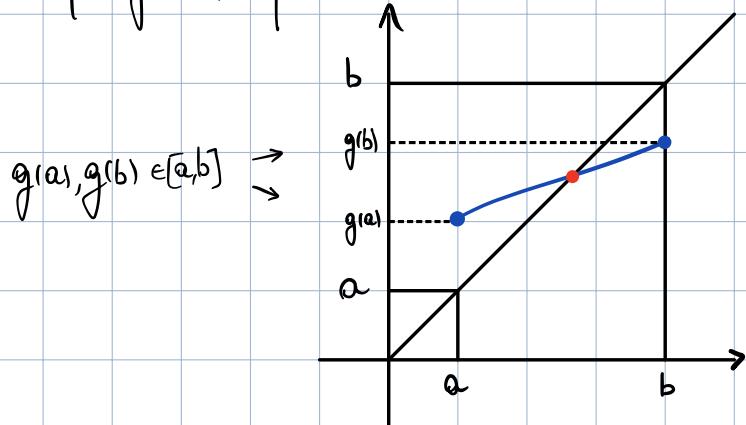
Mandi  $[a, b]$  in se stesso. Allora  $\exists \alpha$

in  $[a, b]$  t.c.  $g(\alpha) = \alpha$ .

Dimostrazione si applichi il Teorema di

Bolzano a  $f(x) = g(x) - x$ .  $\square$

"proof by picture"



Sotto ipotesi aggiuntive possiamo ottenere più informazioni:

Teorema Se  $g: [a,b] \rightarrow \mathbb{R}$  continua,  $g([a,b]) \subset [a,b]$ . Supponiamo esista  $0 < \lambda < 1$

t.c.

$$|g(x) - g(y)| \leq \lambda |x - y|, \quad \forall x, y \in [a, b],$$

(nel qual caso  $g$  è detta "contrattiva" in  $[a, b]$  e  $\lambda$  è detta "costante di contrazione").

Allora:

$$(1) \exists \alpha \in [a, b] \text{ t.c. } g(\alpha) = \alpha$$

$$(2) \forall x^{(0)} \in [a, b], \quad x^{(k+1)} = g(x^{(k)}) \underset{k \rightarrow +\infty}{\longrightarrow} \alpha.$$

$$(3) \forall k \geq 0 : |x^{(k)} - \alpha| \leq \frac{\lambda^k}{1-\lambda} |x^{(1)} - x^{(0)}|.$$

Dimo (cenni)

(1) L'esistenza è conseguenza del Teorema precedente.

L'unicità si dimostra per escludo:

se esistessero  $\alpha \neq \beta$  t.c.  $g(\alpha) = \alpha$  e  $g(\beta) = \beta$ ,

Allora

$$0 < |\alpha - \beta| = |g(\alpha) - g(\beta)| \leq \lambda |\alpha - \beta| < |\alpha - \beta|,$$

assurdo.

(2) Ricordando che  $x^{(k+1)} = g(x^{(k)})$  e  $\alpha = g(\alpha)$ :

$$\begin{aligned} 0 &\leq |x^{(k+1)} - \alpha| = |g(x^{(k)}) - g(\alpha)| \leq \\ &\leq \lambda |x^{(k)} - \alpha| \leq \dots \leq \lambda^k |x^{(0)} - \alpha| \xrightarrow[k \rightarrow +\infty]{} 0. \end{aligned}$$

(3)

$$\begin{aligned} |x^{(0)} - \alpha| &= |x^{(0)} + x^{(1)} - x^{(1)} - \alpha| \leq \\ &\quad \text{aggiungendo e sottraendo} \quad \text{disug. triangolare} \\ &\leq |x^{(0)} - x^{(1)}| + |x^{(1)} - \alpha| \leq \\ &\quad \text{disug. al punto (2)} \\ &\leq |x^{(1)} - x^{(0)}| + \lambda |x^{(0)} - \alpha|. \end{aligned}$$

Ridistribuendo gli addendi:

$$\begin{aligned} (1-\lambda) |x^{(0)} - \alpha| &\leq |x^{(1)} - x^{(0)}| \iff \\ |x^{(0)} - \alpha| &< \frac{1}{1-\lambda} |x^{(1)} - x^{(0)}| \quad \text{perché } 1-\lambda > 0 \end{aligned}$$

Squartando ancora le disug. in (2) :

$$|x^{(k)} - \alpha| \leq \lambda^k |x^{(0)} - \alpha| \leq \frac{\lambda^k}{1-\lambda} |x^{(1)} - x^{(0)}|. \quad \square$$

### OSSERVATIONI

il punto (3) è una stima dell'  
errore assoluto

$$\hookrightarrow |x^{(k)} - \alpha| \leq \frac{\lambda^k}{1-\lambda} |x^{(1)} - x^{(0)}|;$$

per  $k=1$ , diventa

$$|x^{(1)} - \alpha| < \frac{\lambda}{1-\lambda} |x^{(1)} - x^{(0)}|.$$

Dato che  $x^{(1)}$  sta a  $x^{(0)}$  come  $x^{(k+1)}$  sta

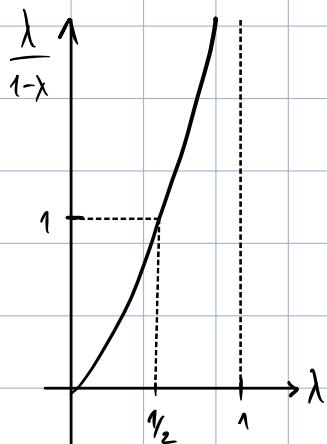
a  $x^{(k)}$ , possiamo scrivere:

$$|x^{(k+1)} - \alpha| < \frac{\lambda}{1-\lambda} |x^{(k+1)} - x^{(k)}|, \quad \forall k \geq 1.$$

↑ errore al  
passo corrente

Questa è una stima utile nella pratica.

Notiamo che



- $\frac{\lambda}{1-\lambda} \leq 1 \quad \text{se} \quad \lambda \in [0, \gamma_2]$
- $\frac{\lambda}{1-\lambda} > 1 \quad \text{se} \quad \lambda > \gamma_2$

Conclusione: se  $0 \leq \lambda \leq \gamma_2$ , allora

$$|x^{(k+1)} - x| < |x^{(k+1)} - x^{(k)}|, \text{ ovvero}$$

la distanza tra i punti consecutivi **sopra stima** l'errore assoluto.

Se, invece,  $\lambda > \gamma_2$ , allora la distanza tra i punti consecutivi può **sotto stima** l'errore assoluto.

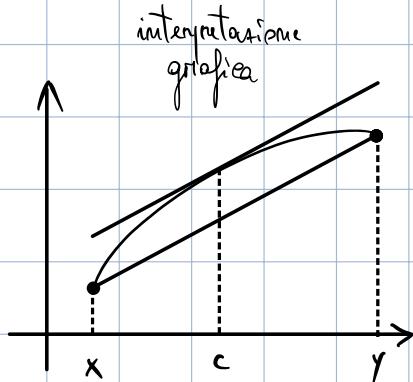
L'ultimo Teorema garantisce la convergenza, ma sotto ipotesi difficili da verifica.

Ricchiamiamo il **Teorema di Lagrange**:

Sia  $g: [a, b] \rightarrow \mathbb{R}$  derivabile in  $[a, b]$ . Allora

per ogni  $x, y \in [a, b]$   $\exists c$  compreso tra  $x$  e  $y$  t.c.

$$\frac{g(x) - g(y)}{x - y} = g'(c)$$



Grafici al Teorema di Lagrange possono ricordarne parte delle ipotesi del Teorema a pag. 1 a proprietà delle derivate prime di  $g$ .

Teorema Se  $g: [a, b] \rightarrow \mathbb{R}$  derivabile con derivata continua, e  $g([a, b]) \subset [a, b]$ . Sappiamo inoltre che

$$\lambda := \max_{x \in [a, b]} |g'(x)| < 1 .$$

Allora vale le Tesi del Teorema a pag. 1,

e inoltre

$$\frac{x^{(k+1)} - \alpha}{x^{(k)} - \alpha} \xrightarrow{k \rightarrow +\infty} g'(\alpha) .$$

Dimostrazione: Per ogni  $x, y$  in  $[a, b]$ , come

Conseguente del Teo. di Lagrange, si ha che

$$|g(x) - g(y)| = |g'(c)| |x - y| \leq \lambda |x - y|,$$

$c$  tra  $a, b$

con  $\lambda < 1$  per ipotesi. Dunque  $g$  è contrattiva in  $[a, b]$  e segue la Tesi. Inoltre, scrivere per il Teorema di Lagrange, si ha:

$$x^{(k+1)} - \alpha = g(x^{(k)}) - g(\alpha) = g'(c^{(k)})(x^{(k)} - \alpha),$$

da cui, per  $k \rightarrow +\infty$ , considerando che  $c^{(k)} \rightarrow \alpha$ ,

segue

$$\frac{x^{(k+1)} - \alpha}{x^{(k)} - \alpha} \xrightarrow{k \rightarrow +\infty} g'(\alpha).$$

□

Corollario: Nelle ipotesi del Teorema alla pagina precedente,

se  $g'(\alpha) \neq 0$ , allora  $x^{(k)} \rightarrow \alpha$  linearmente.

Se, invece  $g'(\alpha) = 0$ , allora  $x^{(k)} \rightarrow \alpha$  in modo superlineare.

Le ipotesi del Teorema precedente restano

dificili da verificare. Si possono modificare come segue.

Teorema. Se  $g: [a,b] \rightarrow \mathbb{R}$  derivabile. Se

a punto fisso per  $g$  in  $[a,b]$  e si suppaga che  $|g'(x)| < 1$ . Allora  $x^{(k+1)} = g(x^{(k)}) \rightarrow x$  per  $x^{(0)}$  suff. vicino a  $x$ .

Dim. Se  $|g'(x)| < 1$ , allora  $\exists \varepsilon > 0$  t.c.

$$|g'(x)| \leq \lambda < 1 \quad \forall x \in I = [x-\varepsilon, x+\varepsilon].$$

Resta da dimostrare che  $g$  manda  $I$  in se stessa. Se  $y \in g(I)$ . Allora  $\exists x \in I$  t.c.

$g(x) = y$ , e si ha:

Teo. di Lagrange

$$|y - x| = |g(x) - g(x)| \stackrel{\text{Teo. di Lagrange}}{\leq} \lambda |x - x| < |x - x| \leq \varepsilon,$$

e dunque  $y \in I$ . Ne segue che  $g(I) \subset I$  e vale

la Tesi del Teorema precedente (pag. 5) per  $x^{(0)}$  in  $I$  ("suff. vicino").



## Osservazioni:

Se  $|g'(x)| < 1$ , allora  $\alpha$  "attrae" i punti

$x^{(0)}$  e i suoi suff. vicini. Al contrario, se

$|g'(\alpha)| > 1$ ,  $x^{(k+1)} = g(x^{(k)})$  non può convergere

e  $\alpha$  poiché per  $k$  suff. grande si avrebbe

$$|x^{(k+1)} - \alpha| \approx |g'(\alpha)| |x^{(k)} - \alpha| > |x^{(k)} - \alpha|,$$

cioè  $\alpha$  "repellerà" i punti a lui suff. vicini. Ciò giustifica la seguente nomenclatura:

$|g'(\alpha)| < 1 \Rightarrow \alpha$  p.t. fisso attrattivo

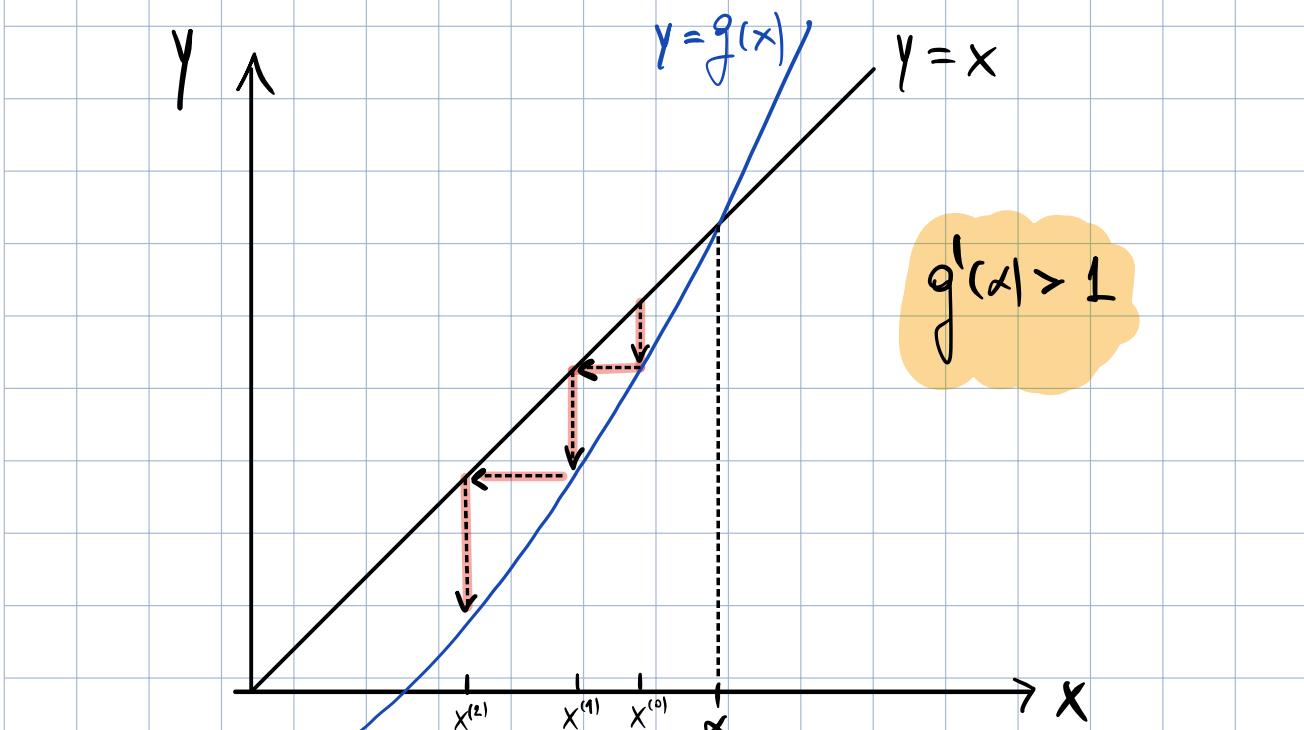
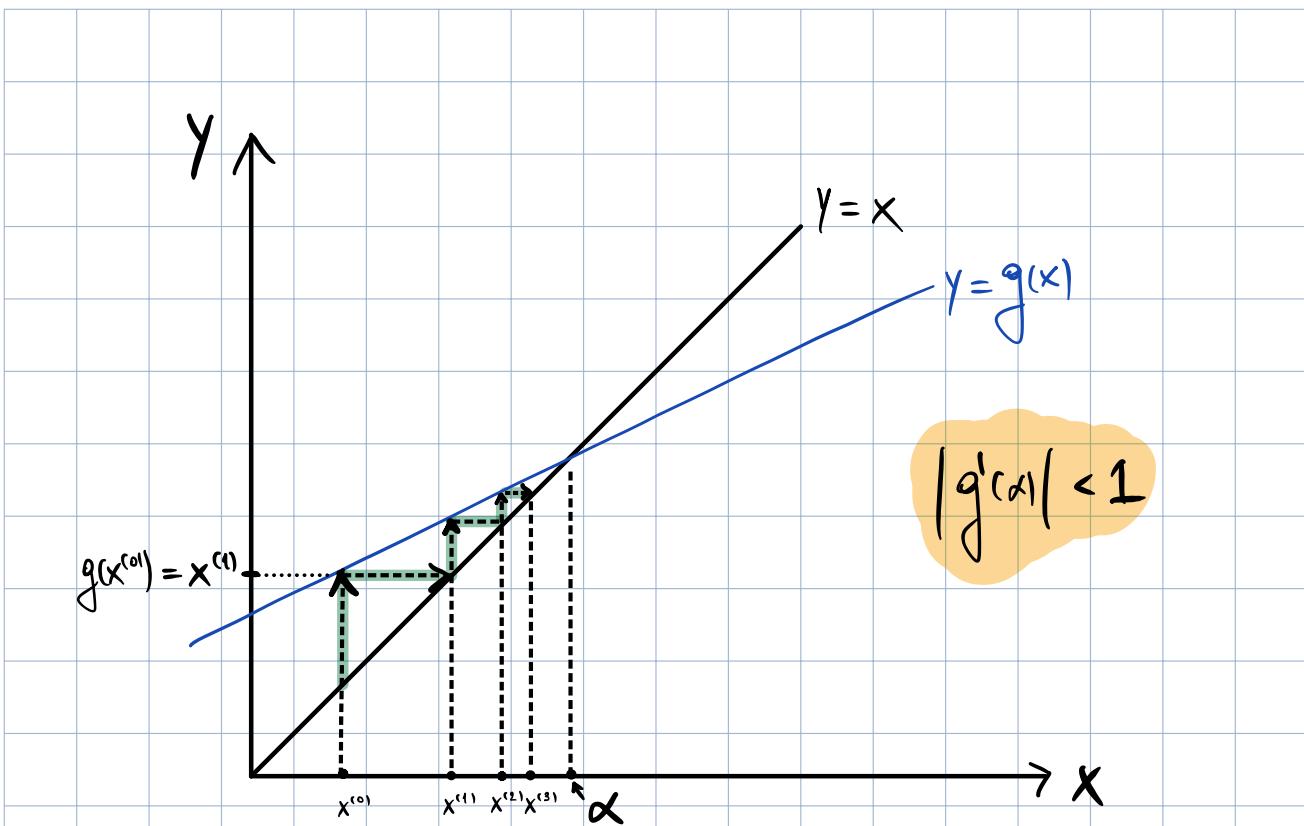


$|g'(\alpha)| > 1 \Rightarrow \alpha$  p.t. fisso repulsivo



N.B.  $|g'(\alpha)| = 1$  indeterminato.

Graficamente (sob il caso  $g'(x) > 0$ )



Vediamo un ultimo Teorema che offre condizioni sufficienti a garantire un certo ordine di convergenza.

Teorema (O.d.C. delle successioni definite per ricorrenza). Se  $g$  derivabile  $p \geq 2$

volte in un intorno di  $\alpha$  punto fisso per

$g$ . Se  $g'(\alpha) = \dots = g^{(p-1)}(\alpha) = 0$

e  $g^{(p)}(\alpha) \neq 0$ , allora, per  $x^{(0)}$

suff. vicino a  $\alpha$ ,  $x^{(k+1)} = g(x^{(k)}) \xrightarrow[k \rightarrow +\infty]{} \alpha$

con O.d.C.  $p$ .

Dimostrazione

Taylor centrato in  $\alpha$  e valutato in  $x^{(k)}$ :

$$g(x^{(k)}) = g(\alpha) + \cancel{g'(\alpha)(x^{(k)} - \alpha)} + \dots + \cancel{\frac{g^{(P-1)}(\alpha)}{(P-1)!}(x^{(k)} - \alpha)^{P-1}} +$$

$\underbrace{+ \frac{g^{(P)}(c^{(k)})}{P!}(x^{(k)} - \alpha)^P}$ , con  $c^{(k)}$  tra  $x^{(k)}$  e  $\alpha$ .

Quindi :

$$x^{(k+1)} - \alpha = \frac{g^{(P)}(c^{(k)})}{P!}(x^{(k)} - \alpha)^P.$$

Per ipotesi,  $|g'(\alpha)| = \sigma < 1$ . Quindi,

dal Teorema precedente,  $x^{(k)} \rightarrow \alpha$  per

$x^{(0)}$  suff. vicino a  $\alpha$ , e anche  $c^{(k)} \rightarrow \alpha$ .

Perciò

$$\frac{|x^{(k+1)} - \alpha|}{|x^{(k)} - \alpha|^P} \rightarrow \frac{|g^{(P)}(\alpha)|}{P!} \neq 0$$

Ne segue che  $x^{(k)} \rightarrow x$  con O.d.c. p.



Osservazione :

Se  $g(x) = x$ ,  $g'(x) = 0$ ,  $g''(x) \neq 0$ .

Allora  $x^{(k+1)} = g(x^{(k)}) \rightarrow x$  con O.d.c. 2.

Esercizio : Applicare il Teorema precedente

al metodo di Newton per ottenere il

relativo Teorema di Convergenza.

Osservazione : L'ordine di convergenza del metodo

iterativi ad un punto è sempre un intero.

Esempio: se  $f'(x) \neq 0$ ,  $f''(x) = 0$ , il metodo di Newton

ha ordine di convergenza (almeno) 3.

Osservazione: Sezioni he O.d.c.  $\approx 1.6$ .

E' possibile perdere a due passi:

$$x^{(k+2)} = g(x^{(k)}, x^{(k+1)}) .$$

Esempio:

(1) Studiare l'O.d.c. del metodo di Newton applicato a  $f(x) = \tan x - e$   
 $x_0 = 0$ .

(Basta considerare  $g(x) = x - \frac{f(x)}{f'(x)}$  e  
evidere il più preciso  $P \geq 2$  t.c.

$$g^{(P-1)}(0) = 0 \text{ ma } g^{(P)}(0) \neq 0)$$

Esempio: determinare le condizioni

per la convergenza del metodo

delle corde. Qual è l'ordine di convergenza atteso? Sotto quali ipotesi la convergenza è quadratica?

Esercizio: determinare l' O.d.c.  
ottese per il metodo quas'-Newton.  
Come mai sperimentalmente il metodo  
risulta una progressione dell'errore  
compatibile con un metodo d'ordine  
2?

Ricchiamo :  $\{x^{(k)}\} \rightarrow x$  con O.d.C.

$P \geq 1$  se

$$\frac{|x^{(k+1)} - x|}{|x^{(k)} - x|^P} \rightarrow C \neq 0$$

Esercizio : data  $f(x) = x^3 - x$ ,

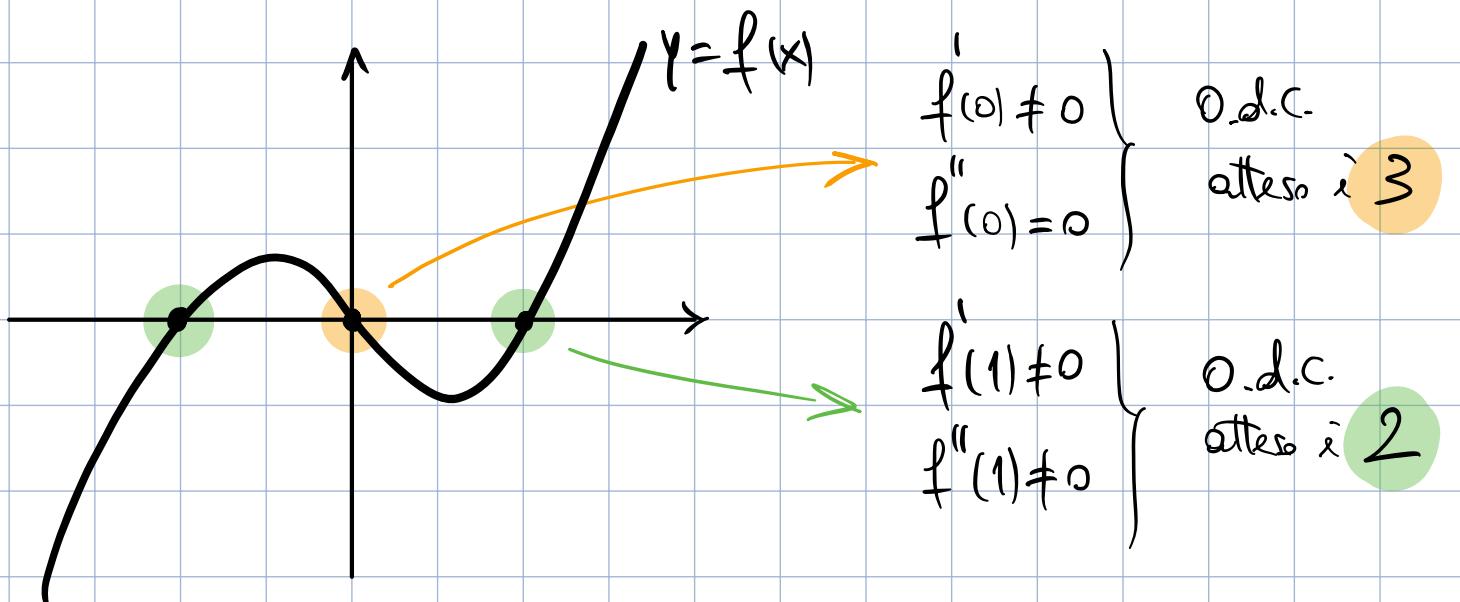
determinare l'O.d.C. delle succ.m.  
generate dal metodo di Newton se non  
ten'  $x_0 = -1, 0, 1$ .

Cominciamo da un grafico di  $f$ .

- $x^3 - x = x(x^2 - 1) = x(x-1)(x+1)$

- $\lim_{x \rightarrow +\infty} f(x) \rightarrow +\infty$ ,  $f(x) \rightarrow +\infty$

- $\lim_{x \rightarrow -\infty} f(x) \rightarrow -\infty$ ,  $f(x) \rightarrow -\infty$



Verifichiamo se puntando le def. di  
O.d.c.

$$\text{Newton: } x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}, \quad k \geq 0.$$

$$f(x) = x^3 - x \Rightarrow f'(x) = 3x^2 - 1$$

Dunque

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

Poniamo da  $\alpha = 1$ .

Sottraggli  $\alpha = 1$  e dividilo per  $(x^{(k)} - \alpha)^p$ ,

dove  $p$  andrà scelto opportunamente.

$$\begin{aligned}
 x^{(k+1)} - 1 &= x^{(k)} - 1 - \frac{(x^{(k)})^3 - x^{(k)}}{3(x^{(k)})^2 - 1} = \\
 &= \frac{3(x^{(k)})^3 - x^{(k)} - 3(x^{(k)})^2 + 1 - (x^{(k)})^3 + x^{(k)}}{3(x^{(k)})^2 - 1} = \\
 &= \frac{2(x^{(k)})^3 - 3(x^{(k)})^2 + 1}{3(x^{(k)})^2 - 1} = \dots
 \end{aligned}$$

Raccogliere al numeratore  
il fattore  $(x^{(k)} - \alpha)$  tante  
volte quanto è possibile

Ruffini su  $2x^3 - 3x^2 + 1$ :

	2	-3	0	1
1	2	-1	-1	
	2	-1	-1	//

$$2x^3 - 3x^2 + 1 = (x-1)(2x^2 - x - 1)$$

Ruffini su  $2x^2 - x - 1$

	2	-1	-1
1	2	1	
	2	1	//

$$2x^2 - x - 1 = (x-1)(2x+1)$$

$$\dots = \frac{(x^{(k)} - 1)^2 (2x^{(k)} + 1)}{3(x^{(k)})^2 - 1}$$

dividiamo per  $(x^{(k)} - 1)^2$  e  
passiamo ai valori assoluti:

$$\frac{|x^{(k+1)} - 1|}{|x^{(k)} - 1|^2} = \frac{|2x^{(k)} + 1|}{|3(x^{(k)})^2 - 1|} \xrightarrow[k \rightarrow +\infty]{x^{(k)} \rightarrow 1} \frac{3}{2},$$

dunque l'O.d.C. è 2.

Passiamo a  $\alpha = 0$ . Considero ancora

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

Sottraggo  $\alpha$  e entro nell'intervallo  
 (questa volta sente effetto perché  $\alpha = 0$ ) e

divido per  $(x^{(k)} - \alpha) = x^{(k)}$ , così

P da segnale quantitativamente.

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

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

$$= \frac{2(x^{(k)})^3}{3(x^{(k)})^2 - 1}$$

Da cui :

$$\frac{x^{(k+1)}}{(x^{(k)})^3} \stackrel{\alpha=0}{=} \frac{x^{(k+1)} - \alpha}{(x^{(k)} - \alpha)^3} = \frac{2}{3(x^{(k)})^2 - 1}.$$

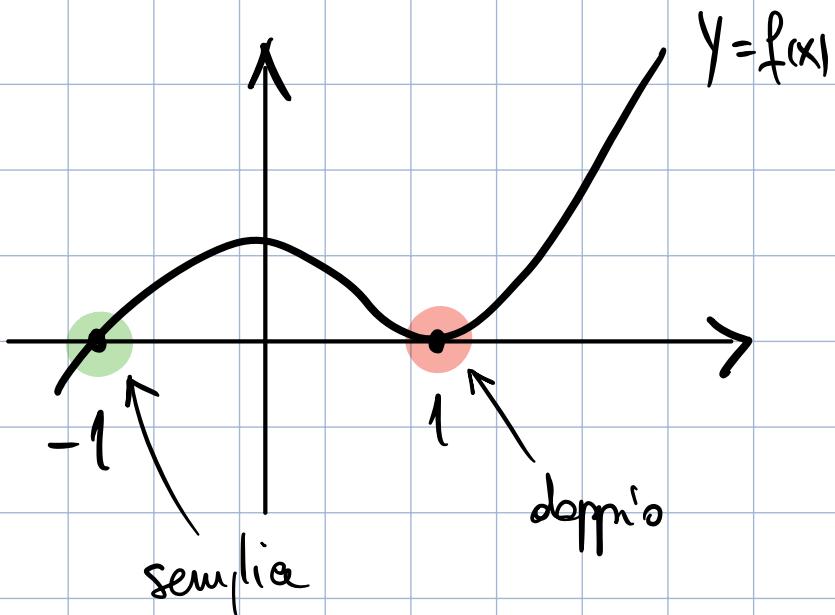
Passo ai valori assoluti :

$$\frac{|x^{(k+1)} - \alpha|}{|x^{(k)} - \alpha|^3} = \frac{2}{|3(x^{(k)})^2 - 1|} \xrightarrow[\substack{k \rightarrow +\infty \\ x^{(k)} \rightarrow 0}]{} 2$$

Quindi l'O.d.C. è 3.

Esercizio : dato  $f(x) = (x-1)(x+1)$

determinare l'O.d.C. delle succ.m.  
generate dal metodo di Newton Qi siano  
tutti  $\alpha = -1, 1$ .



$\alpha = -1$ : o. d. c.  
atteso è 2

$\alpha = 1$ : o. d. c.  
atteso 1 2  
(perché zero doppio)

Ripetere l'analisi fatta sopra ...

# ARITMETICA DI MACCHINA

$\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}$  sono infiniti;  
questi tutti i numeri in  $\mathbb{Q}$  e  $\mathbb{R}$  hanno  
siffatto decimale (o binario) infinito.

Teorema (di rappresentazione in base)

Dati  $x \in \mathbb{R}$ ,  $x \neq 0$  e  $\beta \in \mathbb{N}$ ,  $\beta \geq 2$ ,  
esiste un unico modo di rappresentare

$x$  come

$$x = \sum_{i=0}^{\infty} d_i \beta^{-i} \times \beta^P, \text{ dove}$$

(1)  $d_i \in \mathbb{N}$ ,  $0 \leq d_i \leq \beta - 1$ ,  $d_i > 0$ .

↳ cifre

(2)  $d_0 \neq 0$  "Normalizzazione"

↳ cifre più significative

(3)  $p \in \mathbb{Z}$

(4)  $\{d_i\}_{i \in \mathbb{N}}$  non definitivamente uguali a  $\beta^{-1}$ .

## Osservazioni

---

- sullo (2) :

$$3.\cancel{1}4 = 0.\cancel{3}14 \times 10 = \\ = \cancel{3}14 \times 10^{-2}$$

- sullo (4) (garantisce l'unicità)

Esempio:

$$1 = 1.0 \cancel{\times 10^0} = 0.\overline{9} = 9.\overline{9} \times \cancel{10^{-1}}$$

# Informalmente

$$x = \sum_{i=0}^{\infty} d_i \beta^{-i} \times \beta^p = \boxed{d_0.d_1d_2 \dots \times \beta^p}$$

Per rendere l'insieme finito dovremo limitare il numero di cifre di e il range per l'esponente  $p$ .

Esempio:

$$(0.\overline{1})_{10} = (?)_2$$

$$\begin{array}{l} 0.\overline{1} \times 2 = 0.\underline{2} \quad \text{periodo} \\ \rightarrow 0.\underline{2} \times 2 = 0.\underline{\underline{4}} \quad \text{allora} \\ 0.\underline{\underline{4}} \times 2 = 0.\underline{\underline{8}} \\ 0.\underline{\underline{8}} \times 2 = 1.\underline{6} \\ 0.\underline{\underline{6}} \times 2 = 1.\underline{2} \end{array}$$

$$(0.\overline{1})_{10} =$$

$$(0.\overline{00011})_2 =$$

$$= (1.\overline{1001} \times 2^{-4})_2$$

una cifra in base 10 ma di cifre in base 2!

# INSIEME DEI NUMERI

## DI MACHINA

L'insieme dei numeri d'Machina

in base  $\beta$  e  $t+1$  cifre significative

e range per l'esponente  $(M_1, M_2)$  è

Costituito dai seguenti elementi

$$\left\{ x \in \mathbb{R} : x = \sum_{i=0}^t d_i \beta^{-i} \times \beta^p \right\}$$

dove  $d_i \in \mathbb{N}$ ,  $0 \leq d_i \leq \beta - 1$

per  $i = 0, 1, \dots, t$ ,  $d_0 \neq 0$

e  $p \in \mathbb{Z}$  con  $M_1 + 1 \leq p \leq M_2 - 1$

Esso è dimostrato con

$$F(\beta, t, M_1, M_2)$$

Osservazioni:

(1)  $O \notin F$

(2) informalmente i numeri magline

hanno la seguente forma:

$$X = \pm d_0.d_1d_2 \dots d_t \times \beta^P$$

# Oscar of 'ome :

$$(0.\overline{1})_{10} = (?)_2$$

$$0.\overline{1} \times 2 = 0.\underline{2} \quad \text{periodo}$$

$$\rightarrow 0.\underline{2} \times 2 = 0.\underline{\underline{4}}$$

$$0.\underline{\underline{4}} \times 2 = 0.\underline{\underline{8}}$$

$$0.\underline{\underline{8}} \times 2 = 1.\underline{\underline{6}}$$

$$1.\underline{\underline{6}} \times 2 = 1.\underline{\underline{2}}$$

allora

$$(0.\overline{1})_{10} = (0.\overline{00011})_2 =$$

$$= (1.\overline{1001} \times 2^{-4})_2$$

Sviluppo finito in base 10 ma  $\infty$  in base 2!

# INSIEME DEI NUMERI

## DI MACCHINA

L'insieme dei numeri di Macchina

in base  $\beta$  a  $t+1$  cifre significative

e range per l'esponente  $(M_1, M_2)$  è  
costituito dai seguenti elementi

$$\left\{ x \in \mathbb{R} : x = \pm \sum_{i=0}^t d_i \beta^{-i} \times \beta^p \right\}$$

dove  $d_i \in \mathbb{N}$ ,  $0 \leq d_i \leq \beta - 1$

per  $i = 0, 1, \dots, t$ ,  $d_0 \neq 0$

e  $p \in \mathbb{Z}$  con  $M_1 + 1 \leq p \leq M_2 - 1 \}$

Esso è dimostrato con

$$F(\beta, t, M_1, M_2)$$

Osservazioni :

(1)  $O \notin F$

(2) informalmente i numeri machine

hanno la seguente forma :

$$X = \underbrace{\pm d_0.d_1d_2 \dots d_t}_{\text{cifre}} \times \beta^P$$

base  
eponente

## ESEMPIO

$$3.14 \in \bar{F}(10, 2, -1, 2) =$$

$$= \left\{ \pm d_0.d_1d_2 \times 10^P, P \in \{0, 1\} \right\}$$

$$3.14 \notin F(10, 1, -1, 2)$$

$$3.14 \notin F(10, 2, 0, 2)$$


---

Più piccoli numeri machine strett. portati:

Matlab

$$\text{redmin} := 1.00\dots0 \times \beta^{M_1+1} = \beta^{M_1+1}$$

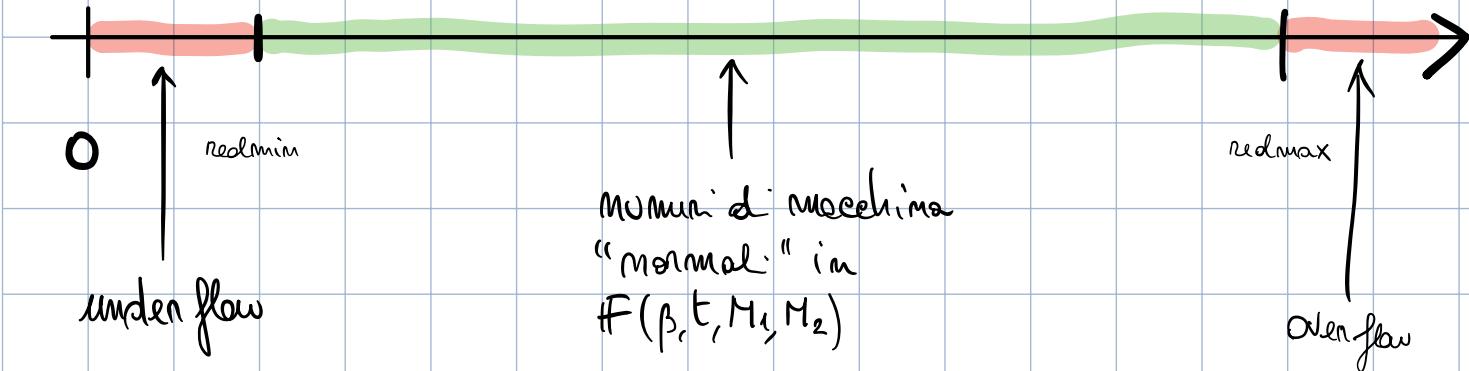
Più grande numero di machine:

$$\text{redmax} := \beta^{-1}, \beta^{-1}, \beta^{-1}, \dots, \beta^{-1} \times \beta^{M_2-1}$$

oggiungo e sottraggio  $0.00 \dots 0^1 \times \beta^{M_2-1}$

$$\begin{array}{r}
 \left| \begin{array}{ccccccc} \beta^{-1}, & \beta^{-1} & \beta^{-1} & \dots & \beta^{-1} & \beta^{-1} & \times \beta^{M_2-1} + \\ 0. & 0 & 0 & \dots & 0 & 1 & \times \beta^{M_2-1} - \\ 0. & 0 & 0 & \dots & 0 & 1 & \times \beta^{M_2-1} = \end{array} \right. \\
 \hline
 10. & 0 & 0 & \dots & 00 & \times \beta^{M_2-1} - \\
 1. & 0 & 0 & \dots & 00 & \times \beta^{M_2-1-t} = 
 \end{array}$$

$$\beta^{M_2} - \beta^{M_2-1-t} = \beta^{M_2-1} (\beta - \beta^{-t})$$



Come vengono gestiti underflow  
e overflow?

Gestione "grado di libere" dell'underflow:

$$\text{minim. demormali} = \left\{ \pm 0.d_1 d_2 \dots d_t \times \beta^{N_1+1} \right. , \quad d_i \neq 0 \\ \left. \uparrow \quad \text{per almeno un } i = 1, 2, \dots, t \right\}$$

Viamo "rileggi" le condizioni di:

Normalizzazione ( $d_0 \neq 0$ )

Più piccolo numero demormali strettamente positivo:

$$\underbrace{0.0 \dots 0}_t \times \beta^{N_1+1} = \beta^{N_1+1-t}$$

ESEMPIO:  $F(2, 2, -2, 3) =$

$$= \left\{ \pm 1.d_1 d_2 \times 2^P, P = -1, 0, 1, 2 \right\}$$

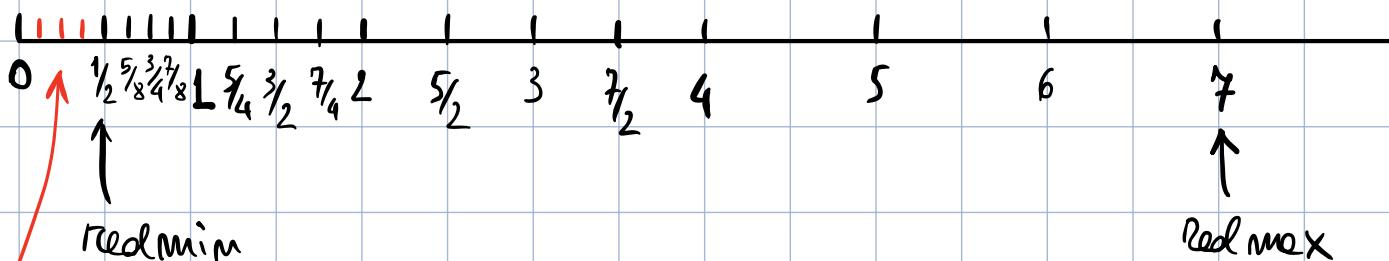
Sono 32 numeri (16 positivi). I positivi sono:

$$P = -1 \quad 1.00 \times 2^{-1} \quad 1.01 \times 2^{-1} \quad 1.10 \times 2^{-1} \quad 1.11 \times 2^{-1}$$

$$P = 0 \quad 1.00 \times 2^0 \quad 1.01 \times 2^0 \quad 1.10 \times 2^0 \quad 1.11 \times 2^0$$

$$\begin{array}{cccc}
 p=1 & 1.00 \times 2^1 & 1.01 \times 2^1 & 1.10 \times 2^1 \\
 p=2 & 1.00 \times 2^2 & 1.01 \times 2^2 & 1.10 \times 2^2
 \end{array}$$

Rappresentazioni sulle rette Reale:



Osserv.: maggiore densità per minori Nodi

OSSERV.: **denormoli** =  $\left\{ \pm 0.d_1 d_2 \times 2^{-1} \right\} =$

$$= \left\{ \pm \frac{1}{8}, \pm \frac{1}{4}, \pm \frac{3}{8} \right\}$$

Distanze tra minori macchine consecutivi:

$$\text{Sia } x = d_0.d_1 d_2 \dots d_t \times \beta^P \in F(\beta, t, \alpha_1, M_2).$$

Allora  $y$ , numero macchina successivo, è dato da

$$y = x + 0.0\dots 01 \times \beta^P \Rightarrow$$

la distanza di  $x$  da  $y$  è  $d = 0.0\dots 01 \times \beta^{P-t}$

Conclusione: se  $x < y \in F(\beta, t, M_1, M_2)$  sono

numeri macchine consecutivi e  $x, y \in [\beta^t, \beta^{t+1}]$ ,

allora la loro distanza è

$$y - x = \beta^{t+1} - \beta^t$$

Esempio: distanza di  $5/4$  dal numero macchina

consecutivo in  $F(2, 2, -2, 3)$ :

$$\frac{5}{4} \in [1, 2] = [2^0, 2^1] \Rightarrow$$

$$\text{distanza} = 2^{1-0} = \frac{1}{4}$$

Standard IEEE 754 single/double precision

Single prec.: il numero è memorizzato sotto

forma di stringa a 32 bit:

S	q	f
1	8	23
segno	esponente	mantissa

Sono i numeri:  $(-1)^S \times 1.F \times 2^P$ , dove

positivo  $\approx S=0$   
negativo  $\approx S=1$

$$S = 0, 1, P = q - \text{"bias"}$$

?

$$q = (00000000)_{\text{2}}, (00000001)_{\text{2}}, \dots,$$

$$(11111110)_{\text{2}}, (11111111)_{\text{2}} =$$

$$= 0, 1, \dots, 254, 255$$

sottraggono 127 per distribuirli  
meglio ottimo a zero

$$P = q - 127 = -126, \dots, 127$$

↑  
bias

Si tratta dell'insieme  $\mathbb{F}(2, 23, -127, 128)$

Esempio:

$$\text{redmin} = \beta^{M_1+1} = 2^{-126} \approx 1.2 \times 10^{-38}$$

$$\text{redmax} = \beta^{n_2-t} (\beta - \beta^{-t}) = 2^{127} (2 - 2^{-23}) \approx$$

$$\approx 2^{128} \approx 1.7 \times 10^{38}$$

più piccolo demone  
positivo  $= \beta^{M_1+1-t} = 2^{-149} \approx 1.4 \times 10^{-45}$

Le stringhe  $q = 00000000, 11111111$  non contribuiscono  
all'esponente perché rivestono un ruolo speciale:

$q$	Mantissa	$m_s$
00000000	0	zero macchina
00000000	$\neq 0$	demoni
11111111	0	$\pm \text{Inf}$
11111111	$\neq 0$	NaN

ESEMPI (Matlab):  $1/0 = \text{Inf}$ ,  $\text{Inf} + \text{Inf} = \text{Inf}$ ,  $\text{Inf} - \text{Inf} = \text{NaN}$

Double prec.: (default in Matlab)

S	q	f
---	---	---

#2<sup>5</sup>: 1      11      52

$$q = \cancel{0}, 1, \dots, 2046, \cancel{2047}$$

bias: 1023

$$P = q - \text{bias} = -1022, \dots, 1023$$

s' tratta di  $\mathbb{F}(2, 52, -1023, 1024)$

Determinare  $\text{redmin}, \text{redmax}$  e

"redmin demoneale".

# Annotandamento

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

Si legge "float di  $x$ " o  
"annotandamento di  $x$ "

numeri  
macchine  
normali

gestione  
overflow

gestione  
underflow

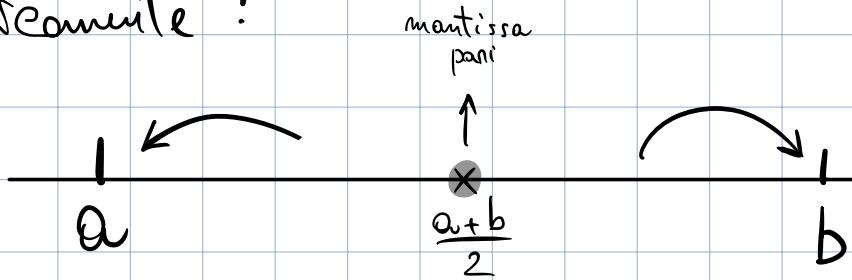


Supponiamo che l'annotandamento non  
coinvolge underflow o overflow,  
ovvero  $|x| \in [\text{redmin}, \text{redmax}]$

L'annotandamento segue la regola

"R N T E": "Round To Nearest  
, Ties To Even"

Graphicamente:



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

Formalmente:

Siamo  $x \in [a, b]$ ,  $a < b \in \mathbb{F}(p, t, m_1, m_2)$  consecutivi,  $\beta$  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:  $B = \max_{i=1}^n |x_i - \bar{x}|$   
nell'intervallo  $[a, b]$ :

$$X \in [\text{realmim}, \text{realmex}] \mapsto fl(x) \in \mathbb{F}$$

(Per completezza, supponiamo  $X > 0$ )

$$\frac{|x - f_l(x_1)|}{|x|} \leq \frac{\gamma_2(b-\alpha)}{\alpha} \leq \frac{\gamma_2 \beta^{-t}}{\beta^p} = \frac{1}{2} \beta^{-t}$$

↑      ↑      ↑

ε-zone  
relativa

$f_l(x)$

$\alpha < b \in [\beta^p, \beta^{p+1}]$  numeri  
massima concentrazione

Definizione: il numero  $\beta^{-t}$  viene detto  
"precisione d'arredamento" e indicato con  
 $\text{eps}$  (anche in Matlab).

Dunque il massimo errore relativo commesso

nell'entroterra mentre RNT è

la metà delle precisione di macchina:

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

## ESEMPI :

$$\text{IEEE single prec. : } \gamma_2 p^{-t} = 2^{-24} \approx 10$$

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

# Condizionamento dell'oriente di macchina

Consideriamo  $\tilde{x}$  approssimazione 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'appross è un eccesso ( $\tilde{x} > x$ ) altrimenti è un difetto.

Si ha:

$$\begin{aligned} x \varepsilon_x = \tilde{x} - x &\Leftrightarrow x + x \varepsilon_x = \tilde{x} \Leftrightarrow \\ &\Leftrightarrow \tilde{x} = x(1 + \varepsilon_x) \end{aligned}$$

## Somma / sottrazione

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

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

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

$\nwarrow$

- "aritmetica di macchina"
- "aritmetica finita"

Abbiamo:

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

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

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

Vogliamo mettere in relazione

$\epsilon_x, \epsilon_y$  (erri. rel. in input) con  
 $\epsilon_{x+y}$  (errore rel. in output)

In aritmetica finita:

$$(x + y)(1 + \epsilon_{x+y}) = x(1 + \epsilon_x) + y(1 + \epsilon_y)$$

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

$$= X + X\varepsilon_X + Y + Y\varepsilon_Y$$

$$(X+Y)\varepsilon_{x+y} = X\varepsilon_X + Y\varepsilon_Y \quad \xleftarrow[X+Y \neq 0]$$

$$\varepsilon_{x+y} = \frac{X\varepsilon_X + Y\varepsilon_Y}{X+Y}$$

$$|\varepsilon_{x+y}| = \frac{|X\varepsilon_X + Y\varepsilon_Y|}{|X+Y|} \leq \frac{|X||\varepsilon_X| + |Y||\varepsilon_Y|}{|X+Y|} \leq$$

disug. triang.

$$\leq \frac{\max\{|\varepsilon_X|, |\varepsilon_Y|\}(|X| + |Y|)}{|X+Y|} \leq \frac{\max\{|\varepsilon_X|, |\varepsilon_Y|\}(|X| + |Y|)}{|X+Y|}$$

con  $\max\{|\varepsilon_X|, |\varepsilon_Y|\}$   
e metto in evidenza

Riscrivo:

Numero di  
aggiornamento  
 $K$

$$|\varepsilon_{x+y}| \leq \frac{(|X| + |Y|)}{|X+Y|} \max\{|\varepsilon_X|, |\varepsilon_Y|\}$$

err. rel. in  
input

err. nel.  
in output

## 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  $X+Y$  è se ne più condizionata solo se gli addendi hanno segno opposto e lo è se  $Y \approx -X$ ;  
E' tanto più  $Y$  è vicino a  $-X$ .

Le perdite di precisione danza alle somme  $X+Y$  quando  
 $Y \approx -X$  è detta errore  
di cancellazione (di cifre ...)

" $X, Y$ , vicini a essere uno l'opposto dell'altro"

Moltiplicazione

Siamo  $X, Y \in \mathbb{R}$ . Consideriamo :

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

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

Abbiamo:

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

e dunque

$$\begin{aligned} \cancel{1 + \varepsilon_{xy}} &= (1 + \varepsilon_x)(1 + \varepsilon_y) = \\ &= 1 + \varepsilon_x + \varepsilon_y + \underbrace{\varepsilon_x \varepsilon_y}_{\text{trascurabile rispetto}} \end{aligned}$$

a  $\varepsilon_x \cdot \varepsilon_y$ , se  
 $|\varepsilon_x|, |\varepsilon_y|$  sono molto  
piccoli (ipotesi  
naturale)

Dunque, trascurando l'ultimo  
addendo:

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

Infine

disug.  $\Delta$

$$|\varepsilon_{xy}| \approx |\varepsilon_x + \varepsilon_y| \leq$$

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

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

allora

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

↑  
err. rel.  
output

↑  
numero d'  
cond. K

↑  
err. rel. input

l'operazione è sempre ben condizionata

# Divisione

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

$$(x, y) \mapsto x_{y} \in \mathbb{R} \quad \text{"lsatta"}$$

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

$$\cancel{x} \quad (1 + \epsilon_{x/y}) = \frac{x(1 + \epsilon_x)}{y(1 + \epsilon_y)}$$

$$1 + \epsilon_{x/y} = \frac{1 + \epsilon_x}{1 + \epsilon_y} \approx \dots$$

Riduciamo (serie geometrica)

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

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

$$\frac{1}{1-x} \approx 1 \pm x \quad \text{per } x \text{ piccolo}$$

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

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

$$\approx 1 + \varepsilon_x - \varepsilon_y (\mu \text{n}(\varepsilon_x), (\varepsilon_y)$$

(recidi) ;

AVVISO

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

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

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

↑  
K

Quindi l'operazione è sempre  
ben condizionata

Riassoltiamo:

Somma  $x+y$

Mol cond.

Se  $y \approx -x$

moltiplicazione

bem cond.

divisione

bem cond.

Condizionamento della valutazione di funzione.

$x \in \mathbb{R}$ ,  $f$  definita in un intorno di  $x$

$$x \mapsto f(x) \in \mathbb{R} \quad \text{"esatta"}$$

$$x \mapsto f(f(f(x))) \in F \quad \text{"di macchina"}$$

Sì ha:

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

Calcoliamo:

$$f(x) + f(x)\varepsilon_{f(x)} = f(x + x\varepsilon_x) \iff$$

$$f(x)\varepsilon_{f(x)} = f(x + x\varepsilon_x) - f(x)$$

dovetiamo che  $h := x\varepsilon_x$ , possiamo supporre che  $|h| \ll |x|$

Alone

$$f(x) \epsilon_{f(x)} = f(x+h) - f(x) =$$

$$\hookrightarrow = \left( \frac{f(x+h) - f(x)}{h} \right) h \approx$$

divido e  
multiplicado por  $h$

$$\approx f'(x)h = f'(x) \times \epsilon_x$$

Therefore

$$f(x) \epsilon_{f(x)} = f'(x) \times \epsilon_x \Rightarrow$$

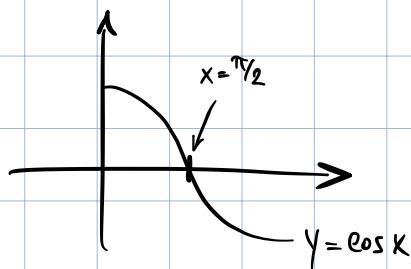
$$|\epsilon_{f(x)}| = \left| \frac{x f'(x)}{f(x)} \right| |\epsilon_x|$$

$$K = K(x) \quad (\text{depends on } x)$$

La volatilità di funzione è una  
condizione ledova  $K(x) \gg 1$ .

Esempio

$$f(x) = \cos x \quad \text{vicino a } x = \pi/2$$



Esercizio (da fare da soli)

definire in Matlab

$$x_0 = \pi/2 \quad e \quad x_1 = 1.57 ;$$

Confrontare  $x_0$  con  $x_1$  e

$\cos(x_0)$  con  $\cos(x_1)$  (err. rel.)

Esercizio (svolto in aula il 3 nov. 2021)

Consideriamo  $\mathbb{F}(10, 6, -11, 11)$  e

$$x = 123457.1467$$

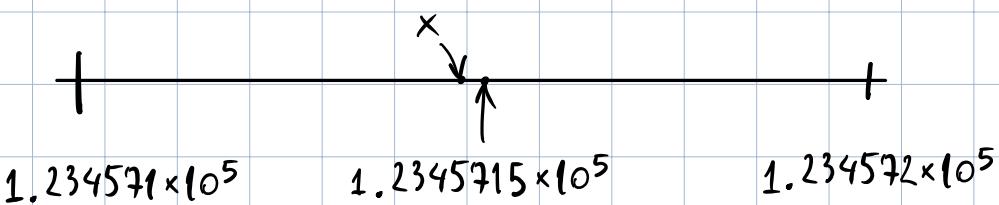
$$y = 123456.659$$

Confrontare  $x-y$  in aritmetica esatta  
con  $x-y$  in aritmetica di macchine.

Aritmetica esatta:  $x-y=0.4877$

Aritm. di macchina:  $x, y \notin \mathbb{F}$

$$x = 1.234571\underset{\substack{\text{e cifre} \\ \text{in eccesso}}}{|}467 \times 10^5$$



dunque  $fl(x) = 1.234571 \times 10^5$

Analogamente

$$y = 1.234566 \cdot 59 \times 10^5$$

dunque  $f(y) = 1.234567 \times 10^5$

Adesso calcoliamo la differenza

$$\begin{aligned} X - Y &= 1.234571 \times 10^5 - \\ &\quad \underline{1.234567 \times 10^5} = \\ &0.000004 \times 10^5 \end{aligned}$$

Ovvero, in arithm. finita:

$$\begin{aligned} X - Y &= 0.4 = \\ &= 4.0 \times 10^{-1} \in F \end{aligned}$$

Risposte labo:

$$X-Y \text{ in antro. esato} : 4.877 \times 10^{-1}$$

$$X-Y \text{ in antro. finito} : 4.0 \times 10^{-1}$$

il risultato in antro. di  
macchina ha solo 1 cifra

corretta! Il max errore

relativo commesso nell'annodamento

$$\text{è } \frac{1}{2} \beta^t = \frac{1}{2} 10^{-6} = 5 \times 10^{-7}$$

Allora per 6 cifre!



7-1

Stimiamo il numero di condizionamento:

$$K = \frac{|X| + |Y|}{|X - Y|} \approx \frac{1.2 \times 10^5 + 1.2 \times 10^5}{4.8 \times 10^{-1}} \approx$$
$$\approx \frac{2.4 \times 10^5}{4.8 \times 10^{-1}} = \frac{1}{2} \times 10^6 =$$
$$= 5 \times 10^5$$

La perdita di precisione è giustificata dal m. condizionamento! Si è verificato il cosiddetto errore di cancellazione

ESERCIZIO (problema 1, primo esame  
2020/2021)

Considero  $\bar{F}(2, 12, -7, 8)$

(1)

$$\text{redmin} = \beta^{M_1+1-6} = 2$$

(2) redmin "demonale" =

$$= \beta^{M_1+1-t-18} = 2$$

(3)

$$a = 0.d_1 \dots d_t \times \beta^{M_1+1} \quad \text{demonale}$$

successivo:  $a + 0.\underbrace{0 \dots 0}_t \times \beta^{M_1+1}$

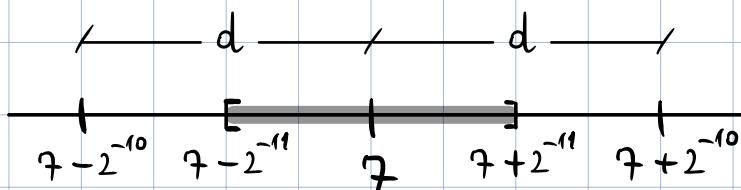
$$\text{dist} = 0.\underbrace{0 \dots 0}_t \times \beta^{M_1+1} = \beta^{M_1+1-t-18} = 2$$

(4)

$$7 \in [4, 8] = [2^3, 2^3] \quad p=2$$

$a < 7 < b$  e f consecutivi

$$\text{allora } d = b - 7 = 7 - a = \beta^{p-t-10} = 2^{-10}$$



Risposta:  $[7 - 2^{-11}, 7 + 2^{-11}]$

(5)

$$\text{---} \mid \quad \mid \quad \mid \quad \mid$$
$$1 - 2^{-13} \quad 1 \quad 1 + 2^{-12}$$

se  $0 \leq p \leq 12$ ,  $1 + 2^{-p} \in F$  e quindi

$$(1 + 2^{-p}) - 2^{-p} = 1$$

se  $p = 13$ ,  $1 + 2^{-13} \notin F$  e  $\text{fl}(1 + 2^{-13}) = 1$ ,

ma  $1 - 2^{-13} \in F$ , per cui

$$(1 + 2^{-p}) - 2^{-p} < 1 \quad (\text{in } F)$$

se  $p \geq 14$ , verifichiamo che

$$(1 + 2^{-p}) - 2^{-p} = 1 \quad (\text{in } F)$$

RISPOSTA:  $p = 13$

(6) Segno: 1 bit

mantissa: 12 bit

esponente: 4 bit ( $16 = 2^4$ )

} 17 bit

(7)

$$\bar{F} = \left\{ \pm 1 \cdot d_1 \dots d_t \times 2^p, p \in \mathbb{Z}, -6 \leq p \leq 7 \right\}$$

$\# F = 2 \cdot 2^{12} \cdot 14$

(8)

$$f = (111)_2 = (1 \cdot 11 \times 2^2) \Rightarrow$$

$$t = 2$$

(9)

redmax per  $\bar{F}(2, 12, -7, M+1)$ :

$$2^M (2 - 2^{-12}) \approx 2^{M+1}$$

$$\text{Se } M = 8 : \text{ redmax} \approx 512$$

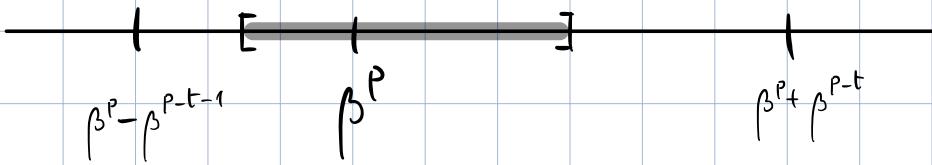
$$\text{Se } M = 9 : \text{ redmax} \approx 1024$$

Risposta:  $M = 8$

Ulteriori considerazioni:

$$\text{in } \mathbb{F}(\beta, t, M_1, M_2)$$

- $\left\{ x \in \mathbb{R} : f(x) = \beta^p \right\}$  non è un intervallo  
simmetrico di  $\beta^p$



$$[\beta^p - \frac{1}{2}\beta^{p-t-1}, \beta^p + \frac{1}{2}\beta^{p-t}]$$

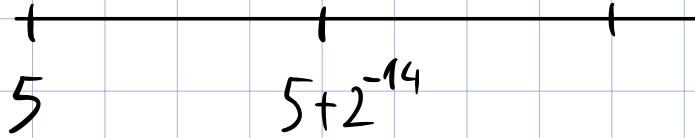
- $\left\{ x \in \mathbb{R} : f(x) = \beta^p + \beta^{p-t} \right\}$  è aperto  
(non contiene gli estremi)

ESERCIZIO In  $\mathbb{F}(2, 16, -63, 64)$  determinare

il più grande  $p \in \mathbb{N}$  per cui

$$f(5 + 2^{-p}) > 5 .$$

$f \in [2^2, 2^3]$ , quindi



$$P \geq 15 \Rightarrow fl(5 + 2^{-P}) = 5$$

$$P = 14 \Rightarrow fl(5 + 2^{-14}) > 5$$

Risposta:  $P = 14$

ESERCIZIO In IEEE 754 dp, uno tra  
 $2, \gamma_2, 3, \gamma_3$  non è numero macchina.

Quale? Perché?

$$\begin{aligned} \text{Risposta: } \gamma_3 &= (0.\overline{01})_2 = \\ &= (1.\overline{01} \times 2^{-2}) \notin \mathbb{F} \quad (\infty \text{ eifre}) \end{aligned}$$

$$\text{Si ha: } fl(\gamma_3) = 1.\underbrace{0101 \dots 01}_{52 \text{ eifre}} \times 2^{-2}$$

# SPAZI VETTORIALI

## Definizioni

Un insieme  $\mathbb{K}$  è detto campo se esistono leggi d'operazione  $+$  e  $\cdot$  interne

$$+, \cdot : \mathbb{K} \times \mathbb{K} \rightarrow \mathbb{K}$$

tali da:

$$1) \exists 0 \in \mathbb{K} \text{ t.c. } x + 0 = 0 + x = x,$$

$$\forall x \in \mathbb{K}$$

$$2) \forall x \in \mathbb{K} \exists (-x) \in \mathbb{K} \text{ t.c.}$$

$$x + (-x) = (-x) + x = 0$$

$$3) \exists 1 \in \mathbb{K} \text{ t.c. } x \cdot 1 = 1 \cdot x = x,$$

$$\forall x \in \mathbb{K}$$

4)  $\forall x \in K, x \neq 0, \exists (x^{-1}) \in K$  t.c.

$$x \cdot (x^{-1}) = (x^{-1}) \cdot x = 1$$

Notazione

$x, y \in K$  : " $xy$ " significa  $x \cdot y$

Nel Corpo deve valere le proprietà associative per

$+ \cdot \circ$ . Non è necessario

che valga la proprietà commutativa.

Un corpo nel quale  $+ \cdot \circ$

siano commutative si dicono

Campo ("field") o Corpo

commutativo.

Esempi :  $K = \mathbb{Q}, \mathbb{R}, \mathbb{C}$

Sono campi ;  $\mathbb{Z}$  non è un campo !

Un insieme  $V$  è detto spazio vettoriale sul campo  $K$

se esistono leggi d'operazione + interna e  $\cdot$  esterna

$$+ : V \times V \rightarrow V$$

$$\cdot : K \times V \rightarrow V$$

tali da :

1)  $+$ ,  $\cdot$  sono commutative

2)  $\exists \mathbf{0} \in V$  t.c.

$$v + \mathbf{0} = \mathbf{0} + v = v \quad \forall v \in V$$

3)  $\forall v \in V \quad \exists (-v) \in V$  t.c.

$$v + (-v) = (-v) + v = \mathbf{0}$$

4)  $\forall \alpha, \beta \in K, \forall v \in V :$

$$(\alpha \beta) v = \alpha (\beta v) \leftarrow \text{associatività}$$

5)  $\forall \alpha, \beta \in K, \forall v \in V :$

$$(\alpha + \beta) v = \alpha v + \beta v \quad \left. \right\} \text{distributività}$$

6)  $\forall \alpha \in K, \forall v, w \in V :$

$$\alpha(v + w) = \alpha v + \alpha w$$

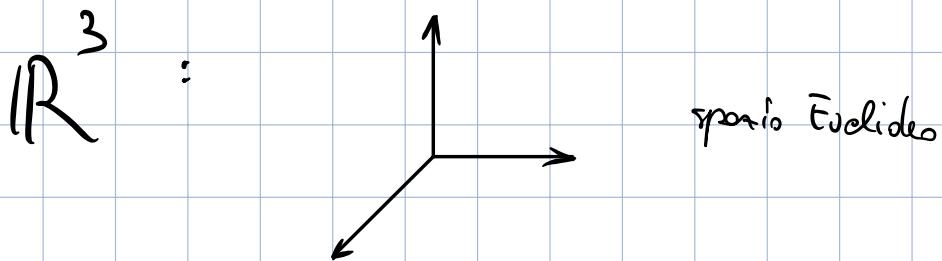
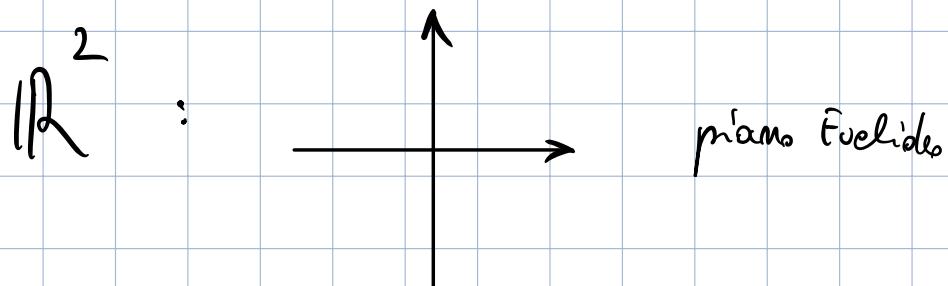
7)  $\forall v \in V : 1 v = v 1 = v$

## Nomenclatura :

Gli elementi di  $V$  sono detti  
vettori; gli elementi di  $K$   
sono detti scalar.

### Esempi

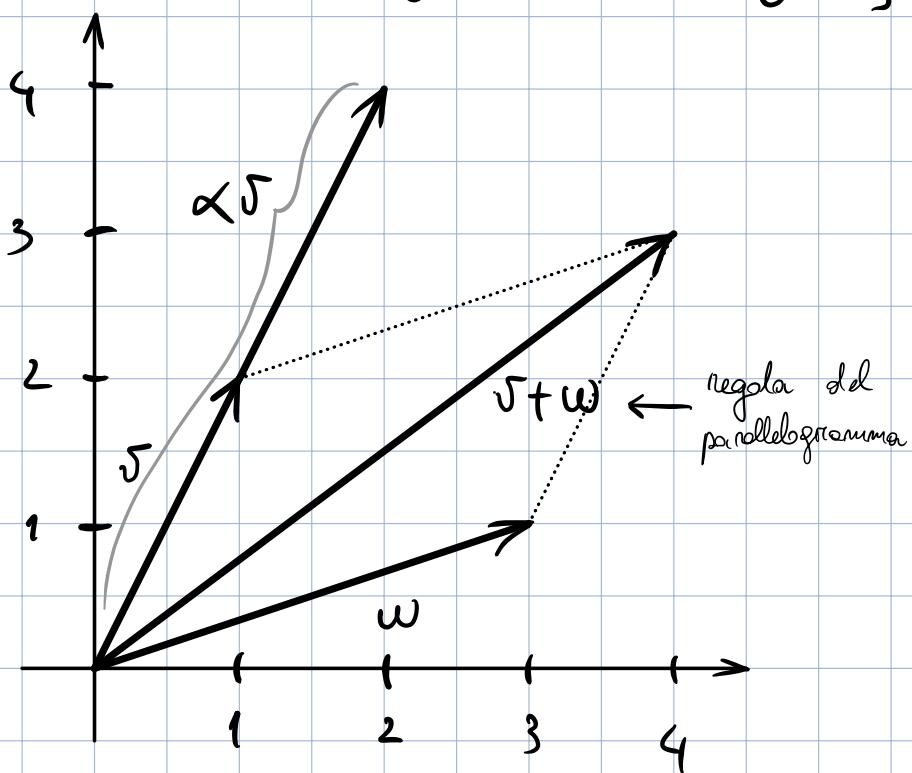
1)  $\mathbb{R}^n = \left\{ \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} : x_i \in \mathbb{R} \text{ } \forall i = 1, 2, \dots, n \right\}$



Esempio

$$\zeta = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \omega = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \alpha = 2$$

$$\zeta + \omega = \begin{bmatrix} 4 \\ 3 \end{bmatrix}; \alpha \zeta = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$



$$2) P_m = \left\{ \sum_{k=0}^m a_k x^k \right\} = \\ = \left\{ \text{polinomi di grado } \leq m \right\}$$

$$3) C([a,b]) = \{ f : [a,b] \rightarrow \mathbb{R} \text{ continua} \}$$

## SOTTOSPazi VETTORIALI

Sia  $W \subset V$ , con  $V$  spazio vettoriale su  $\mathbb{K}$ . Dineamo che

$W$  è sottospazio vettoriale di  $V$

Se:

$$1) \forall \sigma, \omega \in W : \sigma + \omega \in W$$

$$2) \forall \lambda \in \mathbb{K}, \forall \sigma \in W : \lambda \sigma \in W$$

divisione di  $W$   
rispetto a +

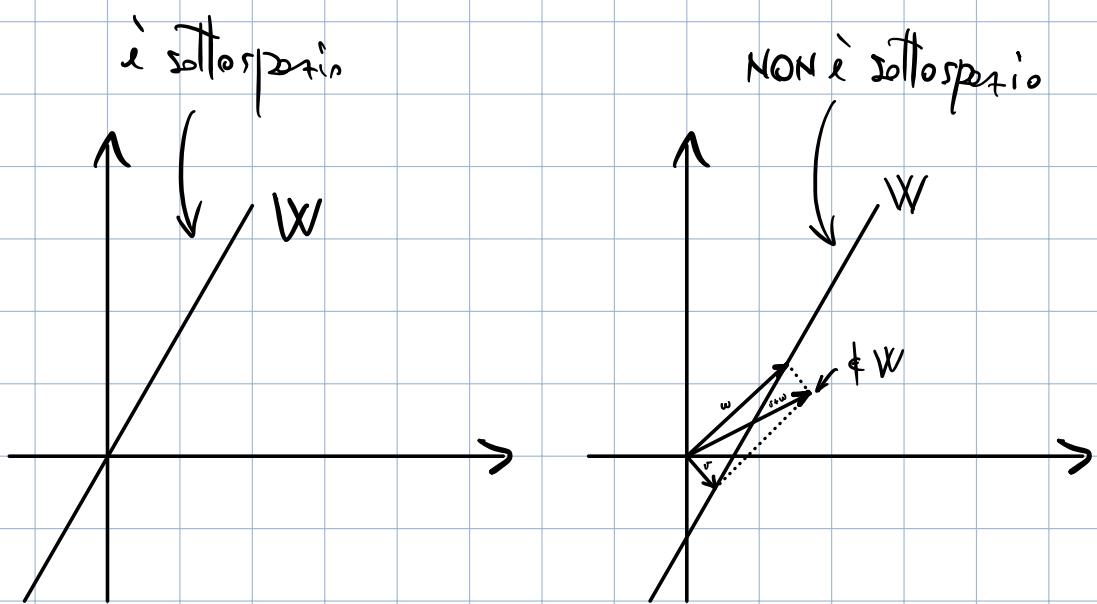


divisione di  $W$   
rispetto a .



Esempio:

$$1) \mathbb{R}^2 : \underbrace{\{(0)\}, \mathbb{R}^2}_{\text{detti "banali"}}, \text{ nette per l'origine}$$



Osservazione : il vettore nullo  $0$

dove necessariamente appartiene a

qualsiasi sotto spazio vettoriale.

2)  $\mathbb{R}^3$  :  $\{(0)\}, \mathbb{R}^3$ , rette per l'origine,  
piani per l'origine}



3)  $P_2$  è sottospazio vettoriale di

$P_3$

4)  $\{f: [a,b] \rightarrow \mathbb{R} \text{ derivabile}\}$  è

sottospazio vettoriale di  $C([a,b])$

## COMBINAZIONI LINEARI

Siano  $v_1, v_2, \dots, v_m \in V$  spazio vett.

su  $K$ , e  $\alpha_1, \alpha_2, \dots, \alpha_m \in K$ .

Allora

$$\alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_m v_m =$$

$$= \sum_{k=1}^m \alpha_k v_k$$

è detta combinazione lineare

dei vettori  $v_1, v_2, \dots, v_m$ . Gli scalari  $\alpha_1, \alpha_2, \dots, \alpha_m$  sono detti coefficienti della combinazione lineare.

Esemp<sup>i</sup>

$$1) \quad v = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad w = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

$$2v - 3w = 2 \begin{bmatrix} 1 \\ 2 \end{bmatrix} - 3 \begin{bmatrix} 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix} - \begin{bmatrix} 9 \\ 3 \end{bmatrix} = \begin{bmatrix} -7 \\ 1 \end{bmatrix}$$

è combinazione  
lineare di  $v$  e  $w$

$$2) \quad f(x) = \sin x, \quad g(x) = \cos x \in C([a, b])$$

allora  $(f + g)(x) = \sin x + \cos x$  è

combinazione lineare di  $f$  e  $g$

Attenzione: sinx cosx non lo è!

## SOTTOSPazi GENERATI

### DA VETTORI

Dono  $v_1, v_2, \dots, v_m$  vettori d'

sotto vett. su  $K$ . Allora

$$\text{Span} \{v_1, v_2, \dots, v_m\} =$$

$$= \left\{ \sum_{k=1}^m \alpha_k v_k : \alpha_i \in K \quad \forall i = 1, 2, \dots, m \right\}$$

$$= \left\{ \alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_m v_m : \alpha_i \text{ sbbri} \right\}$$

$= \{ \text{Tutte le possibili combinazioni lineari dei vettori } v_1, v_2, \dots, v_m \}$

è un sottospazio vettoriale di  $V$ ,

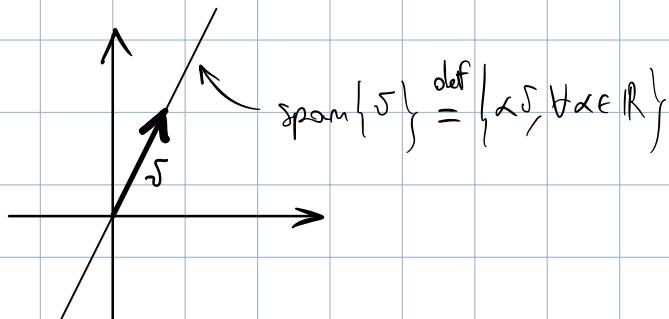
detto sottospazio vettoriale generato

da  $v_1, v_2, \dots, v_m$ .

Sono detti "generatori" del sottospazio

Esemp<sup>2</sup>

1) Se  $v \in \mathbb{R}^2$ ,  $v \neq 0$ .

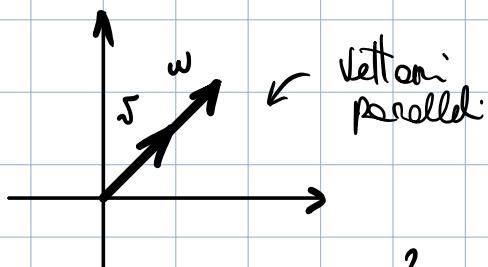


Ovvero,  $\text{span}\{\vec{v}\}$  è l'insieme dei vettori che  
giacciono sulla retta individuata da  $\vec{v}$ .  
(Vedere figura)

Sono  $\vec{v}, \omega \in \mathbb{R}^2$ ,  $\vec{v}, \omega$  non nulli.

DEFINIZIONE :  $\vec{v}, \omega \in \mathbb{R}^m$  si dicono

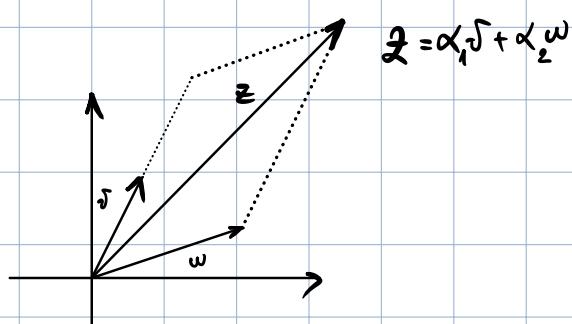
paralleli se  $\vec{v} = \alpha \omega$ , per qualche  $\alpha \in \mathbb{R}$ .



NOTAZIONE :

- $\vec{v} \parallel \omega$  :  $\vec{v}$  e  $\omega$  paralleli
- $\vec{v} \not\parallel \omega$  :  $\vec{v}$  e  $\omega$  NON paralleli

Supponiamo  $\vec{v}, \omega \in \mathbb{R}^2$  non nulli e non  
paralleli

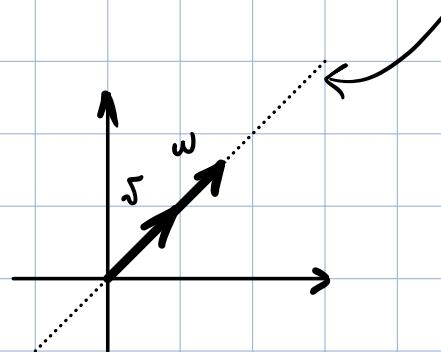


Domanda : di che è  $\text{span}\{\vec{v}, \omega\}$  ?

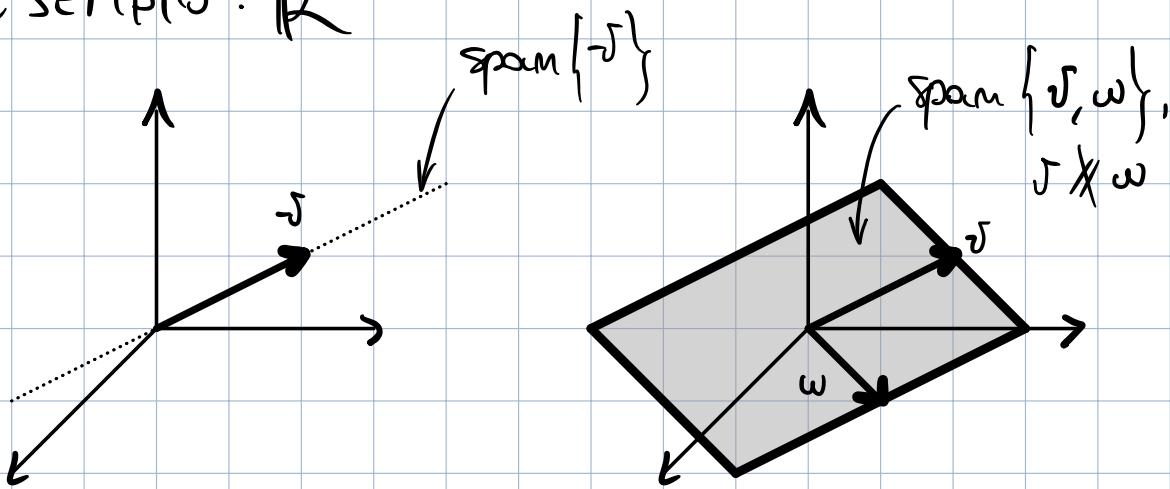
Risposta :  $\text{span}\{\vec{v}, \omega\} = \mathbb{R}^2$

Se invece  $\vec{v}, \omega \in \mathbb{R}^2$  fossero paralleli,

$\text{span}\{\vec{v}, \omega\}$  sarebbe una retta



ESEMPIO :  $\mathbb{R}^3$



E se prendessimo  $v_1, v_2, -v_3$  NON coplanari?

(... e ovviamente non nulli)

Allora  $\text{span}\{v_1, v_2, -v_3\} = \mathbb{R}^3$

Se invece  $v_1, v_2, v_3$  fossero coplanari,

Allora  $\text{Span}\{v_1, v_2, v_3\}$  sarebbe una retta  
oppure un piano.

$$\begin{aligned} \underline{\text{ESEMPIO}} : \quad & \text{Span}\{1, x, x^2\} = \\ &= \left\{ a_0 + a_1 x + a_2 x^2, \quad \forall a_i \in \mathbb{R} \right\} = \\ &= \left\{ \text{polinomi di grado } \leq 2 \right\} = \mathbb{P}_2 \end{aligned}$$

DEFINIZIONE : I vettori  $v_1, v_2, \dots, v_p \in V$

non nulli si dicono linearmente indipendenti:

se

$$\alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_p v_p = 0 \Rightarrow \alpha_1 = \alpha_2 = \dots = \alpha_p = 0$$

ovvero se l'unica combinazione lineare nulla di vettori  $v_i, i=1, 2, \dots, p$ , è data da coefficienti  $\alpha_i$  tutti nulli!

ESEMPIO :  $\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}$  non sono lin.

indip., poiché

$$1 \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 1 \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix} - 3 \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

I vettori  $\vec{v}_i$ ,  $i=1, 2, \dots, p$ , non nulli sono detti **linearmente dipendenti** se NON SONO lin. indipendenti. Che significa?

Supponiamo esistano  $\alpha_1, \alpha_2, \dots, \alpha_p$  non tutti nulli tali che  $\alpha_1 \vec{v}_1 + \alpha_2 \vec{v}_2 + \dots + \alpha_p \vec{v}_p = \vec{0}$ .

Supponiamo (per fissare le idee) sia  $\alpha_1 \neq 0$ . Allora

$$\alpha_1 \vec{v}_1 = -(\alpha_2 \vec{v}_2 + \dots + \alpha_p \vec{v}_p) \iff$$

$$\iff \vec{v}_1 = -\frac{\alpha_2}{\alpha_1} \vec{v}_2 + \dots - \frac{\alpha_p}{\alpha_1} \vec{v}_p.$$

Allora  $\vec{v}_1$  è combinazione lineare dei vettori

$\vec{v}_2, \vec{v}_3, \dots, \vec{v}_p$ . Dunque  $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_p$  sono lin.

indipendenti se nessuno di loro puo' essere

espresso come combinazione lineare degli altri.

ESEMPIO:  $\vec{v}, \omega \in \mathbb{R}^2$ ,  $\vec{v}, \omega$  non nulli

$\vec{v}, \omega$  lin. dipendenti  $\iff \vec{v} \parallel \omega$

$\vec{v}_1, \vec{v}_2, \vec{v}_3 \in \mathbb{R}^3$

$\vec{v}_1, \vec{v}_2, \vec{v}_3$  lin. dipendenti  $\iff \vec{v}_1, \vec{v}_2, \vec{v}_3$  sono  
complementari

OSSERVAZIONE: Se  $\vec{v}_1, \vec{v}_2, \vec{v}_3$  sono lin. dipendenti,

dovendo uno "non dà contributo" a  $\text{span}\{\vec{v}_1, \vec{v}_2, \vec{v}_3\}$ .

Ad esempio, sia  $\vec{v}_3$  comb. lin. di  $\vec{v}_1, \vec{v}_2$ .

Allora  $\text{span}\{\vec{v}_1, \vec{v}_2, \vec{v}_3\} = \text{span}\{\vec{v}_1, \vec{v}_2\}$ .

DEFINIZIONE Sono  $V$  spazi vett. su  $K$ ,

$W \subset V$  sottospazio vett. di  $V$ . I vettori

$\{\vec{v}_1, \vec{v}_2, \dots, \vec{v}_p\} \subset W$  costituiscono una base

di ( $\circ$  per)  $W$  se:

①  $v_1, v_2, \dots, v_p$  sono lin. indipendenti

②  $\text{span}\{v_1, v_2, \dots, v_p\} = W$ .

In tal caso, il numero  $p \in \mathbb{N}$  è

detto dimensione di  $W$ , e scriviamo

$$\dim(W) = p.$$

S'può dimostrare facilmente il seguente

TEOREMA :  $\{v_1, v_2, \dots, v_p\} \subset W$  è base per il  
sottosp. vett.  $W$  se e solo se

$\forall j \in W \quad \exists ! \alpha_1, \alpha_2, \dots, \alpha_p$  scalari t.c.

$$j = \alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_p v_p .$$

esistenza : ②

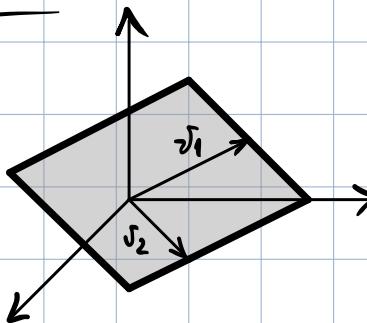
unicità : ①

Dimostrazione

Esistenza : banale

Unicità : per assurdo sfruttando la linearità  
indipendenza (esercizio)

ESEMPIO :



$v_1, v_2 \in \mathbb{R}^3$  non paralleli ;  
consideriamo  $W = \text{span}\{v_1, v_2\}$

Allora  $\{v_1, v_2\}$  i basi per  $V$ , e

$$\dim(V) = 2.$$

ESEMPIO: i vettori  $e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \dots,$

$e_m = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$  costituiscono una base per  $\mathbb{R}^n$ ,

detta base canonica di  $\mathbb{R}^n$ . In particolare,

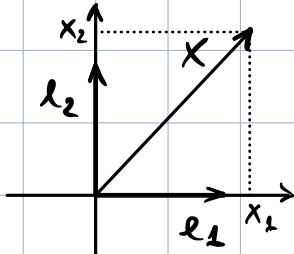
$\left\{ e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$  i basi per  $\mathbb{R}^2$ .

Infatti: se  $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$  arbitrario. Allora

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ x_2 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} =$$

$$= x_1 e_1 + x_2 e_2$$

Sono  
gli  $\alpha_i$



Domanda :  $\left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\}$  è base per  $\mathbb{R}^2$ ?

Ora, preso  $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$  arbitrario, esiste un unico modo di scrivere  $x$  come combinazione lineare di  $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$  e  $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ ?

Sarà

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \alpha_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} =$$

$$= \begin{bmatrix} \alpha_1 \\ \alpha_1 \end{bmatrix} + \begin{bmatrix} \alpha_2 \\ -\alpha_2 \end{bmatrix} = \begin{bmatrix} \alpha_1 + \alpha_2 \\ \alpha_1 - \alpha_2 \end{bmatrix}.$$

Cerchiamo  $\alpha_1, \alpha_2$  t.c.

$$\begin{cases} x_1 = \alpha_1 + \alpha_2 \\ x_2 = \alpha_1 - \alpha_2 \end{cases} \Leftrightarrow \begin{cases} x_1 + x_2 = 2\alpha_1 \\ x_1 - x_2 = 2\alpha_2 \end{cases} \Leftrightarrow$$

Somma / sottrazione

$$\Leftrightarrow \begin{cases} \alpha_1 = \frac{1}{2}(x_1 + x_2) \\ \alpha_2 = \frac{1}{2}(x_1 - x_2) \end{cases}$$

Risposta: sì,  $\left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right\}$  è base per  $\mathbb{R}^2$ !

ESEMPIO:

$$x = \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \text{ allora}$$

$$\begin{pmatrix} 2 \\ 4 \end{pmatrix} = 3 \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

TEOREMA: Se  $W$  spazio vett.,  $\dim(W) = p$ .

Allora:

①  $p$  è il più grande numero d' vettori

lin. indip. d'  $W$ ;

②  $p$  è il più piccolo numero d' generatori

d'  $W$ .

OSSERVAZIONI:

① La base non è unica (Ved. esempio precedente), ma due basi qualsiasi devono essere costituite dallo stesso numero di elementi! (c'è giustificata la def. di dimensione)

②  $\dim(P_2) = 3$  (poiché  $\{1, x, x^2\}$  ne costituisce una base)

$$\dim(\{\text{polinomi di grado arbitrario}\}) = \infty$$

poiché  $\{1, x, x^2, \dots, x^k, \dots\}$  ne è base!

Indichiamo l'insieme dei polinomi di grado arbitrario con  $P_\infty$ .

Siamo  $A \in \mathbb{R}^{m \times m}$ ,  $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$ ,  $b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$ . Allora

Possiamo considerare il sistema di equazioni lineari (anche detto "sistema lineare") espresso dalla relazione matriciale:

$$Ax = b,$$

costituito da  $m$  equazioni in  $m$  incognite.

DEFINIZIONE: La matrice  $\underbrace{[A, b]}_{\text{"a blocchi"}}$   $\in \mathbb{R}^{m \times (m+1)}$

è detta matrice completa del sistema lineare.

DEFINIZIONI: L'insieme

$$\text{"Kernel"} \rightsquigarrow \text{Ker}(A) := \left\{ v \in \mathbb{R}^m \text{ t.c. } A v = \overset{\text{vettore in } \mathbb{R}^m}{\underset{\downarrow}{0}} \right\} \text{ è}$$

detto nucleo di  $A$ .

OSSERVAZIONI

1)  $0 \in \text{Ker}(A)$  (il vettore nullo appartiene sempre al nucleo)

2)  $\text{Ker}(A) \subset \mathbb{R}^n$ , sottosistema.

TEOREMA  $\text{Ker}(A)$  è sottospazio vettoriale di  $\mathbb{R}^m$ .

Dimostrazione.

1)  $\forall \varsigma, \omega \in \text{Ker}(A) \stackrel{\text{def.}}{\Leftrightarrow} A\varsigma = 0 \wedge A\omega = 0$ .

$$\begin{aligned} A(\varsigma + \omega) &= A\varsigma + A\omega = 0 + 0 = 0 \Rightarrow \\ &\Rightarrow \varsigma + \omega \in \text{Ker}(A) \end{aligned}$$

2)  $\varsigma \in \text{Ker}(A), \alpha \in \mathbb{R}$ :

$$A(\alpha \varsigma) = \alpha(A\varsigma) = \alpha 0 = 0$$

↪ "gli scalari li sposta  
dove voglio"



DEFINIZIONE.

L'immagine

$$\text{Im}(A) := \left\{ y \in \mathbb{R}^m : \exists x \in \mathbb{R}^n \text{ t.c. } y = Ax \right\}$$

è detto immagine di  $A$ .

## OSSERVAZIONI

$$1) \quad 0 \in \text{Im}(A)$$

$$2) \quad \text{Im}(A) \subset \mathbb{R}^m, \text{ sottoinsieme}$$

TEOREMA :  $\text{Im}(A)$  è sottospazio vettoriale di  $\mathbb{R}^m$ .

Dimostrazione: per esercizio.

## OSSERVAZIONI

$$1) \quad \text{Im}(A) = \left\{ b \in \mathbb{R}^m \text{ t.c. } Ax=b \text{ ammette soluzione} \right\}$$

Quindi

"se e solo se"

$$Ax=b \text{ ammette soluzione} \Leftrightarrow b \in \text{Im}(A)$$

$$2) \quad \text{siano } b \in \text{Im}(A) \text{ e } \gamma \in \text{Ker}(A), \gamma \neq 0.$$

L'è  $x \in \mathbb{R}^m$  t.c.  $Ax=b$ . Consideriamo  $x+\gamma$ :

$$A(x+\gamma) = Ax + A\gamma = b + 0 = b \Rightarrow$$

$x+\gamma$  risolve il sistema lineare  $Ax=b$ !

DEFINIZIONE: A ha "nucleo banale" se  $\text{Ker}(A) = \{0\}$ .

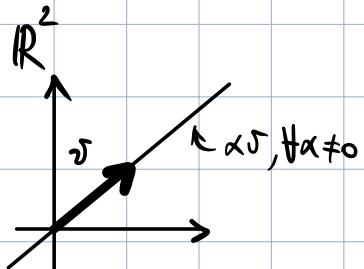
Allora, dato  $b \in \text{Im}(A)$ ,

$Ax=b$  ammette unica soluzione se

e solo se A ha nucleo banale!

Nucleo non banale  $\Leftrightarrow$  "multiple solutions"

Richiamo



Un sottospazio non banale ha  $\infty$  elementi!

Quindi, se  $b \in \text{Im}(A)$  e  $\text{Ker}(A)$  non è banale,  $Ax=b$  ammette infinte soluzioni!

$$\begin{aligned} 3) \quad \text{Im}(A) &= \left\{ \text{Tutti i vettori del tipo } Ax, \text{ con } x \in \mathbb{R}^m \right\} = \\ &= \left\{ \text{Tutte le possibili combinazioni lineari di} \underbrace{\text{colonne di } A}_{\text{combinazioni lin. colonne di } A} \right\} \end{aligned}$$

## DEFINIZIONE

Dato  $A \in \mathbb{R}^{m \times n}$ , si definisce Rango di A

il seguente numero intero:

$$\text{rank}(A) = \dim(\text{Im}(A))$$

## OSSERVAZIONI :

(1)

Essendo  $\text{Im}(A) = \left\{ \begin{array}{l} \text{Tutte le comb. lin. di colonne} \\ \text{di } A \end{array} \right\}'$ ,

s'ha che  $\text{rank}(A) \leq m$

(2)  $\text{Im}(A)$  è sottospazio di  $\mathbb{R}^m \Rightarrow$

$$\text{rank}(A) \leq m$$

Ne deduciamo il

TEOREMA : Se  $A \in \mathbb{R}^{m \times n}$ , allora

$$\text{rank}(A) \leq \min \{ m, n \}$$

# NUCLEO, RANGO E OPERAZIONI ELEMENTARI

TEOREMA Siano  $A \in \mathbb{R}^{M \times M}$  e  $B \in \mathbb{R}^{M \times M}$  invertibile.

$$\text{Allora } \text{Ken}(A) = \text{Ken}(BA)$$

Ovvero, "il nucleo è invariante rispetto a premoltiplicazione per matrici invertibili".

Dimo.

$$\begin{aligned} J \in \text{Ken}(A) &\stackrel{\text{def}}{\Leftrightarrow} AJ = 0 \stackrel{\text{def}}{\Leftrightarrow} BAJ = 0 \stackrel{\text{def}}{\Leftrightarrow} \\ &\Leftrightarrow J \in \text{Ken}(BA) \quad \square \end{aligned}$$

Ora, diamo  $A \in \mathbb{R}^{M \times M}$  e  $B \in \mathbb{R}^{M \times M}$  invertibile.

Sia  $\{v_1, v_2, \dots, v_n\}$  base per  $\text{Im}(A)$ .

Osserviamo che

$$\text{Im}(A) = \left\{ \begin{array}{l} \text{insieme di tutte le combinazioni} \\ \text{lineari di } v_1, v_2, \dots, v_n \end{array} \right\}$$

Ovvero

$y \in \text{Im}(A) \Leftrightarrow$  possibile scrivere in modo unico

$$y = \sum_{k=1}^n \alpha_k v_k$$

Premoltiplicando per  $B$ , deduce che ogni elemento di  $\text{Im}(BA)$  si può scrivere in modo unico come

$$\sum_{k=1}^n \alpha_k B v_k$$

Quindi  $\{Bv_1, Bv_2, \dots, Bv_n\}$  è base per  $\text{Im}(BA)$ .

Conseguenza importante:

TEOREMA Se  $A \in \mathbb{R}^{m \times n}$  e  $B \in \mathbb{R}^{n \times m}$  invertibile, allora  $\text{rank}(A) = \text{rank}(BA)$ .

Dimo. Conseguenza del fatto che  $M(n)$ :

invertibili "moniamo così in basso"  $\square$

Ovvero, "il range è invariante rispetto a premoltiplicazione per matrici invertibili".

ESEMPI DI MATRICI B E OPERAZIONI

AD ESSE COLLEGATE

(1)

$$\underbrace{\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_B \underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}}_A = \begin{bmatrix} 5 & 7 & 9 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

- abbiamo sommato alla prima riga di A la seconda riga di A
- $\det(B) = 1 \Rightarrow B$  è invertibile
- posso operare anche a righe arbitrarie di A di dimensione dimensione arbitraria

(2)

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_B \underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}}_A = \begin{bmatrix} 1 & 2 & 3 \\ 8 & 10 & 12 \\ 7 & 8 & 9 \end{bmatrix}$$

- abbiamo moltiplicato per 2 la seconda riga di A
- $\det(B) = 2 \Rightarrow B$  invertibile
- posso generalizzare a moltiplicazione per uno scalare  $\alpha \neq 0$  arbitrario d'una riga arbitraria d'A ol' dimensione arbitraria

(3)

$$\underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_B \underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}}_A = \begin{bmatrix} 4 & 5 & 6 \\ 1 & 2 & 3 \\ 7 & 8 & 9 \end{bmatrix}$$

- abbiamo scambiato le prime righe d'A

con le seconde righe di A

- $\det(B) = -1$  ( $\begin{array}{l} \text{si ottiene da I sommissione} \\ \text{le prime due righe} \end{array}$ )

dunque B è invertibile

- posso generalizzare allo scambio di righe arbitrarie d'A di dimensione arbitraria

Notiamo che, unendo (1) e (2),

otteniamo l'operazione di sommare

ad una riga un multiplo di un'altra riga.

DEFINIZIONE: "Sommare a una riga un multiplo di un'altra riga" e "sottrarre due righe tra di loro" sono dette operazioni elementari (tre righe).

TEOREMA Se  $A \in \mathbb{R}^{m \times n}$ , le operazioni elementari tra righe di  $A$  consentono molte e varie!

### ESEMPIO

$$A = \begin{bmatrix} 1 & -2 & 2 & 2 \\ -3 & 5 & -5 & -7 \\ 2 & -5 & 5 & 2 \end{bmatrix} = \begin{bmatrix} \text{righe di } A \\ R_1 \\ R_2 \\ R_3 \end{bmatrix}$$

Li voglio annullare

Vogliamo determinare  $\text{rank}(A)$ .

Idea: semplifichiamo  $A$  mediante operazioni elementari.

1

$$\begin{aligned} R_1 &\rightarrow R_1 \\ R_2 + 3R_1 &\rightarrow R_2 \\ R_3 - 2R_1 &\rightarrow R_3 \end{aligned}$$

$$\begin{bmatrix} 1 & -2 & 2 & 2 \\ 0 & -1 & 1 & -1 \\ 0 & -1 & 1 & -2 \end{bmatrix}$$

Lo voglio annullare

2

$$\begin{array}{l}
 R_1 \rightarrow R_1 \\
 R_2 \rightarrow R_2 \\
 R_3 - R_2 \rightarrow R_3
 \end{array}
 \quad
 \left[ \begin{array}{cccc}
 1 & -2 & 2 & 2 \\
 0 & -1 & 1 & -1 \\
 0 & 0 & 0 & -1
 \end{array} \right]$$

Mediante operazioni elementari

abbiamo trasformato  $A$  in

$$U = \left[ \begin{array}{cccc}
1 & -2 & 2 & 2 \\
0 & -1 & 1 & -1 \\
0 & 0 & 0 & -1
\end{array} \right]$$

Si ha  $\text{Rank}(A) = \text{Rank}(U)$ , ma

$U$  ha una struttura detta "a gradini", che mi rivelò facilmente il range.

gradini

$$\xrightarrow{\hspace{2cm}}
 \left[ \begin{array}{cccc}
 1 & -2 & 2 & 2 \\
 0 & -1 & 1 & -1 \\
 0 & 0 & 0 & -1
 \end{array} \right]$$

Ogni colonna corrispondente ad un gradino è lineare indipendente dalle precedenti!

Tre gradini  $\Rightarrow \text{rank}(A) = \text{rank}(U) = 3$ .

Consideriamo

$$A = \begin{bmatrix} 1 & -2 & 2 & 2 \\ 0 & -1 & 1 & -1 \\ 0 & 0 & 0 & -1 \end{bmatrix},$$

mediante operazioni elementari

abbiamo trasformato  $A$  in

$$U = \begin{bmatrix} 1 & -2 & 2 & 2 \\ 0 & -1 & 1 & -1 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

$U$  ha una struttura detta "a gredini".

DEFINIZIONE  $U \in \mathbb{R}^{m \times n}$  è detta "a

gredini" se : (1) il primo elemento non nullo di ogni riga s' trova a destra del primo elemento non nullo delle righe precedenti ; (2) eventuali righe nulle s' trovano in fondo

alla matrice. Formalmente :

$$(1) \forall i \geq 2: \min \{ j : U_{ij} \neq 0 \} > \min \{ j : U_{i-1,j} \neq 0 \}$$

$$(2) \text{ se } \{ j : U_{ij} \neq 0 \} = \emptyset, \text{ allora}$$

$$\{ j : U_{kj} \neq 0 \} = \emptyset \text{ per ogni } k > i$$

DEFINIZIONE. Se  $U$  è in forma a gradini,

il primo elemento non nullo di ciascuna riga è detto pivot oppure elemento pivotale.

Formalmente, per ogni  $i \geq 1$ ,

$U_{ij}$  è pivot se  $j = \min \{ k : U_{kj} \neq 0 \}$ , dove

$$\{ k : U_{kj} \neq 0 \} \neq \emptyset.$$

ESEMPIO :  $U = \begin{bmatrix} 1 & -2 & 2 & 2 \\ 0 & -1 & 1 & -1 \\ 0 & 0 & 0 & -1 \end{bmatrix}$

gradini

pivot

DEFINIZIONI. Se  $U \in \mathbb{R}^{m \times n}$  è a gradini, le colonne di  $U$  che contengono un pivot sono dette colonne pivotali. Le altre sono dette colonne non pivotali.

ESEMPPIO PRECEDENTE

$$\left\{ \begin{array}{l} \text{colonne pivotali: } 1, 2, 4 \\ \text{colonne non pivotali: } 3 \end{array} \right.$$

DEFINIZIONE. Se  $U \in \mathbb{R}^{m \times n}$  è a gradini ed è stata ottenuta da  $A \in \mathbb{R}^{m \times m}$  mediante operazioni elementari (free right), allora diremo che  $U$  è una forma a gradini d'  $A$ .

↪ non è unica

$$A = \begin{bmatrix} x & x & x & x \\ x & x & x & x \\ x & x & x & x \\ x & x & x & x \end{bmatrix} \xrightarrow{\text{operazioni elementari}} U = \begin{bmatrix} x & x & x & x \\ x & x & x & x \\ \cancel{x} & \cancel{x} & \cancel{x} & \cancel{x} \end{bmatrix}$$

pivot

È evidente che il rango di  $U$

è dato dal numero d' pivot di  $U$ .

Ne deriva il seguente:

TEOREMA: Si  $U \in \mathbb{R}^{m \times m}$  è una forma  
a gradini di  $A \in \mathbb{R}^{m \times n}$ , allora

$$\text{rank}(A) = \# \text{ d' pivot di } U.$$

Dimo  $A \xrightarrow{\substack{\text{operazioni} \\ \text{elementari}}} U$ , e le operazioni elementari  
conservano il rango.

ESEMPIO PRECEDENTE :

$$A = \begin{bmatrix} 1 & -2 & 2 & 2 \\ -3 & 5 & -5 & -7 \\ 2 & -5 & 5 & 2 \end{bmatrix} \rightarrow U = \begin{bmatrix} 1 & -2 & 2 & 2 \\ 0 & -1 & 1 & -1 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

Dunque  $\text{rank}(A) = 3$ .

DEFINIZIONE Siamo  $V$  spazio vettoriale su  $\mathbb{R}$ ,

$W \subset V$  sottospazio di  $V$ . Se  $\{\nu_1, \nu_2, \dots, \nu_p\}$  è una base di  $W$ . Allora si ha

$$W = \left\{ \nu \in V : \nu = \alpha_1 \nu_1 + \alpha_2 \nu_2 + \dots + \alpha_p \nu_p, \alpha_k \in \mathbb{R} \right\}.$$

Questa rappresentazione è detta parametrizzazione di ( $p$  per)  $W$ .

Viceversa, se  $W = \left\{ \nu \in V : \nu = \alpha_1 \nu_1 + \alpha_2 \nu_2 + \dots + \alpha_p \nu_p, \alpha_k \in \mathbb{R} \right\}$  e  $\nu_1, \nu_2, \dots, \nu_p$  sono lin. indip., allora  $\{\nu_1, \nu_2, \dots, \nu_p\}$  è una base per  $W$ .



Base per  $\text{Ker}(A)$

ESEMPPIO Siano  $A$  e  $U$  tali che :

$$A = \begin{bmatrix} 1 & -2 & 2 & 2 \\ -3 & 5 & -5 & -7 \\ 2 & -5 & 5 & 2 \end{bmatrix} \xrightarrow{\text{op. elem.}} U = \begin{bmatrix} 1 & -2 & 2 & 2 \\ 0 & -1 & 1 & -1 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

Sappiamo che  $\text{Ker}(A) = \text{Ker}(U)$ , ovvero

$$Ax = 0 \iff Ux = 0, \text{ con } x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

$$\text{Ker}(U) = \left\{ \text{soltuzioni del sist. lin. } Ux = 0 \right\}$$

Dobbiamo risolvere il sistema :

$$\begin{cases} x_1 - 2x_2 + 2x_3 + 2x_4 = 0 \\ -x_2 + x_3 - x_4 = 0 \\ -x_4 = 0 \end{cases}$$

$$U = \left[ \begin{array}{cccc} 1 & -2 & 2 & 2 \\ 0 & -1 & 1 & -1 \\ 0 & 0 & 0 & -1 \end{array} \right]$$

$x_1 \quad x_2 \quad x_3 \quad x_4$

$x_1, x_2, x_4$  variabili pivotali

$x_3$  variabile NON pivotale

La variabile non pivotale  $x_3$  divenuta

parametro arbitrario del problema e le spostiamo a destra del segno di uguaglianza:

$$\begin{cases} X_1 - 2X_2 + 2X_4 = -2X_3 \\ -X_2 - X_4 = -X_3 \\ -X_4 = 0 \end{cases} \quad \begin{array}{l} \text{esimili al} \\ \text{Tennino mto} \end{array}$$

Adessoolviamo il s.t. lin. rispetto a  $X_1, X_2, X_4$  per "sostituzione all'indietro":

$$\begin{array}{l} \xrightarrow{\text{Eq1}} \begin{cases} X_1 = 2X_2 - 2X_3 = 0 \\ X_2 = X_3 \\ X_4 = 0 \end{cases} \quad \begin{array}{l} \text{quindi l'insieme delle} \\ \text{soltuzioni di } \sum x = 0 \text{ è} \end{array} \\ \xrightarrow{\text{Eq2}} \\ \xrightarrow{\text{Eq3}} \end{array}$$

$$\left\{ X \in \mathbb{R}^4 : X = \begin{bmatrix} 0 \\ X_3 \\ X_3 \\ 0 \end{bmatrix}, \forall X_3 \in \mathbb{R} \right\} =$$

$$= \left\{ X \in \mathbb{R}^4 : X = \alpha \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \forall \alpha \in \mathbb{R} \right\} .$$

Conclusioni:  $\left\{ \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \right\}$  è base per  $\text{Ker}(A)$ .

Richiami e considerazioni:

- 1) possiamo trasformare  $A \in \mathbb{R}^{m \times m}$  in  $U \in \mathbb{R}^{m \times m}$  in forma a gradini mediante una sequenza di operazioni elementari
- 2) ogni operazione elementare corrisponde alla premoltiplicazione per un'opportuna matrice invertibile
- 3) il prodotto di matrici invertibili è invertibile.

Conclusione:  $\det A \in \mathbb{R}^{m \times m}, \exists M \in \mathbb{R}^{m \times m}$  invertibile t.c.

$$M A = U$$

Si partitionano  $A$  e  $U$  per colonne,

$$A = \begin{bmatrix} | & | & | \\ a_1 & a_2 & \cdots & a_m \\ | & | & | \end{bmatrix} \text{ e } U = \begin{bmatrix} | & | & | \\ u_1 & u_2 & \cdots & u_m \\ | & | & | \end{bmatrix},$$

Allora  $u_1 = Ma_1, \dots, u_m = Ma_m$

Per determinare una base per  $\text{Im}(A)$ ,  
basta ricordare che le matrici invertibili  
trasformano basi in basi.

Esempio

Le colonne pivotali di  $U = \begin{bmatrix} 1 & -2 & 2 & 2 \\ 0 & -1 & 1 & -1 \\ 0 & 0 & 0 & -1 \end{bmatrix}$

Costituiscono fondamente una base per  $\text{Im}(U)$ .

Avendo le corrispondenti colonne di

$$A = \begin{bmatrix} 1 & -2 & 2 & 2 \\ -3 & 5 & -5 & -7 \\ 2 & -5 & 5 & 2 \end{bmatrix} \quad \text{costituiscono una base}$$

per  $\text{Im}(A)$ . Allora :

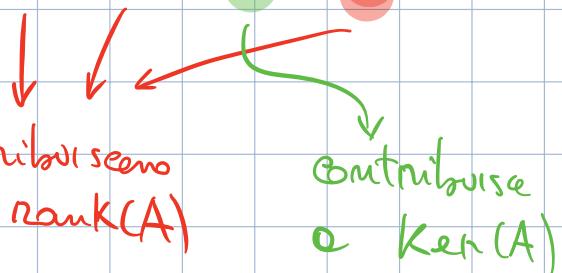
$$\left\{ \begin{bmatrix} 1 \\ -3 \\ 2 \end{bmatrix}, \begin{bmatrix} -2 \\ 5 \\ -5 \end{bmatrix}, \begin{bmatrix} 2 \\ -7 \\ 2 \end{bmatrix} \right\} \text{ è base per } \text{Im}(A).$$

TEOREMA Se  $A \in \mathbb{R}^{m \times m}$ . Allora

$$\dim(\text{Ker}(A)) + \text{rank}(A) = m$$

Dimostrazione: i owing! (esempio precedente :

$$U = \begin{bmatrix} 1 & -2 & 2 & 2 \\ 0 & -1 & 1 & -1 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \text{ ma s' pu' generalizzare}$$

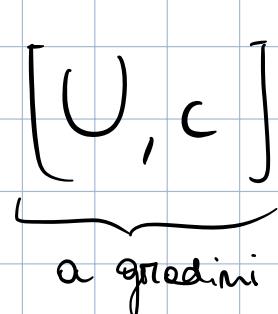
  
Contribuiscono  
a  $\text{rank}(A)$   
Contribuisce  
a  $\text{Ker}(A)$

Obiettivo: risolvere il sistema lineare

$$Ax = b, \quad A \in \mathbb{R}^{m \times n} \quad e \quad b \in \mathbb{R}^m.$$

Strategia:

(1)  $[A, b] \xrightarrow{\text{op. elem.}} [U, c]$

  
a gradini

(2) Risulta rispetto alle Variabili pivotali il sistema  $Ux = c$  per sostituzione all'indietro, dopo aver portato a destra del segno di uguaglianza le Variabili non pivotali (che diventano parametri arbitrari della soluzione).

C' sono 3 possibili scenari:

$$\left[ \begin{array}{ccc|c} x & x & x & x \\ x & x & x & x \\ x & & x & x \end{array} \right]$$

$\underbrace{\quad\quad\quad}_{U}$        $\uparrow$   
                 $c$

Tutte e sole le colonne di  $U$  sono pivotali.



$\exists!$  soluzione  
di  $Ux = c$

$$\text{rank}(U) = \text{rank}([U, c])$$

$$\text{rank}(U) = m$$

$$\left[ \begin{array}{cc|cc} x & x & x & x \\ & & x & x \\ & & x & x \end{array} \right]$$

c' è un pivot nelle colonne  $c$

↓  
ultima riga  
è impossibile

↓  
 $\nexists$  soluzioni

$$\left[ \begin{array}{ccc|c} x & x & x & x \\ x & x & x & x \\ x & x & x & x \end{array} \right]$$

nessun pivot  
in  $c$ , presente  
di una colonna  
non pivotale in  $U$

↓  
 $\exists \infty$  soluzioni

$$\text{rank}(U) = \text{rank}([U, c])$$

$$\text{rank}(U) < m$$

Ricordiamo:

$$\text{rank}(A) = \text{rank}(U)$$

$$\text{rank}([A, b]) = \text{rank}([U, c])$$

Ricordiamo tutto nel seguente

TEOREMA (Rouché-Capelli)

Siano  $A \in \mathbb{R}^{m \times m}$  e  $b \in \mathbb{R}^m$ . Allora:

(1) il sistema  $Ax=b$  ammette soluzione  
se e solo se

$$\text{rank}(A) = \text{rank}([A, b])$$

(2) se  $\text{rank}(A) = \text{rank}([A, b])$ . Allora

Le soluzioni di  $Ax=b$  è unica se e solo se

$$M=1.$$

Altrimenti esistono  $\infty^{M-n}$  soluzioni,

allora l'insieme delle soluzioni è parametrizzato da  $M-n$  parametri arbitri.

Esempio precedente:  $\mathbf{Ux} = \mathbf{0}$  ammette

$\infty^1$  soluzioni perché  $\text{Ker}(\mathbf{U})$  è parametrizzato mediante un solo

parametro. Infatti

$$\# \text{ colonne di } \mathbf{U} - \text{rank}(\mathbf{U}) =$$

$$3 - 2 = 1$$

## ESERCIZIO

Siamo

$$A = \begin{bmatrix} 2 & -1 & 1 \\ -2 & 2 & \alpha \\ 2 & -3 & -1 \end{bmatrix}, \quad b = \begin{bmatrix} \beta \\ -6 \\ 8 \end{bmatrix}, \quad \alpha, \beta \in \mathbb{R}.$$

Determinare i valori di  $\alpha$  e  $\beta$  per i quali il sistema  $Ax=b$ :

- (a) NON ammette soluzione,
- (b) ammette unica soluzione,
- (c) ammette infinite soluzioni.

Per i valori di  $\alpha$  e  $\beta$  del punto (c), determinare:

- (1) Tutte le possibili soluzioni del sist. lin.  $Ax=b$ ,
- (2) una base per  $\text{Im}(A)$ ,
- (3) una base per  $\text{Ker}(A)$ ,
- (4) il rango di  $A$ .

## SOLUZIONE

Forniamo la matrice completa:

$$(A:b) = \left[ \begin{array}{ccc|c} 2 & -1 & 1 & \beta \\ -2 & 2 & \alpha & -6 \\ 2 & -3 & -1 & 8 \end{array} \right] \begin{matrix} \leftarrow R_1 \\ \leftarrow R_2 \\ \leftarrow R_3 \end{matrix}$$

La portiamo in forma a gradini:

$$R_2 + R_1 \rightarrow \left[ \begin{array}{ccc|c} 2 & -1 & 1 & \beta \\ 0 & 1 & \alpha+1 & \beta-6 \\ 0 & -2 & -2 & 8-\beta \end{array} \right]$$

$$R_3 - R_1 \rightarrow \left[ \begin{array}{ccc|c} 2 & -1 & 1 & \beta \\ 0 & 1 & \alpha+1 & \beta-6 \\ 0 & 0 & 2\alpha & \beta-4 \end{array} \right]$$

$$R_3 + 2R_2 \rightarrow \left[ \begin{array}{ccc|c} 2 & -1 & 1 & \beta \\ 0 & 1 & \alpha+1 & \beta-6 \\ 0 & 0 & 2\alpha & \beta-4 \end{array} \right]$$

(a) pivot lungo la colonna dei termini noti:

$$\alpha = 0, \beta \neq 4$$

(b)  $\alpha \neq 0$

(c)  $\alpha = 0, \beta = 4$

S'ha  $[U:c] = \left[ \begin{array}{ccc|c} 2 & -1 & 1 & 4 \\ 0 & 1 & 1 & -2 \\ 0 & 0 & 0 & 0 \end{array} \right]$

(1)  $\begin{cases} 2x_1 - x_2 + x_3 = 4 \\ x_2 + x_3 = -2 \end{cases}$  ;  $x_3$  è non pivotale

$$\begin{cases} 2x_1 - x_2 = 4 - x_3 \\ x_2 = -2 - x_3 \end{cases} \iff \begin{cases} 2x_1 = x_2 + 4 - x_3 = -2 - x_3 + 4 - x_3 \\ x_2 = -2 - x_3 \end{cases} \iff$$

$$\iff \begin{cases} x_1 = 1 - x_3 \\ x_2 = -2 - x_3 \end{cases}$$

Soluzioni :  $\left\{ \begin{bmatrix} 1-x_3 \\ -2-x_3 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \\ 0 \end{bmatrix} + \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix} x_3, \forall x_3 \in \mathbb{R} \right\}.$

parametrizzazione dell'insieme delle soluzioni

(2) base per  $\text{Im}(A)$  :  $\left\{ \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix}, \begin{bmatrix} -1 \\ 2 \\ -3 \end{bmatrix} \right\}$

(3) base per  $\text{Ker}(A)$  :  $\left\{ \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix} \right\}$

(4)  $\text{rank}(A) = 2$

## PREMessa

$\text{Se } Ax = b \text{ sist. lin. con } A \in \mathbb{R}^{m \times m}$ .

Rossibilità :



A "fat",  $m < n$ :

$$\begin{array}{c|c|c} A & = & b \\ \hline & & x \end{array}$$

Aspettativa:

c' attendiamo  
 $\infty$  soluzioni

A quadrata,  $m = n$ :

$$\begin{array}{c|c|c} A & = & b \\ \hline & x & \end{array}$$

c' attendiamo  
UNICA soluzione

A "skinny",  $m > n$ :

$$\begin{array}{c|c|c} A & = & b \\ \hline & x & \end{array}$$

Non c' attendiamo  
soluzioni

Affrontiamo prima il caso delle matrici quadrate, ovvero  $\#\text{incognite} = \#\text{equazioni}$ .

### ESEMPIO

Consideriamo

$$\begin{cases} -x_1 + 2x_2 + x_3 = 5 & \leftarrow \bar{E}_q 1 \\ x_1 - x_2 + x_3 = -6 & \leftarrow \bar{E}_q 2 \\ 4x_1 - 11x_2 - 13x_3 = -11 & \leftarrow \bar{E}_q 3 \end{cases}$$

Applichiamo il metodo di eliminazione di Gauss:

PRIMO PASSO: eliminiamo da  $\bar{E}_q 2$  e  $\bar{E}_q 3$  la variabile  $x_1$  sottraendo a  $\bar{E}_q 2$  e  $\bar{E}_q 3$  un multiplo di  $\bar{E}_q 1$ .

$$\begin{array}{l} \bar{E}_q 1 \\ \bar{E}_q 2 + \bar{E}_q 1 \\ \bar{E}_q 3 + 4\bar{E}_q 1 \end{array} : \quad \begin{cases} -x_1 + 2x_2 + x_3 = 5 \\ x_2 + 2x_3 = -1 \\ -3x_2 - 9x_3 = 9 \end{cases}$$

SECONDO PASSO: eliminiamo da  $\bar{E}_q 3$  la variabile  $x_2$  sottraendo a  $\bar{E}_q 3$  un multiplo di  $\bar{E}_q 2$ .

$$\begin{array}{l} \bar{E}_q 1 \\ \bar{E}_q 2 \\ \bar{E}_q 3 + 3\bar{E}_q 2 \end{array} : \quad \begin{cases} -x_1 + 2x_2 + x_3 = 5 \\ x_2 + 2x_3 = -1 \\ -3x_3 = 6 \end{cases}$$

Aesso posso risolvere per SOSTITUZIONE ALL'INDIETRO:

$$\begin{array}{l} \xrightarrow{\text{Eq.1}} \\ \xrightarrow{\text{Eq.2}} \\ \xrightarrow{\text{Eq.3}} \end{array} \left\{ \begin{array}{l} x_1 = -5 + 2x_2 + x_3 = -5 + 6 - 2 = -1 \\ x_2 = -1 - 2x_3 = 3 \\ x_3 = -2 \end{array} \right.$$

Soluzione:  $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -1 \\ 3 \\ -2 \end{bmatrix}$ .

Algoritmo nel caso generale di  $n$  Eq.mi in  $n$  incognite:

- per  $k = 1, 2, \dots, n-1$

: Eliminare da  $\bar{\text{Eq}}_{k+1}, \bar{\text{Eq}}_{k+2}, \dots, \bar{\text{Eq}}_n$

l'incognita  $x_k$  sommando e sottraendo di queste eq.m.i un multiplo di  $\bar{\text{Eq}}_k$ ,

se possibile.

fine

Qui non è possibile:

$$\begin{cases} x_2 + x_3 = 0 \\ x_1 + 2x_2 + x_3 = 1 \\ x_1 + x_2 - x_3 = 2 \end{cases}$$

- Risolviamo all'indietro  $\bar{\text{Eq}}_n \rightarrow \dots \rightarrow \bar{\text{Eq}}_2 \rightarrow \bar{\text{Eq}}_1$

REINTERPRETIAMO le operazioni dell'esempio precedente come operazioni elementari sulle

Nucleo di  $A$ :

$$A = \begin{bmatrix} -1 & 2 & 1 \\ 1 & -1 & 1 \\ 4 & -11 & -13 \end{bmatrix} \rightarrow R_1 : \begin{bmatrix} -1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & -3 & -9 \end{bmatrix} \rightarrow$$

$$R_2 + R_1 : \begin{bmatrix} -1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & -3 & -9 \end{bmatrix} \rightarrow$$

$$R_3 + 4R_1 : \begin{bmatrix} -1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & -3 \end{bmatrix} \Rightarrow$$

$$\rightarrow R_1 : \begin{bmatrix} -1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & -3 \end{bmatrix} =: U$$

$$R_2 : \begin{bmatrix} -1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & -3 \end{bmatrix}$$

$$R_3 + 3R_2 : \begin{bmatrix} -1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & -3 \end{bmatrix}$$

REINTERPRETIAMO le operazioni elementari

come premoltiplicazione per opportune Matrici:

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 4 & 0 & 1 \end{bmatrix}}_{M_1} \underbrace{\begin{bmatrix} -1 & 2 & 1 \\ 1 & -1 & 1 \\ 4 & -11 & -13 \end{bmatrix}}_{A^{(0)}} = \underbrace{\begin{bmatrix} -1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & -3 & -9 \end{bmatrix}}_{A^{(1)}}$$

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3 & 1 \end{bmatrix}}_{M_2} \underbrace{\begin{bmatrix} -1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & -3 & -9 \end{bmatrix}}_{A^{(1)}} = \underbrace{\begin{bmatrix} -1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & -3 \end{bmatrix}}_{A^{(2)}}$$

dove  $A^{(0)} = A$  e  $U := A^{(2)}$ .

Algebraicamente :

$$M_2 M_1 A = J$$

OSSERVAZIONI :

①  $\det(M_1) = \det(M_2) = 1 \Rightarrow$

$\Rightarrow M_1, M_2$  invertibili

②  $M_2 M_1 A = J \Leftrightarrow M_1 A = M_2^{-1} J \Leftrightarrow$

$$\Leftrightarrow A = M_1^{-1} M_2^{-1} J$$

③ Ottenere  $M_1^{-1}$  e  $M_2^{-1}$  è banale. Infatti :

$$\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -4 & 0 & 1 \end{bmatrix} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 4 & 0 & 1 \end{bmatrix}}_{M_1} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_I$$

e dunque  $M_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -4 & 0 & 1 \end{bmatrix}$

Analogamente

$$M_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3 & 1 \end{bmatrix} \Rightarrow M_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -3 & 1 \end{bmatrix}$$

④ ora moltiplicate  $M_1^{-1}$  per  $M_2^{-1}$  i banchi:

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -4 & 0 & 1 \end{bmatrix}}_{M_1^{-1}} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -3 & 1 \end{bmatrix}}_{M_2^{-1}} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -4 & -3 & 1 \end{bmatrix}$$

Quanto visto nei punti (3) e (4) non accade per es., ma lo chiariremo più avanti.

In conclusione abbiamo ottenuto:

$$\underbrace{\begin{bmatrix} -1 & 2 & 1 \\ 1 & -1 & 1 \\ 4 & -11 & -13 \end{bmatrix}}_A = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -4 & -3 & 1 \end{bmatrix}}_{M_1^{-1} M_2^{-1} =: L} \underbrace{\begin{bmatrix} -1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & -3 \end{bmatrix}}_U$$

$A = L \cup$  è detta "fattorizzazione LU" di A.

Notiamo che

- $L$  è triang. inf. con elem. d'eq. uguali a 1 e riflette le storie di tutte le oper. elementari fatte sulle righe di A per ottenere  $\cup$
- $\cup$  è una forma a gradini di A

DEFINIZIONE  $L \in \mathbb{R}^{m \times n}$  è detta triangolare inferiore speciale se :

$$(L)_{ij} = 0 \quad \text{per } i < j$$

$$(L)_{ij} = 1 \quad \text{per } i = j$$

Visivamente :  $L = \begin{bmatrix} 1 & & & \\ * & 1 & & \\ 1 & / & 1 & \\ * & & & 1 \end{bmatrix}$

Quando una fattorizzazione LU apparirà

Così segue:

$$\underbrace{\begin{bmatrix} x & x & \dots & x \\ x & x & \dots & x \\ \vdots & \vdots & & \vdots \\ x & x & \dots & x \end{bmatrix}}_A = \underbrace{\begin{bmatrix} 1 & & & \\ x & 1 & & \\ \vdots & & \ddots & \\ x & & \dots & 1 \end{bmatrix}}_L \underbrace{\begin{bmatrix} x & x & \dots & x \\ x & x & \dots & x \\ \vdots & \vdots & & \vdots \\ x & x & \dots & x \end{bmatrix}}_U$$

A

L

U

triang. inf.  
speciale  
"lower"

↑  
triang.  
sup.  
"upper"

## TEOREMA (di chiusura)

(1) il prodotto fra metrici triang. sup. (risp. inf.)

è triang. sup. (risp. inf.);

(2) l'inverso di una metrica triang. sup. (risp. inf.)

invertibile è triang. sup. (risp. inf.);

(3) ai punti (1) e (2) restano dei aggiungimenti  
quontonamente la parola "speciale".

RICHIAMO Se  $A \in \mathbb{R}^{n \times n}$ . Si dice Minore  
principale d'ordine  $K = 1, 2, \dots, n$  il determinante

delle sottomatrici principali di Teste di ordine  $K$  di  $A$ .

## ESEMPIO

$$A = \begin{bmatrix} -1 & -3 & 3 \\ -3 & 11 & -7 \\ -1 & 5 & -3 \end{bmatrix}$$

Minore d'ordine 1: -1

Minore d'ordine 2:

Minore d'ordine 3:  $\det(A) =$

$$= 33 - 21 - 45 - (-33 - 27 + 35) =$$

$$= -33 - (-6) = -27$$

Le matrici  $M_1$  e  $M_2$  dell'esempio precedente sono un caso speciale di "matrici elementari di Gauss".

DEFINIZIONE  $M_k \in \mathbb{R}^{n \times n}$  è detta elementare di Gauss se :

(1) è triangolare inferiore speciale

(2) differisce dall'identità lungo al più una colonna :

$$M_k = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & x & 1 & \\ & - & x & 1 \end{bmatrix} \quad \text{colonna } k$$

Formalmente :  $M_k$  triang. inf. speciale  $\Leftrightarrow \exists k \in \mathbb{N}$  t.c.

$$(M_k)_{ij} = 0 \quad \text{se } j \neq k \text{ e } i > j$$

TEOREMA ① Se  $M_k$  elem. di Gauss.

$$M_k = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & M_{k+1,k} & \\ & & M_{n,k} & 1 \end{bmatrix} .$$

Allora

$$M_K^{-1} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & -m_{K+1,K} & 1 & \\ & -m_{n,K} & & 1 \end{bmatrix}.$$

(2)

Siano  $M_K, M_h$  elem. d' Gauss, con  $K < h$ :

$$M_K = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & m_{K+1,K} & 1 & \\ & m_{n,K} & & 1 \end{bmatrix}, \quad M_h = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & m_{h+1,h} & 1 & \\ & m_{n,h} & & 1 \end{bmatrix}.$$

Allora

$$M_K M_h = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & m_{K+1,K} & 1 & \\ & m_{n,K} & m_{h+1,h} & 1 \\ & & m_{n,h} & & 1 \end{bmatrix}$$

prodotto "per  
sovrapposizione"  
delle partizioni  
sotto la diagonale

Le proprietà (2) restano vere anche per il prodotto  
di più di due matrici elem. d' Gauss, nel  
caso

$M_{K_1} M_{K_2} \dots M_{K_p}$  con  $K_1 < K_2 < \dots < K_p$ .

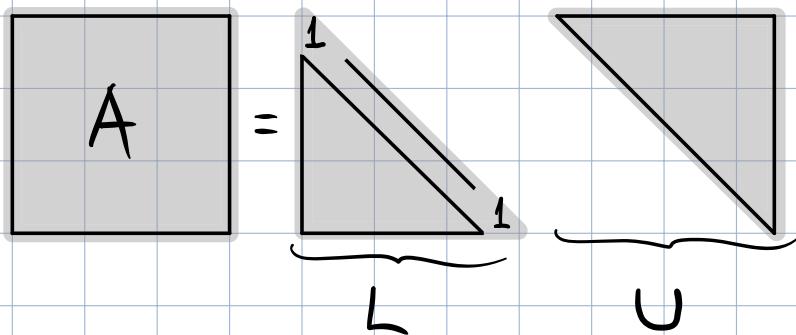
(Bisogna rispettare l'ordine delle colonne!)

## TEOREMA (Esistenza delle fattorizzazioni LU)

Sia  $A \in \mathbb{R}^{M \times N}$  avere i minori principali d'ordine  $1, 2, \dots, M-1$  non nulli. Allora esistono  $L, U \in \mathbb{R}^{M \times M}$ ,  $L$  triang. inf. speciale e  $U$  triang. sup., tali che

$$A = LU.$$

Viz: vementi



Ricordiamo:

## TEOREMA (Esistenza della fattorizzazione LU)

Se  $A \in \mathbb{R}^{M \times M}$  avendo i minori principali d'ordine  $1, 2, \dots, M-1$  non nulli. Allora esistono  $L, U \in \mathbb{R}^{M \times M}$ , con  $L$  triang. inf. simile a  $U$  triang. sup., tali che

$$A = L U.$$

### Dimostrazione / Algoritmo

Poniamo  $A^{(0)} := A$ .

Passo ①: cerca  $M_1 \in \mathbb{R}^{M \times M}$  elementare t.c.

$$\begin{bmatrix} 1 & & & \\ -m_{2,1} & 1 & & \\ & | & & \\ & -m_{n,1} & & \end{bmatrix} \xrightarrow{\text{scopriamo } m_{2,1}} \begin{bmatrix} a_{11}^{(0)} & a_{12}^{(0)} & \cdots & a_{1M}^{(0)} \\ a_{21}^{(0)} & a_{22}^{(0)} & \cdots & a_{2M}^{(0)} \\ \vdots & \vdots & & \vdots \\ a_{M1}^{(0)} & a_{M2}^{(0)} & \cdots & a_{MM}^{(0)} \end{bmatrix} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1M}^{(1)} \\ 0 & a_{22}^{(1)} & \cdots & a_{2M}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{M2}^{(1)} & \cdots & a_{MM}^{(1)} \end{bmatrix} \quad A^{(1)} := A$$

da scegliere in modo da annullare questi elementi

Dunque:

$$Q_{1j}^{(1)} = Q_{1j}^{(0)}, \quad \forall j = 1, 2, \dots, n \quad e$$

$$Q_{ij}^{(1)} = Q_{ij}^{(0)} - M_{i1} Q_{1j}^{(0)}, \quad \forall i = 2, 3, \dots, m \\ \forall j = 1, 2, \dots, n$$

Dobbiamo scegliere  $M_{i1}$  t.c., per  $i \geq 2$ :

$$Q_{i1}^{(1)} = 0 \iff Q_{i1}^{(0)} - M_{i1} Q_{11}^{(0)} = 0 \iff$$

$$\iff M_{i1} = \frac{Q_{i1}^{(0)}}{Q_{11}^{(0)}}, \quad i = 2, 3, \dots, m.$$

minore prima d'ordine 1,  
non nullo per ipotesi

In generale, d per sì  $K = 1, 2, \dots, m-1$

Si costruisce  $M_K$  elementare t.c.

$$M_K A^{(K-1)} = A^{(K)}, \quad \text{dove}$$

$$Q_{ik}^{(K)} = 0 \quad \text{per } i = K+1, \dots, m$$

$\uparrow$   
 $K$ -esima colonna  
sotto la diagonale

N.B. gli elementi sotto la diagonale  
lungo le colonne precedenti sono  
già stati annullati

Facendo i cololi, si ottiene:

- $Q_{ij}^{(k)} = Q_{ij}^{(k-1)}, \quad \forall i = 1, \dots, k \quad \text{e} \\ \forall j = 1, 2, \dots, m; \\$
  - $Q_{ij}^{(k)} = Q_{ij}^{(k-1)} - M_{ik}^{(k-1)} Q_{kj}, \quad i = k+1, \dots, n \\ j = 1, 2, \dots, m$ 

$\uparrow$   
 elementi  
 "non banali"  
 $d: M_k$

(In particolare:  $Q_{ij}^{(k)} = 0$  per  $i > j$  e  $j = 1, \dots, k-1$ )

Dalle condizioni  $Q_{ik}^{(K)} = 0$ ,  $i = k+1, \dots, m$

$$\text{otherwise } Q_{ik}^{(k-1)} - M_{ik} Q_{kk}^{(k-1)} = 0 \iff$$

$$\Leftrightarrow M_{ik} = \frac{a_{ik}}{Q_{kk}^{(K-1)}}$$

dove  $a_{kk}^{(k-1)}$  è diverso da

Aero per ipotesi (perché si muo')

dimostrare che  $Q_{kk}^{(k-1)} = 0$  se e

solo se il minore prima. d'ordine k  
di A è nullo)

C'è sempre un'altra colonna.  $M_k$  annulla gli elementi di  $A^{(k-1)}$  lungo la colonna  $k$  sotto la diagonale, senza alterare gli elementi già annullati da  $M_1, \dots, M_{k-1}$ . Dunque :

$$M_{m-1} \cdots M_2 M_1 A =: U \quad \text{triang. sup.}$$

Premolt. per  $M_{m-1}^{-1}, \dots, M_2^{-1}, M_1^{-1}$  si ha

$$A = \underbrace{M_1^{-1} M_2^{-1} \cdots M_{m-1}^{-1}}_{:= L} U, \quad \text{dove}$$

$$L = \begin{bmatrix} 1 & & & & \\ M_{21} & 1 & & & \\ M_{31} & M_{32} & 1 & & \\ | & | & & & \\ M_{m1} & M_{m2} & \cdots & M_{m,m-1} & 1 \end{bmatrix}$$

gli  $m_{ik}$  sono detti moltiplicatori:

$$M_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}},$$

↖  $k=1, \dots, n-1$  :  
pivot di  $U$



Abbiamo dimostrato l'esistenza delle fatt. LU  
(sotto opportune ipotesi). Domande: unicità?

## TEOREMA (unicità delle fatt. LU)

Se  $A \in \mathbb{R}^{n \times n}$  una matrice che ammette una fattorizzazione LU. Se  $\det(A) \neq 0$ , la fattorizzazione è unica.

Dimostrazione: esercizio.

[Idee: Supponne  $A = L_1 U_1 = L_2 U_2$ ; osservare che  $L_1, L_2, U_1, U_2$  sono invertibili; premoltiplicare  $L_1 U_1 = L_2 U_2$  per  $L_1^{-1}$  e postimoltiplicarla per  $U_2^{-1}$ ; dedurne che  $L_1^{-1} L_2 = U_1 U_2^{-1} = I$ , da cui segue la tesi.]

## PSEUDOCODE dell'Algorithmus

INPUT :  $A \in \mathbb{R}^{M \times M}$  di elem.  $a_{ij}$

per  $K = 1, 2, \dots, M-1$

verifico che  $a_{kk} \neq 0$

per  $i = k+1, \dots, M$

$$m_{ik} = \frac{a_{ik}}{a_{kk}}$$

per  $j = k+1, \dots, M$

$$a_{ij} \leftarrow a_{ij} - m_{ik} a_{kj}$$

fine

fine

fine

OUTPUT  $L, U \in \mathbb{R}^{M \times M}$ , dove :

$$U_{ij} = \begin{cases} a_{ij} & \text{per } i \leq j \\ 0 & \text{altrimenti} \end{cases}$$

$$L_{ij} = \begin{cases} L & \text{per } i=j \\ m_{ij} & \text{per } i > j \\ 0 & \text{altrimenti} \end{cases}$$

## ESEMPIO Calcola la fattorizzazione

LU di

$$A = \begin{bmatrix} 1 & 0 & 3 & -3 \\ -1 & 1 & 0 & 6 \\ 3 & -2 & 0 & -15 \\ -2 & -3 & -21 & -4 \end{bmatrix}$$

Passo per passo, calcolo i moltiplicatori da  
infilare in L effettuo le operazioni elementari.

Passo 1

$$\begin{array}{c} \left[ \begin{array}{cccc} 1 & 0 & 3 & -3 \\ -1 & 1 & 0 & 6 \\ 3 & -2 & 0 & -15 \\ -2 & -3 & -21 & -4 \end{array} \right] \xrightarrow{\quad} \left[ \begin{array}{cccc} 1 & 0 & 3 & -3 \\ 0 & 1 & 3 & 3 \\ 0 & -2 & -9 & -6 \\ 0 & -3 & -15 & -10 \end{array} \right] \\ \text{A}^{(0)} \qquad \qquad \qquad \text{A}^{(1)} \end{array}$$

da annullare

$$L = \begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ 3 & -2 & 1 & & \\ -2 & -3 & 2 & 1 & \end{bmatrix}$$

Passo 2

passo 1      passo 2      passo 3

passo 1:

$$\begin{cases} m_{21} = \frac{a_{21}^{(0)}}{a_{11}^{(0)}} = -1 \\ m_{31} = \frac{a_{31}^{(0)}}{a_{11}^{(0)}} = 3 \\ m_{41} = \frac{a_{41}^{(0)}}{a_{11}^{(0)}} = -2 \end{cases}$$

$$\begin{bmatrix} 1 & 0 & 3 & -3 \\ 0 & 1 & 3 & 3 \\ 0 & -2 & -9 & -6 \\ 0 & -3 & -15 & -10 \end{bmatrix}$$

$A^{(1)}$



$$\begin{bmatrix} 1 & 0 & 3 & -3 \\ 0 & 1 & 3 & 3 \\ 0 & 0 & -3 & 0 \\ 0 & 0 & -6 & -1 \end{bmatrix}$$

$A^{(2)}$

Passo 3

$$\begin{bmatrix} 1 & 0 & 3 & -3 \\ 0 & 1 & 3 & 3 \\ 0 & 0 & -3 & 0 \\ 0 & 0 & -6 & -1 \end{bmatrix}$$

$A^{(2)}$



$$\begin{bmatrix} 1 & 0 & 3 & -3 \\ 0 & 1 & 3 & 3 \\ 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

$A^{(3)} =: U$

Calcoliamo il determinante di A :

$$\det(A) = \det(LU) = \det(L)\det(U) =$$

$\uparrow$        $\uparrow$   
 $A = LU$       teorema di Binet :

$$= 1 \cdot \prod_{k=1}^4 U_{kk} = 3$$

$\uparrow$   
 $\det(L) = 1$

RISOLVIMENTO DI UN SISTEMA LIN.  $Ax = b$

MEMANTE  $A = LU$  :

$$Ax = b \Leftrightarrow LUx = b \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} Ly = b & \leftarrow \text{Sostituzione in} \\ & \text{avanti } y_1, y_2, \dots, y_m \\ Ux = y & \leftarrow \text{Sostituzione} \\ & \text{all'indietro } x_m, x_{m-1}, \dots, x_1 \end{cases}$$

# COSTO COMPUTAZIONALE DI $A = LU$

Ridhiamu :

## PSEUDOCODE dell' Algorithmus

INPUT :  $A \in \mathbb{R}^{m \times n}$  di elem.  $a_{ij}$

pen  $K = 1, 2, \dots, M-1$

Verifiziere der  $Q_{KK} \neq 0$

per  $i = k+1, \dots, M$

$$m_{ik} = \frac{a_{ik}}{a_{kk}} \leftarrow \text{Multiplikation}$$

$$\mu_n \quad j = k+1, \dots, m$$

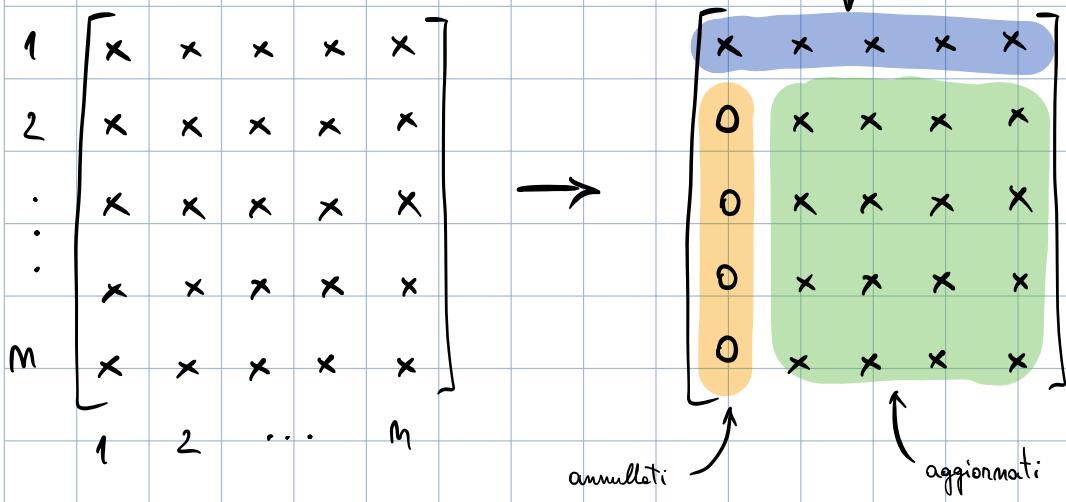
$Q_{ij} \leftarrow Q_{ij} - M_{ik} Q_{kj} \leftarrow$  assignment.

fine

fine

fine

## Exemp's. Passo 1 :

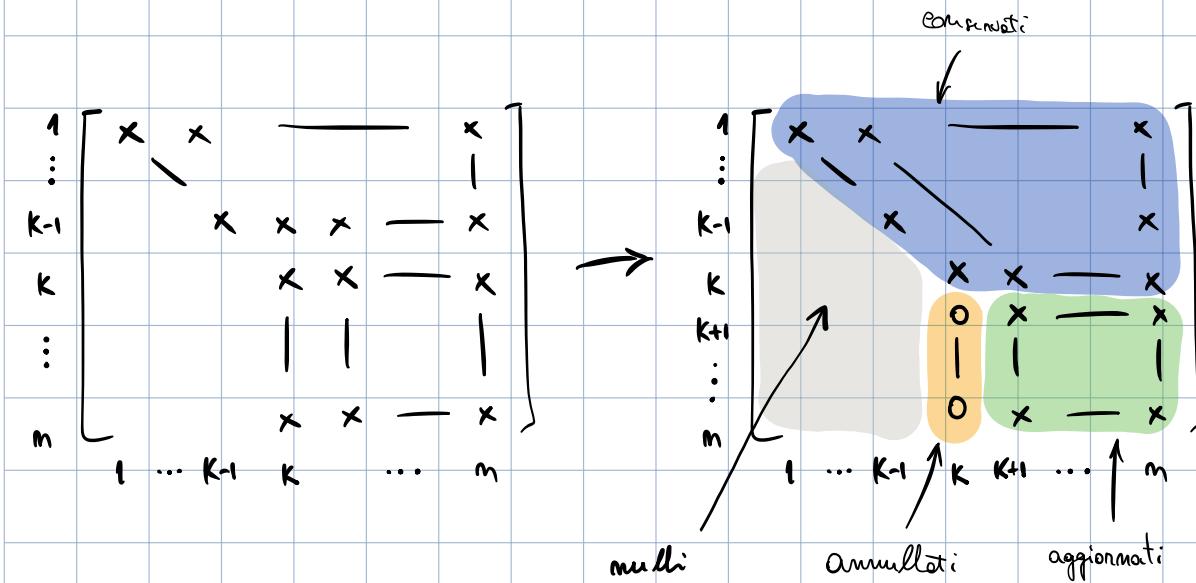


Moltiplicatori :  $m-1$  moltiplicazioni

aggiornamento :  $2(m-1)^2$  operazioni algebriche

(metà moltiplicazioni,  
metà addizioni)

Passo  $K$  :



Moltiplicatori :  $M-K$  moltiplicazioni

aggiornamento :  $2(m-k)^2$  operazioni algebriche

In Totale :

$$(M-1) + (M-2) + \dots + 2 + 1 \leftarrow \text{moltiplicatori}$$

$$2(M-1)^2 + 2(M-2)^2 + \dots + 2 \cdot 4 + 2 \leftarrow \text{aggiornamento}$$

allora :

$$\sum_{k=1}^{m-1} k + 2 \sum_{k=1}^{m-1} k^2$$

## DOMANDA

$$\sum_{k=1}^m k = 1 + 2 + \dots + m = ?$$

$$\sum_{k=1}^m k^2 = 1 + 4 + \dots + m^2 = ?$$

Sapete che

$$\sum_{k=1}^m k = \frac{m(m+1)}{2} \underset{m \text{ grande}}{\sim} \frac{m^2}{2}$$

Lo avete dimostrato per induzione nel Corso di Mat. Discreta.

Come si ottiene le formule?

Definiamo  $T_m := 1 + 2 + \dots + m = \sum_{k=1}^m k$

$$T_m = 1 + 2 + \dots + m-1 + m$$

$$T_m = m + m-1 + \dots + 2 + 1$$

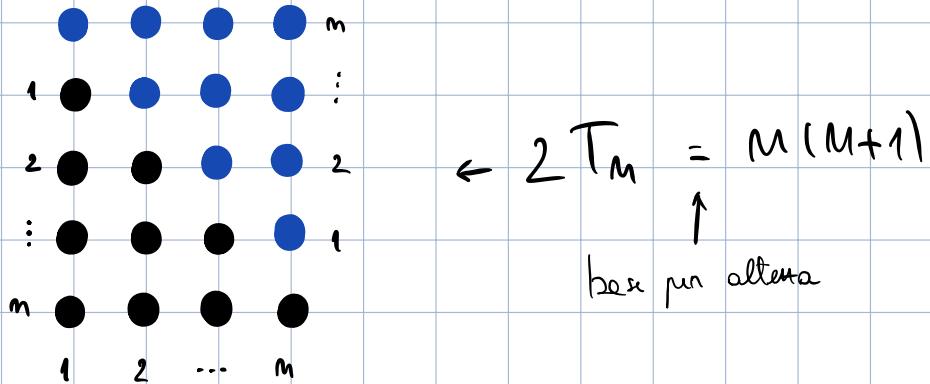
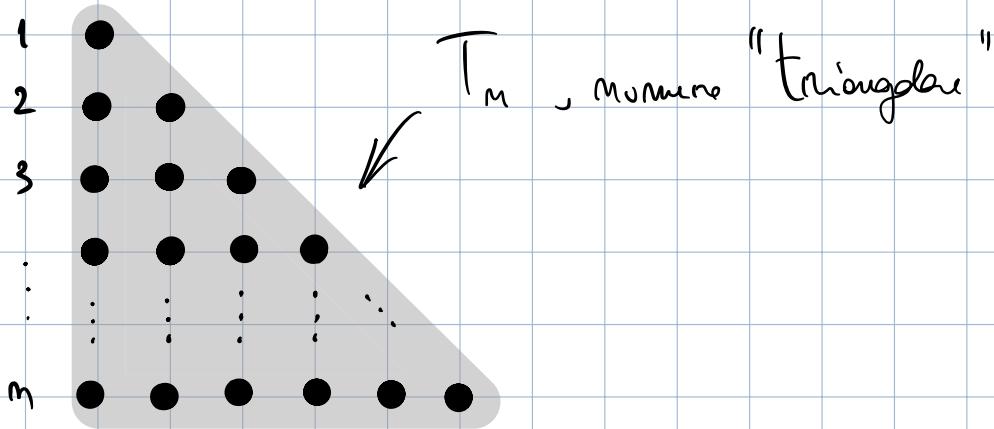
Sommiamo per colonne:

$$2T_m = (M+1) + (M+1) + \dots + (M+1) + (M+1)$$

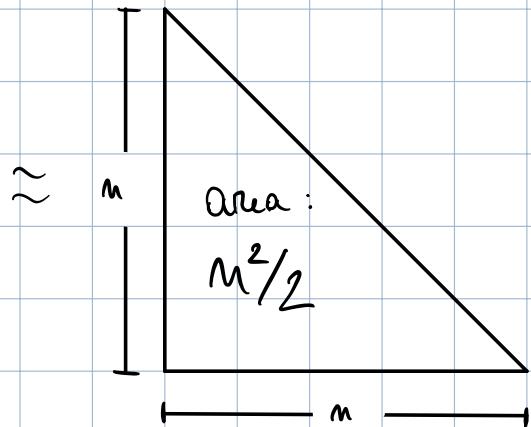
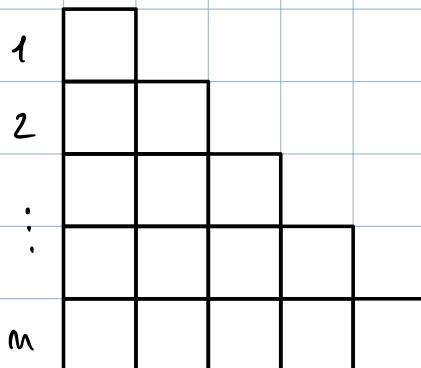
$\underbrace{\hspace{10em}}$   
 $M$  addendi

da cui  $T_m = \frac{M(M+1)}{2}$

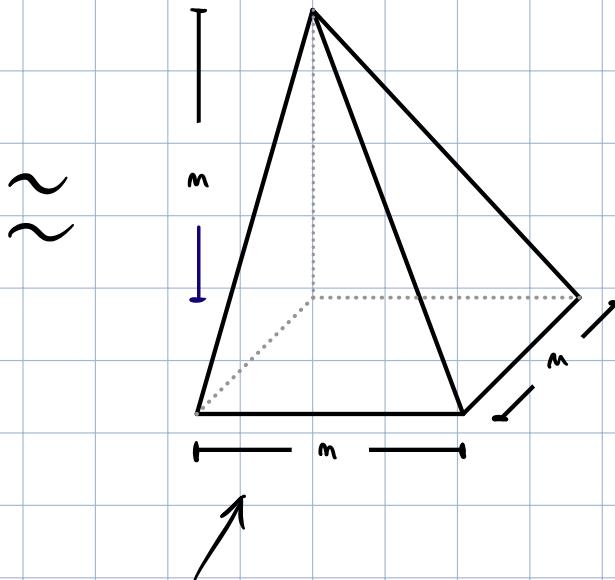
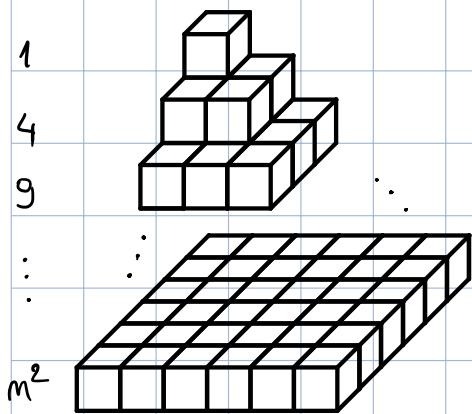
Geometricamente:



Stima geometrica di  $1+2+\dots+m$ :



Analogamente:



$$\text{Volume: } \frac{1}{3} \cdot \text{area di base} \cdot \text{altezza} = \\ = \frac{m^3}{3}$$

Ne deduciamo la

seguente stima:

$$\sum_{k=1}^m k^2 \approx \frac{m^3}{3}$$

S'ha dimostrare che

$$\sum_{k=1}^m k^2 = \frac{m(m+1)(2m+1)}{6}$$

Dimostrazione: esercizio (aggiornamenti in fondo alle note)

Tornando alla fattorizzazione LU:

Costo computazionale:

$$\text{circa } M^2/2 + 2M^3/3 \approx \frac{2}{3} M^3$$

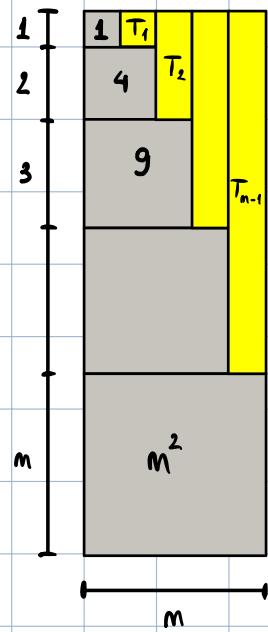
m grande

da confrontare con Laplace ( $m!$ )

La risoluzione dei sistemi  $Ly = b$  e  $Ux = y$  ha un costo di circa  $M^2$  per sistema.

[Dimostrazione: esercizio]

Esercizio:



(a) Dedunne dal disegno a sinistra dei posti:  $T_m := 1 + 2 + \dots + m = \sum_{k=1}^m k$  e  $S_m := 1 + 4 + \dots + m^2 = \sum_{k=1}^m k^2$ , si ha

$$m T_m = S_m + \sum_{k=1}^{m-1} T_k \quad [1]$$

(b) Utilizzare la [1],  $T_k = \frac{k(k+1)}{2}$ .

$S_{m-1} = S_m - m^2$  per ottenere

la formula:

$$S_m = \frac{m(m+1)(2m+1)}{6}.$$

Formulazione tutti i passaggi e

completare la dim. per induzione.

Esercizio: dedurre da disegno sotto che

$$1^3 + 2^3 + 3^3 + \dots + n^3 = \left( \frac{n(n+1)}{2} \right)^2.$$

	9	9
2	4	9
1	2	

Formulazione e dimostrazione per induzione.

## ESEMPIO

Vogliamo risolvere

$$\begin{cases} X_1 + X_2 + 3X_3 = -1 \\ 2X_1 + 2X_2 + 20X_3 = -16 \\ 3X_1 + 6X_2 + 4X_3 = 5 \end{cases}$$

Eliminazione di Gauss : Eq2 - 2Eq1 , Eq3 - 3Eq1

$$\begin{cases} X_1 + X_2 + 3X_3 = -1 \\ 14X_3 = -14 \\ 3X_2 - 5X_3 = 8 \end{cases}$$

Adesso il metodo prevede di eliminare da Eq3 l'incognita  $X_2$  aggiungendo a Eq3 un multiplo di Eq2 :

IMPOSSIBILE !

Se avessimo effettuato l'algo. LU su A :

$$\underbrace{\begin{bmatrix} 1 & 1 & 3 \\ 2 & 2 & 20 \\ 3 & 6 & 4 \end{bmatrix}}_{A^{(0)}}$$

$$R_2 - 2R_1$$

$$R_3 - 3R_1$$

pivot nullo !

$$\underbrace{\begin{bmatrix} 1 & 1 & 3 \\ 0 & 0 & 14 \\ 0 & 3 & -5 \end{bmatrix}}_{A^{(1)}}$$

Soluzione: scambio Eq2 con Eq 3 dopo  
il primo passo di eliminazione:

$$\left\{ \begin{array}{l} x_1 + x_2 + 3x_3 = -1 \\ 14x_3 = -14 \\ 3x_2 - 5x_3 = 8 \end{array} \right. \rightarrow \left\{ \begin{array}{l} x_1 + x_2 + 3x_3 = -1 \\ 3x_2 - 5x_3 = 8 \\ 14x_3 = -14 \end{array} \right.$$

Adesso posso riconoscere (un'eliminazione all'inverso):

$$x_3 = -1 \Rightarrow x_2 = \frac{1}{3}(-5+8) = 1 \Rightarrow \\ \Rightarrow x_1 = -1 + 3 - 1 = 1 .$$

Analogamente, otterremo risolvendo il problema del pivot nella mediana uno scambio fra righe delle matrice dei coefficienti. Formuliammo:

### DEFINIZIONE

$P \in \mathbb{R}^{n \times n}$  è detta "di permutazione" se la si può ottenere dall'identità mediante scambi di righe. Diciamo anche che  $P$  è "una permutazione".

## ESEMPIO

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{è di permutazione.}$$

Sia  $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$ .

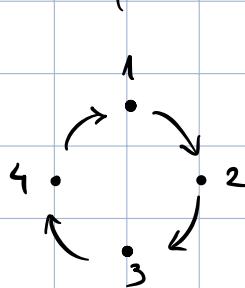
$$PA = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = \begin{bmatrix} 7 & 8 & 9 \\ 4 & 5 & 6 \\ 1 & 2 & 3 \end{bmatrix} \quad \text{da sinistra  
scambia le righe}$$

$$AP = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 3 & 2 & 1 \\ 6 & 5 & 4 \\ 9 & 8 & 7 \end{bmatrix} \quad \text{da destra  
scambia le colonne}$$

## ESEMPIO

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \text{è una permutazione "eclissi":}$$

$$P \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 4 \\ 1 \end{bmatrix} :$$



OSSERVATIONI Se  $P \in \mathbb{R}^{n \times n}$  è permutazione.

(1)  $\det(P) = \pm 1$

(2)  $P$  è invertibile

(3)  $P^T P = P P^T = I \Rightarrow P^{-1} = P^T$

A noi interessano delle particolari matrici d'permutazione:

DEFINIZIONE  $E \in \mathbb{R}^{n \times n}$  è detta "di scambio"

se è di permutazione e si ottiene dall'identità scambiando al più due righe.

NOTA:  $P$  del primo esempio è di scambio,

$P$  del secondo NO!

Imponendo sul procedimento una matrice di scambio, rimoveremmo l'elemento  
nella stessa posizione pivot di:

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}}_E \underbrace{\begin{bmatrix} 1 & 1 & 3 \\ 0 & 0 & 14 \\ 0 & 3 & -5 \end{bmatrix}}_{A^{(1)}} \rightarrow \underbrace{\begin{bmatrix} 1 & 1 & 3 \\ 0 & 3 & -5 \\ 0 & 0 & 14 \end{bmatrix}}_U$$

Nom sono solo i pivot nulli o  
causano problemi, ma anche quelli piccoli!  
Si pensi all'esperimento Matlab.

pivot piccoli  $\rightarrow$  moltiplicatori grandi  $\rightarrow$  perdita di precisione

Soluzione:

**STRATEGIA DEL PIVOTING PER RIGHE  
(O PARZIALE)**

Se  $A \in \mathbb{R}^{n \times n}$  invertibile. Poniamo  $A^{(0)} := A$ .  
per  $k = 1, 2, \dots, n-1$

- scambia le righe  $k$  di  $A^{(k)}$  con le righe  $i \geq k$  di  $A^{(k-1)}$  in modo tale che, dopo lo scambio, si ottiene:

$$|a_{kk}^{(k-1)}| \geq |a_{ik}^{(k-1)}|, i = k, \dots, m$$

- effettua il passo  $k$  dell'eliminazione di Gauß, in modo da ottenere  $A^{(k)}$  t. c.

$$a_{ik}^{(k)} = 0 \text{ per } i = k+1, \dots, m$$

fine

In sostanza, ogni passo d'eliminazione viene preceduto da uno scambio tra righe che ha l'obiettivo di portare in posizione pivot il più grande elemento possibile (in valore assoluto).

### ESEMPIO

$$A^{(0)} = \left( \begin{array}{cccc} -1 & -4 & 2 & 1 \\ 2 & 0 & -4 & 1 \\ -3 & -3 & -3 & -1 \\ 2 & -3 & 1 & -2 \end{array} \right)$$

Scambio

### ESEMPIO

Applichiamo l'algoritmo su

$$A = \begin{bmatrix} -1 & -4 & -2 \\ 3 & 2 & -2 \\ -4 & 1 & 0 \end{bmatrix}$$

Primo passo di scambio:

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \underbrace{\begin{bmatrix} -1 & -4 & -2 \\ 3 & 2 & -2 \\ -4 & 1 & 0 \end{bmatrix}}_{E_1} = \underbrace{\begin{bmatrix} -4 & 1 & 0 \\ 3 & 2 & -2 \\ -1 & -4 & -2 \end{bmatrix}}_{E_1 A^{(0)}}$$

Primo passo di eliminazione:

$$\begin{bmatrix} 1 & 0 & 0 \\ \frac{3}{4} & 1 & 0 \\ -\frac{1}{4} & 0 & 1 \end{bmatrix} \underbrace{\begin{bmatrix} -4 & 1 & 0 \\ 3 & 2 & -2 \\ -1 & -4 & -2 \end{bmatrix}}_{M_1 E_1 A^{(0)}} = \underbrace{\begin{bmatrix} -4 & 1 & 0 \\ 0 & \frac{11}{4} & -2 \\ 0 & -\frac{17}{4} & -2 \end{bmatrix}}_{A^{(1)}}$$

Secondo passo di scambio:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \underbrace{\begin{bmatrix} -4 & 1 & 0 \\ 0 & \frac{11}{4} & -2 \\ 0 & -\frac{17}{4} & -2 \end{bmatrix}}_{E_2 A^{(1)}} = \underbrace{\begin{bmatrix} -4 & 1 & 0 \\ 0 & -\frac{17}{4} & -2 \\ 0 & \frac{11}{4} & -2 \end{bmatrix}}_{E_2 A^{(1)}}$$

Secondo passo di eliminazione:

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{11}{17} & 1 \end{bmatrix}}_{M_2} \underbrace{\begin{bmatrix} -4 & 1 & 0 \\ 0 & -\frac{17}{4} & -2 \\ 0 & \frac{11}{4} & -2 \end{bmatrix}}_{E_2 A^{(1)}} = \underbrace{\begin{bmatrix} -4 & 1 & 0 \\ 0 & -\frac{17}{4} & -2 \\ 0 & 0 & -\frac{56}{17} \end{bmatrix}}_{A^{(2)} =: U}$$

Algebricamente abbiamo:

$$M_2 E_2 M_1 E_1 A = U,$$

o equivalentemente

$$\underbrace{M_2 E_2 M_1}_{=: R} \underbrace{E_2^T E_1}_{=: P} A = U.$$

$P$  è di permutazione (banche)

$R$  è triang. inferiore speciale (mimo banche)

Posto  $L = R^{-1}$ , si ha

$$PA = LU$$

← fattorizzazione LU con pivoting per le banche di  $A$ .

A parole, la strategia del pilotino  
particolare consiste nel far precedere ad ogni  
passo di eliminazione un passo di scambio,  
con l'obiettivo di portare in posizione  
pivotale l'elemento più grande in valore  
assoluto tra i candidati ad occupare la posizione  
pivotale. I candidati sono l'elemento in  
posizione pivotale e gli elementi lungo la  
stessa colonna al di sotto della diagonale.

Più in generale :

Se  $A \in \mathbb{R}^{n \times n}$  invertibile.

per  $k = 1, 2, \dots, n-1$

- determiniamo  $E_k$  di scarto t.c.

$$(E_k A)_{kk} > (E_k A)_{ik}, \forall i \geq k+1$$

- $A \leftarrow E_k A$

- determiniamo  $M_k$  elementare t.c.

$$(M_k A)_{ik} = 0, \forall i \geq k+1$$

- $A \leftarrow M_k A$

fine

Nota L'esistenza di un pivot non nullo ad ogni passo è conseguente dell'invertibilità di  $A$ .

Algebricamente :

$$M_{n-1} E_{n-1} \cdots M_2 E_2 M_1 E_1 A = U$$

$\uparrow$  triang. sup.,

ovvero, equivalentemente :

$$M_{m-1} E_{m-1} \cdots M_2 E_2 M_1 E_2^T \cdots \underbrace{E_{m-1}^T E_{m-1} \cdots E_2 E_1}_I A = U$$

$=: R$        $=: P$

Si può dimostrare che  $R$  è triang. inf. spezzata e invertibile. Posto  $L = R^{-1}$ , si ha:

$$PA = LU$$

→ fatt. LU con pivoting per i valori di  $A$

Si può verificare che, per ottenere  $L$  in questo caso, è sufficiente scomporre ad ogni passo i moltiplicatori già creati in accordo con gli scambi d'righe effettuati su  $A$ .

### ESEMPIO

Se  $A = \begin{bmatrix} 1 & 2 & -2 & -4 \\ -2 & -4 & 0 & -1 \\ -4 & 0 & 1 & 3 \\ -2 & 1 & 2 & 4 \end{bmatrix}$ . Determiniamo la fattorizzazione

$$PA = LU.$$

Initializziamo un vettore per  $\mathbb{R}^4$  e la matrice  $L$ :

$$P = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \xrightarrow{E_1} \begin{bmatrix} 3 \\ 2 \\ 1 \\ 4 \end{bmatrix} \xrightarrow{E_3} \begin{bmatrix} 3 \\ 2 \\ 4 \\ 1 \end{bmatrix}$$

$$L = \begin{bmatrix} 1 & & & \\ \frac{1}{2} & 1 & & \\ -\frac{1}{4} & -\frac{1}{2} & 1 & \\ \frac{1}{2} & -\frac{1}{4} & & 1 \end{bmatrix} \xrightarrow{E_3} \begin{bmatrix} 1 & & & \\ \frac{1}{2} & 1 & & \\ \frac{1}{2} & -\frac{1}{4} & 1 & \\ -\frac{1}{4} & -\frac{1}{2} & -\frac{8}{11} & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & -1 & -4 \\ -2 & -4 & 0 & -1 \\ -4 & 0 & 1 & 3 \\ -2 & 1 & 2 & 4 \end{bmatrix} \xrightarrow{E_1} \begin{bmatrix} -4 & 0 & 1 & 3 \\ -2 & -4 & 0 & -1 \\ 1 & 2 & -1 & -4 \\ -2 & 1 & 2 & 4 \end{bmatrix} \xrightarrow{M_1} \begin{bmatrix} -4 & 0 & 1 & 3 \\ 0 & -4 & -\frac{1}{2} & -\frac{5}{2} \\ 0 & 2 & -\frac{3}{4} & -\frac{13}{4} \\ 0 & 1 & \frac{3}{2} & \frac{5}{2} \end{bmatrix} \xrightarrow{M_2}$$

$$\begin{bmatrix} -4 & 0 & 1 & 3 \\ 0 & -4 & -\frac{1}{2} & -\frac{5}{2} \\ 0 & 0 & -1 & -\frac{9}{2} \\ 0 & 0 & \frac{11}{8} & \frac{15}{8} \end{bmatrix} \xrightarrow{E_3} \begin{bmatrix} -4 & 0 & 1 & 3 \\ 0 & -4 & -\frac{1}{2} & -\frac{5}{2} \\ 0 & 0 & \frac{11}{8} & \frac{15}{8} \\ 0 & 0 & -1 & -\frac{9}{2} \end{bmatrix} \xrightarrow{M_3} \begin{bmatrix} -4 & 0 & 1 & 3 \\ 0 & -4 & -\frac{1}{2} & -\frac{5}{2} \\ 0 & 0 & \frac{11}{8} & \frac{15}{8} \\ 0 & 0 & 0 & -\frac{69}{22} \end{bmatrix}$$

Notare: al passo 2 non è stato necessario estrarre  
scambi ( $E_2 = I$ ).

Dunque si ha:

$$P = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}; L = \begin{bmatrix} 1 & & & \\ \frac{1}{2} & 1 & & \\ \frac{1}{2} - \frac{1}{4} & 1 & & \\ -\frac{1}{4} - \frac{1}{2} - \frac{8}{11} & 1 & & \end{bmatrix}; U = \begin{bmatrix} -4 & 0 & 1 & 3 & 7 \\ 0 & -4 & -\frac{1}{2} & -\frac{5}{2} & \\ 0 & 0 & \frac{11}{8} & \frac{15}{8} & \\ 0 & 0 & 0 & -\frac{69}{22} & \end{bmatrix}$$

Per esempio, verifichiamo che  $PA = LU$ .

RISOLUZIONE DEI SISTEMI LINEARI:

Risolviamo  $Ax = b$  data  $PA = LU$ .

$$Ax = b \Leftrightarrow PAx = Pb \Leftrightarrow LUx = Pb \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} Ly = Pb & \leftarrow \text{sost. in avanti} \\ Ux = y & \leftarrow \text{sost. eliminazione} \end{cases} := Y$$

CALCOLO DEL DETERMINANTE:

$$PA = LU \Rightarrow \det(P) \det(A) = \underbrace{\det(L) \det(U)}_{= 1}$$

Da cui  $\det(A) = \frac{\det(U)}{\det(P)}$ , dove

$$\det(P) = (-1)^{\text{# di scambi effettuati}}$$

ESEMPIO PRECEDENTE :

$$\det(P) = (-1)^2 = 1 \Rightarrow \det(A) = -20$$

COMPLESSITÀ COMPUTAZIONALE :

- calcolo della fattorizzazione  $PA = LU$   $\Theta(m^3)$
- risoluzione di  $Ly = Pb$  in avanti  $\Theta(m^2)$
- risoluzione di  $Ux = y$  all'indietro  $\Theta(m^2)$

Nota: il pivoting parziale aggiunge un sussesso speditio ( $\Theta(m^2)$  confronti in totale)

ad un algoritmo cubico; esso ha, dunque, un impatto trascurabile per  $m$  grande.

## OSSERVAZIONI

(1) l'algoritmo che consente di ottenere la fattorizzazione  $PA = LU$  si applica anche a matrici  $A$  rettangolari.

(2) a seguito del pivoting si ha che

$$M_{ik} = \frac{|a_{ik}^{(k-i)}|}{|a_{kk}^{(k-i)}|} \leq 1$$

Moltiplicatori

TEOREMA Se  $A \in \mathbb{R}^{m \times m}$ . Esistono

$P \in \mathbb{R}^{m \times m}$  di permutazione,  $L \in \mathbb{R}^{m \times m}$  triang.

inf. simile con  $|L_{ij}| \leq 1$  se  $i > j$ ,  
 $U \in \mathbb{R}^{m \times m}$  a quadri tali da

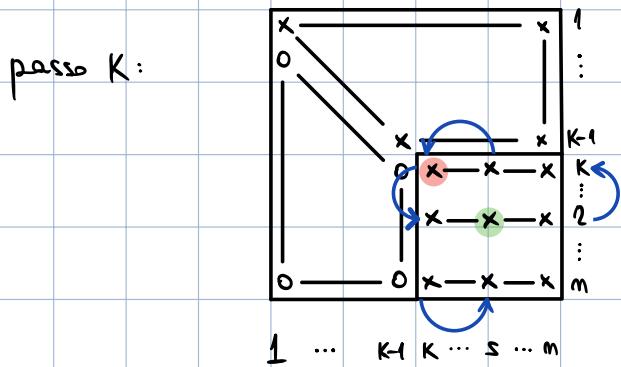
$$PA = LU$$

Visualmente:

$$P \quad A = \begin{array}{c} L \\ \diagdown \end{array} \quad U$$

Nota:  $A$  quadrata e  $\Rightarrow U$  quadrata, triang. sup. con elementi diagonali invertibili

Esiste anche una strategia di pivoting detta pivoting completo, in cui al passo  $K$  si porta in posizione pivotale l'elemento più grande in valore assoluto del blocco  $(m-K+1) \times (m-K+1)$  in basso e dietro della matrice:



L'elemento di posto 25 viene portato sulla diagonale mediante uno scambio di riga e colonna

Si ottengono pivot più grandi, e dunque moltiplicatori più piccoli, a vantaggio della stabilità numerica.

Perciò la complessità computazionale cresce in modo non più trascurabile ( $\Theta(m^3)$  confronti in totale).

In generale si può dire che :

"il pivoting parziale introduce un miglioramento significativo a fronte di un sovraccosto modesto;

il pivoting completo introduce un ulteriore modesto miglioramento a fronte di un sovraccosto significativo"

## IL METODO DEI MINIMI QUADRATI

Consideriamo un sistema lineare  $Ax = b$ ,

$A \in \mathbb{R}^{m \times n}$ ,  $b \in \mathbb{R}^m$ ,  $x \in \mathbb{R}^n$ . Sappiamo che

$Ax = b$  ammette soluzione  $\Leftrightarrow b \in \text{Im}(A)$

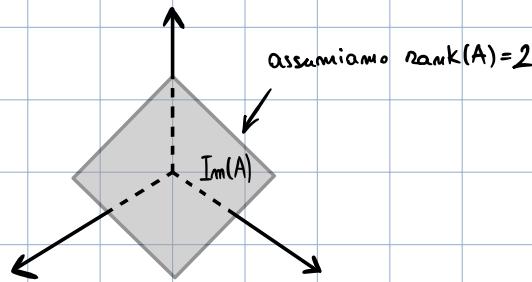
Cose fare se  $b \notin \text{Im}(A)$ ? E' un caso  
molto importante nelle applicazioni, e c'è ottendiamo  
se frequente se  $A$  è skinny ( $m > n$ ).

### ESEMPIO

$A \in \mathbb{R}^{3 \times 2}$ ; Sappiamo che

- $\text{rank}(A) = \dim(\text{Im}(A)) \leq 2$
- $\text{Im}(A)$  sottospazio vett. di  $\mathbb{R}^3$

Dunque



E' molto probabile che  $b \notin \text{Im}(A)$ .

Condurremo d' ora in su e "risolvere"

il sistema  $Ax = b$ " anche in questo caso.

Prime ci servono dei strumenti.

## NORME VETTORIALI

DEFINIZIONE Si dice "Norma Vettoriale" in applicazione

$\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$  che verifica le seguenti:

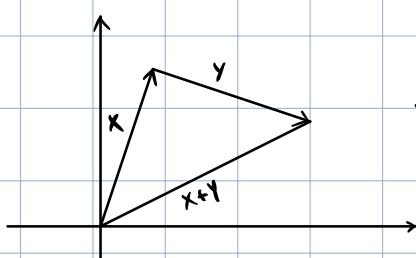
proprietà:

$$(1) \|x\| \geq 0, \forall x \in \mathbb{R}^n; \|x\| = 0 \Leftrightarrow x = 0 \quad \text{vettore nullo}$$

$$(2) \|\alpha x\| = |\alpha| \|x\|, \forall \alpha \in \mathbb{R} \text{ e } x \in \mathbb{R}^n$$

$$(3) \|x+y\| \leq \|x\| + \|y\|, \forall x, y \in \mathbb{R}^n \quad \begin{matrix} \leftarrow & \text{disug.} \\ & \text{triangolare} \end{matrix}$$

Idea:  $\|x\| =$  "giordate di  $x$ "; potrebbe esprimere le lunghezze.



diseg triangolare:

in un triangolo, le lunghezze di un lato non può superare la somma delle lunghezze dei altri due

ESEMPI Se  $x \in \mathbb{R}^n$ ;  $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ .

$$\|x\|_2 = \sqrt{\sum_{k=1}^n x_k^2} \quad \text{Norme euclidea (o Norme 2)}$$

$$\|x\|_1 = \sum_{k=1}^m |x_k| \quad \text{norma 1}$$

$$\|x\|_\infty = \max_{k=1,2,\dots,m} |x_k| \quad \text{norma infinito}$$

$$\|x\|_p = \left( \sum_{k=1}^m |x_k|^p \right)^{1/p} \quad \text{norma p}$$

TEOREMA :  $\|x\|_p \xrightarrow[p \rightarrow +\infty]{} \|x\|_\infty$

### OSSERVAZIONE

$$x \in \mathbb{R}^n, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} :$$

$$x^T x = [x_1 \ x_2 \ \dots \ x_n] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} = x_1^2 + x_2^2 + \dots + x_m^2 \Rightarrow$$

$$\Rightarrow x^T x = \|x\|_2^2$$

FATTO IMPORTANTE : ogni norma induce un concetto di distanza fra vettori di  $\mathbb{R}^n$  :

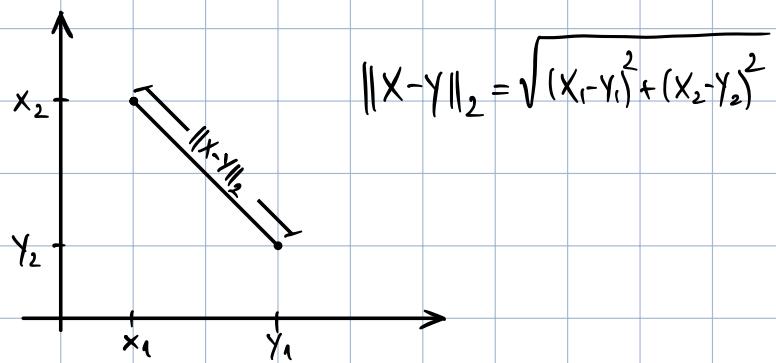
$$\text{dist}(x, y) = \|x - y\|$$

distanza di x da y

(giustificato dalla regola del parallelogramma)

Esempio: per  $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$  e  $y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$  in  $\mathbb{R}^2$ ,

$\|x - y\|_2$  è la distanza tra i punti di coordinate  $(x_1, x_2)$  e  $(y_1, y_2)$  ottenuta attraverso il teorema di Pitagora.



Minimi quadrati :

Idea :  $\left\{ \begin{array}{l} \text{se } b \notin \text{Im}(A), \text{ cerchiamo } \bar{x} \in \mathbb{R}^m \\ \text{t.c. } \|A\bar{x} - b\|_2 \text{ è il valore più vicino possibile!} \text{ ovvero, cerchiamo } \bar{x} \text{ t.c.} \\ A\bar{x} \text{ dista il meno possibile da } b \end{array} \right.$

DEFINIZIONE Sono  $A \in \mathbb{R}^{m \times n}$  e  $b \in \mathbb{R}^m$ .

Diciamo che  $\bar{x} \in \mathbb{R}^n$  è soluzione "nel senso dei minimi quadrati" di  $Ax = b$  se

$$\underbrace{\|A\bar{x} - b\|_2}_{\substack{\downarrow \\ \text{è detto "residuo" del sistema lineare}}} = \min_{x \in \mathbb{R}^n} \|Ax - b\|_2.$$

Donque  $\bar{x}$  minimizza la norma euclidea del residuo! Indicheremo la soluzione nel senso dei minimi quadrati con LSS (Least Squares Solution)

N.B. Residuo  $r := Ax - b$  null  $\Leftrightarrow \bar{x}$  soluzione nel senso classico

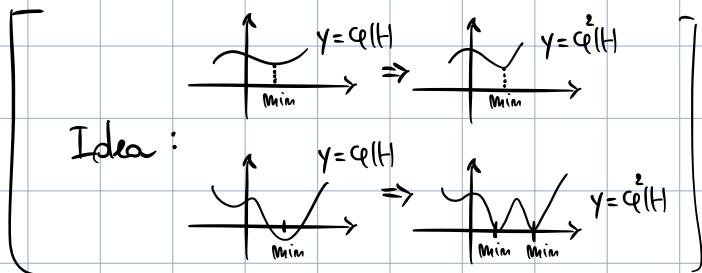
## OSSERVAZIONE

$$\|\bar{A}\bar{x} - b\|_2 = \min_{x \in \mathbb{R}^m} \|Ax - b\|_2$$

è equivalente a

per scappare le  
radice quadrata

$$\|\bar{A}\bar{x} - b\|_2^2 = \min_{x \in \mathbb{R}^m} \|Ax - b\|_2^2$$



Pen trovare  $\bar{x}$  si usa un principio detto  
"Variazionale":

$$\bar{x} \in \text{LSS} \stackrel{\text{def}}{\iff} \|\bar{A}\bar{x} - b\|_2^2 = \min_{x \in \mathbb{R}^m} \|Ax - b\|_2^2 \iff$$

$$\forall v \in \mathbb{R}^m: \|\bar{A}\bar{x} - b\|_2^2 \leq \|\bar{A}(\bar{x} + tv) - b\|_2^2 \quad \forall t \in \mathbb{R}$$

Fissato  $v \in \mathbb{R}^m$  arbitrario, se  $\bar{x}$  è una LSS è definito

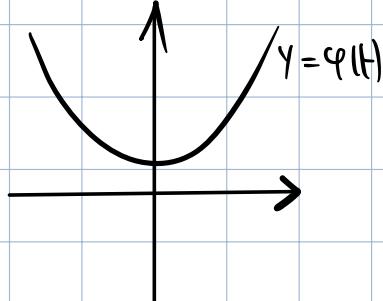
$$q(t) := \|\bar{A}(\bar{x} + tv) - b\|_2^2 =$$

$$= [\bar{A}(\bar{x} + tv) - b]^T [\bar{A}(\bar{x} + tv) - b] \in \mathbb{R}$$

## OSSERVAZIONI

- (1)  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  ;
- (2)  $\varphi(t) = \alpha t^2 + \beta t + \gamma$ , per qualche  $\alpha, \beta, \gamma$  ;
- (3)  $\varphi(t) \geq 0, \forall t$  ;
- (4)  $\varphi$  ha un minimo assoluto per  $t=0$ .

Dunque



Il grafico di  $\varphi$  è una parabola  
con concavità rivolta verso l'alto  
e simmetrica rispetto all'asse delle  
ordinate.

Allora si deve avere  $\beta = 0$ !

Notiamo che ciò equivale a  $\beta = \varphi'(0) = 0$

Imponendo (fare i conti per esercizio) che

$$\varphi(t) = [A(\bar{x} + tv) - b]^T [A(\bar{x} + tv) - b]$$

ottiene termine di primo grado in  $t$  nulla

(essendo  $\beta = 0$ ), si ottiene

$$v^T A^T (A\bar{x} - b) = 0.$$

Essendo  $\bar{x} \in \mathbb{R}^m$  arbitrario, si deve avere

$$A^T(A\bar{x} - b) = 0 \iff$$

$$A^T A \bar{x} = A^T b$$

detto:  
← "sistema delle  
equazioni normali"

TEOREMA  $\bar{x}$  è LSS di  $Ax = b$

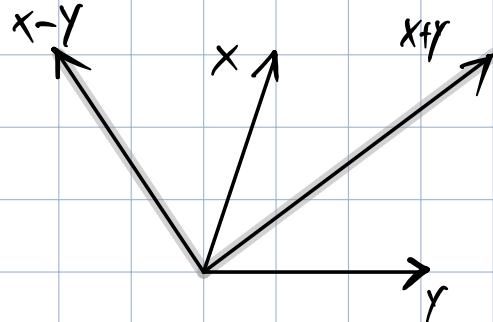
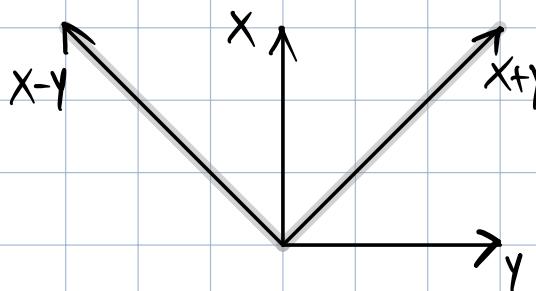
se e solo se  $\bar{x}$  risolve il sistema delle  
equazioni normali

$$A^T A \bar{x} = A^T b$$

(abbiamo dimostrato il "solo se")

## VETTORI ORTOGONALI

$$x, y \in \mathbb{R}^m$$



S'vede che  $\|x+y\|_2 = \|x-y\|_2$  se e solo se  $x$  e  $y$   
sono perpendicolari. Ciò giustifica la definizione

$$\text{Def. } \underset{\substack{\uparrow \\ \text{perpendicolari}}}{X \perp Y} \iff X^T Y = 0$$

$$\text{Infatti } \|X+Y\|_2 = \|X-Y\|_2 \iff \|X+Y\|_2^2 = \|X-Y\|_2^2 \iff \\ (X+Y)^T(X+Y) = (X-Y)^T(X-Y) \iff X^T Y = 0.$$

Torniamo alle eq. m. non nbl.

$\bar{x}$  LSS di  $Ax=b$ ,  $A^T A$  invertibile

$$\text{Allora } \bar{x} = (A^T A)^{-1} A^T b \Rightarrow A \bar{x} = \underbrace{A(A^T A)^{-1} A^T b}_{:= \bar{b}}$$

$$\text{Nota: si può dimostrare che } \text{rank}(A) = \text{rank}(A^T) =$$

$$= \text{rank}(A^T A) = \text{rank}(A A^T), \text{ da cui segue che}$$

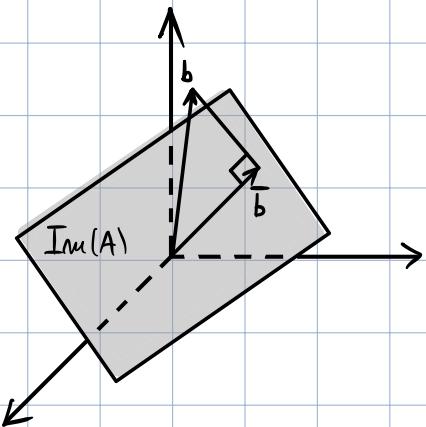
$$\text{Im}(A^T A) = \text{Im}(A^T) \Rightarrow A^T b \in \text{Im}(A^T A), \forall b.$$

TEOREMA:  $A^T A x = A^T b$  ha esclusiva soluzione!

Sia ora  $b \notin \text{Im}(A)$ . Allora  $Ax = b$  non ha soluzione, mentre  $Ax = \bar{b}$  ha soluzione. Inoltre

$$\tilde{b} \perp b - \tilde{b}$$

Graficamente :



Risoluzione nel senso dei minimi quadrati di  $Ax = b$ :

L'soluzione  $\bar{x}$  di  $A\bar{x} = \bar{b}$ , dove  $\bar{b}$  è la proiezione ortogonale di  $b$  su  $Im(A)$ .

Esercizio : Risolvene del senso dei minimi quadrati:

$$\begin{cases} -x_1 - x_2 = 1 \\ -2x_1 + 2x_2 = 1 \\ -x_1 = 0 \\ -2x_2 = -1 \end{cases}$$

N.B. non c'è soluzione classica!

$$A = \begin{bmatrix} -1 & -1 \\ -2 & 2 \\ -1 & 0 \\ 0 & -2 \end{bmatrix} ; b = \begin{bmatrix} 1 \\ 1 \\ 0 \\ -1 \end{bmatrix} .$$

$$A^T A = \begin{bmatrix} 6 & -3 \\ -3 & 9 \end{bmatrix} ; A^T b = \begin{bmatrix} -3 \\ 3 \end{bmatrix} .$$

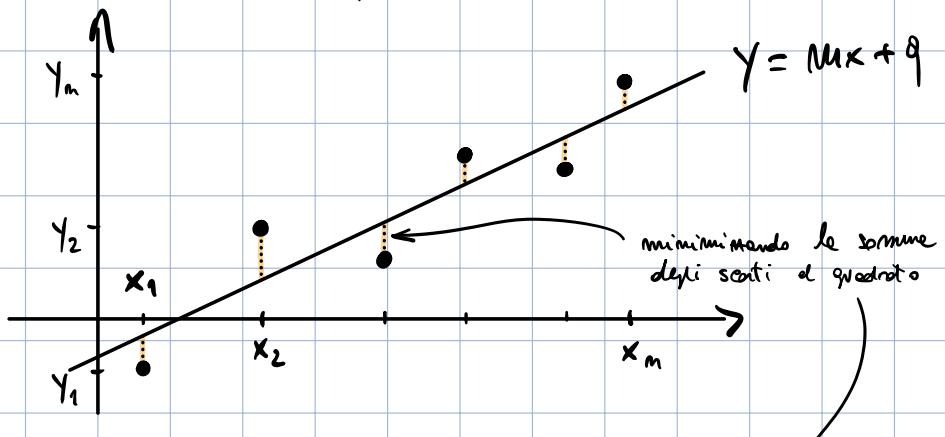
Eq. mi normali:

$$\begin{cases} 6x_1 - 3x_2 = -3 \\ -3x_1 + 9x_2 = 3 \end{cases} \iff \begin{cases} 3x_1 - x_2 = -1 \\ -x_1 + 3x_2 = 1 \end{cases} \iff \dots$$

$$\bar{x} = \begin{bmatrix} -2/5 \\ 1/5 \end{bmatrix} ; \text{ norma del residuo: } \|Ax-b\| = \frac{\sqrt{30}}{5}$$

## APPICAZIONE: REGRESSIONE LINEARE

Idea:  $\{(x_i, y_i)\}_{i=1,2,\dots,m}$  punti del piano cartesiano.



Cosa la retta da meglio si adatta ai dati,

ovvero cosa  $m, q$  che minimizzano:

$$\sum_{k=1}^n \left[ Y_k - (m X_k + q) \right]^2 \quad [\star]$$

Se definisce  $V = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{bmatrix}$ ,  $b = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix}$ ,

allora la quantità  $[\star]$  si può scrivere

come  $\| V \begin{bmatrix} m \\ q \end{bmatrix} - b \|_2^2$ . Dunque trovare

$m$  e  $q$  delle rette che cerchiamo equivale a  
risolvere

$$V \begin{bmatrix} m \\ q \end{bmatrix} = b \quad \text{nel senso dei minimi quadrati!}$$

ESEMPIO Determinare le rette di regressione

per i punti  $(0, 0), (1, -1), (2, 1), (3, 1)$ .

Definiamo  $A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix}$ ;  $b = \begin{bmatrix} 0 \\ -1 \\ 1 \\ 1 \end{bmatrix}$ .

Vogliamo risolvere  $A^T A z = A^T b$ ,  $z = \begin{bmatrix} m \\ q \end{bmatrix}$ .

$$A^T A = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix} = \begin{bmatrix} 14 & 6 \\ 6 & 4 \end{bmatrix}$$

$$A^T b = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$$

Sistema da risolvere

$$\begin{cases} 14m + 6q = 4 \\ 6m + 4q = 1 \end{cases} \iff \begin{cases} 7m + 3q = 2 \\ 6m + 4q = 1 \end{cases} \iff \dots$$

$$\begin{cases} m = \frac{1}{2} \\ q = -\frac{1}{2} \end{cases}; \text{ retta: } y = \frac{1}{2}x - \frac{1}{2}$$

Si può generalizzare:

TEOREMA Sono  $(x_i, y_i)$ ,  $i = 1, \dots, n$  i punti in

$\mathbb{R}^2$ . Il polinomio

$$P(x) = a_0 + a_1 x + \dots + a_{m-1} x^{m-1} + a_m x^m$$

che minimizza

$$\sum_{i=1}^m |y_i - (a_0 + a_1 x_i + \dots + a_m x_i^m)|^2$$

è detto polinomio d'adjion approssimazione

("best fit") per i dati  $(x_i, y_i)$  nel senso dei minimi quadrati. Ricorda

$$V = \begin{bmatrix} x_1^m & x_1^{m-1} & \dots & x_1 & 1 \\ x_2^m & x_2^{m-1} & \dots & x_2 & 1 \\ \vdots & \vdots & & & \\ x_m^m & x_m^{m-1} & \dots & x_m & 1 \end{bmatrix} \leftarrow \text{detta matrice di Vandermonde}$$

$$b = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad z = \begin{bmatrix} a_m \\ a_{m-1} \\ \vdots \\ a_1 \\ a_0 \end{bmatrix}, \quad \text{i coeff. del polinomio}$$

Ricavo otengo calcolando la LSS di  $Vz = y$ .

# CONDIZIONAMENTO DI UN SISTEMA LINEARE (esumi)

Siamo  $A \in \mathbb{R}^{n \times n}$  invertibile e  $b \in \mathbb{R}^n$ .

INPUT

OUTPUT

$$A, b \mapsto x \in \mathbb{R}^n \text{ t.c. } Ax = b$$

$$A + \delta A, b + \delta b \mapsto x + \delta x \in \mathbb{R}^n \text{ t.c. } (A + \delta A)(x + \delta x) = b + \delta b$$

Ci interoghiamo sulla sensibilità di  $x$  a perturbazioni su  $A$  e su  $b$ :

$\delta A, \delta b$  piccole  $\Rightarrow \delta x$  oltrattutto piccole?

semplifichiamo:  $\delta A = 0$

INPUT

OUTPUT

$$b \mapsto x \in \mathbb{R}^n \text{ t.c. } Ax = b$$

$$b + \delta b \mapsto x + \delta x \in \mathbb{R}^n \text{ t.c. } A(x + \delta x) = b + \delta b$$

errore relativo sui dati:  $\|\delta b\| / \|b\|$

errore relativo sulla soluzione:  $\|\delta x\| / \|x\|$

obiettivo: minimizzare  $\frac{\|\delta x\| / \|x\|}{\|\delta b\| / \|b\|}$

N.B.  $Ax = b$  &  $A(x + \delta x) = b + \delta b \Rightarrow$

$$\Rightarrow A \delta x = \delta b$$

$$\frac{\|\delta x\| / \|x\|}{\|\delta b\| / \|b\|} = \frac{\|\delta x\|}{\|\delta b\|} \frac{\|b\|}{\|x\|} = \frac{\|A^{-1} \delta b\|}{\|\delta b\|} \frac{\|Ax\|}{\|x\|}$$

$$\delta x = A^{-1} \delta b$$

$$b = Ax$$

Maggiorazione:

$$\frac{\|\delta x\| / \|x\|}{\|\delta b\| / \|b\|} \leq \max_{\substack{\delta b \in \mathbb{R}^m \\ \delta b \neq 0}} \frac{\|A^{-1} \delta b\|}{\|\delta b\|} \max_{\substack{x \in \mathbb{R}^m \\ x \neq 0}} \frac{\|Ax\|}{\|x\|}$$

Il condizionamento è legato a due quantità

che adesso definiamo.

Considerate una matrice vettoriale  $\|\cdot\|$ ,  
l'applicazione

$$A \in \mathbb{R}^{m \times n} \mapsto \|A\| = \max_{\substack{X \in \mathbb{R}^m \\ X \neq 0}} \frac{\|AX\|}{\|X\|}$$

è detta norma matriciale indotta dalla  
norma vettoriale  $\|\cdot\|$ . Essa soddisfa:

$$1) \|A\| \geq 0, \|A\| = 0 \iff A = 0$$

$$2) \|\alpha A\| = |\alpha| \|A\|, \forall \alpha \in \mathbb{R}, \forall A \in \mathbb{R}^{m \times n}$$

$$3) \|A+B\| \leq \|A\| + \|B\|, \forall A, B \in \mathbb{R}^{m \times n}$$

$$4) \|AB\| \leq \|A\| \|B\|, \forall A, B \in \mathbb{R}^{m \times n}$$

Non è facile calcolare  $\|A\|$  dalla sua

definizione, ma per le norme indotte dalle norme  
vettoriali  $\|\cdot\|$ , e  $\|\cdot\|_\infty$  esiste un'espressione  
semplice da calcolare.

$$\|A\|_1 = \max_{j=1, \dots, m} \sum_{i=1}^n |a_{ij}|$$

$$\|A\|_\infty = \max_{i=1, \dots, n} \sum_{j=1}^m |a_{ij}|$$

ESEMPIO

$$A = \begin{bmatrix} -1 & 2 \\ 3 & -4 \end{bmatrix}$$

$$\|A\|_1 = \max \{4, 6\} = 6$$

$$\|A\|_\infty = \max \{3, 7\} = 7$$

(comando Matlab: `cond(A)`)

Torniamo al condizionamento.

Definiamo, per  $A$  invertibile

$$K(A) = \|A\| \|\bar{A}\|, \text{ detto numero di}$$

condizionamento di  $A$ .

Allora da

$$\frac{\|\delta x\|}{\|x\|} \leq K(A) \frac{\|\delta b\|}{\|b\|}.$$

$K(A) \approx 1 \Rightarrow$  sistema ben condizionato

$K(A) \gg 1 \Rightarrow$  sistema Mal condizionato

Lo stesso fattore determina il condizionamento  
anche in presenza di perturbazioni su  $A$ !