Lezione 02 (ore 01,02) - 24 settembre 2025, 11:30 - 13:30 A13

(1) ZERI DI FUNZIONI E ARITMETICA DEL CALCOLATORE

(1.01) Problema.

Data f:[a,b] → R continua e tale che esiste t in R t.c. f(t) = 0, determinare t. Il numero t si chiama 'zero di f'.

(1.02) Teorema (di esistenza degli zeri)

Sia f:[a,b] → R continua e t.c. f(a)f(b) < 0. Allora: esiste t in (a,b) t.c. f(t) = 0.

(1.03) Osservazione.

La condizione f(a)f(b) < 0 è equivalente alla condizione:

f(a) *non è zero* & f(b) *non è zero* & segno f(a) *diverso da* segno f(b)

(1.04) Metodo di bisezione.

Idea: utilizzare il Teorema di esistenza degli zeri per ottenere una successione di intervalli I(k) = [a(k),b(k)] tale che:

- per ogni k, esiste zero di f in I(k)
- I(k+1) incluso in I(k)
- quando k → ∞ si ha mis I(k) → 0

(1.05) Descrizione del metodo.

z = *Bisezione*(f,a,b)

ingresso: f:(a,b) → R t.c. f(a)f(b) < 0

- a(0) = a; b(0) = b; I(0) = [a(0),b(0)]; x(0) = (a(0) + b(0)) / 2;
- per k = 1,2,3,… ripeti:
  se f(x(k-1)) = 0 allora STOP; altrimenti
  - se f(x(k-1))f(b(k-1)) < 0 allora a(k) = x(k-1); b(k) = b(k-1);
    altrimenti a(k) = a(k-1); b(k) = x(k-1);
  - I(k) = [a(k),b(k)]; x(k) = (a(k) + b(k)) / 2;

uscita: quando un opportuno *criterio d'arresto* è verificato: z = x(k), punto medio dell'ultimo intervallo determinato.

(1.06) Osservazione.

(A) mis I(k) = b(k) - a(k) = mis I(k-1) / 2^1 = mis I(k-2) / 2^2 = ... = mis I(0) / 2^k e quindi:

quando k → ∞ si ha mis I(k) → 0

(B) se f continua allora: per ogni k, I(k) contiene uno zero di f e

quando k → ∞ si ha x(k) → t con f(t) = 0

(Dimostrazione ...)

(1.07) Criterio d'arresto.

Il metodo di bisezione è un *metodo iterativo*, ovvero un metodo che approssima l'oggetto cercato costruendo una *successione*. Poiché è materialmente impossibile costruire *tutti* gli elementi della successione, è *necessario* introdurre un criterio d'arresto, ovvero una condizione che, quando verificata, arresta la costruzione delle successione.

Un esempio di criterio d'arresto è: dato Δ numero reale *positivo* ...

se mis I(k) < Δ allora STOP

Proprietà del criterio d'arresto:

(1) la *condizione* mis I(k) < Δ 'è *calcolabile*'
(2) la condizione è *certamente verificata* dopo un numero *finito* di iterazioni (vedi l'Osservazione (B) in (1.06)): il criterio 'è *efficace*'
(3) se f continua e k è tale che mis I(k) < Δ allora:

- esiste t in I(k) zero di f
- |x(k) - t| < mis I(k) / 2 < Δ/2 < Δ

ovvero la procedura restituisce un valore x(k) che è un'approssimazione di uno zero di f con *errore assoluto* |x(k) - t| minore di Δ: 'la procedura restituisce un'approssimazione *accurata quanto richiesto dall'utilizzatore*'.

Lezione 03 (ore 03,04) - 25 settembre 2025, 8:30 - 10:30 F3

(1.08) Realizzazione *Scilab*.

```
function [z, v, info, k, mis] = bisezione(f, a, b, E, kmax)
//
// Uso:
//     [ z,v,info,[k,[mis]] ] = bisezione(f,a,b,E,kmax)
//
//
// Approssima uno zero della funzione f:[a,b] -> R, che deve
// essere continua, con il metodo di bisezione. La funzione f
// deve assumere valori non nulli e di segno opposto in a e b.
//
// L'iterazione si arresta quando:
// (*) la funzione f ha valore zero nel punto medio x_m
//     dell'intervallo considerato [a(k),b(k)];
// (*) l'intervallo considerato [a(k),b(k)] ha misura minore di
//     E: in tal caso si ha, in teoria, che z approssima uno zero di
//     f con errore assoluto non superiore ad E/2;
// (*) dopo kmax iterazioni.
//
// kmax: valore opzionale (valore predefinito: 50).
//
// z: approssimazione finale (zero di f oppure punto medio
//    dell'ultimo intervallo generato);
// v: valore di f in z;
// info = 0: individuato valore in cui f si annulla (f(z) = 0);
//      = 1: f(z) ~= 0 e l'ultimo intervallo considerato ha misura
//           minore di E (mis < E);
//      = 2: f(z) ~= 0, mis >= E e il numero di iterazioni ha
//           raggiunto il massimo consentito (k = kmax);
// k: numero di iterazioni effettuate;
// mis: ampiezza dell'ultimo intervallo determinato.
//
//
// Inizializzazioni
//
if ~exists('kmax','l') then kmax = 50; end;
k_bis = 0; // contatore delle iterazioni eseguite
//
// Costruzione successioni
//
x_m = (a + b)/2;
f_m = f(x_m);
while (abs(b-a) >= E & f_m ~= 0 & k_bis < kmax),
  k_bis = k_bis+1;
  if sign(f_m) == sign(f(b)) then b = x_m; else a = x_m; end;
  x_m = (a + b)/2;
```

```
  f_m = f(x_m);
end;
//
// Fine costruzione: assegno variabili di uscita
//
z = x_m; v = f_m; k = k_bis; mis = abs(b-a);
if f_m == 0 then info = 0;
   else if abs(b-a) >= E then info = 2; else info = 1; end;
end;
//
endfunction
```

(1.09) Osservazione.

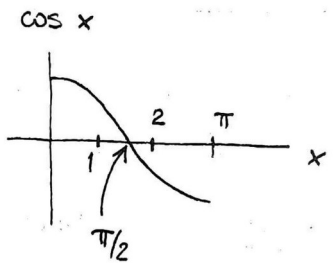
Il *costrutto Scilab*

```
while condizione,
  istruzioni;
end;
```

è equivalente a:

```
ripeti:
  se condizione è vera allora istruzioni;
       altrimenti esci dal ciclo;
```

(1.10) Esempio



Sia f(x) = cos(x).

- La funzione è continua in [a,b] = [1,2] e f(a) > 0, f(b) < 0
- Scelto E > 0 si ha:

$$\text{mis } I(k) < E \quad \Leftrightarrow \quad \text{mis } I(0) / 2^k < E \quad \Leftrightarrow$$

$$\Leftrightarrow \quad 2^k > \text{mis } I(0) / E \quad \Leftrightarrow \quad k > \log_2(\text{mis } I(0) / E)$$

dunque: ci aspettiamo di ottenere un'approssimazione di pi/2 con errore assoluto minore di E in

$$VA = \text{parte intera superiore di } \log_2(\text{mis } I(0) / E)$$

iterazioni.

- Si ottiene (utilizzando il file EsempioBisezione.sce scaricabile dalla sezione 'altro materiale didattico' della pagina web del corso):

E	info	mis	k	VA	kmax
10^{-5}	1	$7.6\ 10^{-6}$	17	17	50
10^{-10}	1	$5.8\ 10^{-11}$	34	34	50
10^{-15}	1	$8.8\ 10^{-16}$	50	50	50
10^{-16}	2	$2.2\ 10^{-16}$	60	54	60
10^{-16}	2	$2.2\ 10^{-16}$	100	54	100

Dalle ultime due righe della tabella si osserva che quando E = 10^{-16} la funzione *bisezione* si arresta perché ha raggiunto il numero massimo di iterazioni consentito ma, mentre nel primo caso (penultima riga) questo *è coerente* con le teoria, nel secondo caso (ultima riga) *non è coerente* con la teoria: la procedura *avrebbe dovuto arrestarsi dopo* 54 *iterazioni con* info = 1.

Per capire come mai accade questo, occorre studiare in maggior dettaglio l'ARITMETICA DEL CALCOLATORE.

(1.11) Domande.

(A) Con *quali numeri* è capace di operare il calcolatore?
(B) *Cosa sa fare* con questi numeri?

(1.12) Osservazione.

Siano x un numero reale *non zero*, β un numero intero maggiore o uguale a due (*base*). Esiste *una sola* fattorizzazione di x nella forma:

$$x = (-1)^s\ \beta^b\ g$$

con:

- s in {0,1}, *segno* di x
- b: numero intero, *esponente* di x in base β
- g: numero reale in [1/β,1), *frazione* di x in base β

(Dimostrazione:

- se x > 0 allora s = 0, se x < 0 allora s = 1;
- b è l'*unico* numero intero tale che

$$\beta^{b-1} < |x| \leq \beta^b$$

- g = |x| / β^b)

(1.13) Esempio.

(1) x = sqrt(5), β = 10 $\Rightarrow$ s = 0, b = 1, g = sqrt(5) / 10
(2) x = sqrt(5), β = 2 $\Rightarrow$ s = 0, b = 2, g = sqrt(5) / 4

(1.14) Osservazione.

La condizione g numero reale in [1/β,1) si traduce così: la scrittura posizionale di g in base β ha la forma:

$$0.c_1c_2c_3\ldots \text{ con } c_1 \textit{ diverso da zero}$$

In particolare: *se* β = 2 *si ha necessariamente* c_1 = 1.

(1.15) Esempio.

(1) x = 1/10, β = 10 $\Rightarrow$ s = 0, b = 0, g = 1/10 = 0.1

(2) x = 1/10, β = 2 $\Rightarrow$ s = 0, b = -3, g = 8/10 = 4/5 = $0.\overline{1100}$

(Ragionamento[1]:

(1)

* 4/5 = $0.c_1c_2c_3...$ $\Rightarrow$ 8/5 = $c_1.c_2c_3...$ e quindi:
* [8/5] = $[c_1.c_2c_3...]$ e {8/5} = $\{c_1.c_2c_3...\}$ ovvero:
* c_1 = 1 e 3/5 = $0.c_2c_3c_4...$

(2)

* 3/5 = $0.c_2c_3c_4...$ $\Rightarrow$ 6/5 = $c_2.c_3c_4...$ e quindi:
* [6/5] = $[c_2.c_3c_4...]$ e {6/5} = $\{c_2.c_3c_4...\}$ ovvero:
* c_2 = 1 e 1/5 = $0.c_3c_4c_5...$

(3)

* 1/5 = $0.c_3c_4c_5...$ $\Rightarrow$ 2/5 = $c_3.c_4c_5...$ e quindi:
* [2/5] = $[c_3.c_4c_5...]$ e {2/5} = $\{c_3.c_4c_5...\}$ ovvero:
* c_3 = 0 e 2/5 = $0.c_4c_5c_6...$

(4)

* 2/5 = $0.c_4c_5c_6...$ $\Rightarrow$ 4/5 = $c_4.c_5c_6...$ e quindi:
* [4/5] = $[c_4.c_5c_6...]$ e {4/5} = $\{c_4.c_5c_6...\}$ ovvero:
* c_4 = 0 e 4/5 = $0.c_5c_6c_7...$

Si osserva adesso che si è ottenuta una nuova scrittura del numero iniziale 4/5. Se ne deduce che 4/5 ha scrittura periodica di periodo quattro.

Fine del ragionamento.)

Si osservi che in entrambi gli esempi si ha x = 1/10 ma nell'esempio (1) la frazione ha scrittura posizionale di *lunghezza finita*, nell'esempio (2) ha *lunghezza infinita*. La lunghezza della scrittura posizionale dipende dalla base.

1 Se q è un numero reale, con [q] si indica la *parte intera* di q e con {q} la *parte frazionaria* di q, ovvero {q} = q - [q].

Lezione 4 (ore 5,6) - 30 settembre 2025, 16:30 - 18:30 F3

(1.16) Definizione (numeri in virgola mobile e precisione finita).

Siano β un numero intero maggiore o uguale a due e m un numero intero maggiore o uguale a 1. L'insieme

$$F(\beta,m) = \{0\} \cup \{x \text{ in } R \text{ t.c. } x = (-1)^s \beta^b\ 0.c_1...c_m \text{ con}$$
$$s \in \{0,1\},\ b \in Z,\ c_1,...,c_m \text{ cifre in base } \beta,\ c_1 \neq 0\}$$

si chiama 'insieme dei numeri in *virgola mobile* e *precisione* m in base β'.

(1.17) Esempio.

Si consideri F(10,1).

- $1/100 \in F(10,1)$: $1/100 = 10^{-2} = 10^{-1}\ 0.1$
- $11/100 \notin F(10,1)$: $11/100 = 0.11 = 10^{0}\ 0.11$ e la frazione 0.11 non è compatibile con la precisione m = 1
- tutti gli elementi di F(10,1) positivi con esponente zero:

  $$B = \{0.1\ ;\ 0.2\ ;\ ...\ ;\ 0.9\}$$

  tutti quelli con esponente $b \in Z$:

  $$10^b\ B \text{ (positivi)} \qquad -10^b\ B \text{ (negativi)}$$

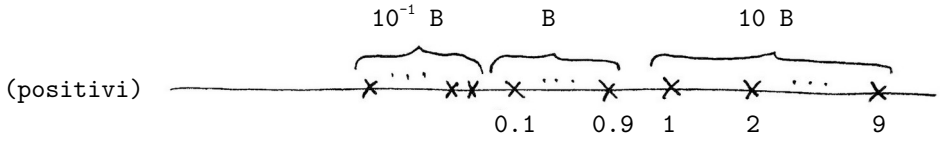
- $F(10,1) = \bigcup_{b \in Z} (-1)\ 10^b\ B \cup \{0\} \cup \bigcup_{b \in Z} 10^b\ B$

(1.18) Osservazione (proprietà di $F(\beta,m)$).

(1) è *sottoinsieme proprio* di Q (dunque numerabile e ordinato)
(2) è *simmetrico* rispetto a zero
(3) zero è (l'unico) punto di accumulazione
(4) $\sup F(\beta,m) = +\infty$, $\inf F(\beta,m) = -\infty$

(1.19) Osservazione (distanza tra elementi consecutivi).

In F(10,1):



Distanza tra consecutivi: $10^{-1}\ 0.1$ (b = -1), $0.1 = 10^{0}\ 0.1$ (b = 0), $1 = 10^{1}\ 0.1$ (b = 1).

- esponente b, distanza tra consecutivi in F(10,1): $10^b\ 0.1 = 10^b\ 10^{-1}$

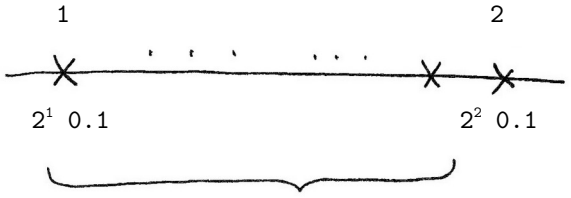
- in F(β,m): dato $\xi = \beta^b$ g e detto $\sigma(\xi)$ il *successore* di ξ si ha:

$$\sigma(\xi) - \xi = \beta^{b-m}$$

- la distanza è tanto maggiore quanto l'esponente è grande ('tanto più ξ è lontano da zero').

(1.20) Osservazione.

Nell'Esempio (1.10) della Lezione 3, la situazione è:



* $\alpha \in (1,2)$

* in *Scilab* (*Octave, Matlab*):

F(2,53)

b = 1 $\Rightarrow$ distanza tra consecutivi = $2^{1-53} = 2^{-52} \approx 2.22\ 10^{-16}$

- Nel caso $E = 10^{-16}$ la *function bisezione* ha trovato l'intervallo (non degenere) più piccolo possibile che contiene lo zero α e di estremi in F(2,53), ma questo intervallo ha misura > E.
- È *inutile* scegliere $E < \beta^{b-m}$.

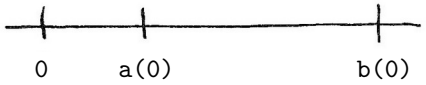
(1.21) Criterio d'arresto (con richiesta sull'errore relativo).

Dato E numero reale *positivo...*

$$\underline{\text{se}}\ \frac{\text{mis } I(k)}{\min\{|a(k)|,|b(k)|\}} < E\ \ \underline{\text{allora}}\ \text{STOP}$$

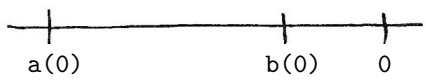
Proprietà del criterio d'arresto:

(1) la condizione è *calcolabile*
(2) se $0 \notin I(0)$ si ha: per ogni k, $0 \notin I(k)$ e

- 

$\Rightarrow$ $\min\{|a(k)|,|b(k)|\} = a(k) > 0$

e $a(0) \leqslant a(k) < b(0)$ $\Rightarrow$ quando $k \to \infty$, mis I(k) / a(k) $\to$ 0

- 

$\Rightarrow$ $\min\{|a(k)|,|b(k)|\} = |b(k)| > 0$

e $|b(0)| \leqslant b(k) < |a(0)|$ $\Rightarrow$ quando $k \to \infty$, mis I(k) / |b(k)| $\to$ 0

quindi: la condizione è *certamente verificata* dopo un numero *finito* di iterazioni (criterio *efficace*).

(3) se f è continua allora:

- esiste $\alpha \in$ I(k) zero di f

- $$\frac{|x(k) - \alpha|}{|\alpha|} \leqslant \frac{\text{mis } I(k) / 2}{|\alpha|} < 1/2 \frac{\text{mis } I(k)}{\min\{|a(k)|,|b(k)|\}} < E/2 < E$$

- x(k) approssima α con *errore relativo* < E: 'la procedura restituisce un'approssimazione accurata *quanto richiesto dall'utilizzatore*'

- è *inutile* scegliere E < β^{1-m}

Lezione 5 (ore 7,8) - 1 ottobre 2025, 11:30 - 13:30 A13

(1.22) Osservazione (conseguenze di F(2,53) $\neq$ R).

Indichiamo con M l'insieme dei numeri che il calcolatore sa manipolare, i '*numeri di macchina*' del calcolatore. Quale insieme sia esattamente M *dipende* dal calcolatore che si considera. Nel caso di *Scilab* (e *Octave* e *Matlab*) l'insieme M è 'sostanzialmente' F(2,53). Riservandoci di chiarire più avanti le differenze tra i due insiemi, assumiamo che:

in *Scilab* si ha M = F(2,53)

Consideriamo i seguenti esempi (il carattere > è il *prompt* della *console* di *Scilab*).

- > x = 0.1;

  Poiché 0.1 = un decimo $\notin$ F(2,53), dopo l'assegnamento il valore di x *non può essere* un decimo.

- > (1 - 9/10) * 10 - 1
  ans = - 2.220D-16

  Si ha: 1, 9, 10 $\in$ F(2,53) ma nove decimi $\notin$ F(2,53). Ovvero:

  esistono x,y $\in$ F(2,53) t.c. x/y $\notin$ F(2,53)

- Sia $f(x) = \dfrac{x(x-1)}{x-\sqrt{x}}$, definita per $x > 0$ e $x \neq 1$.

  (A) Si ha: $f(x) = \dfrac{x^2 - x}{x - \sqrt{x}} = \dfrac{(x+\sqrt{x})(x-\sqrt{x})}{x-\sqrt{x}} = x + \sqrt{x}$

  (B) Per x = 2 $\in$ F(2,53) si ha:

  ```
  > a = 2 * (2 - 1)/(2 - sqrt(2));
  > b = 2 + sqrt(2);
  > a == b
  ans = F
  ```

(1.23) Definizione (funzione arrotondamento).

Il calcolatore usa gli elementi di F(β,m) per *approssimare* numeri reali. L'approssimazione è realizzata dalla *funzione arrotondamento* rd: R $\to$ F(β,m) così definita:

rd(x) = l'elemento di F(β,m) *più vicino* ad x o, in caso di ambiguità, quello dei due elementi di F(β,m) equidistanti da x che ha la frazione che termina con una *cifra pari*.

(1.24) Osservazione.

La definizione è ben posta se β è pari e m $\geqslant$ 2. In tal caso, se l'ultima cifra della frazione di $\xi \in F(\beta,m)$ è *pari* (rispettivamente: *dispari*), l'ultima cifra della frazione del successore di ξ è *dispari* (rispettivamente: *pari*).

Se β è pari e m = 1 oppure β è dispari, invece, la definizione non è ben posta. Ad esempio, in F(3,2) gli elementi positivi con esponente zero sono:

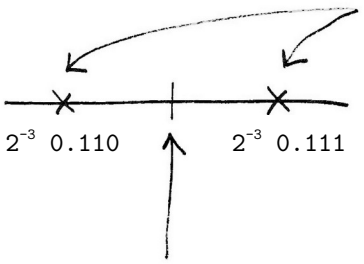
$$3^0\ 0.10\ ;\ 3^0\ 0.11\ ;\ 3^0\ 0.12\ ;\ 3^0\ 0.20\ ;\ \ldots$$

e gli ultimi due elementi scritti sono consecutivi ed hanno *entrambi* l'ultima cifra della frazione *pari*.

(1.25) Esempio.

Sia x = 1/10. Si vuole determinare l'arrotondato di x in F(2,3).

Si è già determinato (Esempio (1.15)) che: $x = 2^{-3}\ 0.\overline{1100}$. Allora si ha la situazione di figura:



elementi di F(2,3) *adiacenti* ad x (quello a sinistra si ottiene *troncando* la scrittura della frazione di x al numero di cifre indicato dalla precisione - in questo caso 3 - quello a destra è il successore)

punto medio = $2^{-3}\ 0.1101 > x \Rightarrow rd(x) = 2^{-3}\ 0.110$ (= 3/32)

(1.26) Osservazione.

La funzione rd *non è* una funzione che il calcolatore mette a disposizione dell'utilizzatore, ma è indispensabile per capire come:

(1) il calcolatore 'legge' i numeri reali;
(2) il calcolatore fa operazioni sugli elementi di $F(\beta,m)$.

(1.27) Esempio.

Riprendiamo il primo esempio dell'Osservazione (1.22). In *Scilab* l'effetto dell'assegnamento:

```
> x = 0.1
```

è: viene assegnata alla variabile x il valore rd(0.1) $\in$ F(2,53) (se al momento dell'assegnamento la variabile x non esistesse, viene creata).

Il calcolatore approssima il numero reale con il suo arrotondato in $F(\beta,m)$. Ci si domanda *quale errore* venga commesso.

(1.28) Teorema (limitazione dell'errore relativo).

Sia rd la funzione arrotondamento in F(β,m). Per ogni numero reale x $\neq$ 0 si ha:

$$\frac{|\mathrm{rd}(x)-x|}{|x|} \leqslant \frac{1}{2}\,\beta^{1-m} = u \quad (\textit{precisione di macchina})$$

(Dimostrazione...)

(1.29) Osservazione.

- La limitazione è *uniforme*, nel senso che la quantità che limita l'errore è *indipendente da* x (dipende solo dai parametri β ed m che definiscono l'insieme dei numeri).

- In F(2,53) si ha $u = \frac{1}{2}\,2^{1-53} = 2^{-53} \approx 1.11\ 10^{-16}$.
- Se si considera l'errore *assoluto,* dal Teorema precedente si ottiene, per ogni numero reale x, la limitazione (*non uniforme*!):

  $$|\mathrm{rd}(x) - x| \leqslant u\,|x|$$

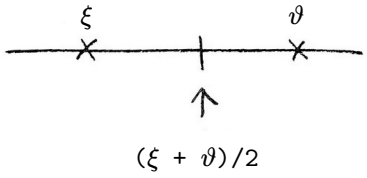
  Se ne deduce che *tanto più lontano da zero* è x *tanto più grande può essere* l'errore assoluto.

  La differenza sostanziale tra le due limitazioni, una è uniforme e l'altra no, è dovuta a come sono distribuiti gli elementi di F(β,m). Questi ultimi sono *pensati appositamente* per ottenere la limitazione uniforme dell'errore relativo.

(1.30) Esempio.

Siano ξ un elemento positivo di F(2,53) e ϑ il successore di ξ. Si ha:

- $\xi/2 \in F(2,53)$, $\vartheta/2 \in F(2,53)$
- $\xi/2 + \vartheta/2 \notin F(2,53)$



Scelto ξ = 1, in *Scilab* si ha il seguente dialogo (per ogni t $\in$ F(2,53), nearfloat('succ',t) è il successore di t):

```
> c = 1/2 + nearfloat('succ',1)/2
c = 1
> c == 1
ans = T
```

Per capire il dialogo è necessario approfondire come *Scilab* esegue la somma di due numeri di macchina. Se $\xi,\vartheta \in$ F(2,53), indichiamo con $\xi \oplus \vartheta$ il valore assegnato da *Scilab*

all'espressione $\xi + \vartheta$. Per definizione si ha:

$$\xi \oplus \vartheta = \mathrm{rd}(\xi + \vartheta)$$

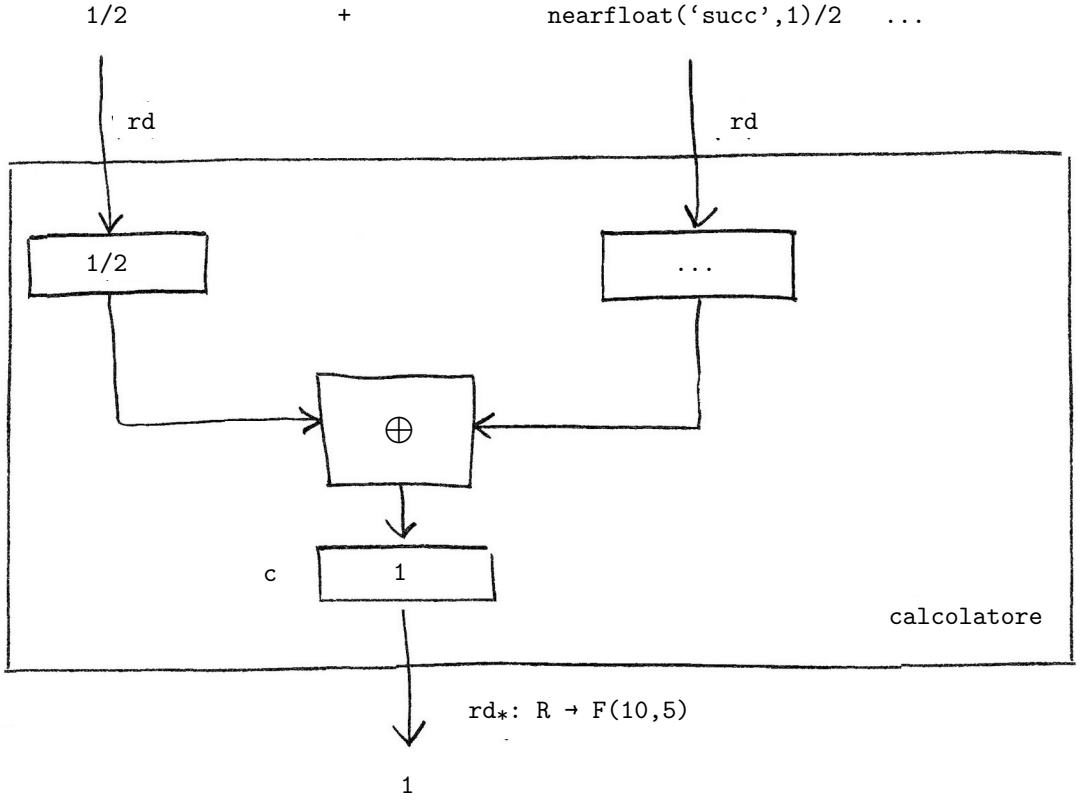
Il valore è definito 'nel modo migliore possibile' nel senso che l'errore tra il valore esatto $\xi + \vartheta$ e quello definito $\xi \oplus \vartheta$ è *il minimo possibile*.

Torniamo all'esempio. Il valore che *Scilab* assegna a c è, allora:

$$1/2 \oplus \mathrm{nearfloat}(\text{'succ'},1)/2 = \mathrm{rd}(1/2 + \mathrm{nearfloat}(\text{'succ'},1)/2)$$

che, secondo la definizione di arrotondamento, vale 1 (quello, tra i due elementi adiacenti al numero da arrotondare, che ha ultima cifra della frazione *pari*).

Quello che accade nel primo assegnamento è:



Lezione 6 (ore 9,10) - 2 ottobre 2025, 8:30 - 10:30 F3

(1.31) Definizione (funzioni predefinite).

Sia M = F(β,m) l'insieme dei numeri di macchina del calcolatore in esame, e rd la funzione arrotondamento in M. L'insieme FP delle *funzioni predefinite*, ovvero delle funzioni che il calcolatore sa calcolare operando con gli elementi di M è costituito da tre classi.

- L'insieme delle funzioni predefinite corrispondenti ad *operazioni aritmetiche*. Se $\cdot$ è una delle operazioni aritmetiche tra numeri reali +, -, $\times$, / allora la funzione predefinita corrispondente si indica con il simbolo $\odot$ (un cerchietto contenente il simbolo dell'operazione considerata) ed è definita, per ogni coppia ξ, ϑ di elementi di F(β,m) facenti parte del dominio dell'operazione $\cdot$, da

$$\xi \odot \vartheta = \mathrm{rd}(\xi \cdot \vartheta)$$

- L'insieme delle funzioni predefinite corrispondenti alle usuali *funzioni elementari* (sen, cos, arcsen, arccos, ln, exp ...). Se f:A $\to$ R è una delle funzioni elementari allora la funzione predefinita corrispondente si indica con il simbolo F ed è definita, per ogni elemento ξ di F(β,m) facente parte del dominio A della funzione elementare f, da

$$F(\xi) = \mathrm{rd}(f(\xi))$$

- L'insieme delle funzioni predefinite corrispondenti ai *confronti* tra numeri reali ($<$, $\leqslant$, $=$, $\neq$, $\geqslant$, $>$). In questo caso, poiché gli elementi di F(β,m) *sono* numeri reali, essi vengono confrontati come tali. Quindi le funzioni predefinite corrispondenti ai confronti sono semplicemente le restrizioni a F(β,m) $\times$ F(β,m) dei confronti tra numeri reali (e non è necessario introdurre simboli nuovi per indicarle).

(1.32) Definizione (algoritmo, algoritmo ingenuo).

Siano $f_1,\ldots,f_k$ funzioni elementari o operazioni aritmetiche e sia f:A $\to$ R, con A un opportuno sottoinsieme di R, la funzione ottenuta *componendo* $f_1,\ldots,f_k$:

$$f(x) = f_1 \circ \ldots \circ f_k(x)$$

(ad esempio: f(x) = sen(x) + cos(x), dove $f_3(x)$ = sen(x), $f_2(x)$ = cos(x) e $f_1(x_1,x_2) = x_1 + x_2$). Se chiediamo a *Scilab* di valutare la funzione f con l'istruzione

```
> f(x)
```

il valore restituito sarà

$$F_1 \circ \ldots \circ F_k(\mathrm{rd}(x))$$

dove $F_1,\ldots,F_k(x)$ sono, rispettivamente, le funzioni predefinite corrispondenti a $f_1,\ldots,f_k(x)$.

L'espressione $F_1 \circ \ldots \circ F_k(\mathrm{rd}(x))$ definisce una funzione φ: A $\to$ M detta *algoritmo ingenuo*

per f (per la funzione dell'esempio: $\varphi(x) = SEN(rd(x)) \oplus COS(rd(x))$, definita per ogni x in R). Con il termine *algoritmo* ci si riferisce, in generale, ad una sequenza *finita* di operazioni di calcolo di funzioni predefinite.

Salvo casi molto particolari, ci saranno valori di x per i quali $f(x) \neq \varphi(x)$. In questi casi si utilizza $\varphi(x)$ per approssimare f(x) ed è interessante avere *informazioni sull'errore commesso*.

Per ottenere queste informazioni introduciamo le nozioni di *algoritmo accurato*, *algoritmo stabile* e di *calcolo ben condizionato del valore di una funzione*.

(1.33) Definizione (algoritmo accurato).

Siano $f:A \to R$ una funzione, $\varphi:A \to M$ l'algoritmo utilizzato per approssimare i valori di f e $x \in A$.

L'algoritmo φ si dice *accurato* (quando utilizzato per approssimare il valore di f in x) se esiste un numero reale ε tale che:

(1) $\varphi(x) = (1 + \varepsilon)\, f(x)$
(2) ε 'piccolo'

Se l'algoritmo è accurato per ogni $x \in B \subset A$, si dirà che l'algoritmo è accurato in B. In tal caso ε dipenderà da x.

(1.34) Osservazione.

- Siano f ed x tali che $f(x) \neq 0$. La (1) della Definizione precedente è *equivalente* alla seguente:

$$\varepsilon = \frac{\varphi(x) - f(x)}{f(x)}$$

  In questo caso dunque, l'algoritmo è accurato equivale a dire che l'errore relativo commesso approssimando f(x) con $\varphi(x)$ è 'piccolo'.

- Se l'algoritmo è accurato si ha: $f(x) = 0 \Leftrightarrow \varphi(x) = 0$.

- La definizione di algoritmo accurato è *qualitativa* perché non si quantifica il termine 'piccolo' relativo ad ε. Il significato concreto del termine 'piccolo' dipende caso per caso. Ad esempio, se, come nel caso del metodo di bisezione, interessa soltanto che $\varphi(x)$ e f(x) abbiano lo stesso segno, ε 'piccolo' significa $\varepsilon > -1$.

  Esercizio: Si approssima una $L > 0$ con λ. Che errore relativo ε si commette utilizzando $\lambda = 0$? Quale valore di λ si deve usare per ottenere un errore relativo $\varepsilon = 1$?

(1.35) Definizione (algoritmo stabile).

Siano $f:A \to R$ una funzione, $\varphi:A \to M$ l'algoritmo utilizzato per approssimare i valori di f e $x \in A$.

L'algoritmo φ si dice *stabile* (quando utilizzato per approssimare il valore di f in x) se esistono numeri reali ε_a, ε_v tali che:

(1) $\varphi(x) = (1 + \varepsilon_v)\, f((1 + \varepsilon_a)x)$

(2) ε_a, ε_v 'piccoli'

Se l'algoritmo è stabile per ogni $x \in B \subset A$, si dirà che l'algoritmo è stabile in B. In tal caso ε_a, ε_v dipenderanno da x.

(1.36) Osservazione.

- Se un algoritmo è accurato allora è stabile ($\varepsilon_a = 0$, $\varepsilon_v = \varepsilon$) ma *non* viceversa.
- Informalmente: un algoritmo stabile restituisce una *buona approssimazione* (ε_v 'piccolo') del valore di f in un punto *vicino* ad x (ε_a 'piccolo').

(1.37) Osservazione (algoritmo 'buono').

La nozione di stabilità formalizza l'idea di algoritmo '*buono*' per approssimare i valori di una data f. Ad esempio, se f è una funzione elementare e φ è l'algoritmo ingenuo per f allora, detta F la funzione predefinita corrispondente ad f, si ha:

$$\varphi(x) = F(\, rd(x)\,) = rd(\, f(\, rd(x)\,)\,)$$

(1.38) Teorema (errore relativo e perturbazione).

Ricordando la definizione di errore relativo commesso approssimando un numero reale t con l'arrotondato rd(t) ed il Teorema (1.28) della Lezione 5 sulla limitazione dell'errore relativo, si ottiene:

Siano x un numero reale e rd la funzione arrotondamento in $F(\beta,m)$. Esiste un numero reale ε tale che:

$$rd(x) = (1 + \varepsilon)x \quad \text{e} \quad |\varepsilon| < u$$

L'uguaglianza esprime l'arrotondato di x come (piccola) *perturbazione moltiplicativa* di x.

(Dimostrazione: se $x \neq 0$ allora ε è l'errore relativo commesso approssimando x con rd(x); se $x = 0$ (e quindi $rd(x) = 0$) l'uguaglianza sussiste, ad esempio, con $\varepsilon = 0$.)

(1.39) Osservazione (continuazione della precedente).

Utilizzando due volte il Teorema precedente si ottiene infine:

$$\varphi(x) = (1 + \varepsilon_2) f(\, (1 + \varepsilon_1)x\,) \quad \text{con} \quad |\varepsilon_1| < u \;\; \text{e} \;\; |\varepsilon_2| < u$$

L'algoritmo φ restituisce la *migliore approssimazione possibile* del valore di f nel punto *più vicino possibile* ad x. In questo senso φ è l'algoritmo 'migliore possibile' che il calcolatore possa utilizzare per approssimare f(x). Da qui, generalizzando, l'idea che un

algoritmo 'buono' per approssimare il valore di una funzione in un punto assegnato sia un algoritmo che restituisce una buona approssimazione del valore della funzione in un punto vicino a quello in cui si voleva calcolarla.

(1.40) Definizione (calcolo ben condizionato del valore di una funzione).

Siano $f:A \to R$ una funzione e $x \in A$. Il calcolo del valore di f in x è *ben condizionato* se: *per ogni* numero reale α 'piccolo' *esiste* un numero reale ε_v 'piccolo' tale che

$$f((1 + \alpha)x) = (1 + \varepsilon_v)f(x)$$

Informalmente: il calcolo del valore di f in x è ben condizionato se il valore di f in *ogni* punto 'vicino' ad x è una 'buona' approssimazione del valore di f in x.

(1.41) Osservazione.

- La proprietà che il calcolo del valore di f in x sia ben condizionato riguarda *esclusivamente* la funzione f. In particolare, non è legata a *quale algoritmo* si sceglie per approssimare i valori di f.
- Se $f(x) \neq 0$, il valore di ε_v, una volta assegnato α, è *determinato*. Precisamente, ε_v risulta:

$$\varepsilon_v = \frac{f((1 + \alpha)x) - f(x)}{f(x)}$$

(1.42) Teorema (stabilità + buon condizionamento => accuratezza).

Siano $f:A \to R$ una funzione, $x \in A$ e φ l'algoritmo utilizzato per approssimare f(x). Se l'algoritmo è *stabile* e il calcolo di f in x è *ben condizionato* allora l'algoritmo è *accurato*.

Dimostrazione. Per la stabilità dell'algoritmo esistono ε_1 e ε_2 tali che:

$$\varphi(x) = (1 + \varepsilon_2)f((1 + \varepsilon_1)x) \quad \text{con} \quad \varepsilon_1 \text{ e } \varepsilon_2 \text{ 'piccoli'}$$

Per il buon condizionamento del calcolo di f in x esiste ε_3 tale che:

$$f((1 + \varepsilon_1)x) = (1 + \varepsilon_3)f(x) \quad \text{e} \quad \varepsilon_3 \text{ 'piccolo'}$$

Allora possiamo riscrivere:

$$\varphi(x) = (1 + \varepsilon_2)(1 + \varepsilon_3)f(x)$$

e, posto $(1 + \varepsilon_2)(1 + \varepsilon_3) = 1 + t$, ovvero $t = \varepsilon_2 + \varepsilon_3 + \varepsilon_2\varepsilon_3$, si ottiene infine:

$$\varphi(x) = (1 + t)f(x) \quad \text{con} \quad t \text{ 'piccolo'}$$

dunque l'algoritmo è accurato.

(1.43) Osservazione (stabilità degli algoritmi ingenui nei casi elementari).

- Per quanto ricavato nelle Osservazioni (1.37) e (1.39), se $f:A \to R$ è una funzione elementare e φ è l'algoritmo ingenuo per f, φ è stabile su A: *l'algoritmo ingenuo per ciascuna funzione elementare è stabile.*
- Sia $f(x_1,x_2) = x_1 + x_2$. L'algoritmo ingenuo per f è:

$$\varphi(x_1,x_2) = rd(x_1) \oplus rd(x_2)$$

Ricordando la definizione di $\oplus$ (vedi Definizione (1.31)) ed utilizzando tre volte il Teorema (1.38) si riscrive:

$$\varphi(x_1,x_2) = (1 + \varepsilon_3)((1 + \varepsilon_1)x + (1 + \varepsilon_2)x) \quad , \quad \text{con } |\varepsilon_j| \leqslant u \ , \ j = 1,2,3$$

Dunque, l'algoritmo ingenuo per la somma è stabile.

Allo stesso modo si dimostra che *l'algoritmo ingenuo per ciascuna delle operazioni aritmetiche è stabile.*

Lezione 7 (ore 11,12) - 7 ottobre 2025, 16:30 - 18:30 F3

(1.44) Osservazione (stabilità, caso non elementare).

Siano f_1, f_2:R → R due funzioni elementari e φ_1, φ_2:R → M gli algoritmi utilizzati per approssimare, rispettivamente, i valori di f_1 ed f_2. Siano poi $x \in R$, $f(x) = f_2(f_1(x))$ e $\varphi(x) = \varphi_2(\varphi_1(x))$. Infine, supponiamo che gli algoritmi φ_1, φ_2 siano *stabili* su R. Ci si domanda se l'algoritmo φ è stabile quando utilizzato per approssimare f in x. Utilizzando la stabilità di φ_1 e φ_2 si ha: esistono numeri reali $\varepsilon_1,\ldots,\varepsilon_4$ tali che $|\varepsilon_j| \leqslant u$, $j = 1,2,3,4$ e:

$$\varphi(x) = \varphi_2(\varphi_1(x)) = (1 + \varepsilon_4)f_2((1 + \varepsilon_3)(1 + \varepsilon_1)f_1((1 + \varepsilon_2)x))$$

Posto $(1 + \varepsilon_3)(1 + \varepsilon_1) = 1 + t$, ovvero $t = \varepsilon_3 + \varepsilon_2 + \varepsilon_2\varepsilon_3$, si ha: $|t| \leqslant 2u + u^2$ (< 1) e

$$\varphi(x) = (1 + \varepsilon_4)f_2((1 + t)f_1((1 + \varepsilon_2)x))$$

Indicato con ϑ l'errore relativo commesso approssimando $f_2(f_1((1 + \varepsilon_2)x))$ con $f_2((1 + t)f_1((1 + \varepsilon_2)x))$ si riscrive:

$$f_2((1 + t)f_1((1 + \varepsilon_2)x)) = (1 + \vartheta)f_2(f_1((1 + \varepsilon_2)x))$$

e quindi:

$$\varphi(x) = (1 + \varepsilon_4)(1 + \vartheta)f_2(f_1((1 + \varepsilon_2)x))$$

Infine, posto $(1 + \varepsilon_4)(1 + \vartheta) = 1 + \varepsilon_v$ e $\varepsilon_2 = \varepsilon_a$, si ottiene:

$$\varphi(x) = (1 + \varepsilon_v)f((1 + \varepsilon_a)x)$$

Per poterne dedurre la stabilità di φ quando utilizzato per approssimare f in x, occorre indagare la grandezza delle perturbazioni ε_v e ε_a. Riguardo ad ε_a si ha $|\varepsilon_a| \leqslant u$, dunque ε_a 'piccolo'. La grandezza di ε_v, invece, *dipende* da quella di ϑ che, a sua volta *dipende* dal condizionamento del calcolo di f_2 in $f_1((1 + \varepsilon_2)x)$. Se quest'ultimo calcolo è *ben condizionato* (dunque ϑ 'piccolo') allora φ è stabile quando utilizzato per approssimare f in x, altrimenti nulla si può dire riguardo alla stabilità di φ.

(1.45) Osservazione (condizionamento del calcolo di funzioni regolari).

Siano f:A → R una funzione *regolare* (ovvero con derivata prima continua), e $x \in A$ tale che $f(x) \neq 0$. Si vuole studiare il condizionamento del calcolo di f in x.

Poiché $f(x) \neq 0$, per quanto detto nell'Osservazione (1.41) della Lezione 6, si deve

studiare, assegnato $\alpha \in R$ 'piccolo', la quantità:

$$\varepsilon_v = \frac{f((1+\alpha)x) - f(x)}{f(x)}$$

Per la regolarità di f, utilizzando il Teorema di Lagrange, si ha:

esiste un numero reale ϑ compreso tra x e $(1+\alpha)x$ tale che

$$f((1+\alpha)x) - f(x) = f'(\vartheta)\,\alpha\,x$$

Quindi si riscrive:

$$\varepsilon_v = \frac{f'(\vartheta)\,\alpha\,x}{f(x)}$$

Per l'ipotesi α 'piccolo' si può ragionevolmente approssimare $\vartheta \approx x$ e riscrivere infine:

$$\varepsilon_v \approx \frac{f'(x)}{f(x)}\,x\,\alpha$$

Introdotto il *numero di condizionamento* del calcolo di f in x:

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

si ha allora:

$$|\varepsilon_v| \approx c(x)\,|\alpha|$$

e il condizionamento del calcolo di f in x dipende solo dalla grandezza del numero di condizionamento c(x).

(1.46) Esempio.

Sia $f(x) = \mathrm{sen}(x)$ e $x \in (0, \pi/2)$. Il numero di condizionamento del calcolo di f in x è:

$$c(x) = \left|\frac{\cos(x)}{\mathrm{sen}(x)}x\right| = \left|\frac{x}{\tan(x)}\right| = \frac{x}{\tan(x)} < 1$$

Dunque in questo caso il calcolo di sen(x) è *ben condizionato*. Ma se consideriamo x *vicino* (ma non uguale) a π, tenuto conto che:

$$\lim_{t\to\pi} c(x) = \lim_{t\to\pi}\left|\frac{x}{\tan(x)}\right| = +\infty$$

il calcolo di sen(x) *non è* ben condizionato.

(1.47) Osservazione (condizionamento delle operazioni aritmetiche).

Siano $f(x_1,x_2) = x_1 + x_2$ e x_1, x_2 tali che $f(x_1,x_2) \neq 0$. Si vuole studiare il condizionamento del calcolo di f in x_1, x_2.

Poiché $f(x_1, x_2) \neq 0$, per quanto detto nell'Osservazione (1.41) della Lezione 6, si deve studiare, assegnati numeri reali α_1 e α_2 'piccoli', la quantità:

$$\varepsilon_v = \frac{(1+\alpha_1)\,x_1 + (1+\alpha_2)\,x_2 - (x_1+x_2)}{x_1+x_2} = \frac{x_1}{x_1+x_2}\,\alpha_1 + \frac{x_2}{x_1+x_2}\,\alpha_2$$

Introdotti i numeri di condizionamento:

$$c_1(x_1,x_2) = \left|\frac{x_1}{x_1+x_2}\right| \quad \text{e} \quad c_1(x_1,x_2) = \left|\frac{x_2}{x_1+x_2}\right|$$

si ha:

se $x_1x_2 > 0$ (ovvero i due addendi hanno lo stesso segno) allora:

$$c_1(x_1,x_2) < 1 \quad \text{e} \quad c_2(x_1,x_2) < 1$$

e il condizionamento del calcolo della somma è *buono*. Invece, se $x_1x_2 < 0$ (ovvero i due addendi hanno lo segno opposto), il condizionamento del calcolo può essere *tanto peggiore quanto più piccolo è* $x_1 + x_2$. Si ha infatti, assegnato $x_1 \neq 0$ e posto $x_2 = y - x_1$ (ovvero $x_1 + x_2 = y$) con $y \neq 0$:

$$c_1(x_1,x_2) = \left|\frac{x_1}{y}\right| \quad , \quad c_1(x_1,x_2) = \left|1-\frac{x_1}{y}\right|$$

e:

$$\lim_{y\to 0} c_1(x_1,x_2) = +\infty \quad , \quad \lim_{y\to 0} c_2(x_1,x_2) = +\infty$$

Nel caso delle altre operazioni aritmetiche si ha:

$$\varepsilon_v = \alpha_1 + \alpha_2 + \alpha_1\alpha_2 \quad \text{(moltiplicazione)}$$

$$\varepsilon_v = \frac{\alpha_1-\alpha_2}{1-\alpha_2} \quad \text{(divisione)}$$

e in entrambi i casi il calcolo è *sempre ben condizionato.*

Lezione 8 (ore 13,14) - 8 ottobre 2025, 11:30 - 13:30 A13

(1.48) Esempio (approssimazione numerica della derivata).

Si supponga di conoscere, agli istanti t_1 e t_2, le posizioni x_1 e x_2 di un punto in moto su una retta. La quantità:

$$\bar{v} = (x_2 - x_1) / (t_2 - t_1)$$

è la velocità media del punto tra i due istanti. Se le quantità x_1 e x_2 sono note soltanto con errore relativo ε_1 e ε_2, ad esempio perché ottenute tramite misurazioni, potremo ottenere di $\bar{v}$ soltanto un'approssimazione:

$$w = \frac{(1 + \varepsilon_2)x_2 - (1 + \varepsilon_1)x_1}{t_2 - t_1}$$

L'errore relativo commesso approssimando $\bar{v}$ con w è:

$$\frac{w - \bar{v}}{\bar{v}} = \frac{x_2}{x_2 - x_1}\,\varepsilon_2 + \frac{x_1}{x_2 - x_1}\,\varepsilon_1$$

Nel caso in cui la differenza $x_2 - x_1$ sia piccola (ad esempio quando $\bar{v}$ sia utilizzato come stima della velocità istantanea di un punto mobile con velocità elevata), per quanto mostrato nell'Osservazione precedente, il calcolo risulta *mal condizionato* e l'errore commesso approssimando $\bar{v}$ con w risulterà molto maggiore dei singoli errori ε_1 e ε_2.

(1.49) Esercizio.

La scrittura:

$$\text{(A)} \qquad x = a + \delta \quad \text{con} \quad |\delta| \leqslant d$$

è *equivalente* alla scrittura:

$$\text{(B)} \qquad x \in [a - d\ ,\ a + d]$$

Si vogliono determinare y e E in modo che anche la scrittura:

$$\text{(*)} \qquad x = (1 + \varepsilon)y \quad \text{con} \quad |\varepsilon| \leqslant E$$

risulti equivalente ad (A) e (B).

La scrittura (*) equivale a:

$$x \in [(1 - E)y\ ,\ (1 + E)y]$$

Quest'ultima scrittura è equivalente alla (B) se e solo se:

$$(1 - E)y = a - d \quad \text{e} \quad (1 + E)y = a + d$$

Risolvendo il sistema si determina:

$$y = a \quad e \quad E = d / a$$

Quindi le scritture (A) e (B) sono equivalenti alla scrittura:

$$(C) \qquad x = (1 + \varepsilon)a \quad \text{con} \quad |\varepsilon| \leqslant d / a$$

(1.50) Teorema (stabilità della procedura *bisezione*).

Si consideri la realizzazione in *Scilab*[1], della procedura *bisezione*.

Se l'assegnamento

```
[z,v,info] = bisezione(f,a,b,delta)
```

termina con info = 0 oppure info = 1, allora:

$$|z - \alpha^*| \leqslant \text{delta}$$

dove α^* è uno zero di una funzione g 'vicina' alla funzione f nel senso che:

per ogni x in [a,b] si ha |f(x) - g(x)| 'piccolo'

Informalmente: se info = 0 oppure info = 1 allora la procedura restituisce una *buona approssimazione* di uno zero di una funzione *vicina* a quella in esame.

(Dimostrazione omessa.)

(1.51) Osservazione (condizionamento degli zeri di una funzione regolare).

Siano f:[a,b] → R regolare (derivabile con f' continua) con $f' \neq 0$ e f(a)f(b) < 0, α l'unico zero di f in [a,b], g:[a,b] → R continua e 'vicina' ad f, precisamente tale che:

per ogni x in [a,b] si ha $|f(x) - g(x)| \leqslant d$ con d 'piccolo' e $d < \min\{|f(a)|, |f(b)|\}$

Per le ipotesi fatte, g ha *almeno uno* zero in [a,b]. Si vuole sapere *quanto distante può essere lo zero* α *di* f *da uno zero di* g.

Sia α^* uno zero di g in [a,b]. Allora si ha (utilizzando il Teorema di Lagrange):

$$f(\alpha^*) = f(\alpha^*) - f(\alpha) = f'(t)(\alpha^* - \alpha) \quad \text{con} \quad t \text{ tra } \alpha^* \text{ e } \alpha$$

Dunque, posto $m = \min\{ |f'(x)|, x \text{ in } [a,b] \}$, si ha:

$$|\alpha^* - \alpha| = \frac{|f(\alpha^*)|}{|f'(t)|} \leqslant \frac{|f(\alpha^*)|}{m}$$

Infine, essendo:

---

1 Asserto (1.08) nella Lezione 2.

$$|f(\alpha^*)| = |f(\alpha^*) - g(\alpha^*)| \leqslant d$$

si ottiene:

$$|\alpha^* - \alpha| \leqslant \frac{d}{m}$$

La quantità $1/m$ ha il ruolo di *numero di condizionamento*: tanto più è grande tanto più gli zeri di g possono essere lontani dallo zero di f.

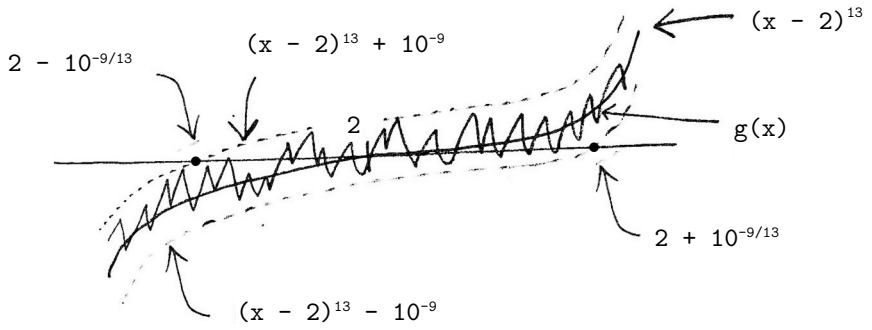
Se f'(x) = 0 per qualche x in [a,b], in particolare <u>se</u> f'(α) = 0, il condizionamento è certamente *cattivo*, come evidenziato nell'esempio seguente.

(1.52) <u>Esempio</u>.

Sia $f(x) = (x - 2)^{13}$. La funzione ha un solo zero, $\alpha = 2$, ed è regolare nell'intervallo [1,3]. Si consideri poi g:[1,3] → R continua tale che:

$$\text{per ogni } x \text{ in } [1,3] \text{ si ha } |f(x) - g(x)| \leqslant 10^{-9}$$

Un esempio di grafico di g è rappresentato in figura.



Nel caso *peggiore* la distanza tra lo zero α di f e uno zero α^* di g è $10^{-9/13} \approx 0.2$, *molto più grande* della distanza 10^{-9} tra f e g.

(1.1) METODI AD UN PUNTO

Il punto di forza del metodo di bisezione è la sua generalità: può essere applicato a *qualunque* funzione che sia semplicemente continua e che assuma valori di segno opposto agli estremi di un intervallo. Per contro, in alcune applicazioni il metodo richiede un *numero eccessivo di iterazioni* per ottenere l'accuratezza richiesta dall'utilizzatore. Per ovviare a questo inconveniente, analizziamo altri metodi per approssimare lo zero di una funzione: i *metodi ad un punto*.

(1.53) <u>Definizione</u> (metodo ad un punto).

Sia h:[a,b] → R una funzione continua. Il *metodo ad un punto definito da* h è la seguente

procedura:

z = *MetodoUnPunto*(h,a,b,γ)

ingresso: h:[a,b] → R continua, γ in [a,b]

- x(0) = γ;
- per k = 1,2,3,... ripeti
  se x(k-1) in [a,b] allora x(k) = h(x(k-1)) altrimenti STOP

uscita: quando un opportuno *criterio d'arresto* è verificato: z = x(k).

(1.54) Osservazione.

Se omettiamo il criterio d'arresto e per ogni k si x(k-1) in [a,b], il metodo ad un punto definisce una successione x(0), x(1), x(2),... Se la successione è *convergente*, il limite è un *punto unito di* h.[2]

(Dimostrazione. La successione x(0), x(1), x(2),... è identica alla successione h(x(0)), h(x(1)), h(x(2)),... Quindi quest'ultima è convergente e, detto α il limite della successione x(k):

$$\lim_{k\to\infty} h(x(k)) = \alpha$$

Poiché h è una funzione continua e la successione x(k) converge ad α, si ha:

$$\lim_{k\to\infty} h(x(k)) = h(\lim_{k\to\infty} x(k)) = h(\alpha)$$

Per l'unicità del limite di una successione convergente, si deduce che α = h(α).)

(1.55) Osservazione.

Sia f la funzione continua della quale si è interessati ad approssimare qualche zero. Per quanto detto nell'Osservazione precedente, il metodo ad un punto definito da h è utilizzabile, 'se tutto va bene', per approssimare *un punto unito* di h. Perché il metodo ad un punto possa essere utilizzato per approssimare qualche zero di f occorre *scegliere* la funzione h che lo definisce in modo che:

(#) {zeri di f} = {punti uniti di h}

Ci si domanda *se esistono* funzioni (continue) h con la proprietà richiesta.

Si consideri la funzione h così definita:

$$h(x) = f(x) + x$$

Se α è zero di f, ovvero f(α) = 0, si ha:

$$h(\alpha) = f(\alpha) + \alpha = \alpha \quad \Rightarrow \quad \alpha \text{ è punto unito di h}$$

---

2 Il numero reale α è un *punto unito di* h significa che α = h(α).

Viceversa, se α è punto unito di h (ovvero α = h(α)), si ha:

$$h(\alpha) = f(\alpha) + \alpha \quad \Rightarrow \quad f(\alpha) = 0 \quad \Rightarrow \quad \alpha \text{ zero di } f$$

La funzione h è quindi *una* funzione che verifica la proprietà (#).

Si verifica facilmente che, se g è una funzione continua tale che g(x) $\neq$ 0 per ogni x, la funzione h definita da:

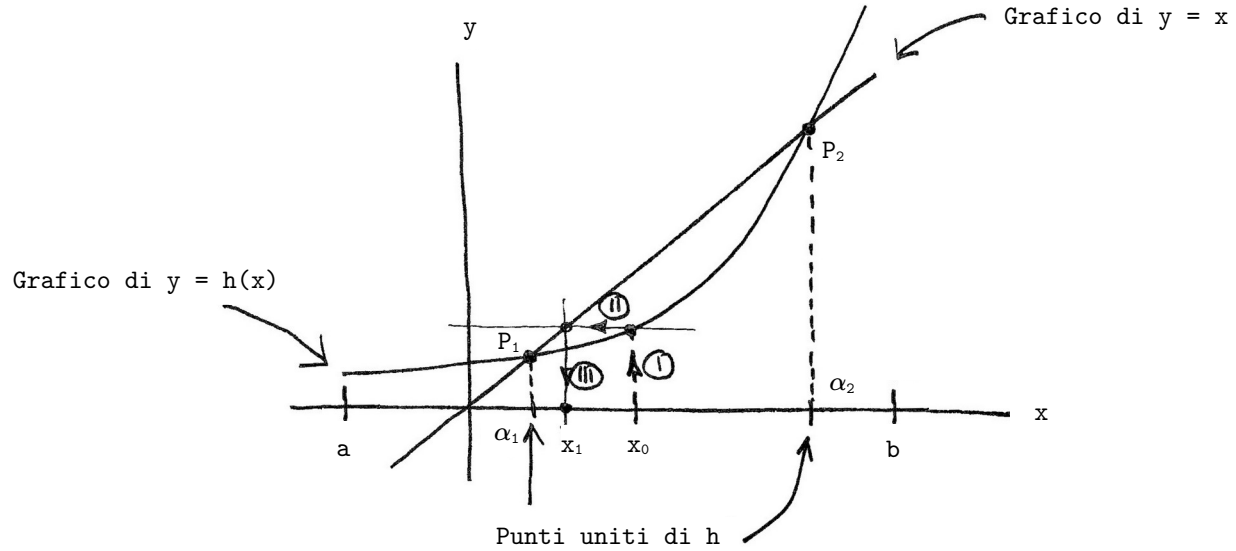
$$h(x) = g(x)f(x) + x$$

è continua ed ha la proprietà (#). Dunque esistono *infinite* funzioni h che hanno come punti uniti tutti e soli gli zeri di f.

Si pone adesso il problema di scegliere, tra tutte le possibili funzioni che hanno la proprietà (#), una h in modo che il metodo da essa definito generi una successione *convergente*.

Lezione 9 (ore 15,16) - 9 ottobre 2025, 8:30 - 10:30 F3

(1.56) Osservazione (costruzioni grafiche).



Si rappresentino su uno stesso piano cartesiano le porzioni del grafico della funzione y = h(x) che definisce il metodo ad un punto da esaminare e della retta grafico della funzione y = x, su un intervallo [a,b].

I punti uniti di h sono le ascisse (α_1 e α_2) dei punti P_1 e P_2 comuni ai due grafici.

Assegnato il punto dell'asse delle ascisse che rappresenta x_0, possiamo costruire il punto dello stesso asse che rappresenta x_1 in tre passaggi: (I) si determina il punto $(x_0,h(x_0)) = (x_0,x_1)$ intersezione tra il grafico di y = h(x) e la retta verticale per $(x_0,0)$; (II) si determina il punto $(h(x_0),h(x_0)) = (x_1,x_1)$ intersezione tra il grafico di y = x e la retta orizzontale per il punto $(x_0,h(x_0))$ determinato al passaggio precedente; (III) si determina il punto $(h(x_0),0) = (x_1,0)$ intersezione tra l'asse delle ascisse e la retta verticale passante per (x_1,x_1).

(1.57) Teorema (di convergenza).

Siano $h:[a,b] \to R$ una funzione con derivata prima continua e γ un punto di [a,b] tali che:

(1) esiste un punto unito α di h in [a,b];
(2) esiste un numero reale $L \in [0,1)$ tale che: per ogni $x \in [a,b]$ si ha $|h'(x)| \leqslant L$;
(3) la procedura MetodoUnPunto(h,a,b,γ) definisce una successione x_k.[1]

Allora si ha:

(A) α è l'*unico* punto unito di h in [a,b];
(B) la successione x_k è *convergente* al limite α.

(1.58) Dimostrazione (del Teorema (1.57)).

1 Ovvero, per ogni k si ha: se $x_k \in [a,b]$ allora $x_{k+1} \in [a,b]$.

(A) Per assurdo. Se β è un altro punto unito di h in [a,b] si ha (utilizzando prima la definizione di punto unito e poi il Teorema di Lagrange):

$$\beta - \alpha = h(\beta) - h(\alpha) = h'(t)(\beta - \alpha) \quad , \quad \text{con t numero reale compreso tra } \alpha \text{ e } \beta$$

Infine, ricordando che $\beta - \alpha \neq 0$, si ottiene:

$$(\#) \qquad h'(t) = 1$$

Ma, siccome α e β sono punti in [a,b], anche t lo è. Allora, per l'ipotesi (2), l'uguaglianza (#) è assurda.

Si osservi che per questa dimostrazione si sono utilizzate *solo* le ipotesi (1) e (2).

(B) Si deve dimostrare che la successione x_k tende ad α, ovvero che la successione $x_k - \alpha$ tende a zero. Si ha, utilizzando il Teorema di Lagrange per la seconda uguaglianza:

$$x_k - \alpha = h(x_{k-1}) - h(\alpha) = h'(t_{k-1})(x_{k-1} - \alpha) \quad \text{con} \quad t_{k-1} \text{ tra } x_{k-1} \text{ e } \alpha$$

Passando ai valori assoluti si ha (la disuguaglianza si ottiene utilizzando l'ipotesi (2)):

$$|x_k - \alpha| = |h'(t_{k-1})|\ |x_{k-1} - \alpha| \leqslant L\ |x_{k-1} - \alpha|$$

Se $k - 1 > 0$ si può ripetere il ragionamento a partire da $x_{k-1} - \alpha$ per ottenere:

$$|x_{k-1} - \alpha| = |h'(t_{k-2})|\ |x_{k-2} - \alpha| \leqslant L\ |x_{k-2} - \alpha|$$

e, sostituendo nella precedente:

$$|x_k - \alpha| \leqslant L^2\ |x_{k-2} - \alpha|$$

Iterando all'indietro fino al primo elemento della successione si ricava:

$$|x_k - \alpha| \leqslant L^k\ |x_0 - \alpha|$$

Ricordando che $0 \leqslant L < 1$ si ottiene il risultato cercato:

$$\lim_{k \to \infty} |x_k - \alpha| = 0$$

Lezione 10 (ore 17,18) - 14 ottobre 2025, 16:30 - 18:30 F3

(1.58) Osservazione.

L'uso del Teorema di convergenza (Teorema (1.57) della Lezione 9) richiede la verifica delle ipotesi (1) - (3). Per le ipotesi (1) e (2) occorre decidere se esiste, ed eventualmente determinare, un intervallo [a,b] che contiene un solo punto unito di h e in tutti i punti x del quale $|h'(x)| \leqslant L$ con $0 \leqslant L < 1$. Una volta determinato un intervallo [a,b] con le proprietà richieste, occorre decidere se sia verificata l'ipotesi (3), ovvero se a partire da γ il metodo definito da h genera una successione in [a,b].

Il teorema e l'osservazione seguenti forniscono criteri concreti riguardo la verifica delle ipotesi.

(1.59) Teorema (utilizzabilità del metodo definito da h).

Sia $h:[a,b] \to \mathbb{R}$ una funzione con derivata prima continua e α un punto unito di h in [a,b]. *Condizione necessaria e sufficiente* affinché esista un intervallo $I \subset [a,b]$ contenente α e in tutti i punti x del quale si abbia $|h'(x)| \leqslant L$ con $0 \leqslant L < 1$ è:

$$|h'(\alpha)| < 1$$

Dimostrazione.

La condizione è *necessaria*: se esiste un intervallo $I \subset [a,b]$ contenente α in tutti i punti x del quale $|h'(x)| \leqslant L$ con $0 \leqslant L < 1$, certamente si ha $|h'(\alpha)| < 1$.

La condizione è *sufficiente*: se $|h'(\alpha)| < 1$, per la continuità della funzione h' esistono un numero reale L con $0 \leqslant L < 1$ e un intervallo $I \subset [a,b]$ tali che $\alpha \in I$ e in tutti i punti $x \in I$ si ha $|h'(x)| \leqslant L$.

(1.60) Osservazione (criterio di scelta del punto iniziale).

Sia $h:[a,b] \to \mathbb{R}$ una funzione con derivata prima continua che verifica le ipotesi (1) e (2) del Teorema di convergenza e sia α l'unico punto unito di h in [a,b]. Allora:

a partire da γ = *l'estremo di* [a,b] *più vicino ad* α, il metodo definito da h genera una successione in [a,b] - dunque convergente ad α.

Dimostrazione.

Posto $x_0 = \gamma$, sia $d = |x_0 - \alpha|$. Indicato con $I(\alpha, d)$ l'intorno di centro α e raggio d, si ha $I(\alpha, d) \subset [a,b]$. Per quanto mostrato nel punto (B) della dimostrazione del Teorema di convergenza, si ha $|x_1 - \alpha| < |x_0 - \alpha| = d$, quindi $x_1 \in I(\alpha, d)$. Allo stesso modo si dimostra che per ogni k si ha $x_k \in I(\alpha, d) \subset [a,b]$.

(1.61) Osservazione.

Siano $h:[a,b] \to \mathbb{R}$ una funzione con derivata prima continua, α un punto unito di h e x_k una successione generata dal metodo definito da h. Se $|h'(\alpha)| > 1$ allora uno soltanto dei seguenti asserti sussiste:

- esiste $\bar{k}$ tale che per ogni $k \geqslant \bar{k}$ si ha $x_k = \alpha$
- $x_k \nrightarrow \alpha$

(Dimostrazione solo in un caso particolare. Sia $h(x) = A(x - \alpha) + \alpha$ con $A > 1$. Si ha: α è l'unico punto unito di h, $h'(x) = A$ e

$$x_k - \alpha = A^k(x_0 - \alpha)$$

Allora: se $x_0 \neq \alpha$, per ogni $M > 0$ esiste n tale che $k \geqslant n \Rightarrow |x_k - \alpha| \geqslant M$. Dunque per ogni $x_0 \neq \alpha$ si ha $x_k \nrightarrow \alpha$.)

L'eventualità di riuscire a determinare *concretamente* un punto iniziale a partire dal quale risulti $x_k = \alpha$ dopo un numero *finito* di termini è estremamente remota. Per questo motivo, se $|h'(\alpha)| > 1$ il metodo definito da h si dichiara *non utilizzabile* per approssimare α.

Resta da chiarire cosa accade se $|h'(\alpha)| = 1$. Vedremo che anche in questo caso il metodo definito da h si dichiara *non utilizzabile* per approssimare α.

Si osservi, infine, che la condizione $|h'(\alpha)| < 1$, necessaria e sufficiente per l'utilizzabilità del metodo per approssimare il punto unito α, è verificabile *graficamente* confrontando la pendenza $(h'(\alpha))$ della retta tangente al grafico di $y = h(x)$ in $x = \alpha$ con quella (1) della retta grafico di $y = x$ e con quella (-1) della retta $y = \alpha - x$.

(1.62) Esercizio.

Per ogni $x > 0$, sia $f(x) = x + \log(x)$. Si vuole (i) sapere se f ha qualche zero e, in caso affermativo: (ii) separare gli zeri e, infine, (iii) decidere se ciascuno dei metodi definiti da

$$h_1(x) = -\log(x) \quad ; \quad h_2(x) = \exp(-x) \quad ; \quad h_3(x) = (\exp(-x) + x)/2$$

sia utilizzabile per approssimare gli zeri di f.

Soluzione.

(i) La funzione f(x) è continua, $f(x) \to -\infty$ quando $x \to 0$ e $f(x) \to +\infty$ quando $x \to +\infty$. Se ne deduce che f ha *almeno uno* zero. La funzione f(x) è anche derivabile e per ogni $x > 0$ risulta $f'(x) \neq 0$. Allora f ha *al più uno* zero. Dunque f ha *uno* zero, α.[1]

(ii) Si ha: $f(1) = 1$, dunque $\alpha \in [0,1]$, ovvero l'intervallo [0,1] *separa* lo zero di f.

(iii) Si consideri la funzione $h_1(x)$. Si verifica facilmente che gli zeri di f sono tutti e soli i punti uniti di h_1. Inoltre, h_1 è derivabile e per ogni $x > 0$ si ha $h_1'(x) = 1/x$. Essendo $\alpha \in (0,1)$ si ha certamente $|h_1'(\alpha)| > 1$. Per l'Osservazione (1.61) il metodo definito da h_1 *non è utilizzabile* per approssimare α.

Si consideri la funzione $h_2(x)$. Si verifica facilmente che gli zeri di f sono tutti e soli i

---

1 Sia $f:[a,b] \to R$ una funzione sufficientemente regolare. Se per ogni x in [a,b] si ha

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

allora f ha *al più* k *zeri distinti* nell'intervallo [a,b].

punti uniti di h_2. Inoltre, h_2 è derivabile e per ogni x si ha $|h_2'(x)| = \exp(-x)$. Essendo $\alpha \in (0,1)$ si ha certamente $|h_2'(\alpha)| < 1$ e, per il Teorema (1.59), il metodo definito da h_2 *è utilizzabile* per approssimare α. In base all'Osservazione (1.60), per determinare un punto iniziale a partire dal quale il metodo definisce una successione convergente ad α è sufficiente determinare un intervallo chiuso I che verifica le ipotesi (1) e (2) del Teorema di convergenza. L'intervallo [0,1] *non va bene* perché l'ipotesi (2) non è verificata: per ogni x in (0,1] si ha $0 \leqslant |h_2'(x)| = \exp(-x) < 1$ <u>ma</u> $|h_2'(0)| = 1$. Allora, un intervallo che verifica anche l'ipotesi (2) è [t,1] con $t \in (0,\alpha)$. Per determinare t si utilizza il Teorema di esistenza degli zeri. Siccome $f(1/2) < 0$, si pone $t = 1/2$ e $I = [1/2, 1]$. A questo punto è sufficiente decidere quale dei due estremi di I è più vicino allo zero. Si utilizza ancora il Teorema di esistenza degli zeri. Siccome $f(3/4) > 0$, si sceglie $x_0 = 1/2$.

Si osservi che, in questo caso, per ogni x in $I = [1/2, 1]$ la derivata prima della funzione che definisce il metodo è *negativa*. Poiché, si riveda la dimostrazione dell'asserto (B) del Teorema di convergenza, per ogni k si ha:

$$x_k - \alpha = h'(t_{k-1})(x_{k-1} - \alpha)$$

per qualche numero reale t_{k-1} in I, allora per ogni k è $h'(t_{k-1}) < 0$ e le differenze $x_k - \alpha$ e $x_{k-1} - \alpha$ hanno *segno opposto*. Ne segue che gli elementi della successione si trovano, alternativamente, a destra e a sinistra di α: la successione 'oscilla' intorno allo zero. La successione delle *distanze* $|x_k - \alpha|$ è comunque *monotona decrescente* come mostrato nella dimostrazione del Teorema di convergenza.

Si consideri infine la funzione $h_3(x)$. Si verifica facilmente che gli zeri di f sono tutti e soli i punti uniti di h_3. Inoltre, h_3 è derivabile e per ogni x si ha:

$$|h_3'(x)| = (1 - \exp(-x))/2$$

Essendo $\alpha \in (1/2,1)$ si ha certamente $|h_3'(\alpha)| < 1$ e, per il Teorema (1.59), il metodo definito da h_3 *è utilizzabile* per approssimare α. In base all'Osservazione (1.60), per determinare un punto iniziale a partire dal quale il metodo definisce una successione convergente ad α è sufficiente determinare un intervallo chiuso I che verifica le ipotesi (1) e (2) del Teorema di convergenza. L'intervallo $I = [1/2,1]$ va bene, infatti per ogni x in I si ha $0 \leqslant |h_3'(x)| < 1$. A questo punto è sufficiente decidere quale dei due estremi di I è più vicino allo zero. Procedendo come nel caso precedente, si sceglie $x_0 = 1/2$.

Si osservi che, in questo caso, per ogni x in $I = [1/2, 1]$ la derivata prima della funzione che definisce il metodo è *positiva*. Ragionando come nel caso precedente, le differenze $x_k - \alpha$ e $x_{k-1} - \alpha$ hanno *lo stesso segno*. Ne segue che gli elementi della successione si trovano tutti dalla stessa parte rispetto ad α. Inoltre, anche in questo caso, la successione delle *distanze* $|x_k - \alpha|$ è *monotona decrescente*, e quindi la successione x_k risulta *monotona* (*crescente* se x_0 è a sinistra di α, *decrescente* nel caso opposto). Infine, si osservi che poiché per ogni x in $I = [1/2, 1]$ la derivata prima della funzione che definisce il metodo è positiva, dalla dimostrazione del criterio di scelta del punto iniziale (Osservazione (1.60)) si deduce che *per ogni* x_0 *in* I la successione x_k converge ad α.

(1.63) <u>Esercizio</u> (per casa).

Per ogni $x \in R$ sia: $h(x) = 2 \operatorname{arctg}(x)$.

(1) Determinare il numero di punti uniti di h e separarli.
(2) Per ciascuno dei punti uniti, decidere se il metodo iterativo definito da h sia utilizzabile per l'approssimazione e, in caso affermativo, indicare un punto iniziale a partire dal quale la successione generata converge al punto unito in esame.
(3) Rispondere alle domande precedenti utilizzando i metodi grafici, aiutandosi con *Scilab*.

Lezione 11 (ore 19,20) - 15 ottobre 2025, 11:30 - 13:30 A13

(1.2) METODO DI NEWTON

(1.64) Definizione (metodo di Newton).

Sia $f:[a,b] \to R$ una funzione con derivata prima tale che $f'(x) \neq 0$ per ogni x in $[a,b]$.

Il *metodo di Newton* applicato alla funzione f è il metodo ad un punto definito dalla funzione $h_N:[a,b] \to R$ tale che:

$$h_N(x) = x - (f'(x))^{-1} f(x) = x - \frac{f(x)}{f'(x)}$$

Si osservi che i *punti uniti di* h_N sono tutti e soli gli *zeri di* f.

(1.65) Osservazione (utilizzabilità del metodo di Newton).

Sia $f:[a,b] \to R$ una funzione con derivata *seconda* continua e con $f'(x) \neq 0$ per ogni x in $[a,b]$. Sia poi α uno zero di f in $[a,b]$. Si ha:

$$h_N'(x) = 1 - \frac{(f'(x))^2 - f''(x)f(x)}{(f'(x))^2} = \frac{f''(x)f(x)}{(f'(x))^2}$$

La funzione h_N' è continua e, essendo $f(\alpha) = 0$ e $f'(\alpha) \neq 0$, si ha

$$h_N'(\alpha) = 0$$

Per il Teorema (1.59) della Lezione 10, il metodo di Newton è *utilizzabile* per approssimare α.

(1.66) Osservazione (criterio di utilizzabilità per il metodo di Newton).

Siano $f:[a,b] \to R$ una funzione con derivata seconda continua e α uno zero di f in $[a,b]$. Condizione *sufficiente* perché il metodo di Newton applicato ad f sia *utilizzabile* per approssimare α è:

$$f'(\alpha) \neq 0$$

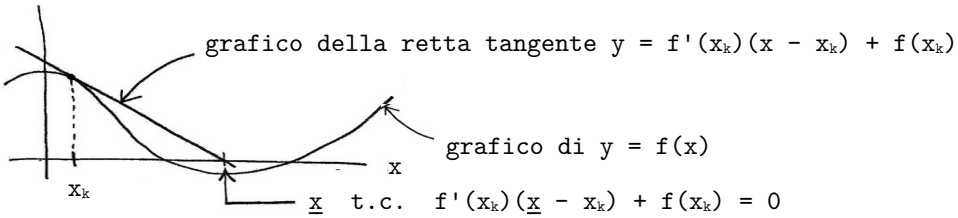
(1.67) Osservazione (interpretazione grafica del metodo di Newton).

Sia $f:[a,b] \to R$ una funzione con derivata prima e sia x_k un numero reale tale che $f'(x_k) \neq 0$. Si disegnino su uno stesso piano cartesiano il grafico della funzione f e quello della retta tangente al grafico di f in x_k (vedi figura). Poiché $f'(x_k) \neq 0$, la retta tangente non è orizzontale e quindi interseca l'asse delle ascisse nel punto $\underline{x}$ tale che:

$$f'(x_k)(\underline{x} - x_k) + f(x_k) = 0$$

ovvero in

$$\underline{x} = x_k - \frac{f(x_k)}{f'(x_k)} = h_N(x_k)$$



(1.68) Osservazione (criterio di scelta del punto iniziale per il metodo di Newton).

Sia f:[a,b] → R con derivata seconda continua tale che:

(1) esiste α zero di f in [a,b]

(2) per ogni x ∈ [a,b] si ha f'(x) ≠ 0 (e quindi α è *l'unico* zero di f in [a,b])

(3) f''(x) ≠ 0 (f è *convessa* in [a,b])

Allora: a partire da γ = *l'estremo di* [a,b] *in cui* f *e* f'' *hanno lo stesso segno*, il metodo di Newton genera una successione in [a,b] *convergente ad* α *e monotona*.

(Dimostrazione. Utilizzando le ipotesi, e ragionando graficamente, si mostra che la successione generata a partire da γ è *monotona e limitata*, e quindi convergente. Il limite non può che essere un punto unito di h_N in [a,b], dunque α.)

(1.69) Osservazione.

Siano f:[a,b] → R una funzione con derivata seconda continua e α uno zero di f in [a,b]. Se f'(α) ≠ 0 (dunque il metodo di Newton applicato ad f è utilizzabile per approssimare α) allora esiste un intervallo I che verifica le ipotesi del criterio di scelta (1.67) se e solo se f''(α) ≠ 0.

(1.70) Osservazione (ordine di convergenza di un metodo ad un punto).

Siano h:[a,b] → R, α un punto unito di h e x_k una successione convergente ad α generata dal metodo definito da h.

(1) Sia h con h' continua e 0 < |h'(α)| < 1. Allora:

- Sia d > 0 tale che h'(x) ≠ 0 per ogni x ∈ I(α,d). Detti λ_d e L_d, rispettivamente, il minimo ed il massimo di |h'(x)| su I(α,d) e $y_{n,d}$ la successione costituita dagli elementi di x_k in I(α,d), per ogni x in I(α,d) si ha:

$$\lambda_d \leqslant |h'(x)| \leqslant L_d$$

- Per ogni n si ha allora:

$$\lambda_d^n \, |y_{0,d} - \alpha| \leqslant |y_{n,d} - \alpha| \leqslant L_d^n \, |y_{0,d} - \alpha|$$

ovvero:

la successione $y_{n,d} - \alpha$ converge a zero *più rapidamente* della successione

$L_d^n\ |y_{0,d} - \alpha|$ ma *meno rapidamente* della successione $\lambda_d^n\ |y_{0,d} - \alpha|$

Scelto d molto piccolo si avrà $\lambda_d \approx L_d \approx |h'(\alpha)|$. Dunque

$$|y_{n,d} - \alpha| \approx |h'(\alpha)|^n\ |y_{0,d} - \alpha|$$

Questa proprietà della successione x_k si esprime dicendo che 'x_k *converge ad* α *in modo esponenziale*'.

(2) Sia $h(x) = \alpha + A(x - \alpha)^2$ con $A \neq 0$. Allora: α è punto unito di h e $h'(\alpha) = 0$. Inoltre, dato un numero reale x_0, per ogni k si ha:

$$x_k - \alpha = A^{-1}\ (\ A(x_0 - \alpha)\ )^{2^k}$$

Se $|A(x_0 - \alpha)| < 1$, la successione x_k converge ad α e, per ogni t in (0,1) si ha

$$\frac{|x_k - \alpha|}{t^k} \longrightarrow 0 \quad \text{per } k \to \infty$$

ovvero: la successione $x_k - \alpha$ tende a zero *più rapidamente* di *qualsiasi* successione esponenziale.

In generale, se h ha derivata seconda continua e $h'(\alpha) = 0$, la successione x_k tende ad α *più rapidamente* di *qualsiasi* successione di tipo esponenziale.

Il sussistere della condizione 'h con h' continua e $0 < |h'(\alpha)| < 1$' si esprime con la frase *l'ordine di convergenza ad* α *del metodo definito da* h *è* <u>*uno*</u>. Il sussistere della condizione 'h con h'' continua, $h'(\alpha) = 0$ e $h^{(2)}(\alpha) \neq 0$' si esprime con la frase *l'ordine di convergenza ad* α *del metodo definito da* h *è* <u>*due*</u>. In generale:

l'ordine di convergenza ad α *del metodo definito da* h *è* p

significa

h ha derivata di ordine p continua, $h^{(m)}(\alpha) = 0$ per $m = 1,\ldots,p - 1$ e $h^{(p)}(\alpha) \neq 0$

Tanto più elevato è l'ordine di convergenza ad α del metodo, tanto più rapidamente convergono ad α le successioni generate dal metodo.

Lezione 12 (ore 21,22) - 16 ottobre 2025, 8:30 - 10:30 F3

(1.71) Esercizio.

Siano t un numero reale positivo, n un numero intero $\geqslant 2$ e $f(x) = x^n - t$. La funzione f ha un solo zero, la radice n-esima di t: $t^{1/n}$.

Decidere se il metodo di Newton sia applicabile per approssimare lo zero e, in caso affermativo, determinare x_0 in modo che il metodo generi una successione convergente allo zero.

(1.72) Osservazione (criteri d'arresto).

I criteri d'arresto presentati per il metodo di bisezione *non* sono utilizzabili per i metodi ad un punto: questi ultimi metodi, contrariamente al metodo di bisezione, non generano una successione di intervalli di misura tendente a zero e ciascuno contenente uno zero della funzione. Occorrono dunque criteri diversi. Discutiamo i due più utilizzati, entrambi di tipo *assoluto*.

Siano f la funzione della quale si vuole approssimare uno zero, $h:[a,b] \to R$ e γ che verificano le ipotesi del Teorema di convergenza, α il punto unito di h (e zero di f) in [a,b] e x_k la successione generata dal metodo definito da h a partire da γ. La successione x_k converge ad α.

(1) Dato un numero reale positivo E (l'errore massimo richiesto dall'utilizzatore) e inserito E tra le variabili di ingresso della procedura:

$$\underline{\text{se}}\ |x_{k+1} - x_k| < E\ \underline{\text{allora}}\ \text{STOP}$$

Il criterio è *calcolabile*: a ciascuna iterazione la procedura conosce x_k, determina $x_{k+1} = h(x_k)$ e verifica la condizione del criterio.

Il criterio è *efficace*: sia la successione x_k che la successione $x_{k+1} = h(x_k)$ convergono ad α (la funzione h è continua e α è punto unito di h), quindi la differenza tende a zero. La condizione del criterio è certamente soddisfatta dopo un numero finito di iterazioni.

Per capire quanto buona sia x_k come approssimazione di α quando la condizione è verificata, si osservi che:

$$|x_{k+1} - x_k| = |h(x_k) - x_k| = |(h(x_k) - \alpha) + (\alpha - x_k)| = |(h(x_k) - h(\alpha)) + (\alpha - x_k)|$$

Utilizzando il Teorema di Lagrange:

$$h(x_k) - h(\alpha) = h'(t_k)(x_k - \alpha) \quad \text{con} \quad t \text{ tra } x_k \text{ e } \alpha$$

dunque:

$$|x_{k+1} - x_k| = |h'(t_k)(x_k - \alpha) + (\alpha - x_k)| = |h'(t_k) - 1|\ |x_k - \alpha| = |1 - h'(t_k)|\ |x_k - \alpha|$$

L'accuratezza di x_k come approssimazione di α *dipende* dal valore di $h'(t_k)$. Precisamente:

- se $h'(t_k) \approx 0$ si ha $|x_{k+1} - x_k| \approx |x_k - \alpha|$ e il criterio d'arresto interrompe la costruzione della successione *non appena* l'approssimazione è accurata (si osservi

che *se* f *è sufficientemente regolare e* f'(α) $\neq$ 0 *il Metodo di Newton rientra in questo caso*);

- se h'(t_k) $\approx$ 1 si ha 1 - h'(t_k) $\approx$ 0 e il criterio d'arresto interrompe la costruzione della successione *prima* che l'approssimazione sia accurata;
- se h'(t_k) < 0 si ha |1 - h'(t_k)| > 1 e quindi $|x_{k+1} - x_k| < E \Rightarrow |x_k - \alpha| < E$ (ma il criterio d'arresto potrebbe interrompere la costruzione della successione *in ritardo*: l'approssimazione potrebbe essere buona già da qualche iterazione).

Esempio: Sia h(x) = α + A(x - α) e h'(x) = A. Per ogni k si ha: $x_k - \alpha = A^k(x_0 - \alpha)$.

Se A = 0.9 ($\approx$ 1) *e* k è tale che $|x_{k+1} - x_k|$ = 0.99 E (criterio d'arresto verificato), *allora* $|x_k - \alpha|$ = (0.99 / 0.1) E = 9.9 E > E e l'accuratezza dell'approssimazione non verifica la richiesta dell'utilizzatore.

Se A = -0.9 *e* k tale che $|x_{k+1} - x_k|$ = 0.99 E (criterio d'arresto verificato), *allora* $|x_k - \alpha|$ = E / 1.9 $\approx$ 0.5 E < E e l'accuratezza dell'approssimazione verifica la richiesta dell'utilizzatore. *Però*: 6 iterazioni prima si aveva già $|x_{k-6} - \alpha| = |x_k - \alpha| / |A|^6$ = E / (1.9 0.9^6) = E / 1.009... < E, ovvero *già 6 iterazioni prima* l'accuratezza dell'approssimazione verificava la richiesta dell'utilizzatore.

(2) Dato un numero reale positivo E (l'errore massimo richiesto dall'utilizzatore) ed inserite tra le variabili di ingresso della procedura sia E che f:

se $|f(x_k)| < E$ allora STOP

Il criterio è *calcolabile*: a ciascuna iterazione la procedura conosce x_k, determina $f(x_k)$ e verifica la condizione del criterio.

Il criterio è *efficace*: la successione x_k converge ad α e la successione $f(x_k)$ converge a $f(\alpha) = 0$ (la funzione f è continua e α è zero di f). La condizione del criterio è quindi certamente soddisfatta dopo un numero finito di iterazioni.

Per capire quanto buona sia x_k come approssimazione di α quando la condizione è verificata, si supponga f regolare e si osservi che:

$$f(x_k) = f(x_k) - f(\alpha)$$

Utilizzando il Teorema di Lagrange:

$$f(x_k) - f(\alpha) = f'(t_k)(x_k - \alpha) \quad \text{con} \quad t_k \text{ tra } x_k \text{ e } \alpha$$

dunque:

$$|f(x_k)| = |f'(t_k)|\ |x_k - \alpha|$$

L'accuratezza di x_k come approssimazione di α *dipende* dal valore di $|f'(t_k)|$. Precisamente:

- se $|f'(t_k)| \approx 1$ si ha $|f(x_k)| \approx |x_k - \alpha|$ e il criterio d'arresto interrompe la costruzione della successione *non appena* l'approssimazione è accurata;
- se $|f'(t_k)| \approx 0$ il criterio d'arresto interrompe la costruzione della successione *prima* che l'approssimazione sia accurata;
- se $|f'(t_k)| > 1$ si ha $|f(x_k)| < E \Rightarrow |x_k - \alpha| < E / |f'(t_k)| < E$ (ma il criterio d'arresto potrebbe interrompere la costruzione della successione *in ritardo*: l'approssimazione potrebbe essere buona già da qualche iterazione).

Lezione 13 (ore 23,24) - 21 ottobre 2025, 16:30 - 18:30 F3

(1.73) Osservazione (criteri d'arresto, continuazione).

Entrambi i criteri d'arresto considerati nell'Osservazione (1.72) della Lezione 12 presentano il problema che, in alcuni casi, x_k è un'approssimazione di α non sufficientemente buona. Questo nasce dal fatto che, nel criterio d'arresto, stimando l'errore assoluto commesso approssimando α con l'ultimo elemento della successione calcolato ($|x_k - \alpha|$) utilizzando la quantità scelta ($|x_{k+1} - x_k|$ in un caso, $|f(x_k)|$ nell'altro), si commette un errore relativo che *non tende a zero* quando $k \to \infty$.

I due criteri si possono modificare in modo da ottenere stime migliori. Ponendosi nel medesimo contesto utilizzato per i due criteri precedenti:

(1-bis) Dato un numero reale positivo E (l'errore massimo richiesto dall'utilizzatore) e inseriti E e la derivata h' tra le variabili di ingresso della procedura:

$$\underline{\text{se}}\ |x_{k+1} - x_k| \ /\ |1 - h'(x_k)| < E\ \underline{\text{allora}}\ \text{STOP}$$

Il criterio è *calcolabile* ed *efficace*.

Per capire quanto buona sia x_k come approssimazione di α quando la condizione è verificata, si osservi che, procedendo come in (1) dell'Osservazione (1.72):

$$\left|\frac{x_{k+1}-x_k}{1-h'(x_k)}\right| = \left|\frac{1-h'(t_k)}{1-h'(x_k)}\right| |x_k - \alpha| = (1 + \varepsilon_k)\ |x_k - \alpha|$$

con

$$\varepsilon_k = \frac{h'(x_k)-h'(t_k)}{1-h'(x_k)}$$

In questo caso, quando $k \to \infty$ si ha $x_k \to \alpha$, $t_k \to \alpha$ e quindi $\varepsilon_k \to 0$.

(2-bis) Dato un numero reale positivo E (l'errore massimo richiesto dall'utilizzatore) ed inserite E, f ed f' tra le variabili di ingresso della procedura sia:

$$\underline{\text{se}}\ |f(x_k)|/|f'(x_k)| < E\ \underline{\text{allora}}\ \text{STOP}$$

Il criterio è *calcolabile* ed *efficace*.

Per capire quanto buona sia x_k come approssimazione di α quando la condizione è verificata, si osservi che, procedendo come in (2) dell'Osservazione (1.72):

$$\left|\frac{f(x_k)}{f'(x_k)}\right| = \left|\frac{f'(t_k)}{f'(x_k)}\right| |x_k - \alpha| = (1 + \varepsilon_k)\ |x_k - \alpha|$$

con

$$\varepsilon_k = \left| \frac{f'(t_k)}{f'(x_k)} \right| - 1$$

Anche in questo caso, quando $k \to \infty$ si ha $x_k \to \alpha$, $t_k \to \alpha$ e quindi $\varepsilon_k \to 0$.

(1.74) Osservazione (metodi ad un punto in $F(\beta,m)$).

Siano:

- $h:[a,b] \to R$ e γ in $[a,b]$ che verificano le ipotesi del Teorema di convergenza
- $\varphi:[a,b] \to F(\beta,m)$ l'algoritmo usato per approssimare i valori di h, tale che:

$$\text{per ogni } \theta \text{ in } [a,b] \cap F(\beta,m) \ , \ |\varphi(\theta) - h(\theta)| \leqslant d_\varphi$$

Siano poi x_k la successione generata dal metodo definito da h a partire da γ, convergente ad α per ipotesi, e ξ_k la successione definita da $\xi_0 = \gamma$, $\xi_{k+1} = \varphi(\xi_k)$. Si supponga che per ogni k sia ξ_k in $[a,b]$.

Si ha:

(1.75) Teorema (stabilità dei metodi ad un punto, parte I).

Sia $\delta > 0$. Se MetodoUnPunto(h,a,b,δ) eseguito in $F(\beta,m)$ definisce ξ in $F(\beta,m)$ tale che

$$|\xi_{k+1} \ominus \xi_k| < rd(\delta)$$

allora ξ è punto unito di una funzione $h^*:[a,b] \to R$ tale che:

$$\text{per ogni } x \text{ in } [a,b] \ , \ |h^*(x) - h(x)| \leqslant d_\varphi + \delta$$

Informalmente: se d_φ 'piccolo', la procedura restituisce un punto unito di una funzione h^* 'vicina' ad h.

(1.76) Teorema (stabilità dei metodi ad un punto, parte II).

Siano inoltre $f:[a,b] \to R$ una funzione regolare tale che $f(\alpha) = 0$, e $\psi:[a,b] \to F(\beta,m)$ l'algoritmo usato per approssimare i valori di f tale che:

$$\text{per ogni } \theta \text{ in } [a,b] \cap F(\beta,m) \ , \ |\psi(\theta) - f(\theta)| \leqslant d_\psi$$

Sia $\delta > 0$. Se MetodoUnPunto(h,a,b,f,δ) eseguito in $F(\beta,m)$ definisce ξ in $F(\beta,m)$ tale che

$$|\psi(\xi_k)| < rd(\delta)$$

allora ξ è zero di una funzione $f^*:[a,b] \to R$ tale che:

$$\text{per ogni } x \text{ in } [a,b] \ , \ |f^*(x) - f(x)| \leqslant d_\psi + \delta$$

Informalmente: se d_ψ 'piccolo', la procedura restituisce uno zero di una funzione f^* 'vicina' ad f.

(1.77) Osservazione (efficacia dei criteri d'arresto in F(β,m)).

I due teoremi precedenti stabiliscono che se in F(β,m) la procedura definisce ξ allora... Questo lascia supporre che la procedura *potrebbe non definire* ξ. La supposizione è corretta: come già sappiamo, in F(β,m) i criteri d'arresto possono risultare *non efficaci*.

Esempio.

> Sia [a,b] non contenente 0. Allora $A = [a,b] \cap F(\beta,m)$ contiene un numero finito di elementi. Sia $\Delta > 0$ la minima distanza tra due elementi consecutivi di A. Se φ non ha punti uniti in [a,b], si ha allora:
>
> $$|\xi_{k+1} - \xi_k| \geqslant \Delta \qquad \text{e quindi} \qquad |\xi_{k+1} \ominus \xi_k| \geqslant \Delta$$
>
> Se l'utilizzatore sceglie $\delta < \Delta$, la condizione $|\xi_{k+1} \ominus \xi_k| < rd(\delta)$ non può essere verificata.
>
> Nell'altro caso, Se ψ non ha zeri in [a,b], detto $\Gamma > 0$ il valore minimo di ψ in A, si ha:
>
> $$|\psi(\xi_k)| \geqslant \Gamma$$
>
> Se l'utilizzatore sceglie $\delta < \Gamma$, la condizione $|\psi(\xi_k)| < rd(\delta)$ non può essere verificata.

(1.78) Esempio.

Sia $f(x) = (x - 2)^2$. La funzione ha un solo zero, $\alpha = 2$ e $f'(\alpha) = 0$. Scelto $x_0 > 2$, per la successione generata dal metodo di Newton applicato ad f si ha:

$$x_{k+1} = (x_k + 2) / 2$$

da cui:

$$x_k - 2 = (1/2)^k (x_0 - 2)$$

La successione converge ad α ma è una successione di *tipo esponenziale*. In questo caso si ha:

$$h_N(x) = (x + 2) / 2$$

dunque $h'(\alpha) = 1/2 \neq 0$. In questo caso, il metodo di Newton risulta avere *ordine di convergenza ad* α pari a *uno*.

(1.3) METODO DI NEWTON PER FUNZIONI DA R^n IN R^n

(1.79) Osservazione.

Se $f:R \to R$ è una funzione regolare, ciascuna iterazione del metodo di Newton costruisce, a partire da un valore x_k noto, il numero reale x_{k+1} determinando lo zero (se esiste) della funzione affine (si veda l'Osservazione (1.67) nella Lezione 11):

$$A_k(x) = f(x_k) + f'(x_k)\ (x - x_k)$$

La funzione $A_k: R \to R$ è lo sviluppo di Taylor di f(x) di ordine uno in x_k (graficamente: la retta di equazione $y = A_k(x)$ è la tangente al grafico di f(x) in x_k).

L'idea del metodo di Newton nel caso in cui $f: R^n \to R^n$ sia regolare è la stessa: a ciascuna iterazione, a partire da un valore noto $x_k \in R^n$, si costruisce lo zero (se esiste) dello sviluppo di Taylor di f(x) di ordine uno in x_k:

$$A_k(x) = f(x_k) + J_f(x_k)\ (x - x_k)$$

dove $J_f(x) \in R^{n \times n}$ è la *matrice jacobiana* di f in x, ovvero la matrice di elemento i,j dato da:

$$\frac{\partial f_i}{\partial x_j}(x)$$

Lezione 14 (ore 25,26) - 22 ottobre 2025, 11:30 - 13:30 A13

(1.80) Esempio.

Sia $f:R^2 \to R^2$ definita da:[1]

$$f(x) = [\ f_1(x_1,x_2)\ ;\ f_2(x_1,x_2)\ ] = [\ x_1^{\,2} - x_2\ ;\ -\ x_1 + x_2^{\,2}\ ]$$

La matrice jacobiana di f in x è:

$$J_f(x) = [\ 2\ x_1\ ,\ -1\ ;\ -1\ ,\ 2\ x_2\ ] : R^2 \to R^{2\times 2}$$

(1.81) Osservazione.[2]

Noto un elemento x(k) in R^n, il metodo di Newton per la funzione $f:R^n \to R^n$ determina l'elemento x(k+1) risolvendo l'equazione:

$$f(x(k)) + J_f(x(k))\ (x - x(k)) = 0$$

ovvero:

$$J_f(x(k))\ (x - x(k)) = -\ f(x(k))$$

Quest'ultima equazione è un *sistema di equazioni lineari*. Se la matrice $J_f(x(k))$ è invertibile allora si ottiene:

$$x - x(k) = -\ J_f(x(k))^{-1}\ f(x(k))$$

L'elemento x(k+1) è quindi:

$$x(k+1) = x(k) - J_f(x(k))^{-1}\ f(x(k))$$

(1.82) Esempio.

Si consideri la funzione $f:R^2 \to R^2$ dell'Esempio (1.80) e sia x(0) = [1 ; -1]. Per determinare x(1) occorre calcolare $J_f(x(0))$, f(x(0)) e poi risolvere il sistema

$$J_f(x(0))\ z = -\ f(x(0))$$

Si ha:

$$J_f(x(0)) = [\ 2\ ,\ -1\ ;\ -1\ ,\ -2\ ] \quad , \quad f(x(0)) = [\ 2\ ;\ 0\ ]$$

Si osserva che $J_f(x(0))$ è invertibile. La soluzione del sistema risulta:

$$p = [\ -\ 4/5\ ;\ 2/5\ ]$$

Allora:

$$x(1) = x(0) + p = [\ 1/5\ ;\ -3/5\ ]$$

---

1 Per le matrici utilizzeremo la notazione di *Scilab*.

2 Per le successioni di elementi in R^n, useremo la notazione x(0), x(1), x(2), ...

(1.83) Definizione.

Il *metodo di Newton* applicato alla funzione $f:R^n \to R^n$, con matrice jacobiana $J_f(x)$ invertibile, è il metodo ad un punto definito dalla funzione:

$$N(x) = x - J_f(x)^{-1}\, f(x) : R^n \to R^n$$

(1.84) Teorema (di convergenza locale per metodi ad un punto in R^n).

Siano $h:R^n \to R^n$ sufficientemente regolare e α punto unito di h.

Se *tutti* gli autovalori di $J_h(\alpha)$ hanno *modulo* < 1 allora esiste un numero reale $\rho > 0$ tale che:

$$\| x(0) - \alpha \| < \rho \quad \Rightarrow$$ la successione x(k) generata dal metodo iterativo definito da h a partire da x(0) *converge ad* α

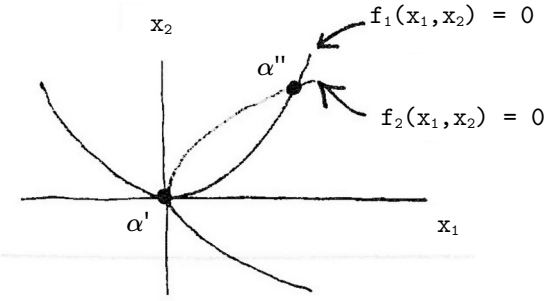
(Dimostrazione omessa.)

Questo teorema fornisce una *condizione sufficiente per l'utilizzabilità* del metodo definito da h per approssimare α. Per un metodo ad un punto in R^n, essere utilizzabile significa che *per ogni* x(0) *sufficientemente vicino ad un punto unito* α *di* h, *la successione generata dal metodo definito da* h *a partire da* x(0) *converge ad* α.

(1.85) Esempio (prima parte).

Si consideri ancora la funzione $f:R^2 \to R^2$ dell'Esempio (1.80).
La funzione ha *due zeri*:

$$\alpha' = [0 ; 0] \quad , \quad \alpha'' = [1 ; 1]$$



Per approssimare i due zeri si considera il metodo definito dalla funzione

$$h(x) = x + f(x) = [\, x_1 + x_2^2 - x_2 \; ; \; x_2 - x_1 + x_2^2 \,]$$

Si verifica facilmente che i punti uniti di h sono tutti e soli gli zeri di f.

Per la matrice jacobiana si ha:

$$J_h(x) = I + J_f(x) = [\, 1 + 2\,x_1 \; , \; -1 \; ; \; -1 \; , \; 1 + 2\,x_2 \,]$$

da cui:

$$J_h(\alpha') = [\, 1 \; , \; -1 \; ; \; -1 \; , \; 1 \,]$$

Gli autovalori sono le radici del polinomio caratteristico:

$$p(\lambda) = \det(\, J_h(\alpha') - \lambda I) = (1 - \lambda)^2 - 1 \quad \text{ovvero} \quad \lambda_1 = 0 \; , \; \lambda_2 = 2$$

Il Teorema di convergenza locale non è applicabile. Sussiste però la seguente

(1.86) Osservazione.

Nelle ipotesi del Teorema di convergenza locale: se *almeno uno* degli autovalori di $J_h(\alpha)$ ha *modulo* > 1 allora il metodo iterativo definito da h *non è utilizzabile* per approssimare α.

(1.87) Esempio.

Per giustificare l'asserto precedente, si consideri il seguente caso particolare.

Siano $h(x) = [\ h_1(x_1)\ ;\ h_2(x_2)\ ]: R^2 \to R^2$ con h_1 e h_2 regolari, α_1 punto unito di h_1 e α_2 punto unito di h_2. Ne segue che $\alpha = [\alpha_1\ ;\ \alpha_2]$ è punto unito di h. La matrice jacobiana di h in α è:

$$J_h(\alpha) = [\ h_1'(\alpha_1)\ ,\ 0\ ;\ 0\ ,\ h_2'(\alpha_2)\ ]$$

i cui autovalori sono:

$$\lambda_1 = h_1'(\alpha_1) \quad \text{e} \quad \lambda_2 = h_2'(\alpha_2)$$

Sia x(k) una successione generata dal metodo definito da h. Allora $x_1(k)$ e $x_2(k)$ sono, rispettivamente, una successione generata dal metodo definito da h_1 e, rispettivamente, dal metodo definito da h_2. Se, ad esempio, $|\lambda_1| = |h_1'(\alpha_1)| > 1$, per la successione $x_1(k)$ si ha (Osservazione (1.61) della Lezione 10): o $x_1(k) = \alpha_1$ per un valore finito di k o $x_1(k)$ non converge ad α_1. Come già osservato a suo tempo, l'eventualità che accada la prima condizione è molto remota. Dunque ci si aspetta che la successione non sia convergente. Se in questa situazione il metodo iterativo definito da h fosse utilizzabile per approssimare α allora per qualunque x(0) sufficientemente vicino ad α la successione x(k) risulterebbe convergere al punto unito di h. Ne seguirebbe che per qualunque $x_1(0)$ sufficientemente vicino ad α_1 la successione $x_1(k)$ risulterebbe convergere al punto unito di h_1. Ma questo, per quanto osservato sopra, non è possibile.

(1.88) Esempio (seconda parte).

Dal risultato finale della prima parte dell'esempio si deduce che il metodo definito da h *non è utilizzabile* per approssimare α'.

Per α'' si ha:

$$J_h(\alpha'') = [\ 3\ ,\ -1\ ;\ -1\ ,\ 3\ ]$$

e quindi:

$$p(\lambda) = \det(\ J_h(\alpha'') - \lambda I) = (3 - \lambda)^2 - 1 \quad \text{ovvero} \quad \lambda_1 = 2\ ,\ \lambda_2 = 4$$

e il metodo definito da h *non è utilizzabile* neppure per approssimare α''.

(1.89) Esercizio (per casa).

Sia f la funzione dell'Esempio (1.85). Determinare la funzione $N: R^2 \to R^2$ che definisce il metodo di Newton applicato ad f e verificare (con tanta pazienza) che si ha: $J_N(\alpha') = 0$ e $J_N(\alpha'') = 0$.

(1.90) Osservazione (utilizzabilità del metodo di Newton).

Quanto mostrato nell'esercizio precedente vale in generale. Si ha infatti:

Se f ha derivate seconde continue, J_f è non singolare e α è uno zero di f, allora $J_N(\alpha) = 0$ e il metodo di Newton *è utilizzabile* per approssimare α. Si ha inoltre che, analogamente a quanto accade nel caso di funzioni di una variabile, l'*ordine di convergenza* ad α del metodo di Newton è *almeno due*.

Lezione 15 (ore 27,28) - 23 ottobre 2025, 8:30 - 10:30 F3

(2) SISTEMI DI EQUAZIONI LINEARI

(2.01) Esempio.

Esempi di contesti in cui si devono risolvere sistemi di equazioni lineari:

- ad ogni iterazione del metodo di Newton per funzioni da R^n in R^n;
- risoluzione di reti elettriche resistive lineari
- risoluzione di reti elettriche RLC lineari in regime sinusoidale

(2.02) Problema.

Dati $A \in R^{n \times n}$ invertibile e $b \in R^n$, determinare $x^* \in R^n$ t.c. $A\,x^* = b$. La colonna x^* si chiama *soluzione* del sistema $A\,x = b$.

(2.03) Osservazione.

Una matrice $A \in R^{n \times n}$ è invertibile se verifica una delle seguenti proprietà *equivalenti*:

- esiste una matrice $M \in R^{n \times n}$ t.c. $AM = MA = I$ (la matrice M si chiama matrice *inversa* di A e si indica con A^{-1})
- $A\,x = 0 \Leftrightarrow x = 0$ (questa proprietà si esprime anche con $\ker A = \{ 0 \}$)
- per *ogni* colonna $b \neq 0$ in R^n, esiste *una sola* soluzione x^* del sistema $A\,x = b$
- $\det A \neq 0$

(2.04) Osservazione (casi semplici).

Decidere se la matrice A del sistema è invertibile e, in caso affermativo, determinare la soluzione del sistema $A\,x = b$ è *semplice* quando la *struttura* di A ricade in uno dei seguenti casi:

(D) *diagonale* (A è diagonale se $i \neq j \Rightarrow a_{i,j} = 0$)

- Si ha: $\det A = a_{1,1} \cdots a_{n,n}$, quindi: $\det A = 0 \Leftrightarrow$ esiste k t.c. $a_{k,k} = 0$. Dunque: A *invertibile se e solo se per ogni* k *si ha* $a_{k,k} \neq 0$.

- Se A è invertibile, le componenti della soluzione x^* del sistema $A\,x = b$ si determinano con:

$$x_k^* = b_k \,/\, a_{k,k}$$

  Il numero di operazioni necessario per determinare la soluzione è:

n *divisioni*.

(T) *triangolare* (A è triangolare *superiore* se $i > j \Rightarrow a_{i,j} = 0$; è triangolare *inferiore* se $i < j \Rightarrow a_{i,j} = 0$)

- Anche in questo caso si ha: $\det A = a_{1,1} \cdots a_{n,n}$. Dunque: A *invertibile se e solo se*

per ogni k *si ha* $a_{k,k} \neq 0$.

- Se A è *triangolare superiore* invertibile, le componenti della soluzione x^* del sistema A x = b si determinano con la seguente procedura di *sostituzione all'indietro*:

```
z = SI(T,c)
      se T non è triangolare superiore invertibile allora STOP;
         altrimenti
            z_n = c_n / t_{n,n};
            per k = n-1,...,1 ripeti
                s = t_{k,k+1} * x_{k+1} + ... + t_{k,n} * x_n;
                x_k = (b_k - s) / t_{k,k};
```

Il numero di operazioni necessario per determinare la soluzione è:

$$n \textit{ divisioni} + \frac{n(n-1)}{2} \ (\textit{moltiplicazioni} + \textit{somme})$$

(2.05) Esercizio (per casa).

Descrivere la procedura di *sostituzione in avanti* di intestazione

```
z = SA(T,c)
```

che, dati una matrice triangolare inferiore invertibile T ed una colonna c, determina la soluzione del sistema T x = c. Determinare anche il numero di operazioni necessario per determinare la soluzione.

Lezione 16 (ore 29,30) - 28 ottobre 2025, 16:30 - 18:30 F3

(2.06) Osservazione (casi semplici, conclusione).

(O) *ortogonale* (A è ortogonale se sussiste una delle tre condizioni *equivalenti*:

(1) le colonne (o le righe) di A sono una *base ortonormale* di R^n con prodotto scalare canonico;
(2) A è invertibile e $A^{-1} = A^t$;
(3) $A^t A = A A^t = I$)

- A è *certamente invertibile.*

- La soluzione x^* del sistema $A x = b$ si determina con:

$$x^* = A^t b$$

Il numero di operazioni necessario per determinare la soluzione è quello delle operazioni necessarie per effettuare il prodotto di una matrice per una colonna:

n^2 *moltiplicazioni* + $n(n-1)$ *somme*

(P) *di permutazione* (A è di permutazione se si ottiene dalla matrice identità I *permutando* le colonne).

Le colonne di una matrice di permutazione sono quindi quelle della matrice identità (a parte l'ordine). Dunque costituiscono una base ortonormale di R^n con prodotto scalare canonico (la base canonica). Se ne deduce che *una matrice di permutazione è ortogonale.*

- Anche in questo caso si ha: A è *certamente invertibile.*

- La soluzione x^* del sistema $A x = b$ si determina con:

$$x^* = A^t b$$

Il numero di operazioni necessario per determinare la soluzione è, questa volta, *zero* perché A^t, come A, è di permutazione e il prodotto $P v$ di una matrice di permutazione P per una colonna v produce una colonna che ha le *stesse componenti* di v ma in ordine diverso.

(2.07) Osservazione (caso generale).

Quando la matrice A del sistema *non* ha struttura tale da ricadere in un caso semplice, il problema si affronta in due passi:

Primo Passo:

Si *fattorizza* A in prodotto di *fattori semplici.*

Esempio: $A = F_1 F_2 F_3$, con F_1 ortogonale, F_2 triangolare superiore e F_3 di permutazione.

Secondo Passo:

Si utilizza la fattorizzazione per decidere se A è invertibile e, in caso affermativo, per determinare la soluzione x^*.

Esempio:

$$A = F_1 F_2 F_3 \Rightarrow \det A = \det F_1 \ \det F_2 \ \det F_3$$

quindi: A *è invertibile* $\Leftrightarrow$ *ciascun fattore è invertibile.* Poi:

(1) $$A\,x = b \ \equiv \ F_1 F_2 F_3\,x = b \ \equiv \ F_2 F_3\,x = F_1^{-1} b = c_1$$

e c_1 si ottiene risolvendo il sistema semplice $F_1\,x = b$.

(2) $$F_2 F_3\,x = c_1 \ \equiv \ F_3\,x = F_2^{-1} c_1 = c_2$$

e c_2 si ottiene risolvendo il sistema semplice $F_2\,x = c_1$.

(3) $$F_3\,x = c_2 \ \equiv \ x^* = F_3^{-1} c_2$$

e x^* si ottiene risolvendo il sistema semplice $F_3\,x = c_2$.

In generale, se A è invertibile, la soluzione si determina risolvendo tanti *sistemi semplici* quanti sono i fattori di A.

(2.08) Definizione (fattorizzazione LR, LR con pivoting e QR).

Sia $A \in R^{n\times n}$.

Una *fattorizzazione LR di* A è una coppia S,D tale che:

- $S \in R^{n\times n}$ è una matrice triangolare inferiore con $s_{kk} = 1$ per $k = 1,\ldots,n$
- $D \in R^{n\times n}$ è una matrice triangolare superiore
- $S\,D = A$

Si osservi che il fattore sinistro S è invertibile. Allora: A *è invertibile se e solo se lo è il fattore destro* D.

Una *fattorizzazione LR con pivoting di* A è una terna P,S,D tale che:

- $P \in R^{n\times n}$ è una matrice di permutazione
- la coppia S,D è una fattorizzazione LR di $P\,A$

La relazione tra A,P,S e D è:

$$P\,A = S\,D \quad \text{ovvero} \quad A = P^t\,S\,D$$

Si osservi che sia P che il fattore sinistro S sono invertibili. Di nuovo: A *è invertibile se e solo se lo è il fattore destro* D.

Una *fattorizzazione QR di* A è una coppia U,T tale che:

- $U \in R^{n \times n}$ è una matrice ortogonale
- $T \in R^{n \times n}$ è una matrice triangolare superiore
- $UT = A$

Si osservi che il fattore sinistro U è invertibile. Anche in questo caso: A *è invertibile se e solo se lo è il fattore destro* T.

(2.09) Definizione (matrice elementare di Gauss).

Data $A \in R^{n \times n}$, per cercare una fattorizzazione LR con pivoting si utilizza la procedura EGP che si basa sul procedimento di *eliminazione di Gauss*. Per descrivere la procedura, occorre la nozione di matrice elementare di Gauss.

$H \in R^{n \times n}$ è una *matrice elementare di Gauss* se: esistono un indice $k \in \{1,\ldots,n-1\}$ e numeri reali $\lambda_{k+1},\ldots,\lambda_n$ tali che H si ottiene dalla matrice identità $I \in R^{n \times n}$ sostituendo alla colonna k-esima e_k (le cui componenti sono tutte uguali a zero ed eccezione della k-esima che vale uno) la colonna:

$$[0\ ;\ldots;\ 0\ ;\ 1\ ;\ \lambda_{k+1}\ ;\ldots;\ \lambda_n]$$
$$\uparrow$$
k-esima componente

Esempi:

- la matrice $I \in R^{n \times n}$ è elementare di Gauss;
- la matrice:

```
[1,0,0;
 1,1,0;
-2,0,1]
```

  è elementare di Gauss;
- la matrice:

```
[1,0,1;
 1,1,0;
-2,0,1]
```

  non è elementare di Gauss.

(2.10) Proprietà (delle matrici elementari di Gauss).

Sia H una matrice elementare di Gauss. Allora:

- H è *triangolare inferiore* con $h_{kk} = 1$ per ogni k (dunque *invertibile*)
- H^{-1} si ottiene da H *cambiando segno* agli elementi al di sotto della diagonale principale

  (ad esempio:

```
H = [1,0,0;      H^-1 = [1,0,0;
     1,1,0;             -1,1,0;
    -2,0,1]              2,0,1]  )
```

(2.11) Definizione (procedura EGP).

La seguente procedura EGP opera su una matrice $A \in R^{n \times n}$, e determina una terna P,S,D che è una fattorizzazione LR con pivoting di A.

```
(P,S,D) = EGP(A)

  A_1 = A;
  per k = 1,...,n-1 ripeti:
    determina opportunamente P_k di permutazione, H_k elementare di Gauss e pone A_{k+1} = H_k P_k A_k;
  D = A_n;
  P = P_{n-1} ... P_1;
  S = P (P_1^t H_1^{-1} … P_{n-1}^t H_{n-1}^{-1})
```

Le matrici P_k e H_k sono determinate in modo da ottenere A_n *triangolare superiore*.

Si osservi che:

$$D = A_n = H_{n-1}\, P_{n-1}\, A_{n-1} = \ldots = H_{n-1}\, P_{n-1} \ldots H_1\, P_1\, A$$

da cui, ricavando A:

$$A = (P_1^t H_1^{-1} \ldots P_{n-1}^t H_{n-1}^{-1})\, D$$

La matrice $P_1^t H_1^{-1} \ldots P_{n-1}^t H_{n-1}^{-1}$ ***non è*** triangolare inferiore con elementi uguali ad uno sulla diagonale ma la matrice

$$P\, (P_1^t H_1^{-1} \ldots P_{n-1}^t H_{n-1}^{-1})$$

lo è. Quindi, la coppia $S = P\, (P_1^t H_1^{-1} \ldots P_{n-1}^t H_{n-1}^{-1})$, D è una fattorizzazione LR di P A, come si voleva.

Resta da chiarire come, ad ogni iterazione, di determinano le matrici P_k e H_k.

Lezione 17 (ore 31,33) - 04 novembre 2025, 15:30 - 18:30 F3

(2.12) Esempio.

Calcolo di EGP(A) con:

$$A = \begin{bmatrix} 1, & 1, & 0, & 0; \\ 2, & 2, & 1, & 0; \\ -2, & 0, & 0, & -1; \\ -1, & 1, & 2, & -1 \end{bmatrix}$$

(*) $A_1 = A$;

(*) $k = 1$; $A_1(1,1) \neq 0 \Rightarrow P_1 = I$; $T_1 = P_1 A_1$;

$$H_1 = \begin{bmatrix} 1, & 0, & 0, & 0; \\ \lambda_2, & 1, & 0, & 0; \\ \lambda_3, & 0, & 1, & 0; \\ \lambda_4, & 0, & 0, & 1 \end{bmatrix}$$

I valori λ_2,λ_3,λ_4 sono determinati dalla richiesta che nella matrice $H_1 T_1$ gli elementi di posto (2,1),(3,1) e (4,1) - ovvero gli elementi della k-esima colonna al di sotto della diagonale - siano *uguali a zero*:

$$\lambda_2 T_1(1,1) + T_1(2,1) = 0 \quad ; \quad \lambda_3 T_1(1,1) + T_1(3,1) = 0 \quad ; \quad \lambda_4 T_1(1,1) + T_1(4,1) = 0$$

Tenuto conto che $T_1(1,1) \neq 0$, i valori λ_2,λ_3,λ_4 sono *univocamente determinati*:

$$\lambda_2 = -\frac{T_1(2,1)}{T_1(1,1)} = -2 \quad ; \quad \lambda_3 = -\frac{T_1(3,1)}{T_1(1,1)} = 2 \quad ; \quad \lambda_4 = -\frac{T_1(4,1)}{T_1(1,1)} = 1$$

Infine:

$$\begin{bmatrix} 1, & 0, & 0, & 0; \\ -2, & 1, & 0, & 0; \\ 2, & 0, & 1, & 0; \\ 1, & 0, & 0, & 1 \end{bmatrix} \begin{bmatrix} 1, & 1, & 0, & 0; \\ 2, & 2, & 1, & 0; \\ -2, & 0, & 0, & -1; \\ -1, & 1, & 2, & -1 \end{bmatrix} = \begin{bmatrix} 1, & 1, & 0, & 0; \\ 0, & 0, & 1, & 0; \\ 0, & 2, & 0, & -1; \\ 0, & 2, & 2, & -1 \end{bmatrix}$$

$$H_1 \qquad T_1 \qquad = \qquad A_2$$

(*) $k = 2$; $A_2(2,2) = 0 \Rightarrow$ essendo $A_2(3,2) \neq 0$, scambio la seconda riga con la terza: $P_2 = P_{2,3}$;

$$\begin{bmatrix} 1, & 0, & 0, & 0; \\ 0, & 0, & 1, & 0; \\ 0, & 1, & 0, & 0; \\ 0, & 0, & 0, & 1 \end{bmatrix} \begin{bmatrix} 1, & 1, & 0, & 0; \\ 0, & 0, & 1, & 0; \\ 0, & 2, & 0, & -1; \\ 0, & 2, & 2, & -1 \end{bmatrix} = \begin{bmatrix} 1, & 1, & 0, & 0; \\ 0, & 2, & 0, & -1; \\ 0, & 0, & 1, & 0; \\ 0, & 2, & 2, & -1 \end{bmatrix}$$

$$P_{2,3} \qquad A_2 \qquad = \qquad T_2$$

Si ha così $T_2(2,2) \neq 0$.

Poi:

$$H_2 = \begin{bmatrix} 1, & 0, & 0, & 0; \\ 0, & 1, & 0, & 0; \\ 0, & \lambda_3, & 1, & 0; \\ 0, & \lambda_4, & 0, & 1 \end{bmatrix}$$

I valori λ_3,λ_4 sono determinati dalla richiesta che nella matrice $H_2\,T_2$ gli elementi di posto (3,2), e (4,2) - ovvero gli elementi della k-esima colonna al di sotto della diagonale - siano *uguali a zero*:

$$\lambda_3\,T_2(2,2) + T_2(3,2) = 0 \quad ; \quad \lambda_4\,T_2(2,2) + T_2(4,2) = 0$$

Tenuto conto che $T_2(2,2) \neq 0$, i valori λ_3,λ_4 sono *univocamente determinati*:

$$\lambda_3 = -\frac{T_2(3,2)}{T_2(2,2)} = 0 \quad ; \quad \lambda_4 = -\frac{T_2(4,2)}{T_2(2,2)} = -1$$

Infine:

$$\begin{bmatrix} 1, & 0, & 0, & 0; \\ 0, & 1, & 0, & 0; \\ 0, & 0, & 1, & 0; \\ 0, & -1, & 0, & 1 \end{bmatrix} \begin{bmatrix} 1, & 1, & 0, & 0; \\ 0, & 2, & 0, & -1; \\ 0, & 0, & 1, & 0; \\ 0, & 2, & 2, & -1 \end{bmatrix} = \begin{bmatrix} 1, & 1, & 0, & 0; \\ 0, & 2, & 0, & -1; \\ 0, & 0, & 1, & 0; \\ 0, & 0, & 2, & 0 \end{bmatrix}$$

$$H_2 \qquad T_2 \qquad = \qquad A_3$$

(*) $k = 3$; $A_3(3,3) \neq 0 \Rightarrow P_3 = I$; $T_3 = A_3$;

$$H_3 = \begin{bmatrix} 1, & 0, & 0, & 0; \\ 0, & 1, & 0, & 0; \\ 0, & 0, & 1, & 0; \\ 0, & 0, & \lambda_4, & 1 \end{bmatrix}$$

Il valore λ_4 è determinato dalla richiesta che nella matrice $H_3\,T_3$ l'elemento di posto (4,3) - ovvero gli elementi della k-esima colonna al di sotto della diagonale - sia *uguale a zero*:

$$\lambda_4\,T_3(3,3) + T_3(4,3) = 0$$

Tenuto conto che $T_3(3,3) \neq 0$, il valore λ_4 è *univocamente determinato*:

$$\lambda_4 = -\frac{T_3(4,3)}{T_3(3,3)} = -2$$

Infine:

$$\begin{bmatrix} 1, & 0, & 0, & 0; \\ 0, & 1, & 0, & 0; \\ 0, & 0, & 1, & 0; \\ 0, & 0, & -2, & 1 \end{bmatrix} \begin{bmatrix} 1, & 1, & 0, & 0; \\ 0, & 2, & 0, & -1; \\ 0, & 0, & 1, & 0; \\ 0, & 0, & 2, & 0 \end{bmatrix} = \begin{bmatrix} 1, & 1, & 0, & 0; \\ 0, & 2, & 0, & -1; \\ 0, & 0, & 1, & 0; \\ 0, & 0, & 0, & 0 \end{bmatrix}$$

$$H_3 \qquad T_3 \qquad = \qquad A_4$$

(*) $D = A_4$; $P = P_3\,P_2\,P_1 = P_{2,3}$;

Poi:

$$\begin{bmatrix} 1,0,0,0; \\ 2,1,0,0; \\ -2,0,1,0; \\ -1,0,0,1 \end{bmatrix} \begin{bmatrix} 1,0,0,0; \\ 0,0,1,0; \\ 0,1,0,0; \\ 0,0,0,1 \end{bmatrix} \begin{bmatrix} 1,0,0,0; \\ 0,1,0,0; \\ 0,0,1,0; \\ 0,1,0,1 \end{bmatrix} \begin{bmatrix} 1,0,0,0; \\ 0,1,0,0; \\ 0,0,1,0; \\ 0,0,2,1 \end{bmatrix} = \begin{bmatrix} 1,0,0,0; \\ 2,0,1,0; \\ -2,1,0,0; \\ -1,1,2,1 \end{bmatrix}$$

H_1^{-1} $P_{2,3}^{t}$ H_2^{-1} H_3^{-1} Σ

Infine:

$$S = P\,\Sigma = \begin{bmatrix} 1,0,0,0; \\ -2,1,0,0; \\ 2,0,1,0; \\ -1,1,2,1 \end{bmatrix}$$

Gli elementi $T_1(1,1)$, $T_2(2,2)$ e $T_3(3,3)$ utilizzati per ricavare le matrici elementari di Gauss H_1, H_2 e H_3 (in generale l'elemento $T_k(k,k)$ utilizzato per ricavare la matrice H_k) si chiamano *pivot*. Il termine *pivoting* si riferisce agli scambi effettuati alla k-esima iterazione per ottenere $T_k(k,k) \neq 0$.

(2.13) Esempio.

Calcolo di EGP(A) con:

$$A = \begin{bmatrix} 1,1,0,0; \\ 2,2,1,0; \\ -2,-2,0,-1; \\ -1,-1,2,-1 \end{bmatrix}$$

(*) $A_1 = A$;

(*) $k = 1$; $A_1(1,1) \neq 0 \Rightarrow P_1 = I$; $T_1 = P_1\,A_1$;

$$H_1 = \begin{bmatrix} 1,0,0,0; \\ \lambda_2,1,0,0; \\ \lambda_3,0,1,0; \\ \lambda_4,0,0,1 \end{bmatrix}$$

I valori $\lambda_2,\lambda_3,\lambda_4$ sono determinati dalla richiesta che nella matrice $H_1\,T_1$ gli elementi di posto (2,1),(3,1) e (4,1) - ovvero gli elementi della k-esima colonna al di sotto della diagonale - siano *uguali a zero*:

$$\lambda_2\,T_1(1,1) + T_1(2,1) = 0 \quad ; \quad \lambda_3\,T_1(1,1) + T_1(3,1) = 0 \quad ; \quad \lambda_4\,T_1(1,1) + T_1(4,1) = 0$$

Tenuto conto che $T_1(1,1) \neq 0$, i valori $\lambda_2,\lambda_3,\lambda_4$ sono *univocamente determinati*:

$$\lambda_2 = -\frac{T_1(2,1)}{T_1(1,1)} = -2 \quad ; \quad \lambda_3 = -\frac{T_1(3,1)}{T_1(1,1)} = 2 \quad ; \quad \lambda_4 = -\frac{T_1(4,1)}{T_1(1,1)} = 1$$

Infine:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 2 & 2 & 1 & 0 \\ -2 & 0 & 0 & -1 \\ -1 & 1 & 2 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 2 & -1 \end{bmatrix}$$

$$H_1 \qquad T_1 \qquad = \qquad A_2$$

(*) $k = 2$; $A_2(2,2) = 0 \Rightarrow$ essendo *anche* $A_2(3,2) = A_2(4,2) = 0$, gli elementi della k-esima colonna al di sotto della diagonale *sono già uguali a zero* si pone: $P_2 = I$ e $H_2 = I$, da cui $T_2 = P_2 A_2 = A_2$ e $A_3 = H_2 T_2 = H_2 A_2 = A_2$;

(*) $k = 3$; $A_3(3,3) = 0 \Rightarrow$ essendo $A_3(4,3) \neq 0$ scambio la terza riga con la quarta: $P_3 = P_{3,4}$, quindi:

$$T_3 = P_3 A_3 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

Questa matrice è già triangolare superiore, quindi $H_3 = I$ e $A_4 = T_3$;

(*) $D = A_4$; $P = P_3 P_2 P_1 = P_{3,4}$;

Poi:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -2 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -2 & 0 & 0 & 1 \\ -1 & 0 & 1 & 0 \end{bmatrix}$$

$$H_1^{-1} \qquad P_{3,4}^{t} \qquad \Sigma$$

Infine:

$$S = P\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -2 & 0 & 0 & 1 \end{bmatrix}$$

(2.14) Teorema (esistenza della fattorizzazione LR con pivoting).

Sia $A \in \mathbb{R}^{n \times n}$. La procedura EGP applicata ad A restituisce *una* fattorizzazione LR con pivoting di A. Ovvero: *per ogni* $A \in \mathbb{R}^{n \times n}$ *esiste almeno una fattorizzazione* LR *con pivoting*.

(Dimostrazione: segue dai due esempi precedenti.)

(2.15) Esercizio (uso della fattorizzazione LR con pivoting).

Siano:

$$EGP(A) = \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right), \quad b = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Senza determinare A, decidere se A è invertibile e, in caso affermativo, determinare la soluzione del sistema A x = b.

(2.16) Procedura (studio di un sistema di equazioni lineari con EGP).

// $A \in R^{n \times n}$, $b \in R^n$.

(S,D,P) = EGP(A);
se $d_{kk} = 0$ per qualche k allora STOP; altrimenti
    c = SA(S,Pb);
    x^* = SI(D,c)

La procedura è *soddisfacente* nel senso che *comunque* assegnati i dati, decide se la matrice è invertibile e, in caso affermativo, determina la soluzione.

(2.17) Definizione (procedura GS).

Una procedura per la ricerca di una fattorizzazione QR di una matrice $A \in R^{n \times n}$ è la seguente procedura GS,[1] descritta nel caso particolare di n = 3.

Sia $A = [a_1,a_2,a_3] \in R^{3 \times 3}$.

Passo uno.

Cerchiamo $\Omega = [\omega_1,\omega_2,\omega_3]$ a colonne ortogonali e Θ triangolare superiore con $\theta_{kk} = 1$ tali che $\Omega\Theta = A$. *Se matrici siffatte esistono*, riscrivendo l'ultima uguaglianza *per colonne* si ha:

$$\omega_1 = a_1 \quad , \quad \omega_1 \theta_{1,2} + \omega_2 = a_2 \quad , \quad \omega_1 \theta_{1,3} + \omega_2 \theta_{2,3} + \omega_3 = a_3 \qquad (*)$$

La prima uguaglianza determina ω_1. Dalla seconda segue che:[2]

$$(\omega_1 \theta_{1,2}) \bullet \omega_1 + \omega_2 \bullet \omega_1 = a_2 \bullet \omega_1$$

Poiché ω_1 e ω_2 sono ortogonali, si ha $\omega_2 \bullet \omega_1 = 0$. Allora, se $\omega_1 \neq 0$, si ha *necessariamente*:

$$\theta_{1,2} = (a_2 \bullet \omega_1) \,/\, (\omega_1 \bullet \omega_1)$$

e quindi:

$$\omega_2 = a_2 - \omega_1 \theta_{1,2}$$

Dalla terza uguaglianza delle (*) si ha poi:

$$(\omega_1 \theta_{1,3}) \bullet \omega_1 + (\omega_2 \theta_{2,3}) \bullet \omega_1 + \omega_3 \bullet \omega_1 = a_3 \bullet \omega_1$$

e

$$(\omega_1 \theta_{1,3}) \bullet \omega_2 + (\omega_2 \theta_{2,3}) \bullet \omega_2 + \omega_3 \bullet \omega_2 = a_3 \bullet \omega_2$$

Poiché $\omega_2 \bullet \omega_1 = 0$ e, analogamente, $\omega_3 \bullet \omega_1 = 0$, allora si ha *necessariamente*:

$$\theta_{1,3} = (a_3 \bullet \omega_1) \,/\, (\omega_1 \bullet \omega_1)$$

---

1 Il nome GS della procedura deriva da quello della *procedura di ortonormalizzazione di Gram-Schmidt*, da cui concettualmente deriva.

2 Date due colonne $v,w \in R^n$, si indica con $v \bullet w$ il loro prodotto scalare canonico: $v \bullet w = v_1 w_1 + \ldots + v_n w_n$.

Essendo anche $\omega_3 \bullet \omega_2 = 0$, se $\omega_2 \neq 0$, si ha *necessariamente*:

$$\theta_{2,3} = (a_3 \bullet \omega_2) \,/\, (\omega_2 \bullet \omega_2)$$

e, infine:

$$\omega_3 = a_3 - \omega_1 \theta_{1,3} - \omega_2 \theta_{2,3}$$

Passo due.

La fattorizzazione di A trovata al passo precedente *non è* una fattorizzazione QR perché le colonne di Ω non hanno norma unitaria. Questo secondo passo determina, se possibile, una fattorizzazione QR normalizzando le colonne di Ω.

Sia: $\Delta = \mathrm{diag}(\|\omega_1\|, \|\omega_2\|, \|\omega_3\|)$.[3] Se anche $\omega_3 \neq 0$, la matrice Δ è invertibile e si verifica facilmente che la coppia

$$U = \Omega \Delta^{-1} \quad , \quad T = \Delta \Theta \qquad (**)$$

è una fattorizzazione QR di A. Si osservi che per la matrice T, triangolare superiore, si ha:

$$T_{k,k} = \|\omega_k\| > 0$$

(2.18) Teorema (procedura GS e fattorizzazione QR).

La procedura GS descritta nella definizione precedente determina *una* fattorizzazione QR di $A \in R^{n \times n}$ se e solo se A è invertibile.

(Dimostrazione. Se la procedura non si interrompe prematuramente perché $\omega_k = 0$ per qualche k, allora la coppia U,T determinata da (**) è costituita da due matrici invertibili (U perché ortogonale, T perché triangolare con sulla diagonale le norme, non nulle, delle colonne ω_k). Viceversa, se fosse $\omega_1 = 0$ allora sarebbe $a_1 = 0$ e quindi A non invertibile. Se fosse $\omega_1 \neq 0$ e $\omega_2 = 0$ allora sarebbe $0 = a_2 - \omega_1 \theta_{1,2} = a_2 - a_1 \theta_{1,2}$, dunque a_1 e a_2 sarebbero linearmente dipendenti, quindi A non invertibile. Se fosse $\omega_1 \neq 0$, $\omega_2 \neq 0$ e $\omega_3 = 0$ …)

(2.19) Osservazione (non unicità della fattorizzazione QR).

Siano $A \in R^{n \times n}$ e U,T una fattorizzazione QR di A. Se $E \in R^{n \times n}$ è una matrice diagonale con, ad esempio, E(1,1) = -1 e E(k,k) = 1 per k = 2,...,n, allora la coppia:

$$U' = U E \quad , \quad T' = E T$$

è una fattorizzazione QR di A *diversa* da U,T.

(2.20) Procedura (studio di un sistema di equazioni lineari con GS).

// $A \in R^{n \times n}$, $b \in R^n$.

Se GS(A) determina $\omega_k = 0$ per qualche k allora STOP; altrimenti
    (U,T) = GS(A);
    $x^* = SI(T, U^t b)$

3 Mutuando la simbologia da *Scilab*, con $\mathrm{diag}(v_1,...,v_n)$ si indica la matrice diagonale di dimensione $n \times n$ che ha sulla diagonale principale gli elementi $v_1,...,v_n$.

Anche questa procedura è *soddisfacente* nel senso che *comunque* assegnati i dati, decide se la matrice è invertibile (utilizzando il Teorema (2.18)) e, in caso affermativo, determina la soluzione.

(2.21) Osservazione (metodo di Householder).

Esistono procedure che determinano una fattorizzazione QR di una *qualsiasi* $A \in R^{n \times n}$ (anche non invertibile). Ad esempio la seguente:

```
(U,T) = Householder(A)

  \\ A ∈ R^{n×n}
  A_1 = A;
  per k = 1,...,n-1 ripeti:
       determina X_k ∈ R^{n×n} ortogonale tale che gli elementi sotto la diagonale principale
       delle prime k colonne di X_k A_k sono nulli e pone: A_{k+1} = X_k A_k;
  T = A_n;
  U = X_1^t … X_{n-1}^t
```

La funzione predefinita qr di Scilab realizza questa procedura.

(2.22) Procedura (studio di un sistema di equazioni lineari con Householder).

// $A \in R^{n \times n}$, $b \in R^n$.

(U,T) = Householder(A);
se $t_{kk} = 0$ per qualche k allora STOP; altrimenti $x^* = SI(T,U^t\, b)$

Anche questa procedura è *soddisfacente*.

(2.1) CONDIZIONAMENTO DELLA SOLUZIONE DI UN SISTEMA DI EQUAZIONI LINEARI

Siano:

- $A \in R^{n \times n}$ invertibile, $b \in R^n$ e x^* la soluzione del sistema $A\, x = b$
- $A' \in R^{n \times n}$ invertibile, $b' \in R^n$ e $\hat{x}$ la soluzione del sistema $A'\, x = b'$

(2.23) Definizione (perturbazioni, scostamento).

Siano:

$$\delta A = A' - A \in R^{n \times n} \quad , \quad \delta b = b' - b \in R^n$$

le *perturbazioni dei dati* e:

$$\delta x = \hat{x} - x^* \in R^n$$

lo *scostamento della soluzione*.

(2.24) Problema (condizionamento della soluzione di un sistema di equazioni lineari).

Assegnato un modo di *misurare* le perturbazioni dei dati e lo scostamento della soluzione, determinare *quanto grande può essere* lo scostamento della soluzione in funzione di *quanto grandi sono* le perturbazioni dei dati.

Lezione 18 (ore 34,35) - 5 novembre 2025, 11:30 - 13:30 A13

(2.25) Definizione (norma in uno spazio vettoriale).

Sia V uno spazio vettoriale su R. Una funzione N:V → R è una *norma* in V se verifica le seguenti condizioni:

(1) per ogni $v \in V$, $N(v) \geqslant 0$ e $N(v) = 0 \Leftrightarrow v = 0$;
(2) per ogni $v \in V$ ed ogni $\alpha \in R$ si ha: $N(\alpha v) = |\alpha| N(v)$;
(3) per ogni $v,w \in V$ si ha: $N(v + w) \leqslant N(v) + N(w)$.

La coppia V,N si chiama *spazio normato*.

(2.26) Esempio (norme usuali in R^n).

Siano $V = R^n$ e $v = [v_1,\ldots,v_n] \in V$. Le funzioni:

- $N_1:R^n \to R$ definita da $N_1(v) = |v_1| + \ldots + |v_n|$
- $N_2:R^n \to R$ definita da $N_2(v) = \mathrm{sqrt}(v_1^2 + \ldots + v_n^2)$
- $N_\infty:R^n \to R$ definita da $N_\infty(v) = \max\{ |v_1|,\ldots,|v_n| \}$

sono norme in R^n.

(2.27) Esercizio (per casa).

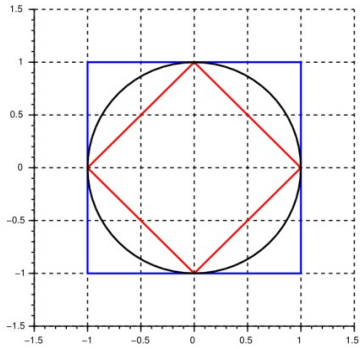
Dimostrare che le funzioni N_1 ed N_∞ verificano le proprietà della Definizione (2.25).

(2.28) Definizione (intorno sferico).

Siano R^n,N uno spazio normato, $v \in R^n$ e $r \in R$. L'insieme:

$$I_N(v,r) = \{ x \in R^n \text{ tali che } N(x - v) \leqslant r \}$$

si chiama *intorno sferico di centro* v *e raggio* r. Nella figura seguente sono rappresentati in nero l'intorno $I_2(0,1)$, in blu $I_\infty(0,1)$, in rosso $I_1(0,1)$, nel caso n = 2.



(2.29) Definizione (norma di matrice).

Siano R^n,N uno spazio normato e $A \in R^{n \times n}$. La quantità:

$$\| A \|_N = \max\{ N(A\,v),\ N(v) = 1 \}$$

si chiama *norma* N *di* A.

(2.30) Proprietà (della norma di matrice).

(I) Si osservi che la norma N di A è ben definita: il sottoinsieme S dei vettori v di R^n definito da N(v) = 1 è *chiuso e limitato* e la funzione $v \to N(A\,v)$ è *continua*. Per il Teorema di Weierstrass, quest'ultima *ha massimo e minimo* su S. In particolare:

$$\text{esiste } y \in R^n \text{ tale che } N(y) = 1 \text{ e } \| A \|_N = N(A\,y)$$

(IIa) Per ogni $A \in R^{n\times n}$ e $v \in R^n$ si ha:

$$N(A\,v) \leqslant \| A \|_N\, N(v)$$

Infatti: La relazione è certamente vera se v = 0. Se $v \neq 0$ si ha:

$$N(A\,v) = N(\ A\,N(v)\ \mathrm{vers}(v)\ )^{1} = N(\ N(v)\ A\ \mathrm{vers}(v)\ ) = N(v)\ N(A\ \mathrm{vers}(v))$$

Inoltre, per la definizione di norma N di A: $N(A\ \mathrm{vers}(v)) \leqslant \| A \|_N$, dunque:

$$N(A\,v) \leqslant \| A \|_N\, N(v)$$

(IIb) Esiste $w \in R^n$ tale che:

$$N(A\,w) = \| A \|_N\, N(w)$$

Per la proprietà (I), esiste $y \in R^n$ tale che N(y) = 1 e $\| A \|_N = N(A\,y)$. Se vers(w) = y si ha l'asserto.

(III) Per ogni $A,B \in R^{n\times n}$ si ha:

$$\| A\,B \|_N \leqslant \| A \|_N \| B \|_N$$

Infatti: per la proprietà (I) esiste $y \in R^n$ tale che N(y) = 1 e $\| A\,B \|_N = N(A\,B\,y)$. Allora, utilizzando due volte la proprietà (II):

$$\| A\,B \|_N = N(A\,B\,y) \leqslant \| A \|_N\, N(B\,y) \leqslant \| A \|_N \| B \|_N\, N(y) = \| A \|_N \| B \|_N$$

---

1 Siano R^n,N uno spazio normato e $v \in R^n$ un vettore non nullo. Allora:

$$\mathrm{vers}(v) = \frac{1}{N(v)}\ v$$

è il *versore* di v. Ovviamente si ha N(vers(v)) = 1.

(2.31) Osservazione.

L'insieme $R^{n\times n}$ è, con le usuali operazioni di somma di matrici e multiplo, uno spazio vettoriale su R. Introdotta in R^n una norma N, la funzione $A \to \| A \|_N$ da $R^{n\times n}$ in R *è una norma in* $R^{n\times n}$ (questo spiega il nome dato alla funzione). Dunque, sussistono le proprietà della norma (Definizione (2.25)):

(1) per ogni $A \in R^{n\times n}$, $\| A \|_N \geqslant 0$ e $\| A \|_N = 0 \Leftrightarrow A = 0$;
(2) per ogni $A \in R^{n\times n}$ ed ogni $\alpha \in R$ si ha: $\| \alpha A \|_N = |\alpha| \, \| A \|_N$;
(3) per ogni $A,B \in R^{n\times n}$ si ha: $\| A + B \|_N \leqslant \| A \|_N + \| B \|_N$.

(2.32) Osservazione (formule di calcolo della norma di una matrice).

Sia $A \in R^{n\times n}$ e siano $a_1,\ldots,a_n$ le colonne di A. Si ha:

- $\| A \|_1 = \max\{ N_1(a_1),\ldots,N_1(a_n) \}$
- $\| A \|_2 = \text{sqrt(massimo degli autovalori di } A^tA \,)$[2]
- $\| A \|_\infty = \| A^t \|_1$ ovvero, dette $r_1,\ldots,r_n$ le *righe* di A: $\| A \|_\infty = \max\{ N_1(r_1),\ldots,N_1(r_n) \}$

Si osservi che mentre il calcolo di $\| A \|_1$ e $\| A \|_\infty$ è elementare, quello di $\| A \|_2$ in generale *non lo è*.

(2.33) Esempio (condizionamento nel caso $\delta A = 0$, $\delta b \neq 0$).

Torniamo al condizionamento della soluzione del sistema $A x = b$. Sia N una norma in R^n.

Supponiamo che sia $\delta A = 0$ e $\delta b \neq 0$. Allora i vettori x^* e $\hat{x}$ verificano:

$$A x^* = b \quad , \quad A \hat{x} = b + \delta b$$

perciò, ricordando l'invertibilità di A, per lo scostamento δx si ha:

$$\delta x = \hat{x} - x^* = A^{-1} (b + \delta b) - A^{-1} b = A^{-1} \delta b$$

Introducendo la *misura assoluta* dello scostamento $N(\delta x)$ e quella della perturbazione $N(\delta b)$, utilizzando la proprietà (IIa) si ottiene:

$$\forall \ \delta b \ , \ N(\delta x) = N(A^{-1} \delta b) \leqslant \| A^{-1} \|_N \, N(\delta b)$$

La precedente è la *migliore limitazione possibile* per la misura assoluta dello scostamento in funzione della misura assoluta della perturbazione. La proprietà (IIb) mostra infatti che:

$$\exists \ \delta b : N(\delta x) = \| A^{-1} \|_N \, N(\delta b)$$

Se $b \neq 0$ (e quindi $x^* \neq 0$), possiamo introdurre anche le *misure relative* dello scostamento $\varepsilon_x = N(\delta x)/N(x^*)$ e della perturbazione $\varepsilon_b = N(\delta b)/N(b)$. Per tali misure si ha:

---

2 La matrice A^tA è *simmetrica* e *semidefinita positiva*. I suoi autovalori sono tutti non negativi.

$$\varepsilon_x = \frac{N(\delta x)}{N(x^*)} \leqslant \frac{\| A^{-1} \|_N \, N(\delta b)}{N(x^*)}$$

Ma:

$$A x^* = b \;\Rightarrow\; N(b) = N(A x^*) \leqslant \| A^{-1} \|_N \, N(x^*) \;\Rightarrow\; \frac{1}{N(x^*)} \leqslant \frac{\| A^{-1} \|_N}{N(b)}$$

da cui:

$$\forall\ \delta b\ ,\ \forall\ b \neq 0 : \varepsilon_x \leqslant \| A^{-1} \|_N \| A \|_N \, \varepsilon_b$$

La precedente è la *migliore limitazione possibile* per la misura relativa dello scostamento in funzione della misura relativa della perturbazione. La proprietà (IIb) mostra infatti che:

$$\exists\ \delta b \text{ e } \exists\ b \neq 0 : \varepsilon_x = \| A^{-1} \|_N \| A \|_N \, \varepsilon_b$$

(2.34) Definizione (numero di condizionamento di una matrice).

Sia $A \in R^{n \times n}$ una matrice *invertibile* e N una norma in R^n. Il numero:

$$c_N(A) = \| A^{-1} \|_N \| A \|_N$$

si chiama *numero di condizionamento di* A (in norma N).

(2.35) Osservazione.

Poiché $A^{-1} A = I$, si ha (usando la proprietà (III) di (2.30)):

$$\| I \|_N = \| A^{-1} A \|_N \leqslant \| A^{-1} \|_N \| A \|_N$$

Per definizione si ha poi:

$$\| I \|_N = \max\{ N(I v),\ N(v) = 1 \} = \max\{ N(v),\ N(v) = 1 \} = 1$$

e quindi:

$$c_N(A) = \| A^{-1} \|_N \| A \|_N \geqslant 1$$

(2.36) Teorema (di condizionamento).

Siano $A \in R^{n \times n}$ una matrice *invertibile* e N una norma in R^n. Allora: per ogni $b \neq 0$, ogni δb tale che $b + \delta b \neq 0$ e ogni δA tale che $c_N(A)\, \varepsilon_A < 1$ si ha:

$$\varepsilon_x \leqslant \frac{c_N(A)}{1 - c_N(A)\, \varepsilon_A} (\varepsilon_A + \varepsilon_b)$$

Lezione 19 (ore 36,37) - 6 novembre 2025, 8:30 - 10:30 F3

(2.36) Esercizio (svolto in classe).

Siano $V = \mathbb{R}^2$ con norma 2 e $v \in \mathbb{R}^2$ tale che $\| v \|_2 = 2$.

- Sia $x^* = v$. Disegnare l'insieme degli $\hat{x}$ tali che $\varepsilon_x \leqslant 1/4$.
- Sia $x^* = v/2$. Disegnare l'insieme degli $\hat{x}$ tali che $\varepsilon_x \leqslant 1/4$.

(2.37) Esercizio (svolto in classe).

Siano $V = \mathbb{R}^2$ con norma 2 e $v \in \mathbb{R}^2$ tale che $\| v \|_2 = 2$.

- Sia $x^* = v$. Disegnare l'insieme degli $\hat{x}$ tali che $\| \delta x \|_2 \leqslant 1/2$.
- Sia $x^* = v/2$. Disegnare l'insieme degli $\hat{x}$ tali che $\| \delta x \|_2 \leqslant 1/2$.

(2.38) Esercizio.

In $\mathbb{R}^2$ con norma 2 si siano: $x^* = [\ 2;\ 0{,}1]$ e $\hat{x}$ tali che $\varepsilon_x \leqslant L$. Determinare: $\max |\delta x_1 / x^*_1|$ e $\max |\delta x_2 / x^*_2|$.

Soluzione: $\varepsilon_x \leqslant L \Rightarrow \| \delta x \|_2 \leqslant L \| x^* \|_2$. Allora, per $k = 1,2$ si ha:

$$\max |\delta x_k / x^*_k| = \max |\delta x_k| / |x^*_k| = \max \| \delta x \|_2 / |x^*_k| \leqslant L \| x^* \|_2 / |x^*_k|$$

Dunque:

$$\max |\delta x_1 / x^*_1| \leqslant L \| x^* \|_2 / |x^*_1| = \text{sqrt}(4 + 0.01) / 2 \approx L$$

e:

$$\max |\delta x_2 / x^*_2| \leqslant L \| x^* \|_2 / |x^*_2| = \text{sqrt}(4 + 0.01) / 0{,}1 \approx 20\ L$$

Per la prima componente l'errore relativo ha una limitazione simile a quella dello scostamento; per la seconda, invece, la limitazione è *peggiore*. Questo accade perché mentre $\| x^* \|_2 / |x^*_1| \approx 1$, $\| x^* \|_2 / |x^*_2|$ è *molto maggiore* di 1.

(2.39) Osservazione.

Siano $A \in \mathbb{R}^{n \times n}$ una matrice invertibile, $b \in \mathbb{R}^n$, x^* la soluzione del sistema $A\,x = b$ e $\hat{x} \in \mathbb{R}^n$. Si usa $\hat{x}$ per approssimare x^*. Ci si domanda quanto è accurata l'approssimazione. Scelta una norma in $\mathbb{R}^n$, per misurare l'accuratezza si utilizza la quantità $N(\hat{x} - x^*)/N(x^*)$.

(A) Per ottenere informazioni sull'accuratezza, si introduce il vettore *residuo* del sistema $A\,x = b$ associato a $\hat{x}$ definito da:

$$r = A\,\hat{x} - b$$

e si *interpreta* $\hat{x}$ come soluzione del sistema perturbato:

$$A\,x = b + r$$

ottenuto con le perturbazioni $\delta A = 0$ e $\delta b = r$. Con questa interpretazione di $\hat{x}$ la quantità

$N(\hat{x} - x^*)/N(x^*)$ risulta essere la misura relativa ε_x dello scostamento della soluzione dovuto alla perturbazione. Applicando il Teorema di condizionamento (2.36) della Lezione 18 si ottiene la limitazione:

$$N(\hat{x} - x^*)/N(x^*) = \varepsilon_x \leqslant c_N(A)\, \varepsilon_b \quad \text{con} \quad \varepsilon_b = N(r)/N(b)$$

(B) Per ottenere informazioni sull'accuratezza, si cerca una matrice $M \in R^{n \times n}$ tale che:

$$M\,\hat{x} = -r$$

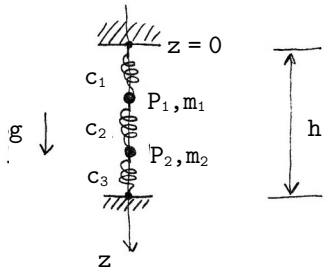
e, posto $\delta A = M$ si *interpreta* $\hat{x}$ come soluzione del sistema perturbato:

$$(A + \delta A)\, x = b$$

Con questa interpretazione di $\hat{x}$ la quantità $N(\hat{x} - x^*)/N(x^*)$ risulta essere la misura relativa ε_x dello scostamento della soluzione dovuto alla perturbazione. Se $c_N(A)\, \varepsilon_A < 1$, dal Teorema di condizionamento (2.36) della Lezione 18 si ottiene la limitazione:

$$N(\hat{x} - x^*)/N(x^*) = \varepsilon_x \leqslant c_N(A)\, \varepsilon_A \,/\, (1 - c_N(A)\, \varepsilon_A)$$

(2.40) Esempio.



Si consideri il sistema di figura, composto da due punti pesanti, P_1 di massa m_1, e P_2 di massa m_2, liberi di scorrere lungo una guida verticale e collegati da tre molle ideali e con lunghezza a riposo 0 come nel disegno.

Scelto l'asse z verticale discendente, per determinare le configurazioni di equilibrio, per ciascuno dei punti si scrivono le equazioni della statica:

$$m_1 g - c_1 z_1 + c_2 (z_2 - z_1) = 0$$
$$m_2 g - c_2 (z_2 - z_1) + c_3 (h - z_2) = 0$$

che, sotto forma di sistema, si riscrivono:

$$\begin{bmatrix} c_1 + c_2 & -c_2 \\ -c_2 & c_2 + c_3 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} m_1 g \\ m_2 g + c_3 h \end{bmatrix}$$

$$A \qquad z \quad = \quad b$$

Scelti i valori dei parametri:

$$c_1 = c_2 = c_3 = 100 \text{ N/m} \quad , \quad m_1 = m_2 = 1 \text{ kg} \quad , \quad h = 5 \text{ m} \quad , \quad g = 9.81 \text{ m/s}^2$$

la soluzione z^* del sistema è:

$$z_1^* \approx 1.76 \text{ m} \quad , \quad z_2^* \approx 3.43 \text{ m}$$

Se adesso assumiamo come valore dell'accelerazione di gravità un valore g' tale che:[1]

$$|g' - g| = |\,\delta g\,| < 10^{-2}$$

1 Si ricordi che il valore dell'accelerazione di gravità è noto solo con approssimazione.

il sistema A z = b si trasforma nel *sistema perturbato* A z = b + δb con:

$$\delta b = [\ m_1\, \delta g\ ;\ m_2\, \delta g\ ]$$

Scelta poi la *norma uno* in R^2 si ha:

$$\varepsilon_b = N_1(\delta b)/N_1(b) < 4 \times 10^{-5} \qquad e \qquad c_1(A) = 3$$

In base al Teorema di condizionamento, per lo scostamento della soluzione $\hat{z}$ del sistema perturbato dalla soluzione z^* si ha la limitazione:

$$\varepsilon_x \leqslant c_N(A)\, \varepsilon_b < 1.2 \times 10^{-4}$$

Infine, essendo:

$$\| z^* \|_1 / |z_1^*| \approx 3 \qquad e \qquad \| z^* \|_1 / |z_2^*| \approx 1.5$$

si ottengono *stime simili* anche per quanto riguarda lo scostamento delle componenti (vedere l'Esercizio (2.38)).

Lezione 20 (ore 38,39,40) - 11 novembre 2025, 15:30 - 18:30 F3

(2.41) Esempio (continuazione).

Supponiamo che le costanti elastiche c_k siano note con incertezza. Assumiamo, ad esempio, che, per k = 1,2,3, sia:

$$c_k' = c_k + \delta c_k \qquad \text{con} \qquad |\delta c_k| < 1 \text{ N/m}$$

Il sistema A z = b si trasforma nel *sistema perturbato* $(A + \delta A)\, z = b + \delta b$ con:

$$\delta A = \begin{bmatrix} \delta c_1 + \delta c_2 & -\delta c_2 \\ -\delta c_2 & \delta c_2 + \delta c_3 \end{bmatrix} \quad , \quad \delta b = \begin{bmatrix} 0 \\ h\, \delta c_3 \end{bmatrix}$$

Per le perturbazioni dei dati si ha:

$$\varepsilon_A = N_1(\delta A)/N_1(A) < 10^{-2} \quad , \quad \varepsilon_b = N_1(\delta b)/N_1(b) < 10^{-2}$$

Inoltre:

$$c_1(A)\, \varepsilon_A \leqslant 3 \times 10^{-2}$$

In base al Teorema di condizionamento, per lo scostamento della soluzione $\hat{z}$ del sistema perturbato dalla soluzione z^* si ha la limitazione:

$$\varepsilon_x \leqslant \frac{c_1(A)}{1 - c_1(A)\, \varepsilon_A} (\varepsilon_A + \varepsilon_b) \approx 6.2 \times 10^{-2}$$

Per quanto riguarda lo scostamento delle componenti si ha, questa volta:

$$\varepsilon_{x,1} \leqslant 0.19 \ (19\,\%) \quad , \quad \varepsilon_{x,2} \leqslant 0.09 \ (9\,\%)$$

(2.42) Esempio (continuazione).

Sia adesso $\hat{z}$ una colonna (ad esempio ottenuta dal calcolatore utilizzando una procedura per la soluzione del sistema A z = b) da usare come *approssimazione* di z^*. Per ottenere una limitazione dell'errore commesso si procede come nell'Osservazione (2.39) della Lezione 19.

Il vettore residuo è:

$$r = A\, \hat{z} - b$$

(1) Si interpreta $\hat{z}$ come soluzione del sistema perturbato A z = b + r. Per il Teorema di condizionamento:

$$\frac{N_1(\hat{z} - z^*)}{N_1(z^*)} \leqslant c_1(A)\, \frac{N_1(r)}{N_1(b)}$$

Domanda: esistono perturbazioni dei parametri δg, δc_k, δm_k, δh che generano perturbazioni dei dati $\delta A = 0$ e $\delta b = r$ (ovvero: si riesce ad 'interpretare fisicamente' il sistema perturbato A z = b + r) ?

Osservazione: la limitazione trovata è valida *indipendentemente dalla risposta* alla domanda: il sistema perturbato *non deve necessariamente essere fisicamente significativo*.

Risposta: si. Ad esempio: $\delta g = 0$, $\delta c_k = 0$, $\delta h = 0$ e $\delta m_1 = r_1/g$, $\delta m_2 = r_2/g$.

(2) Si cerca $M \in R^{2\times 2}$ tale che $M\hat{z} = -r$, e si interpreta $\hat{z}$ come soluzione del sistema perturbato $(A + M)\, z = b$. Per il Teorema di condizionamento, posto $\varepsilon_A = \| M \|_1 / \| A \|_1$:

$$\text{se} \quad c_1(A)\varepsilon_A \leqslant 1 \text{ allora } \frac{N_1(\hat{z}-z^*)}{N_1(z^*)} \leqslant \frac{c_1(A)\varepsilon_A}{1-c_1(A)\varepsilon_A}$$

Domanda: esistono perturbazioni dei parametri δg, δc_k, δm_k, δh che generano perturbazioni dei dati $\delta A = M$ e $\delta b = 0$ (ovvero: si riesce ad 'interpretare fisicamente' il sistema perturbato $A\, z = b + r$) ?

(2.43) Esempio.

Sia:

$$\hat{z} = \begin{bmatrix} 1.8 \\ 3.4 \end{bmatrix} \text{ m}$$

Allora:

$$r = A\,\hat{z} - b = \begin{bmatrix} 10.19 \\ -9.81 \end{bmatrix} \text{ N}$$

Cerchiamo α e β in modo che, posto:

$$M = \begin{bmatrix} \alpha+\beta & -\beta \\ -\beta & \beta \end{bmatrix}$$

si abbia:

$$M\hat{z} = -r$$

Si ottiene un sistema di due equazioni nelle incognite α e β la cui unica soluzione è:

$$\alpha = -(r_1 + r_2)/\hat{z}_1 \approx -0.21 \text{ N/m} \qquad \text{e} \qquad \beta = -r_2/(\hat{z}_2 - \hat{z}_1) \approx -6.13 \text{ N/m}$$

Si ottiene allora:

$$\varepsilon_A \approx 4.1 \times 10^{-2} \quad \text{e} \quad c_1(A)\,\varepsilon_A \approx 0.12 < 1$$

da cui, per il Teorema di condizionamento:

$$\varepsilon_x \leqslant \text{circa } 0.14$$

Infine, la risposta è si: $\delta g = 0$, $\delta m_k = 0$, $\delta h = 0$ e $\delta c_1 = \alpha$ N/m, $\delta c_2 = \beta$ N/m, $\delta c_3 = 0$.

(2.2) STUDIO DI UN SISTEMA DI EQUAZIONI LINEARI IN F(β,m)

(2.44) Osservazione (studio con EGP).

Siano $A \in R^{n \times n}$ e $b \in R^n$. Il procedimento per lo studio del sistema $A\,x = b$ che usa la procedura EGP è:

```
(S,D,P) = EGP(A);
      se esiste k tale che d_kk = 0 allora STOP;
        altrimenti
             c = SA(S,Pb);
             x* = SI(D,c)
```

in R

Quando si utilizza un calcolatore, con insieme di numeri di macchina F(β,m), la procedura si trasforma in:

```
(Ŝ,D̂,P̂) = EGP_M(Â);
      se esiste k tale che d̂_kk = 0 allora STOP;
        altrimenti
             ĉ = SA_M(S,Pb̂);
             x̂ = SI_M(D,c)
```

in F(β,m)

dove:

- EGP_M, SA_M e SI_M sono, rispettivamente, la procedura EGP, SA ed SI in cui ciascuna operazione aritmetica è sostituita dalla corrispondente funzione predefinita,
- $\hat{A}$ e $\hat{b}$ sono, rispettivamente, la matrice rd(A) e la colonna rd(b) di elementi gli arrotondati in F(β,m) dei corrispondenti elementi di A e b.

(2.45) Esempio.

Ricordando il Teorema (1.38) della Lezione 6, per ciascuna componente della matrice $\hat{A} = rd(A)$ e della colonna $\hat{b} = rd(b)$ si ha:

$$\hat{a}_{ij} = rd(a_{ij}) = (1 + \varepsilon_{ij})\, a_{ij} \quad , \quad \hat{b}_i = rd(b_i) = (1 + \varepsilon_i)\, b_i$$

con $|\varepsilon_{ij}| \leqslant u$ e $|\varepsilon_i| \leqslant u$ per ogni i e j. Ne segue che, utilizzando ad esempio la norma uno in R^n, per le misure assolute delle perturbazioni si ha:

$$\| \delta A \|_1 \leqslant u \| A \|_1 \quad , \quad N_1(\delta b) \leqslant u\, N_1(\delta b)$$

e quindi, per le misure relative:

$$\varepsilon_A \leqslant u \quad \text{e} \quad \varepsilon_b \leqslant u$$

Se $c_1(A)\, u \leqslant 1$ allora $c_1(A)\, \varepsilon_A \leqslant 1$ e, per il Teorema di condizionamento (Teorema (2.36) della Lezione 18) si ha:

$$\varepsilon_x \leqslant 2\,\frac{c_1(A)\,u}{1-c_1(A)\,u} \equiv \Lambda$$

Quando il calcolatore *legge i dati* A e b, li cambia (salvo il caso in cui le componenti dei dati siano in $F(\beta,m)$) e il sistema $A\,x = b$ è sostituito dal sistema $\hat{A}\,x = \hat{b}$. Questa sostituzione, nel caso migliore possibile in cui sia trascurabile l'effetto delle sostituzioni di EGP, SA ed SI con EGP_M, SA_M e SI_M, *può generare* uno scostamento della soluzione x^* di misura relativa Λ. Dunque, nel caso usuale in cui l'effetto delle sostituzioni di EGP, SA ed SI con EGP_M, SA_M e SI_M non è trascurabile, *non è ragionevole* aspettarsi uno scostamento tra x^* e l'approssimazione $\hat{x}$ ottenuta dal calcolatore *minore di* Λ.

(2.46) Esempio.

Si consideri la seguente situazione 'quasi ideale':

- $\hat{A} = A$, $\hat{b} = b$ - i dati hanno componenti in $F(\beta,m)$);
- $EGP_M(A) = EGP(A) = (S,D,P)$ - la fattorizzazione EGP_M è esatta, con D invertibile;
- $SA_M(S,Pb) = \hat{c} = rd(c)$ - il risultato di SA_M è 'quasi ideale';
- $SI_M(D,\hat{c}) = SI(D,\hat{c})$ - il risultato di SI_M è esatto.

Sotto queste ipotesi si ha: $x^* = SI(D,c)$ è la soluzione del sistema $D\,x = c$, $\hat{x} = SI(D,\hat{c})$ è la soluzione del sistema $D\,x = \hat{c}$. Introdotta la perturbazione $\delta c = \hat{c} - c$ si ha, utilizzando la norma uno (vedi l'esempio precedente):

$$N_1(\delta c) \leqslant u\,N_1(c) \qquad \text{e quindi} \qquad \varepsilon_c \leqslant u$$

Per il Teorema di condizionamento si ha allora:

$$\varepsilon_x \leqslant c_1(D)\,\varepsilon_c \leqslant c_1(D)\,u$$

La limitazione della misura relativa dello scostamento dipende da $c_1(D)$ ovvero, posto:

$$c_1(D) = c_1(A)\,\frac{c_1(D)}{c_1(A)}$$

dal *fattore di amplificazione del numero di condizionamento* $c_1(D)/c_1(A)$.

(2.47) Esempio.

Siano $\gamma \in (0,1)$ e $A = \begin{bmatrix} \gamma & 1 \\ 1 & 0 \end{bmatrix}$. Si ha:

- $\| A \|_1 = 1 + \gamma < 2$
- $A^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & -\gamma \end{bmatrix}$ da cui $\| A^{-1} \|_1 = 1 + \gamma$ e $c_1(A) = (1 + \gamma)^2 < 4$
- $EGP(A) = (S,D,P) = \left(\begin{bmatrix} 1 & 0 \\ 1/\gamma & 1 \end{bmatrix}, \begin{bmatrix} \gamma & 1 \\ 0 & -1/\gamma \end{bmatrix}, I \right)$ e $\| D \|_1 = 1 + 1/\gamma$
- $D^{-1} = \begin{bmatrix} 1/\gamma & 1 \\ 0 & -\gamma \end{bmatrix}$ da cui $\| D^{-1} \|_1 = \max\{ 1/\gamma, 1 + \gamma \}$ e $c_1(D) = (1 + 1/\gamma)\max\{ 1/\gamma, 1 + \gamma \}$

Per il fattore di amplificazione del numero di condizionamento si ha allora:

$$\lim_{\gamma \to 0} \frac{c_1(D)}{c_1(A)} = +\infty$$

Dunque: scegliendo γ *sufficientemente piccolo* è possibile ottenere un fattore di amplificazione del numero di condizionamento *grande quanto si vuole*: il procedimento di soluzione del sistema di equazioni che usa EGP trasforma il sistema A x = b nel sistema *equivalente* D x = c ma mentre le proprietà di condizionamento di A sono *buone* ($c_1(A) < 4$) quelle di D, scelto γ opportunamente piccolo, sono *pessime* ($c_1(D)$ enorme).

Mentre il procedimento di soluzione del sistema di equazioni che usa EGP è *soddisfacente* quando si opera in R (si veda (2.16) della Lezione 17), il procedimento può risultare *non soddisfacente* quando si opera in F(β,m).

(2.48) Definizione (procedura EGPP).

Per ovviare al potenziale pericolo evidenziato nell'esempio precedente, si ricorre ad una modifica della procedura EGP che porta alla definizione della procedura EGPP (Eliminazione di Gauss con Pivoting Parziale). La differenza con EGP consiste *solo* nella *scelta della matrice di permutazione* P_k. Nella procedura EGP si ha:

se $A_k(k,k) \neq 0$ allora $P_k = I$ altrimenti
se esiste $i > k$ tale che $A_k(i,k) \neq 0$ allora $P_k = P_{k,i}$ altrimenti $P_k = I$

Nella procedura EGPP si pone:

se per ogni $i \geqslant k$ si ha $A_k(i,k) = 0$ allora $P_k = I$ altrimenti
scelto i tale che $|A_k(i,k)| = \max \{ |A_k(j,k)| , j \geqslant k \}$ si pone $P_k = P_{k,i}$

La scelta nella procedura EGP ha lo scopo di assicurarsi che il pivot sia *diverso da zero*, nella procedura EGPP lo scopo è quello di avere come pivot *l'elemento della colonna* k*-esima di modulo massimo possibile* tra tutti quelli con indice di riga $j \geqslant k$.

(2.49) Esempio.

Calcolo di EGPP(A) con:

```
A = [ 1, 0, 1;
      2, 1,-1;
      1, 2, 1 ]
```

(*) $A_1 = A$;

(*) $k = 1$; $|A_1(2,1)| = \max \{ |A_1(j,1)| , j \geqslant 1 \} \Rightarrow P_1 = P_{1,2}$;

$$T_1 = P_1 A_1 = \begin{bmatrix} 2 & 1 & -1 \\ 1 & 0 & 1 \\ 1 & 2 & 1 \end{bmatrix} \quad , \quad H_1 = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_2 & 1 & 0 \\ \lambda_3 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{bmatrix}$$

I valori λ_2, λ_3 sono determinati come nella procedura EGP.

Infine:

$$H_1 T_1 = \begin{bmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & -1 \\ 1 & 0 & 1 \\ 1 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 & -1 \\ 0 & -1/2 & 3/2 \\ 0 & 3/2 & 3/2 \end{bmatrix} = A_2$$

(*) $k = 2$; $|A_2(3,2)| = \max \{ |A_2(j,2)| \ , \ j \geqslant 2 \} \Rightarrow P_2 = P_{2,3}$;

$$T_2 = P_2 A_2 = \begin{bmatrix} 2 & 1 & -1 \\ 0 & 3/2 & 3/2 \\ 0 & -1/2 & 3/2 \end{bmatrix} \quad , \quad H_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_3 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/3 & 1 \end{bmatrix}$$

Il valore λ_3 è determinato come nella procedura EGP.

Infine:

$$H_2 T_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/3 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & -1 \\ 0 & 3/2 & 3/2 \\ 0 & -1/2 & 3/2 \end{bmatrix} = \begin{bmatrix} 2 & 1 & -1 \\ 0 & 3/2 & 3/2 \\ 0 & 0 & 2 \end{bmatrix} = A_3$$

(*) $D = A_3$; $P = P_2 P_1$; $S = \begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/2 & -1/3 & 1 \end{bmatrix}$ (ricavata come in EGP).

(2.50) Osservazione.

Per ogni $A \in R^{n \times n}$ invertibile, posto (S,D,P) = EGPP(A), si ha: $c_1(D)/c_1(A) \leqslant F(n)$. La funzione F dipende *solo* dalla dimensione n della matrice e dalla norma scelta, in particolare *non dipende* da A. Dunque, il fattore di crescita del numero di condizionamento è *limitato*.

Tornando all'Esempio (2.47) si ha:

$$\text{EGPP}\left(\begin{bmatrix} \gamma & 1 \\ 1 & 0 \end{bmatrix} \right) = (S,D,P) \text{ con } D = I \Rightarrow c_1(D) = 1$$

(2.51) Osservazione (studio con qr).

Siano $A \in R^{n \times n}$ e $b \in R^n$. Il procedimento per lo studio del sistema $Ax = b$ che usa la procedura qr è:

(U,T) = qr(A);
se esiste k tale che $t_{kk} = 0$ allora STOP;
altrimenti
$c = U^t b$;
$x^* = SI(D,c)$

in R

Quando si utilizza un calcolatore, con insieme di numeri di macchina $F(\beta,m)$, la procedura si trasforma in:

$(\hat{U},\hat{T}) = qr_M(\hat{A})$;
  se esiste k tale che $\hat{t}_{kk} = 0$ allora STOP;
    altrimenti
      $\hat{c} = \hat{U}^t \otimes \hat{b}$;
      $\hat{x} = SI_M(\hat{T},\hat{c})$

in $F(\beta,m)$

dove $\hat{U}^t \otimes b$ è la colonna che si ottiene sostituendo in $U^t b$ le operazioni aritmetiche con le corrispondenti funzioni predefinite in $F(\beta,m)$.

(2.52) Esempio.

Analogamente a quanto fatto per il procedimento che usa EGP, si consideri la seguente situazione 'quasi ideale':

- $\hat{A} = A$, $\hat{b} = b$ - i dati hanno componenti in $F(\beta,m)$);
- $qr_M(A) = qr(A) = (U,T)$ - la fattorizzazione qr_M è esatta, con T invertibile;
- $U^t \otimes b = \hat{c} = rd(c)$ - il risultato di $U^t \otimes b$ è 'quasi ideale';
- $SI_M(T,\hat{c}) = SI(T,\hat{c})$ - il risultato di SI_M è esatto.

Sotto queste ipotesi si ha: $x^* = SI(T,c)$ è la soluzione del sistema $T\,x = c$, $\hat{x} = SI(T,\hat{c})$ è la soluzione del sistema $T\,x = \hat{c}$. Introdotta la perturbazione $\delta c = \hat{c} - c$ si ha, utilizzando la norma due (la norma 'naturale' da utilizzare in R^n quando si utilizza la fattorizzazione QR che fa entrare in gioco la nozione di ortogonalità, dunque il prodotto scalare in R^n, è la norma due: quella indotta dal prodotto scalare):

$$N_2(\delta c) \leqslant u\, N_2(c) \qquad \text{e quindi} \qquad \varepsilon_c \leqslant u$$

Per il Teorema di condizionamento si ha ancora:

$$\varepsilon_x \leqslant c_2(T)\, \varepsilon_c \leqslant c_2(T)\, u$$

e la limitazione della misura relativa dello scostamento dipende dal fattore di amplificazione del numero di condizionamento $c_2(T)/c_2(A)$.

Però in questo caso si ha:

- $A = U\,T \Rightarrow \| A \|_2 = \| U\,T \|_2 = \max \{ N_2(U\,T\,v),\ N_2(v) = 1 \} =^1 \max \{ N_2(T\,v),\ N_2(v) = 1 \} = \| T \|_2$
- $T^{-1} = A^{-1}\,U \Rightarrow \| T^{-1} \|_2 = \| A^{-1}\,U \|_2 = \max \{ N_2(\,A^{-1}\,U\,v),\ N_2(v) = 1 \} =^2 \max \{ N_2(A^{-1}\,w),\ N_2(U^t\,w) = 1 \} = \max \{ N_2(A^{-1}\,w),\ N_2(w) = 1 \} = \| A^{-1} \|_2$

Ne segue che $c_2(T) = c_2(A)$, ovvero il fattore di amplificazione del numero di condizionamento è $c_2(T)/c_2(A) = 1$.

Il procedimento di soluzione del sistema di equazioni che usa qr è soddisfacente *anche* quando si opera in $F(\beta,m)$.

---

1 Poiché U è ortogonale si ha: $N_2(U\,T\,v) = \text{sqrt}(\ v^t\,T^t\,U^t\,U\,T\,v\ ) = \text{sqrt}(\ v^t\,T^t\,T\,v\ ) = N_2(T\,v)$.

2 Cambio di variabile: $w = U\,v$. Essendo U ortogonale si ha poi $v = U^t\,w$ e $N_2(U^t\,w) = N_2(w)$.

(2.3) COSTO DELLA SOLUZIONE DI UN SISTEMA DI EQUAZIONI LINEARI

(2.53) Definizione (costo aritmetico).

Un metodo per confrontare i due procedimenti descritti per ottenere un'approssimazione della soluzione di un sistema di equazioni lineari (quello che usa EGPP e quello che usa qr) è di considerare il tempo necessario per il calcolo dell'approssimazione.

Nel contesto della risoluzione dei sistemi di equazioni lineari, si introduce la seguente nozione di *costo* del calcolo di $\varphi(x)$, $C(\varphi)$, dove φ è l'*algoritmo ingenuo* (si veda Definizione (1.32), Lezione 6) per f:

$C(\varphi)$ = il numero di operazioni aritmetiche necessario per calcolare f

(2.54) Osservazione (ragionevolezza della definizione di costo).

Perché $C(\varphi)$ sia indicativo del *tempo* necessario per il calcolo di $\varphi(x)$ è necessario che siano soddisfatte le seguenti due condizioni:

(1) Durante il calcolo di $\varphi(x)$, il tempo impiegato in attività che *non* siano l'esecuzione di operazioni aritmetiche (ovvero: nel calcolo di funzioni predefinite corrispondenti a funzioni elementari o confronti) *deve essere trascurabile* (un esempio di algoritmo in cui questa condizione *non è* verificata è quello che calcola la norma infinito di un vettore: in questo caso l'algoritmo esegue solo *confronti* tra le componenti del vettore);
(2) Il tempo di calcolo di ciascuna delle funzioni predefinite corrispondenti ad operazioni aritmetiche deve essere *indipendente dagli operandi.*

La seconda condizione *non è verificata*, ad esempio, nel caso della moltiplicazione tra due elementi di $F(\beta,m)$: per calcolare $\xi_1 \otimes \xi_2$ occorre *moltiplicare le frazioni* - e questo avviene in un tempo indipendente dai fattori perché le frazioni hanno sempre *lo stesso numero di cifre* - e *sommare gli esponenti*; è quest'ultima operazione che non può essere ritenuta indipendente dai fattori perché gli esponenti sono numeri interi qualsiasi che hanno *un numero di cifre che dipende da quali elementi di* $F(\beta,m)$ *si considerano.* In particolare, perché la nozione di costo aritmetico sia indicativa del tempo necessario per il calcolo occorre che *l'insieme dei numeri di macchina del calcolatore non sia* $F(\beta,m)$.

Lezione 21 (ore 41,42) - 12 novembre 2025, 11:30 - 13:30 A13

(2.55) Osservazione.

Analizziamo il costo del procedimento di soluzione del sistema $A\,x = b$, con $A \in R^{n\times n}$ invertibile, che utilizza la procedura EGPP. Le procedure eseguite sono EGPP, SA ed SI. Si ha:

$$C(EGPP) = \frac{2}{3}n^3+\ldots \quad , \quad C(SA) = C(SI) = n^2$$

Si osservi che mentre nel calcolo di SA ed SI si eseguono *solo* operazioni aritmetiche, nel calcolo di EGPP si eseguono *anche* confronti, ma il loro numero è *trascurabile* rispetto a quello delle operazioni aritmetiche.[1]

(Esercizio: determinare il numero di confronti eseguito da EGPP.)

Complessivamente:

$$C(EGPP) + C(SA) + C(SI) = \frac{2}{3}n^3+\ldots$$

Nel procedimento che utilizza la procedura qr si eseguono le procedure qr, prodotto matrice per colonna (indicato con: pmc) e SI. Si ha:

$$C(qr) = \frac{4}{3}n^3+\ldots \quad , \quad C(pmc) = 2\,n^2 + \ldots \quad , \quad C(SI) = n^2$$

Si osservi che nella procedura qr (come in quella GS) si esegue anche il calcolo di radici quadrate ma il loro numero (n) è *trascurabile* rispetto a quello delle operazioni aritmetiche.

(Esercizio: determinare il numero di operazioni aritmetiche eseguito da pmc.)

Complessivamente:

$$C(qr) + C(pmc) + C(SI) = \frac{4}{3}n^3+\ldots$$

Il termine dominante nel costo aritmetico della procedura che usa qr è dunque *doppio* rispetto a quello della procedura che usa EGPP.

1 Si può ragionevolmente ritenere che il tempo necessario per confrontare due numeri di macchina sia *simile* a quello necessario per eseguire un'operazione aritmetica sugli stessi numeri.

(2.4) METODI ITERATIVI PER LA SOLUZIONE DI UN SISTEMA DI EQUAZIONI LINEARI

(2.56) Definizione (metodo iterativo per la soluzione di un sistema di equazioni lineari).

Siano $H \in R^{n\times n}$ e $c \in R^n$. Il *metodo iterativo definito da* H *e* c è l'applicazione che a ciascun vettore $g \in R^n$ associa la successione di vettori x(k) definita da:

$$x(0) = g \quad , \quad x(k) = H\,x(k-1) + c \quad \text{per } k = 1,2,\ldots$$

(2.57) Osservazione.

- Il metodo iterativo definito da H e c è il metodo iterativo definito dalla funzione $h: R^n \to R^n$ tale che:

  $$h(x) = H\,x + c$$

  La funzione h è *continua* perciò (si veda l'Osservazione (1.54) nella Lezione 8) se la successione x(k) generata dal metodo è convergente, allora il suo limite $v \in R^n$ è un punto unito di h, ovvero verifica la relazione:

  $$v = H\,v + c \quad \text{equivalente a} \quad (I - H)\,v = c$$

  e quest'ultima relazione significa che:

  v è *soluzione* del sistema di equazioni lineari $(I - H)\,x = c$

- Sia $A \in R^{n\times n}$ invertibile. Il metodo iterativo definito da H e c è *utilizzabile* per approssimare la soluzione del sistema A x = b se:

  (1) i sistemi A x = b e $(I - H)\,x = c$ sono *equivalenti* (in particolare: I - H è invertibile) e

  (2) è (praticamente) possibile determinare $g \in R^n$ a partire dal quale la successione generata dal metodo è convergente.

(2.58) Esempio.

(1) Siano:

$$A = \begin{bmatrix} 1/2 & 0 \\ 0 & -1 \end{bmatrix} \quad , \quad b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- Posto: $H = I - A = \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix}$ e c = b, i sistemi A x = b e $(I - H)\,x = c$ sono equivalenti;
- Sia $g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$. La successione generata dal metodo definito da H e c è allora:

  $x(0) = g \quad , \quad x(1) = H\,x(0) + c = H\,g \quad , \quad x(2) = H\,x(1) + c = H^2\,g \quad , \quad \ldots$

  e quindi:

$$x(k) = H^k g = \begin{bmatrix} (1/2)^k & 0 \\ 0 & 2^k \end{bmatrix} g = \begin{bmatrix} (1/2)^k g_1 \\ 2^k g_2 \end{bmatrix}$$

La successione è convergente *se e solo se* $g_2 = 0$. In tal caso si ha:

$$\lim_{k \to \infty} x(k) = 0$$

e la successione converge all'unica soluzione del sistema $A\,x = b$.

(2) Siano:

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \quad , \quad b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- Posto: $J = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ si riscrive $A = 2\,I + J$. Allora:

$$A\,x = b \quad \text{è equivalente a} \quad x = -\,(1/2)\,J\,x + (1/2)\,b$$

ovvero, posto $H = -(1/2)\,J$ e $c = (1/2)\,b$:

$$A\,x = b \quad \text{è equivalente a} \quad (I - H)\,x = c$$

- Gli autovalori della matrice H sono: $\lambda_1 = -\,1/2$ e $\lambda_2 = 1/2$, quindi H è diagonalizzabile.[2] Si ha:

$$H = S \begin{bmatrix} -1/2 & 0 \\ 0 & 1/2 \end{bmatrix} S^{-1} \quad \text{con} \quad S = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

- Posto $g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$, la successione generata dal metodo definito da H e c è:

$$x(k) = H^k g = S \begin{bmatrix} (-1/2)^k & 0 \\ 0 & (1/2)^k \end{bmatrix} S^{-1} g$$

ovvero, posto $y = S^{-1} g$:

$$x(k) = \begin{bmatrix} (-1/2)^k y_1 \\ (1/2)^k y_2 \end{bmatrix}$$

In questo caso si ha:

$$\text{per ogni } g \in R^2 : \lim_{k \to \infty} x(k) = 0$$

ovvero: per ogni $g \in R^2$ la successione converge all'unica soluzione del sistema $A\,x = b$.

---

2 Si ricordi che (1) una matrice $M \in R^{n \times n}$ è diagonalizzabile se esistono una matrice diagonale Λ e una matrice invertibile S tali che: $M\,S = S\,\Lambda$, ovvero $M = S\,\Lambda\,S^{-1}$; gli elementi $\lambda_1, \ldots, \lambda_n$ sulla diagonale di Λ sono gli autovalori di M, la k-esima colonna di S è un autovettore associato all'autovalore λ_k; (2) se una matrice ha autovalori distinti allora è diagonalizzabile.

(3) Siano $A = -I$ e $b = 0$.

- Posto $H = I - A$ e $c = b$, i sistemi $A\,x = b$ e $(I - H)\,x = c$ sono equivalenti.
- La successione generata dal metodo iterativo definito da H e c a partire da $g \in R^n$ è:

$$x(k) = H^k g = 2^k g$$

La successione è convergente *se e solo se* $g = 0$ e, in tal caso, converge all'unica soluzione del sistema $A\,x = b$.

Lezione 22 (ore 43,44) - 13 novembre 2025, 8:30 - 10:30 F3

(2.59) Esempio (numeri di macchina di Scilab).

In *Scilab* l'insieme dei numeri di macchina è:

$$M = F_d(2,53,-1021,1024)$$

ovvero l'insieme dei numeri in virgola mobile, base due, precisione 53, *esponente limitato* (tra -1021 e 1024) e *con elementi denormalizzati*.

Gli elementi di M sono:

- zero;
- gli *elementi normalizzati*:

  $$(-1)^s\ 2^b\ 0.c_1 \ldots c_{53}$$

  con $s \in \{0,1\}$, $-1021 \leqslant b \leqslant 1024$, ogni c_k cifra in base due e $c_1 \neq 0$;
- gli *elementi denormalizzati*:

  $$(-1)^s\ 2^{-1024}\ 0.c_1 \ldots c_{53}$$

  con $s \in \{0,1\}$, ogni c_k cifra in base due e $c_1 = 0$.

L'insieme M ha un numero *finito* di elementi. Inoltre:

- $\max M = \xi_{max} = 2^{1024}\ 0.1\cdots1 = 2^{1024}\ (1 - 2^{-53})$
- $\min\{\ \xi \in M,\ \xi > 0\ \} = \xi_{min} = 2^{-1021}\ 0.0\cdots01 = 2^{-1021}\ 2^{-53} = 2^{-1074}$
- il *successore di zero* è definito e: $\sigma(0) = \xi_{min}$
- $\min\{\ \xi \in M,\ \xi > 0$ e ξ normalizzato $\} = 2^{-1021}\ 0.10\cdots0 = 2^{-1021}\ 2^{-1} = 2^{-1022}$
- M contiene *elementi simbolici*: Nan (utilizzato quando al risultato di una funzione predefinita non è assegnabile un valore numerico 'sensato'), Inf (quando una funzione predefinita restituisce un valore numerico positivo 'troppo grande'), -Inf (quando una funzione predefinita restituisce un valore numerico negativo 'troppo grande'); in *Scilab* le costanti %nan e %inf hanno valore, rispettivamente, Nan e Inf
- detta $rd:R \to M$ l'usuale funzione arrotondamento in M, la funzione $rd^*:R \to M$ che *Scilab* utilizza per arrotondare i numeri reali è definita così:

  se $|rd(x)| \leqslant \xi_{max}$ allora $rd^*(x) = rd(x)$

  se $rd(x) > \xi_{max}$ allora $rd^*(x) = Inf$

  se $rd(x) < -\xi_{max}$ allora $rd^*(x) = -Inf$

La funzione predefinita *number_properties* di *Scilab* restituisce informazioni sull'insieme M. Precisamente:

number_properties(<stringa>)

restituisce:

- la *base* dell'insieme M quando <stringa> = 'radix'
- la *precisione* dell'insieme M quando <stringa> = 'digits'

- l'*esponente minimo* dell'insieme M quando <stringa> = 'minexp'
- l'*esponente massimo* dell'insieme M quando <stringa> = 'maxexp'
- la presenza di *elementi denormalizzati* quando <stringa> = 'denorm'
- il *massimo* elemento di M quando <stringa> = 'huge'
- il *minimo* elemento *positivo* di M quando <stringa> = 'tiniest'
- il *minimo* elemento *positivo normalizzato* di M quando <stringa> = 'tiny'
- la *precisione di macchina* in M quando <stringa> = 'eps'

La funzione predefinita *log2* di Scilab restituisce la frazione e l'esponente di un elemento di M. Precisamente, se $\xi = (-1)^s\ 2^b\ g$, l'assegnamento:

```
[f,e] = log2(ξ)
```

assegna ad f il valore $(-1)^s\ g$ e ad e il valore b.

La funzione predefinita *nearfloat* di Scilab restituisce il predecessore o il successore di un elemento di M. Precisamente:

```
nearfloat(<stringa>,ξ)
```

restituisce:

- il *successore* di ξ quando <stringa> = 'succ'
- il *predecessore* di ξ quando <stringa> = 'pred'

(2.60) Esercizio (per casa).

Eseguire e discutere (utilizzando opportune rappresentazioni grafiche) i seguenti dialoghi in *Scilab*:

```
> xi_min = number_properties('tiniest')
> xi_min == 2^(-1074)
> [f,e] = log2(xi_min)
> y = xi_min / 2
> y == 0
> z = 2^(-1075) * (3 / 2)
> z == 0
> z = xi_min * (3 / 4)
> z == xi_min
> xi_max = number_properties('huge')
> [f,e] = log2(xi_max)
> f == 1 - 2^(-53)
> xi_max + 2^971[1]
> nearfloat('succ',xi_max)
> xi_max + 2^970
> xi_max + 2^969 == xi_max
```

(2.61) Esercizio (per casa).

La funzione predefinita *bitstring* di *Scilab* restituisce la stringa di cifre in base due che rappresenta la codifica usuale di un numero di macchina nel calcolatore. Consultare la pagina di Wikipedia: Double-precision floating-point format per 'decifrare' il risultato

1 La distanza tra xi_max e il suo successore in F(2,53) è $2^{1024-53} = 2^{971}$.

del seguente dialogo in *Scilab*:
```
> bitstring(1)
> bitstring(xi_min)
> bitstring(0)
> bitstring(%inf)
```

(2.62) Osservazione.

In generale, assegnata $H \in \mathbb{R}^{n\times n}$ tale che $I - H$ invertibile e posto:

$$C = \{ g \in \mathbb{R}^n \text{ tali che } x(k) \text{ è convergente}\}$$

sussiste una ed una sola delle seguenti eventualità:

(1) C ha *un solo* elemento (la soluzione del sistema $(I - H)\, x = c$)

(2) C è *un sottospazio vettoriale di* $\mathbb{R}^n$ *di dimensione* $\leqslant n$ (determinato dagli autovettori di H)

(3) $C = \mathbb{R}^n$

Se sussiste uno dei casi (1) o (2), è *praticamente impossibile* determinare g tale che la successione x(k) risulti convergente: il metodo *non è utilizzabile* per approssimare la soluzione di $A\, x = b$.

Se sussiste il caso (3), qualunque g genera una successione convergente alla soluzione del sistema $A\, x = b$: il metodo *è utilizzabile* per approssimare la soluzione di $A\, x = b$.

(2.63) Definizione (metodo convergente).

Siano $H \in \mathbb{R}^{n\times n}$ e $c \in \mathbb{R}^n$. *Il metodo* iterativo definito da H e c *è convergente* se:

(1) per ogni $g \in \mathbb{R}^n$, *la successione* x(k) generata dal metodo a partire da g *è convergente*;

(2) tutte le successioni generate dal metodo hanno *lo stesso limite*.

(2.64) Osservazione.

Nel caso (usuale) in cui il metodo iterativo sia utilizzato per approssimare la soluzione del sistema $A\, x = b$ con A invertibile, i sistemi $A\, x = b$ e $(I - H)\, x = c$ sono equivalenti, e quindi il metodo definito da H e c ha *un solo punto unito*. In questo caso (si veda l'Osservazione (2.57) della Lezione 21) si ha che (1) $\Rightarrow$ (2), ovvero: metodo convergente significa che tutte le successioni generate dal metodo sono convergenti.

(2.65) Definizione (spettro e raggio spettrale).

Sia $A \in \mathbb{R}^{n\times n}$. Si chiama *spettro di* A l'insieme degli autovalori di A:

$$\sigma(A) = \{ \lambda \in C \text{ tali che } \lambda \text{ è autovalore di } A \}$$

Si chiama *raggio spettrale di* A il numero:

$$\rho(A) = \max \{\ |\lambda|\ \text{tali che } \lambda \text{ è autovalore di A}\ \}^2$$

(2.66) Teorema (caratterizzazione dei metodi convergenti).

Siano $H \in R^{n\times n}$ e $c \in R^n$. Il metodo iterativo definito da H e c è convergente *se e solo se* $\rho(H) < 1$.

(2.67) Esempi.

(1) Siano $H = \begin{bmatrix} 1/2 & 0 \\ 0 & -1 \end{bmatrix}$, $c = 0$ e $g \in R^2$. La successione generata dal metodo iterativo definito da H e c a partire da g è:

$$x(k) = H^k g = \begin{bmatrix} (1/2)^k & 0 \\ 0 & (-1)^k \end{bmatrix} = \begin{bmatrix} (1/2)^k g_1 \\ (-1)^k g_2 \end{bmatrix}$$

La successione è convergente (all'unico punto unito del metodo: 0) se e solo se $g_2 = 0$. Dunque il metodo *non è* convergente. Infatti: $\sigma(H) = \{ 1/2, -1 \}$ e $\rho(H) = 1$.

(2) Siano $H = \begin{bmatrix} 1/2 & 0 \\ 0 & 1 \end{bmatrix}$, $c = 0$ e $g \in R^2$. La successione generata dal metodo iterativo definito da H e c a partire da g è:

$$x(k) = H^k g = \begin{bmatrix} (1/2)^k & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} (1/2)^k g_1 \\ g_2 \end{bmatrix}$$

La successione è convergente per ogni g e:

$$\lim_{k\to\infty} \begin{bmatrix} (1/2)^k g_1 \\ g_2 \end{bmatrix} = \begin{bmatrix} 0 \\ g_2 \end{bmatrix}$$

Il valore del limite *dipende da* g, dunque il metodo *non è* convergente. Infatti: $\sigma(H) = \{ 1/2, 1 \}$ e $\rho(H) = 1$.

---

2 Si rappresentino gli autovalori di A, cioè $\sigma(A)$, sul piano di Gauss. Scelto un numero reale positivo r sufficientemente grande, l'insieme $I(0,r) = \{ z \in C : |z| \leqslant r \}$ - il cerchio di centro l'origine e raggio r - include $\sigma(A)$. Il raggio spettrale di A è il *minimo* valore di r tale che $I(0,r) \supset \sigma(A)$.

Lezione 23 (ore 45,46,47) - 18 novembre 2025, 15:30 - 18:30 F3

(2.68) Definizione (metodo di Jacobi).

Sia $A \in \mathbb{R}^{n\times n}$ invertibile con elementi diagonali A(k,k) tutti diversi da zero. Posto:[1]

$$D = \mathrm{diag}(A) \quad , \quad M = A - D$$

la matrice D risulta invertibile e: $A\,x = b$ è equivalente a $x = -D^{-1}M\,x + D^{-1}b$.
Il *metodo di Jacobi* (applicato al sistema $A\,x = b$) è il metodo iterativo definito da:
$H_J = -D^{-1}M$ e $c_J = D^{-1}b$.

(2.69) Definizione (matrice a predominanza diagonale forte).

Sia $A \in \mathbb{R}^{n\times n}$. La matrice A è *a predominanza diagonale forte per righe* se

$$\text{per ogni k:} \quad |A(k,k)| > \sum_{i\neq k} |A(k,i)|$$

(2.70) Teorema (predominanza diagonale forte $\Rightarrow$ invertibilità).

Sia $A \in \mathbb{R}^{n\times n}$. Se A è a predominanza diagonale forte per righe allora A è invertibile.

(Dimostrazione: Per assurdo, se A fosse a predominanza diagonale forte per righe e non invertibile allora esisterebbe una colonna $y \neq 0$ tale che $A\,y = 0$. Detta y_j la componente di y di massimo modulo (certamente diversa da zero), si avrebbe allora:

$$A(j,1)\,y_1 + \ldots + A(j,j)\,y_j + \ldots + A(j,n)\,y_n = 0 \quad \text{ovvero} \quad A(j,j)\,y_j = -\sum_{i\neq j} A(j,i)y_i$$

da cui:

$$|A(j,j)\,y_j| = \left|\sum_{i\neq j} A(j,i)y_i\right| \quad \Rightarrow \quad |A(j,j)|\,|y_j| \leqslant \sum_{i\neq j} |A(j,i)||y_i|$$

Poiché per definizione $y_j \neq 0$ e per ogni $i \neq j$ è $|y_i| / |y_j| \leqslant 1$ si avrebbe infine:

$$|A(j,j)| \leqslant \sum_{i\neq j} |A(j,i)| \left|\frac{y_i}{y_j}\right| \leqslant \sum_{i\neq j} |A(j,i)|$$

assurdo.)

(2.71) Esempio.

Siano:[2]

$$A = \begin{bmatrix} 3 & & & 1 \\ 1 & 3 & & 1 \\ 1 & & 3 & 1 \\ 1 & & & 3 \end{bmatrix} \quad , \quad b = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

1 Se $A \in \mathbb{R}^{n\times n}$, si indica con diag(A) la matrice [A(1,1);...;A(n,n)]. La notazione è mutuata da *Scilab*.

2 Se nella scrittura di una matrice un elemento *non è indicato*, il suo valore è *zero*.

- La matrice A risulta a predominanza diagonale forte per righe, e quindi invertibile, e con elementi diagonali tutti diversi da zero. Il metodo di Jacobi è definito e si ha:

$$H_J = -\frac{1}{3}\begin{bmatrix} 0 & & & 1 \\ 1 & 0 & & 1 \\ 1 & & 0 & 1 \\ 1 & & & 0 \end{bmatrix} \quad , \quad c_J = \frac{1}{3}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

- Gli autovalori di H_J ($\lambda_1 = \lambda_2 = 0$, $\lambda_3 = 1/3$, $\lambda_4 = -1/3$) hanno tutti modulo minore di uno. Per il Teorema di caratterizzazione (2.66) della Lezione 22 il metodo risulta convergente. Per ogni g in R^4 la successione generata dal metodo a partire da g è convergente alla soluzione x^* del sistema A x = b.

(2.72) Teorema (condizione sufficiente di convergenza per il metodo di Jacobi).

Siano $A \in R^{n\times n}$ a predominanza diagonale forte per righe e $b \in R^n$. Allora il metodo di Jacobi applicato al sistema A x = b è convergente.

Il risultato è una semplice conseguenza del teorema e dell'osservazione seguenti.

(2.73) Teorema (norma e raggio spettrale).

Siano $A \in R^{n\times n}$ e N una norma in R^n. Allora: $\rho(A) \leqslant \| A \|_N$.

Dimostrazione. Per definizione: $\| A \|_N = \max\{ N(A\,v),\ N(v) = 1 \}$. Siano poi $\lambda \in C$ un autovalore di A e $w \in R^n$ un autovettore associato. Allora, posto w' = w / N(w) si ha:

$$N(w') = 1 \quad \text{e} \quad N(A\,w') = N\left(A\frac{w}{N(w)}\right) = \frac{N(A\,w)}{N(w)} = \frac{N(\lambda\,w)}{N(w)} = |\lambda|\frac{N(w)}{N(w)} = |\lambda|$$

quindi $|\lambda| \in \{ N(A\,v),\ N(v) = 1 \}$. Allora:

$$\rho(A) = \max\{ |\lambda| \text{ t.c } \lambda \in \sigma(A) \} \leqslant \max\{ N(A\,v),\ N(v) = 1 \} = \| A \|_N$$

(2.74) Osservazione.

Siano $A \in R^{n\times n}$ e $b \in R^n$. Se A è a predominanza diagonale forte per righe allora per la matrice H_J del metodo di Jacobi applicato al sistema A x = b si ha $\| H_J \|_\infty < 1$.

(Esercizio: dimostrare che l'asserto è conseguenza immediata della definizione di matrice a predominanza diagonale forte per righe.)

(2.75) Scilab (esempio precedente).

Si consideri l'Esempio (2.71). Per costruire la matrice A in *Scilab*, si utilizzano i seguenti assegnamenti:[3]

3 In *Scilab*: per ogni numero intero *n*, eye(*n*,*n*) è la matrice identica di ordine *n*; se A è una matrice e *m*,*k*,*l* sono numeri interi allora: A(*m*:*k*,*l*) = [A(*m*,*l*);...;A(*k*,*l*)].

```
--> A = 3 * eye(4,4)

 A = [4x4 double]

   3.   0.   0.   0.
   0.   3.   0.   0.
   0.   0.   3.   0.
   0.   0.   0.   3.

--> A(2:4,1) = 1

 A = [4x4 double]

   3.   0.   0.   0.
   1.   3.   0.   0.
   1.   0.   3.   0.
   1.   0.   0.   3.

--> A(1:3,4) = 1

 A = [4x4 double]

   3.   0.   0.   1.
   1.   3.   0.   1.
   1.   0.   3.   1.
   1.   0.   0.   3.

--> b = [1;1;1;1]

 b = [4x1 double]

   1.
   1.
   1.
   1.
```

Per costruire la matrice H_J e la colonna c_J:[4]

```
--> D = diag(diag(A))

 D = [4x4 double]

   3.   0.   0.   0.
   0.   3.   0.   0.
   0.   0.   3.   0.
   0.   0.   0.   3.
```

4 In *Scilab*, se A è una matrice n $\times$ n allora diag(A) = [A(1,1);...;A(n,n)]; se v = $[v_1;\ldots;v_n] \in \mathbb{R}^n$ allora diag(v) è la matrice $M \in \mathbb{R}^{n\times n}$ diagonale tale che $M(1,1) = v_1, \ldots, M(n,n) = v_n$.

```
--> M = A - D

 M = [4x4 double]

   0.   0.   0.   1.
   1.   0.   0.   1.
   1.   0.   0.   1.
   1.   0.   0.   0.

--> HJ = - diag(1./diag(A)) * M

 HJ = [4x4 double]

   0.          0.   0.  -0.3333333
  -0.3333333   0.   0.  -0.3333333
  -0.3333333   0.   0.  -0.3333333
  -0.3333333   0.   0.   0.

--> cJ = diag(1./diag(A)) * b

 cJ = [4x1 double]

   0.3333333
   0.3333333
   0.3333333
   0.3333333
```

Un'approssimazione della soluzione del sistema A x = b, calcolata utilizzando la funzione predefinita *backslash* (\)[5] è:

```
--> y = A\b

 y = [4x1 double]

   0.25
   0.1666667
   0.1666667
   0.2500000
```

Per ottenere un'approssimazione della soluzione con il metodo di Jacobi, si calcolano dieci elementi della successione generata dal metodo a partire dal vettore 0.[6] Ad ogni iterazione l'istruzione disp(norm(x - y,%inf)) mostra $\| x - y \|_\infty$ ovvero la distanza tra l'ultimo elemento calcolato, x, della successione e y.

```
--> x = zeros(4,1); for k = 1:10, x = HJ * x + cJ; disp(norm(x - y,%inf)); end

   0.1666667
```

5 L'assegnamento y = A\b è equivalente alla sequenza: (S,D,P) = $EGPP_M$(A); w = SA_M(S,P b); y = SI_M(D,w).

6 Se m,n sono numeri interi, zeros(m,n) è la matrice di ordine m $\times$ n di elementi tutti uguali a zero.

0.0555556

0.0185185

0.0061728

0.0020576

0.0006859

0.0002286

0.0000762

0.0000254

0.0000085

Si osservi che, come ci si doveva aspettare dalla convergenza della successione, la distanza $\| x - y \|_\infty$ è decrescente.

(2.76) Definizione (metodo di Gauss-Seidel).

Sia $A \in R^{n \times n}$ invertibile con elementi diagonali A(k,k) tutti diversi da zero. Posto:[7]

$$T = \text{tril}(A) \ , \quad M = A - T$$

la matrice T risulta invertibile e: $A\,x = b$ è equivalente a $x = -T^{-1} M\,x + T^{-1} b$.
Il *metodo di Gauss-Seidel* (applicato al sistema $A\,x = b$) è il metodo iterativo definito da: $H_{GS} = -T^{-1} M$ e $c_{GS} = T^{-1} b$.

(2.77) Esempio.

Siano A e b come nell'Esempio (2.71).

- La matrice A risulta a predominanza diagonale forte per righe, e quindi invertibile, e con elementi diagonali tutti diversi da zero. Il metodo di Gauss-Seidel è definito e si ha:

$$H_{GS} = \begin{bmatrix} & -1/3 \\ & -2/9 \\ & -2/9 \\ & 1/9 \end{bmatrix} \in R^{4 \times 4} \quad , \quad c_{GS} = \begin{bmatrix} 1/3 \\ 2/9 \\ 2/9 \\ 2/9 \end{bmatrix}$$

- Gli autovalori di H_{GS} ($\lambda_1 = \lambda_2 = \lambda_3 = 0$, $\lambda_4 = 1/9$) hanno tutti modulo minore di uno. Per il Teorema di caratterizzazione (2.66) della Lezione 22 il metodo risulta convergente. Per ogni g in R^4 la successione generata dal metodo a partire da g è convergente alla soluzione x^* del sistema $A\,x = b$.

---

7 Se $A \in R^{n \times n}$, si indica con tril(A) la *parte strettamente triangolare inferiore* di A, ovvero la matrice B (triangolare inferiore) tale che: $i \leqslant j \Rightarrow B(i,j) = A(i,j)$ e $i > j \Rightarrow B(i,j) = 0$. La notazione è mutuata da *Scilab*.

(2.78) Teorema (condizione sufficiente di convergenza per il metodo di Gauss-Seidel).

Siano $A \in R^{n \times n}$ e $b \in R^n$. Se:

(1) A è a predominanza diagonale forte per righe

oppure:

(2) A è simmetrica definita positiva

allora il metodo di Gauss-Seidel applicato al sistema $A\,x = b$ è convergente.

(2.4) COSTO DELLA SOLUZIONE DI UN SISTEMA DI EQUAZIONI LINEARI CON UN METODO ITERATIVO

(2.79) Osservazione.

Siano $A \in R^{n \times n}$, $b \in R^n$ e x', x" le approssimazioni della soluzione x^* del sistema $A\,x = b$ ottenute, rispettivamente, con un *metodo diretto* e con un *metodo iterativo* (dotato, come vedremo, di un opportuno criterio d'arresto). Vogliamo confrontare x' ed x" dal punto di vista del *costo aritmetico*.

Supponiamo x' calcolata con il procedimento che utilizza EGPP. Il costo asintotico del calcolo è allora: $(2/3)\,n^3$.

Il costo del calcolo di x" è:

(costo per iterazione) * (numero di iterazioni)

Dobbiamo quindi determinare il costo di una *singola* iterazione.

Consideriamo, ad esempio, il metodo di Gauss-Seidel. Per calcolare la colonna x(k+1) si hanno (almeno) due alternative:

(1) calcolare $-T^{-1} M\, x(k) + T^{-1} b$;
(2) calcolare la soluzione del sistema $T\,x = -M\,x(k) + b$.

Per il costo della prima alternativa si ha:

(1.a) $2\,n^2 - 3\,n$ operazioni per calcolare $-T^{-1} M\, x(k)$
(1.b) n operazioni per calcolare la somma $-T^{-1} M\, x(k) + T^{-1} b$

in totale: $2\,n^2 - 2\,n$ operazioni.

Per il costo della seconda alternativa si ha:

(2.a) $n^2 - 2\,n + 1$ operazioni per calcolare $-M\,x(k)$
(2.b) n operazioni per calcolare la somma $-M\,x(k) + b$

(2.c) n^2 operazioni per calcolare la soluzione del sistema

in totale: $2\,n^2 - n + 1$ operazioni.

In entrambi i casi il *costo asintotico* è $2\,n^2$. Dunque: se x" è stata calcolata con k iterazioni dal metodo di Gauss-Seidel, il costo asintotico del calcolo è $2\,k\,n^2$. Il metodo di Gauss-Seidel risulta più economico del metodo diretto che usa EGPP se $k < n/3$.

(Esercizio: verificare i costi per entrambe le alternative.)

Occorre studiare la *rapidità di convergenza* di un metodo iterativo.

(2.80) Esempio.

Siano $H = \mathrm{diag}(s_1, s_2)$ con $|s_2| < |s_1| < 1$ e $c = 0$. Per il Teorema di caratterizzazione dei metodi convergenti (vedi Teorema (2.66) della Lezione 22), il metodo iterativo definito da H e zero è convergente: per ogni g in R^2 la successione x(k) generata converge a zero. Quanto rapidamente?

Sia:

$$g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \neq 0.$$

Allora:

$$x(k) = H^k g = \mathrm{diag}(s_1^k,\ s_2^k)\, g = \begin{bmatrix} s_1^k g_1 \\ s_2^k g_2 \end{bmatrix}$$

e, utilizzando la norma uno:

$$\| x(k) \|_1 = |s_1^k g_1| + |s_2^k g_2|$$

- Se $g_1 \neq 0$:.

  $$\| x(k) \|_1 = |s_1|^k |g_1| \ (1 + |s_2/s_1|^k |g_2/g_1|)$$

  da cui:

  $$\frac{\| x(k) \|_1}{|s_1|^k} \ \to \ |g_1| \neq 0$$

  e:

  $\| x(k) \|_1$ tende a zero *con la stessa rapidità di* $|s_1|^k$

- Se $g_1 = 0$, invece:

  $$\frac{\| x(k) \|_1}{|s_2|^k} \ \to \ |g_2| \neq 0$$

  e:

  $\| x(k) \|_1$ tende a zero *con la stessa rapidità di* $|s_2|^k$

  dunque, essendo $|s_2| < |s_1|$, *più rapidamente* di $|s_1|^k$.

(2.81) Teorema (sulla rapidità di convergenza).

Quanto accade nell'Esempio (2.80) si ritrova in generale.

Si consideri il metodo iterativo convergente definito da $H \in R^{n \times n}$ e $c \in R^n$. Detta x^* la soluzione del sistema $(I - H)\, x = c$ e detta $x(k)$ la successione generata dal metodo a partire da $g \in R^n$, allora, indicato con $\rho(H)$ il raggio spettrale di H:[8]

$$\| x(k) - x^* \| \text{ converge a zero } \textit{almeno con la stessa rapidità di } \rho(H)^k$$

Inoltre, se il vettore iniziale g è scelto in modo aleatorio, la probabilità che la successione converga a zero più rapidamente di $\rho(H)^k$ è nulla.

(2.82) Esempio.

In base a quanto ottenuto negli esempi (2.71) e (2.77) in cui $\rho(H_J) = 1/3$ e $\rho(H_{GS}) = 1/9$: scelto $g \in R^2$ in modo aleatorio, la successione generata dal metodo di Gauss-Seidel converge a x^* *più rapidamente* di quella generata dal metodo di Jacobi.

8 Vedere la Definizione (2.65) della Lezione 22.

Lezione 24 (ore 48,49) - 19 novembre 2025, 11:30 - 13:30 A13

(2.83) Osservazione (criteri d'arresto).

Siano $A \in \mathbb{R}^{n\times n}$ invertibile e $b \in \mathbb{R}^n$ non zero. Si utilizza il metodo iterativo convergente definito da $H \in \mathbb{R}^{n\times n}$ e $c \in \mathbb{R}^n$ per approssimare la soluzione x^* del sistema $A\,x = b$. Scelto $g \in \mathbb{R}^n$, il metodo iterativo genera la successione $x(k)$, convergente ad x^*. Descriviamo due possibili criteri d'arresto.

(a) Assegnato $E > 0$ e posto $r(k) = b - A\,x(k)$ (vettore residuo associato ad $x(k)$):

$$\text{se } \|r(k)\| \,/\, \|b\| < E \text{ allora STOP}$$

- Il criterio è *calcolabile*;
- Il criterio è *efficace* (infatti: se $x(k) \to x^*$ allora $x(k) - x^* \to 0$ e quindi $r(k) = A\,(x^* - x(k)) \to 0$);
- Quando il criterio è verificato si ha, interpretando $x(k)$ come soluzione del sistema perturbato $A\,x = b - r(k)$ ed utilizzando i risultati della teoria del condizionamento:

$$\|x(k) - x^*\| \,/\, \|x^*\| \leqslant c(A)\,\|r(k)\| \,/\, \|b\| < c(A)\,E$$

Il criterio risulta dunque di *tipo relativo*. Si osservi che se il numero di condizionamento di A è molto grande, l'approssimazione restituita può non essere accurata quanto richiesto dall'utilizzatore.

(b) Assegnato $E > 0$:

$$\text{se } \|x(k) - x(k-1)\| < E \text{ allora STOP}$$

- Il criterio è *calcolabile*;
- Il criterio è *efficace* (infatti: se $x(k) \to x^*$ allora $x(k-1) - x^* \to 0$ e quindi $x(k) - x(k-1) \to 0$);
- Quando il criterio è verificato: se $\|H\| < 1$ allora, posto $F(H) = \|H\| \,/\, (1 - \|H\|)$ si ha:[1]

$$\|x(k) - x^*\| \leqslant F(H)\,\|x(k) - x(k-1)\| < F(H)\,E$$

Il criterio risulta dunque di *tipo assoluto*. Si osservi che se $\|H\|$ vale poco meno di uno allora $F(H)$ è molto grande e l'approssimazione restituita può non essere accurata quanto richiesto dall'utilizzatore.

(2.84) Esercizio (per casa).

Scrivere una function *Scilab*, di intestazione

```
x = GaussSeidel(A,b,E)
```

che, dopo aver verificato che gli elementi sulla diagonale di A sono tutti diversi da zero, applica il metodo di Gauss-Seidel al sistema $A\,x = b$ utilizzando come criterio d'arresto

1 Dimostrazione omessa.

quello esposto in (b) dell'Osservazione (2.83).

(3) INTERPOLAZIONE E MINIMI QUADRATI

(3.01) Problema.

Siano assegnate k+1 coppie di numeri reali (dette *dati*):

$$(x_0,y_0) \ , \ \dots \ , \ (x_k,y_k)$$

con $x_0,\dots,x_k$ *distinti*, e un sottospazio vettoriale F dello spazio vettoriale su R delle funzioni continue da $I \subset R$ in R tale che:

$$\dim F = m$$

- Il *problema dell'interpolazione* consiste nel determinare gli elementi $g \in F$ tali che:

  $$g(x_0) = y_0 \ , \ \dots \ , \ g(x_k) = y_k$$

  Ciascuno degli elementi g che verifica le condizioni si chiama un elemento di F che *interpola i dati*.

  La condizione che $x_0,\dots,x_k$ siano distinti è *necessaria* affinché il problema dell'interpolazione possa avere *almeno una* soluzione.

- Sia $m < k+1$. Il *problema dei minimi quadrati* consiste nel determinare gli elementi $g \in F$ punti di *minimo assoluto* della funzione $SQ: F \to R$ definita da:

  $$SQ(f) = (f(x_0) - y_0)^2 + \dots + (f(x_k) - y_k)^2$$

  Ciascuno degli elementi g che verifica la condizione si chiama *un elemento di* F *che meglio approssima i dati nel senso dei minimi quadrati*.

  Si osservi che in questo problema *non si richiede* la condizione che $x_0,\dots,x_k$ siano distinti.

(3.02) Osservazione (interpretazione geometrica dei due problemi).

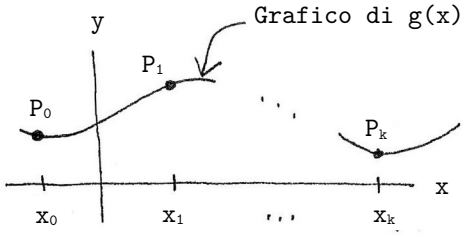
Si rappresentino in un piano cartesiano i k+1 punti:[2]

$$P_0 \equiv (x_0,y_0), \ \dots \ ,P_k \equiv (x_k,y_k)$$

Il problema dell'interpolazione consiste nel *determinare gli elementi* $g \in F$ *il cui grafico contiene tutti i* k+1 *punti* (vedere la figura seguente).

---

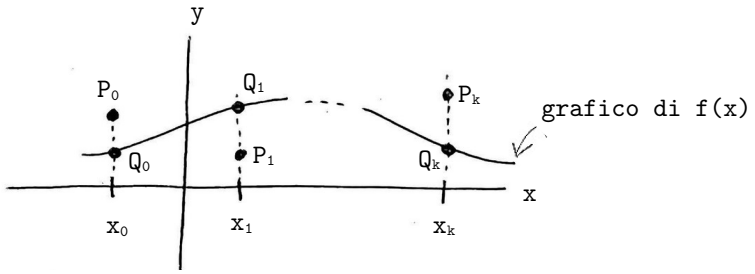
2 Assegnato un piano cartesiano, la scrittura $P \equiv (x,y)$ significa che la coppia di numeri reali (x,y) costituisce le *coordinate* del punto P del piano.



Per ogni $f \in F$, siano:

$$Q_0 \equiv (x_0, f(x_0)), \ \ldots \ , Q_k \equiv (x_k, f(x_k))$$

Il valore SQ(f) è *la somma dei quadrati delle lunghezze dei segmenti* $P_0Q_0, \ \ldots \ , P_kQ_k$. Questo valore può essere pensato come 'distanza' del grafico di f dai dati (vedere la figura seguente).



(3.03) Esempio (riformulazione del problema dell'interpolazione).

Si considerino i dati (k = 2): $(x_0,y_0),(x_1,y_1),(x_2,y_2)$, con x_0,x_1,x_2 distinti, e lo spazio vettoriale $F = \text{span}\{ f_1(x), f_2(x) \} = \{ a_1 f_1(x) + a_2 f_2(x) \text{ con } a_1, a_2 \in R \}$. Le condizioni di interpolazione:

$$g(x_0) = y_0 \ , \ g(x_1) = y_1 \ , \ g(x_2) = y_2$$

si riscrivono (utilizzando l'espressione di g(x) in termini dei generatori $f_1(x), f_2(x)$):

$$a_1 f_1(x_0) + a_2 f_2(x_0) = y_0 \ , \ a_1 f_1(x_1) + a_2 f_2(x_1) = y_1 \ , \ a_1 f_1(x_2) + a_2 f_2(x_2) = y_2$$

Dunque:

$$g(x) = a_1 f_1(x) + a_2 f_2(x) \text{ interpola i dati}$$

$$\Updownarrow$$

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \text{ è soluzione del sistema } \begin{bmatrix} f_1(x_0) & f_2(x_0) \\ f_1(x_1) & f_2(x_1) \\ f_1(x_2) & f_2(x_2) \end{bmatrix} z = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix}$$

Il sistema ha *tante equazioni quanti sono i dati da interpolare, tante incognite quanti sono i generatori di* F *assegnati.*

(3.04) Osservazione (interpolazione polinomiale).

Si considerino i dati: $(x_0,y_0),\ldots,(x_k,y_k)$, con $x_0,\ldots,x_k$ distinti, e lo spazio vettoriale F =

$P_k(R)$.[3] Si osservi che in questo caso la dimensione di F è *uguale* al numero di dati. Il problema di determinare gli elementi di $P_k(R)$ che interpolano i dati si chiama *problema dell'interpolazione polinomiale*. Per studiare il problema si introduce una base in $P_k(R)$. La scelta della base influisce sulla forma del sistema da risolvere e sull'espressione degli eventuali elementi di F determinati. Vediamo tre possibili scelte.

(1) Si consideri la base:

$$1, x, x^2, \ldots, x^k$$

detta *base di Vandermonde* di $P_k(R)$. Il sistema che traduce le condizioni di interpolazione è:

$$\begin{bmatrix} 1 & x_0 & \ldots & x_0^k \\ \vdots & \vdots & & \vdots \\ 1 & x_k & \ldots & x_k^k \end{bmatrix} c = \begin{bmatrix} y_0 \\ \vdots \\ y_k \end{bmatrix}$$

la cui matrice si chiama *matrice di Vandermonde*. Se $c^* = (c_0, \ldots, c_k)^t$ è una soluzione del sistema, il polinomio:

$$p_k(x) = c_0 + c_1 x + \cdots + c_k x^k$$

interpola i dati e l'espressione ottenuta si chiama *forma di Vandermonde* del polinomio.

(2) Si consideri la base:

$$1, (x - x_0), (x - x_0)(x - x_1), \ldots, (x - x_0)\cdots(x - x_{k-1})$$

detta *base di Newton* di $P_k(R)$. Si verifica facilmente che il sistema che traduce le condizioni di interpolazione, ad esempio nel caso k = 3, è:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & x_1 - x_0 & 0 & 0 \\ 1 & x_2 - x_0 & (x_2 - x_0)(x_2 - x_1) & 0 \\ 1 & x_3 - x_0 & (x_3 - x_0)(x_3 - x_1) & (x_3 - x_0)(x_3 - x_1)(x_3 - x_2) \end{bmatrix} c = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

La matrice del sistema è *triangolare inferiore* e invertibile (si ricordi che i numeri $x_0, \ldots, x_k$ sono distinti). La base di Newton è *costruita appositamente* affinché la matrice del sistema risulti triangolare inferiore. Se $c^* = (c_0, \ldots, c_k)^t$ è una soluzione del sistema, il polinomio:

$$p_k(x) = c_0 + c_1 (x - x_0) + \cdots + c_k (x - x_0)\cdots(x - x_{k-1})$$

interpola i dati e l'espressione ottenuta si chiama *forma di Newton* del polinomio.

(3) Si considerino i k+1 elementi di $P_k(R)$ definiti da:

---

3 Si indica con $P_k(R)$ lo spazio vettoriale su R dei polinomi a coefficienti reali di grado *al più* k.

$$l_0(x) = \frac{(x-x_1)\cdots(x-x_k)}{(x_0-x_1)\cdots(x_0-x_k)},\ l_1(x) = \frac{(x-x_0)(x-x_2)\cdots(x-x_k)}{(x_1-x_0)(x_1-x_2)\cdots(x_1-x_k)},\ \ldots,\ l_k(x) = \frac{(x-x_0)\cdots(x-x_{k-1})}{(x_k-x_0)\cdots(x_k-x_{k-1})}$$

Questi elementi sono costruiti in modo tale che per i = 0,...,k si abbia: $l_i(x_i) = 1$ e $l_i(x_j) = 0$ per $j \neq i$. Inoltre, sono elementi *linearmente indipendenti* di $P_k(R)$ (infatti: *se* $A(x) = a_0\, l_0(x) + \cdots + a_k\, l_k(x) = 0$ per ogni $x \in R$ *allora* per i = 0,...,k si ha: $A(x_i) = a_i = 0$) e quindi, poiché dim $P_k(R) = k+1$, sono una *base* di $P_k(R)$, detta *base di Lagrange* di $P_k(R)$.

Si verifica facilmente che il sistema che traduce le condizioni di interpolazione è:

$$c = \begin{bmatrix} y_0 \\ \vdots \\ y_k \end{bmatrix}$$

infatti la matrice del sistema è la *matrice identità*. La base di Lagrange è *costruita appositamente* affinché accada questo. Infine, il polinomio:

$$p_k(x) = y_0\, l_0(x) + \cdots + y_k\, l_k(x)$$

interpola i dati e l'espressione ottenuta si chiama *forma di Lagrange* del polinomio.

(3.05) Teorema (esistenza ed unicità del polinomio interpolante)

Assegnati i dati: $(x_0,y_0),\ldots,(x_k,y_k)$, con $x_0,\ldots,x_k$ distinti, esiste *un solo* elemento $p(x) \in P_k(R)$ che li interpola. Il polinomio p(x) si chiama *il polinomio interpolante*.

(Dimostrazione. Per quanto mostrato nel punto (3) dell'osservazione precedente, esiste una sola combinazione lineare degli elementi della base di Lagrange che interpola i dati. Dunque esiste un solo elemento di $P_k(R)$ che interpola i dati.)

(3.06) Osservazione.

Per risolvere un problema di interpolazione polinomiale si *sceglie* una base di $P_k(R)$, si determina *la* soluzione del sistema che traduce le condizioni di interpolazione e si individua *la* combinazione lineare degli elementi della base scelta che interpola i dati. A seconda della base scelta si ottiene una *forma diversa dell'unico polinomio interpolante*.

(3.07) Esercizio (per casa).

Si risolvano i seguenti problemi, *nessuno dei quali è di interpolazione polinomiale* (perché?).

(1) Assegnati i dati (-1,1), (0,0), (1,0), determinare gli elementi $g \in P_1(R)$ che interpolano i dati.
(2) Assegnati i dati (-1,0), (0,0), (1,0), determinare gli elementi $g \in P_1(R)$ che interpolano i dati.
(3) Assegnato il dato (0,0), determinare gli elementi $g \in P_1(R)$ che interpolano i dati.

Lezione 25 (ore 50,51) - 20 novembre 2025, 8:30 - 10:30 F3

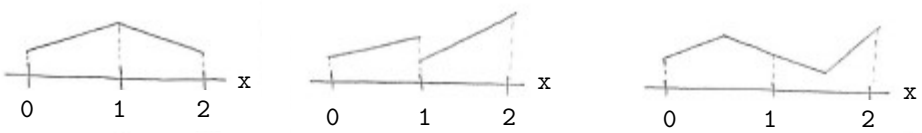
(3.08) Definizione (funzioni continue e lineari a tratti).

Assegnati numeri reali *ordinati* $x_0 < x_1 < \cdots < x_k$ e posto, per $j = 1,\ldots,k$: $I_j = [x_{j-1},x_j]$, una funzione $f: [x_0,x_k] \to R$ si dice *continua e lineare a tratti su* $x_0,\ldots,x_k$ se:

- f è continua;
- f è lineari a tratti su $x_0,\ldots,x_k$ ovvero: detta $p_j(x)$ la restrizione di f a I_j si ha: $p_j \in P_1(R)$.

L'insieme delle funzioni continue e lineari a tratti su $x_0,\ldots,x_k$ si indica con $C\text{-}LAT(x_0,\ldots,x_k)$.

(3.09) Esempio.



(1) (2) (3)

La figura (1) rappresenta il grafico di una funzione continua e lineare a tratti su 0,1,2. La figura (2) rappresenta il grafico di una funzione lineare a tratti su 0,1,2 ma *non* continua. La figura (3) rappresenta il grafico di una funzione continua e lineari a tratti ma *non* su 0,1,2.

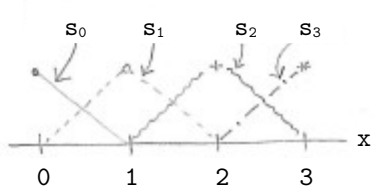
(3.09) Osservazione.[1]

(1) L'insieme $C\text{-}LAT(x_0,\ldots,x_k)$ è un *sottospazio vettoriale* dello spazio delle funzioni continue su $[x_0,x_k]$, di *dimensione* k+1. I k+1 elementi $s_0(x),\ldots,s_k(x)$ di $C\text{-}LAT(x_0,\ldots,x_k)$ definiti da:

$$s_i(x_j) = 1 \text{ se } i = j \quad , \quad s_i(x_j) = 0 \text{ se } i \neq j$$

sono la '*base canonica*' di $C\text{-}LAT(x_0,\ldots,x_k)$.

Ad esempio, i grafici degli elementi della base canonica di C-LAT(0,1,2,3) sono:



(2) Assegnati numeri reali $y_0,\ldots,y_k$, la combinazione lineare $y_0\, s_0(x) + \cdots + y_k\, s_k(x)$ è *l'unico* elemento di $C\text{-}LAT(x_0,\ldots,x_k)$ che interpola i dati $(x_0,y_0),\ldots,(x_k,y_k)$.

1 La dimostrazione degli asserti di questa Osservazione è omessa.

(3.10) Applicazioni.

(1) Grafici in *Scilab*.

Sia f: [a,b] → R una funzione continua. La sequenza di istruzioni:

```
> x = linspace(a,b,n)';
> plot(x,f(x));
```

genera, in una finestra grafica, il grafico della spezzata di vertici i punti di coordinate (x(1),f(x(1))),...,(x(n),f(x(n)). Questa spezzata è il grafico dell'unico elemento $\sigma_n(x) \in$ C-LAT(x(1),...,x(n)) che interpola i dati (x(1),f(x(1))),..., (x(n),f(x(n)). Il grafico di $\sigma_n(x)$ è utilizzato come approssimazione di quello della funzione f(x). Vedremo tra poco quanto sia accurata l'approssimazione.

(2) Formula dei trapezi.

Sia f: [a,b] → R una funzione continua. Si vuole conoscere il valore (certamente esistente per la continuità di f):

$$I = \int_a^b f(x)dx$$

Un procedimento che fornisce un'approssimazione di I è il seguente:

- scelto k, si suddivide l'intervallo [a,b] in k sottointervalli di uguale ampiezza:

$$\frac{b-a}{k}$$

  individuati dai k+1 punti $x_0 = a, x_1, \ldots, x_{k-1}, x_k = b$ (detti *nodi*):
- si considera l'unico elemento $\sigma_k(x) \in$ C-LAT($x_0,\ldots,x_k$) che interpola i dati $(x_0,f(x_0)),\ldots,(x_k,f(x_k))$;
- si approssima I con:

$$J_k = \int_a^b \sigma_k(x)dx$$

Il valore J_k si calcola facilmente. Introdotta la base canonica $s_0(x),\ldots,s_k(x)$ di C-LAT($x_0,\ldots,x_k$) si ha:

$$J_k = \int_a^b \sigma_k(x)dx = \int_a^b (f(x_0)s_0(x) + \cdots + f(x_k)s_k(x))dx = f(x_0)\int_a^b s_0(x)dx + \cdots + f(x_k)\int_a^b s_k(x)dx$$

e quindi:

$$J_k = h\left[\frac{f(x_0)}{2} + f(x_1) + \cdots + f(x_{k-1}) + \frac{f(x_k)}{2}\right]$$

Quest'ultima espressione si chiama *formula dei trapezi*. Vedremo tra poco quanto sia accurata l'approssimazione.

(3.11) Teorema (errore nell'interpolazione polinomiale).

Assegnati numeri reali $x_0 < x_1 < \cdots < x_k$, e posto $I = [x_0,x_k]$, siano $f: I \to R$ con derivata di ordine k+1 continua e $p_k \in P_k(R)$ il polinomio che interpola i dati $(x_0,f(x_0)),\ldots,(x_k,f(x_k))$. Allora, per ogni $x \in I$ esiste $\theta \in I$ tale che:

$$f(x) - p_k(x) = \frac{f^{(k+1)}(\theta)}{(k+1)!}(x-x_0)\cdots(x-x_k)$$

Inoltre, posto:

$$M_j = \max_{x \in I} |f^{(j)}(x)|$$

si ottiene facilmente la limitazione:

$$\max_{x \in I} |f(x) - p_k(x)| \leqslant \frac{M_{k+1}}{(k+1)!}(\text{mis } I)^{k+1}$$

Dimostrazione: omessa.

(3.12) Osservazione (approssimazione con elementi di C-LAT).

Siano $f: [a,b] \to R$ con derivata seconda continua e, per $j = 0,\ldots,k$:

$$x_j = a + \frac{b-a}{k} j$$

I punti $x_0,\ldots,x_k$ dividono l'intervallo [a,b] in k intervalli di uguale ampiezza.

Si consideri l'intervallo $[x_0,x_1]$. Detto $p_1 \in P_1(R)$ il polinomio che interpola i dati $(x_0,f(x_0)),(x_1,f(x_1))$, utilizzando la limitazione mostrata nel teorema precedente, si ha:

$$\max_{x \in [x_0,x_1]} |f(x) - p(x)| \leqslant \frac{M_2}{2}\left(\frac{b-a}{k}\right)^2$$

Ripetendo il ragionamento si ottiene la stessa limitazione per ciascuno dei k sottointervalli di [a,b]. Perciò, detto $\sigma_k(x)$ l'elemento di C-LAT$(x_0,\ldots,x_k)$ che interpola i dati $(x_0,f(x_0)),\ldots,(x_k,f(x_k))$, si ha:

$$\max_{x \in [a,b]} |f(x) - \sigma_k(x)| \leqslant \frac{M_2}{2}\left(\frac{b-a}{k}\right)^2$$

(3.13) Osservazione (accuratezza delle approssimazioni nelle applicazioni).

(1) Scelta come misura (assoluta) dell'errore commesso approssimando il grafico di f(x) con quello di $\sigma_n(x)$ la quantità:

$$e_n(f) = \max_{x \in [a,b]} |f(x) - \sigma_n(x)|$$

il risultato dell'osservazione precedente mostra che: se f ha derivata seconda continua allora:

$$\lim_{n \to \infty} e_n(f) = 0$$

e l'approssimazione può essere resa *accurata quanto si vuole* scegliendo n opportunamente grande.

(2) Scelta come misura (assoluta) dell'errore commesso approssimando I con J_k la quantità $|J_k - I|$, il risultato dell'osservazione precedente mostra che: se f ha derivata seconda continua allora:

$$|J_k - I| = \left| \int_a^b (\sigma_k(x) - f(x))\,dx \right| \leqslant \int_a^b |\sigma_k(x) - f(x)|\,dx \leqslant \int_a^b \frac{M_2}{2}\left(\frac{b-a}{k}\right)^2 dx = \frac{M_2}{2}\frac{(b-a)^3}{k^2}$$

Anche in questo caso si ha dunque:

$$\lim_{k\to\infty} |J_k - I| = 0$$

e l'approssimazione può essere resa *accurata quanto si vuole* scegliendo k opportunamente grande.

(3.14) Esempio (riformulazione del problema dei minimi quadrati).

Si considerino i dati (k = 2): $(x_0,y_0),(x_1,y_1),(x_2,y_2)$ e lo spazio vettoriale F = span{ $f_1(x),f_2(x)$ } = { $a_1 f_1(x) + a_2 f_2(x)$ con $a_1, a_2 \in R$ }. Lo scarto quadratico SQ(f) si riscrive, utilizzando l'espressione di f(x) in termini dei generatori $f_1(x),f_2(x)$:

$$SQ(f) = (a_1 f_1(x_0) + a_2 f_2(x_0) - y_0)^2 + (a_1 f_1(x_1) + a_2 f_2(x_1) - y_1)^2 + (a_1 f_1(x_2) + a_2 f_2(x_2) - y_2)^2$$

Osservando che se $v = (v_1,\ldots,v_n)^t \in R^n$ si ha $(N_2(v))^2 = v_1^2 + \cdots + v_n^2$, l'ultima somma può essere riscritta come:[2]

$$\left\| \begin{bmatrix} a_1 f_1(x_0) + a_2 f_2(x_0) - y_0 \\ a_1 f_1(x_1) + a_2 f_2(x_1) - y_1 \\ a_1 f_1(x_2) + a_2 f_2(x_2) - y_2 \end{bmatrix} \right\|^2 = \left\| \begin{bmatrix} f_1(x_0) & f_2(x_0) \\ f_1(x_1) & f_2(x_1) \\ f_1(x_2) & f_2(x_2) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} - \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix} \right\|^2$$

e quindi, posto:

$$A = \begin{bmatrix} f_1(x_0) & f_2(x_0) \\ f_1(x_1) & f_2(x_1) \\ f_1(x_2) & f_2(x_2) \end{bmatrix} \quad , \quad x = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix}$$

si ha infine:

$$SQ(f) = \| A\,x - b \|^2$$

(3.15) Osservazione.

Il sistema A x = b ottenuto nell'esempio precedente è *il sistema che traduce le condizioni di interpolazione* $f(x_0) = y_0$, $f(x_1) = y_1$, $f(x_2) = y_2$.

---

2 Per alleggerire la notazione, per ogni $v \in R^n$, in questa parte indicheremo con $\| v \|$ la norma due di v.

Lezione 26 (ore 52,53) - 25 novembre 2025, 16:30 - 18:30 F3

(3.16) Definizione (soluzione nel senso dei minimi quadrati di un sistema).

Siano $A \in \mathbb{R}^{r \times c}$ con $r > c$, e $b \in \mathbb{R}^r$. Un elemento $x^* \in \mathbb{R}^c$ si chiama *soluzione del sistema* $A\,x = b$ *nel senso dei minimi quadrati* se x^* è punto di minimo assoluto della funzione $SQ: \mathbb{R}^c \to \mathbb{R}$ definita da:

$$SQ(x) = \| A\,x - b \|^2$$

Si osservi che: se $y \in \mathbb{R}^c$ è una soluzione di $A\,x = b$ allora y è *anche* una soluzione di $A\,x = b$ nel senso dei minimi quadrati (come mai?) ma, salvo casi particolari, una soluzione di $A\,x = b$ nel senso dei minimi quadrati *non è* una soluzione di $A\,x = b$.

Vediamo *come si determinano* le soluzioni di $A\,x = b$ nel senso dei minimi quadrati.

(3.17) Osservazione (scomposizione ortogonale di un vettore).

Siano $A = (a_1,\ldots,a_c) \in \mathbb{R}^{r \times c}$ con $r > c$, e $b \in \mathbb{R}^r$. Detta b_* la *proiezione ortogonale*[1] di b su $\text{span}\{ a_1,\ldots,a_c \} = C(A)$[2], e posto $b_\perp = b - b_*$ si ottiene la scomposizione ortogonale:

$$b = b_* + b_\perp$$

Si osservi che:

(1) Poiché $b_* \in C(A)$, esiste $y \in \mathbb{R}^c$ tale che $b_* = A\,y$;

(2) Per definizione di proiezione ortogonale, la colonna $b_\perp = b - b_*$ è ortogonale a tutti gli elementi di $C(A)$.

(3.18) Osservazione.

Per determinare le soluzioni di $A\,x = b$ nel senso dei minimi quadrati si osservi che, utilizzando la scomposizione ortogonale di b introdotta nell'osservazione precedente, per ogni $x \in \mathbb{R}^c$ si ha:

$$SQ(x) = \| A\,x - b \|^2 = \| A\,x - b_* + b_\perp \|^2 = \| A\,x - A\,y + b_\perp \|^2 = \| A\,(x - y) + b_\perp \|^2$$

Poiché $A\,(x - y) \in C(A)$ e $b_\perp$ è ortogonale a tutti gli elementi di $C(A)$, per il Teorema di Pitagora[3] si ha:

$$\| A\,(x - y) + b_\perp \|^2 = \| A\,(x - y) \|^2 + \| b_\perp \|^2$$

Allora:

- Per ogni $x \in \mathbb{R}^c$ si ha: $SQ(x) = \| A\,(x - y) \|^2 + \| b_\perp \|^2 \geqslant \| b_\perp \|^2$

---

1 La proiezione ortogonale di $v \in \mathbb{R}^n$ su un sottospazio $W \subset \mathbb{R}^n$ è l'*unico* elemento $v_* \in W$ tale che la differenza $v - v_*$ è ortogonale a tutti gli elementi di W.

2 $C(A)$ si chiama anche *spazio delle colonne di* A e coincide con l'immagine dell'applicazione lineare da $\mathbb{R}^c$ in $\mathbb{R}^r$ definita da $x \to A\,x$.

3 Siano a,b elementi di $\mathbb{R}^n$, e sia $\langle a,b \rangle = b^t a$ il prodotto scalare di a e b. Se a e b sono ortogonali (ovvero, se $\langle a,b \rangle = 0$) allora si ha:

$$\| a + b \|^2 = \langle a + b, a + b \rangle = \langle a,a \rangle + 2\langle a,b \rangle + \langle b,b \rangle = \langle a,a \rangle + \langle b,b \rangle = \| a \|^2 + \| b \|^2$$

- $SQ(x) = \| b_\perp \|^2 \Leftrightarrow \| A\,(x - y) \|^2 = 0 \Leftrightarrow A\,(x - y) = 0$, ovvero $x - y \in \ker A$.[4]

Dunque, l'insieme $S_{MQ}(A,b)$ delle soluzioni di $A\,x = b$ nel senso dei minimi quadrati è dato da:

$$S_{MQ}(A,b) = y + \ker A$$

(3.19) Osservazione (equazioni normali).

Per determinare tutte le colonne $y \in R^c$ tali che $b_* = A\,y$ si osservi che, per definizione di proiezione ortogonale su C(A):

$y \in R^c$ è tale che $A\,y = b_* \Leftrightarrow b - b_* = b - A\,y$ è ortogonale a *tutti* gli elementi di C(A)

Ma: perché una colonna $v \in R^r$ sia ortogonale a tutti gli elementi di C(A) è *necessario e sufficiente* che v sia ortogonale alle colonne di A (dimostrarlo!). Dunque: v ortogonale a tutti gli elementi di C(A) $\Leftrightarrow \langle v,a_1\rangle = a_1^t\,v = 0,\ldots,\langle v,a_c\rangle = a_c^t\,v = 0 \Leftrightarrow A^t\,v = 0$. Allora:

$$y \in R^c \text{ è tale che } A\,y = b_* \;\Leftrightarrow\; A^t\,(b - A\,y) = 0 \;\Leftrightarrow\; A^t\,A\,y = A^t\,b$$

Il sistema $A^t\,A\,x = A^t\,b$ si chiama *sistema delle equazioni normali* associato al sistema $A\,x = b$.

Si osservi che: $\ker A = \ker A^t\,A$ (infatti: $x \in \ker A \Rightarrow A\,x = 0 \Rightarrow A^t\,(A\,x) = 0 \Rightarrow A^t\,A\,x = 0 \Rightarrow x \in \ker A^t\,A$; viceversa: $x \in \ker A^t\,A \Rightarrow A^t\,A\,x = 0 \Rightarrow x^t\,(A^t\,A\,x) = 0 \Rightarrow (x^t\,A^t)(A\,x) = 0 \Rightarrow (A\,x)^t\,(A\,x) = 0 \Rightarrow \| A\,x \|^2 = 0 \Rightarrow A\,x = 0 \Rightarrow x \in \ker A$). Allora:

$$S_{MQ}(A,b) = y + \ker A = y + \ker A^t\,A = \{ \text{soluzioni del sistema delle equazioni normali}\}$$

Inoltre:

- La matrice $A^t\,A \in R^{c\times c}$ è *simmetrica e semidefinita positiva* (infatti, per ogni colonna $x \neq 0$ di R^c si ha: $x^t\,(A^t\,A)\,x = (x^t\,A^t)(A\,x) = (A\,x)^t\,(A\,x) = \| A\,x \|^2 \geqslant 0$) ed è *definita positiva* se e solo se le colonne di A sono *linearmente indipendenti* (dimostrarlo!).
- Le colonne di A sono *linearmente indipendenti* $\Leftrightarrow \ker A = \ker A^t\,A = \{0\} \Leftrightarrow$ il sistema delle equazioni normali ha *una sola soluzione* $\Leftrightarrow$ la matrice $A^t\,A$ *è invertibile*.
- Le colonne di A sono *linearmente dipendenti* $\Leftrightarrow \dim \ker A = \dim \ker A^t\,A > 0 \Leftrightarrow$ il sistema delle equazioni normali ha *infinite soluzioni* $\Leftrightarrow$ la matrice $A^t\,A$ *non è invertibile*.

4 Se $A \in R^{r\times c}$, si indica con $\ker A$ il sottospazio vettoriale di R^c delle soluzioni del sistema omogeneo $A\,z = 0$.

Lezione 27 (ore 54,55) - 26 novembre 2025, 11:30 - 13:30 A13

(3.20) Esempi.

(1) Determinare le soluzioni nel senso dei minimi quadrati del sistema $A\,x = b$:

$$\begin{bmatrix}1\\1\\1\end{bmatrix} x = \begin{bmatrix}1\\0\\0\end{bmatrix}$$

e la proiezione ortogonale di b su C(A).

Soluzione: Il sistema delle equazioni normali è: $3\,x = 1$ e quindi l'unica soluzione nel senso dei minimi quadrati è: $x^* = 1/3$. La proiezione ortogonale di b su C(A) è:

$$b_* = 1/3 \begin{bmatrix}1\\1\\1\end{bmatrix}$$

(2) Determinare le soluzioni nel senso dei minimi quadrati del sistema $\underline{A}\,x = \underline{b}$:

$$\begin{bmatrix}1 & 2\\1 & 2\\1 & 2\end{bmatrix} x = \begin{bmatrix}1\\0\\0\end{bmatrix}$$

e la proiezione ortogonale di $\underline{b}$ su C($\underline{A}$).

Soluzione: Il sistema delle equazioni normali è:

$$\begin{bmatrix}3 & 6\\6 & 12\end{bmatrix} x = \begin{bmatrix}1\\2\end{bmatrix}$$

e:

$$\ker\, \underline{A}^t\, \underline{A} = \text{span}\{ \begin{bmatrix}2\\-1\end{bmatrix} \}$$

Posto:

$$y = \begin{bmatrix}1/3\\0\end{bmatrix}$$

si ottiene:

$$S_{MQ}(\underline{A},\underline{b}) = \begin{bmatrix}1/3\\0\end{bmatrix} + \text{span}\{ \begin{bmatrix}2\\-1\end{bmatrix} \} = \{ \begin{bmatrix}1/3+2\lambda\\-\lambda\end{bmatrix} ,\ \lambda \in R \}$$

In questo caso, l'insieme $S_{MQ}(\underline{A},\underline{b})$ ha *infiniti elementi* perché le colonne di $\underline{A}$ sono *linearmente dipendenti*.

La proiezione ortogonale di $\underline{b}$ su C($\underline{A}$) è:

$$\underline{b}_* = 1/3 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

La proiezione è la stessa dell'esempio precedente perché $\underline{b} = b$ e $C(\underline{A}) = C(A)$.

(3.21) Esempio (minimi quadrati pesati).

Si considerino il sistema $A\,x = b$:

$$\begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} x = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

ed il sistema $\underline{A}\,x = \underline{b}$, ottenuto moltiplicando per due la prima e la terza equazione del sistema $A\,x = b$:

$$\begin{bmatrix} 2 & -2 \\ 1 & 0 \\ 2 & 2 \end{bmatrix} x = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

I due sistemi sono *equivalenti* ma: $S_{MQ}(A,b) \neq S_{MQ}(\underline{A},\underline{b})$. Infatti:

$$S_{MQ}(A,b) = \begin{bmatrix} 2/3 \\ 1/2 \end{bmatrix} \quad , \quad S_{MQ}(\underline{A},\underline{b}) = \begin{bmatrix} 5/9 \\ 1/2 \end{bmatrix}$$

Questo *non deve sorprendere*, infatti per i due sistemi si ha $SQ(x) \neq \underline{SQ}(x)$:

$$SQ(x) = (x_1 - x_2)^2 + (x_1 - 1)^2 + (x_1 + x_2 - 1)^2$$

e:

$$\underline{SQ}(x) = 4\,(x_1 - x_2)^2 + (x_1 - 1)^2 + 4\,(x_1 + x_2 - 1)^2$$

La funzione $\underline{SQ}$ si ottiene *pesando* gli addendi della funzione SQ con 'pesi' positivi.

(3.22) Esempio.

Si considerino i dati (-1,0), (0,1), (1,1). Determinare gli elementi di $F = \text{span}\{ 1,x \}$ che meglio approssimano i dati nel senso dei minimi quadrati.

Si osservi che, scelto un piano cartesiano, ciascuno degli elementi di F ha per grafico una retta non verticale. Il problema si può quindi riformulare in: Determinare *le rette* che meglio approssimano i dati nel senso dei minimi quadrati.

Per quanto mostrato nell'Esempio (3.14) e nell'Osservazione (3.15) della Lezione 25, il problema si risolve determinando le soluzioni nel senso dei minimi quadrati del sistema (che traduce le condizioni di interpolazione):

$$\begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} x = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

Il sistema delle equazioni normali ha una sola soluzione (infatti le colonne...):

$$x^* = \begin{bmatrix} 2/3 \\ 1/2 \end{bmatrix}$$

e l'unica retta che meglio approssima i dati nel senso dei minimi quadrati è il grafico dell'unico elemento di F = span{ 1,x } che meglio approssima i dati nel senso dei minimi quadrati:

$$p(x) = \frac{2}{3} + \frac{1}{2}\, x$$

(3.23) Esempio.

Si considerino i dati (1,0), (1/2,1), (1/3,2). Determinare gli elementi di F = span{ 1,1/x } che meglio approssimano i dati nel senso dei minimi quadrati.

Procedendo come nell'esempio precedente, il problema si risolve determinando le soluzioni nel senso dei minimi quadrati del sistema (che traduce le condizioni di interpolazione):

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} x = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

Il sistema delle equazioni normali ha una sola soluzione:

$$x^* = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

e l'unico elemento di F che meglio approssima i dati nel senso dei minimi quadrati è:

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

(3.24) Osservazione.

Siano $A \in R^{r \times c}$ con r > c, $b \in R^r$ e $S_{MQ}(A,b)$ l'insieme delle soluzioni di A x = b nel senso dei minimi quadrati. Si ha:[1]

esiste una sola colonna $y_* \in S_{MQ}(A,b)$ di *norma minima*[2]

(3.25) Esempio.

Sia A x = b il sistema:

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Risulta:

$$S_{MQ}(A,b) = \begin{bmatrix} 1/2 \\ 0 \end{bmatrix} + \text{span}\{ \begin{bmatrix} -1 \\ 1 \end{bmatrix} \}$$

e, disegnando l'insieme $S_{MQ}(A,b)$ su un piano cartesiano, si verifica facilmente che

1 Dimostrazione dell'asserto omessa.

2 Più formalmente: la funzione $\| x \|$ ha un solo punto di minimo assoluto su $S_{MQ}(A,b)$.

l'elemento di norma minima (ovvero quello più vicino all'origine) è:

$$y_* = \begin{bmatrix} 1/4 \\ 1/4 \end{bmatrix}$$

(3.26) Osservazione (matrice pseudoinversa).

Sia $A \in R^{r \times c}$ con $r > c$. Per ogni $b \in R^r$, sia $S_{MQ}(A,b)$ l'insieme delle soluzioni di $A\,x = b$ nel senso dei minimi quadrati.

La funzione $F: R^r \to R^c$ definita da:

$$F(b) = \text{l'elemento di } S_{MQ}(A,b) \text{ di norma minima}$$

è un'*applicazione lineare* da R^r in R^c.[3] Quindi esiste una matrice di dimensione $c \times r$, che si indica con A^+, tale che:

$$F(b) = A^+ b$$

La matrice A^+ si chiama *matrice pseudoinversa di* A.

Se le colonne di A sono linearmente indipendenti allora (si veda l'Osservazione (3.19) della Lezione 26) $S_{MQ}(A,b)$ ha un solo elemento, che è quello di norma minima, e dalle equazioni normali si ottiene, essendo $A^t A$ invertibile:

$$F(b) = (A^t A)^{-1} A^t b$$

In questo caso si ha allora:

$$A^+ = (A^t A)^{-1} A^t$$

Si osservi che se $A \in R^{n \times n}$ è una matrice invertibile, allora risulta $A^+ = A^{-1}$. Questo spiega perché A^+ si chiami matrice pseudoinversa.

(3.27) Esempio.

Determinare la matrice pseudoinversa di:

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Per definizione $A^+ \in R^{2 \times 3}$ è l'unica matrice tale che: per ogni $b \in R^r$, $A^+ b$ = l'elemento di $S_{MQ}(A,b)$ di norma minima. Le tre colonne di A^+ sono allora, dette e_1, e_2, e_3 le colonne della base canonica di R^3:

$$F(e_1)\ ,\ F(e_2)\ ,\ F(e_3)$$

Si ha:

$$S_{MQ}(A,e_1) = \begin{bmatrix} 1/3 \\ 0 \end{bmatrix} + \text{span}\{ \begin{bmatrix} -1 \\ 1 \end{bmatrix} \}$$

e quindi, ragionando come nell'Esempio (3.25):

---

3 Dimostrazione omessa.

$$F(e_1) = \begin{bmatrix} 1/6 \\ 1/6 \end{bmatrix}$$

Allo stesso modo si determinano:

$$F(e_2) = F(e_3) = \begin{bmatrix} 1/6 \\ 1/6 \end{bmatrix}$$

Infine:

$$A^+ = \begin{bmatrix} 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 \end{bmatrix}$$

(3.28) Definizione (fattorizzazione QR caso non quadrato).

Sia $A \in R^{r \times c}$ con $r > c$. Una *fattorizzazione QR* di A è una coppia U,T tale che:

- $U \in R^{r \times c}$ è una matrice a colonne ortonormali
- $T \in R^{c \times c}$ è una matrice triangolare superiore
- $U\,T = A$

(3.29) Esempio (di calcolo di una fattorizzazione QR nel caso non quadrato, con GS).

Sia:

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \in R^{3 \times 2}$$

Per cercare una fattorizzazione QR di A possiamo utilizzare una ovvia variante della procedura GS. Dette a_1 e a_2 le colonne di A:

Passo uno.

Cerchiamo $\Omega = [\omega_1,\omega_2] \in R^{3 \times 2}$ a colonne ortogonali e $\Theta \in R^{2 \times 2}$ triangolare superiore con $\theta_{kk} = 1$ tali che $\Omega\,\Theta = A$. *Se matrici siffatte esistono*, riscrivendo l'ultima uguaglianza *per colonne* si ha:

$$\omega_1 = a_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad , \quad \omega_1\,\theta_{1,2} + \omega_2 = a_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

La prima uguaglianza determina ω_1. Dalla seconda segue che:

$$(\omega_1\,\theta_{1,2}) \bullet \omega_1 + \omega_2 \bullet \omega_1 = a_2 \bullet \omega_1 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \bullet \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = 2$$

Poiché ω_1 e ω_2 sono ortogonali, si ha $\omega_2 \bullet \omega_1 = 0$. Allora, essendo $\omega_1 \neq 0$, si ha *necessariamente*:

$$\theta_{1,2} = (a_2 \bullet \omega_1) \,/\, (\omega_1 \bullet \omega_1) = 2/3$$

e quindi:

$$\omega_2 = a_2 - \omega_1\,\theta_{1,2} = \begin{bmatrix} -2/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

Dunque:

$$\Omega = \begin{bmatrix} 1 & -2/3 \\ 1 & 1/3 \\ 1 & 1/3 \end{bmatrix} \quad \text{e} \quad \Theta = \begin{bmatrix} 1 & 2/3 \\ 0 & 1 \end{bmatrix}$$

Passo due.

La fattorizzazione di A trovata al passo precedente *non è* una fattorizzazione QR perché le colonne di Ω non hanno norma unitaria. Questo secondo passo determina, se possibile, una fattorizzazione QR normalizzando le colonne di Ω.

Sia: Δ = diag($\|\omega_1\|$,$\|\omega_2\|$) = diag(sqrt(3),sqrt(2/3)). Si verifica facilmente che la coppia

$$U = \Omega \Delta^{-1} = \begin{bmatrix} 1/\sqrt{3} & -\sqrt{2/3} \\ 1/\sqrt{3} & 1/\sqrt{6} \\ 1/\sqrt{3} & 1/\sqrt{6} \end{bmatrix} \quad , \quad T = \Delta \Theta = \begin{bmatrix} \sqrt{3} & 2/\sqrt{3} \\ 0 & \sqrt{2/3} \end{bmatrix}$$

è una fattorizzazione QR di A. Si osservi che per la matrice T, triangolare superiore, si ha:

$$T_{k,k} = \|\omega_k\| > 0$$

(3.30) Osservazione (fattorizzazione QR e minimi quadrati).

Sia U,T una fattorizzazione QR della matrice $A \in R^{r \times c}$ con r > c. Si ha:

$$A^t A = (U T)^t (U T) = (T^t U^t) (U T) = T^t (U^t U) T$$

Poiché la matrice $U \in R^{r \times c}$ ha colonne ortonormali, allora si ha $U^t U = I \in R^{c \times c}$. Allora:

$$A^t A = T^t T$$

Inoltre:

$$A^t b = (U T)^t b = T^t U^t b$$

Se la matrice T è invertibile, ovvero se le colonne di A sono linearmente indipendenti, allora anche T^t è invertibile e i sistemi

$$A^t A x = A^t b \quad \text{e} \quad T x = U^t b$$

sono *equivalenti*. I due sistemi, però, *non hanno le stesse proprietà di condizionamento*. Infatti si ha:[4]

$$c_2(A^t A) = (c_2(T))^2$$

ovvero: il sistema $A^t A x = A^t b$ ha proprietà di condizionamento *peggiori* di quelle del sistema $T x = U^t b$. Per determinare le soluzioni di A x = b nel senso dei minimi quadrati utilizzando un calcolatore si determina una fattorizzazione QR di A e si risolve il sistema $T x = U^t b$.

4 Dimostrazione omessa.

Lezione 28 (ore 56,57) - 27 novembre 2025, 8:30 - 10:30 F3

(3.31) Scilab.

La funzione predefinita *pinv* di *Scilab* restituisce la matrice pseudoinversa di una matrice. Ad esempio (si veda l'Esempio (3.27) della Lezione 27):

```
--> A = [1,1;1,1;1,1]

 A = [3x2 double]

   1.   1.
   1.   1.
   1.   1.

--> pinv(A)

 ans = [2x3 double]

   0.1666667   0.1666667   0.1666667
   0.1666667   0.1666667   0.1666667
```

La funzione predefinita *backslash* (\) è utilizzata per risolvere un sistema di equazioni lineari. Precisamente, se $A \in \mathbb{R}^{r \times c}$ è una matrice e $b \in \mathbb{R}^r$ è una colonna, dopo l'assegnamento:

```
x = A\b
```

si ha:[1]

- se r = c e $c_1(A) \leqslant \frac{1}{10\,u}$

  allora:

  x è un'approssimazione della soluzione del sistema A x = b calcolata con un procedimento equivalente all'applicazione delle procedure EGPP, SA, SI;

- se r = c e $c_1(A) > \frac{1}{10\,u}$ oppure r > c

  allora:

  x è un'approssimazione di un elemento di $S_{MQ}(A,b)$ - di solito *non* quello di norma minima - calcolato con un procedimento che utilizza una fattorizzazione QR di A.

Ad esempio (vedere l'Eesempio (3.25) della Lezione 27):

```
--> A = [1,1;1,1]

 A = [2x2 double]

   1.   1.
   1.   1.
```

1 Sia N una norma in $\mathbb{R}^n$. In *Scilab*, quando $A \in \mathbb{R}^{n \times n}$ è una matrice *non invertibile*, si pone: $c_N(A) = +\infty$.

```
--> b = [1;0]

 b = [2x1 double]

   1.
   0.

--> x = A\b

 x = [2x1 double]

   0.5000000
   0.

--> y = pinv(A) * b

 y = [2x1 double]

   0.2500000
   0.2500000
```

Le funzione predefinita *qr* restituisce un'approssimazione di una fattorizzazione QR di una matrice, anche non quadrata. Precisamente, se $A \in R^{r \times c}$ con r > c, dopo l'assegnamento:

```
[Q,R] = qr(A)
```

la matrice $Q \in R^{r \times r}$ è un'approssimazione della matrice ortogonale calcolata con il metodo di Householder (Osservazione (2.21) della Lezione 17) applicato ad A e $R \in R^{r \times c}$ è una matrice con elementi nulli sotto la diagonale principale. Ad esempio:

```
--> A = [1,0;1,1;1,1]

 A = [3x2 double]

   1.   0.
   1.   1.
   1.   1.

--> [Q,R] = qr(A)

 Q = [3x3 double]

  -0.5773503   0.8164966  -8.756D-17
  -0.5773503  -0.4082483  -0.7071068
  -0.5773503  -0.4082483   0.7071068

 R = [3x2 double]

  -1.7320508  -1.1547005
   0.         -0.8164966
   0.          0.
```

Per ottenere un'approssimazione di una fattorizzazione QR di A come definita nella Definizione (3.28) della Lezione 27 si può utilizzare la funzione qr come segue:

```
--> [U,T] = qr(A,'e')

 U = [3x2 double]

  -0.5773503   0.8164966
  -0.5773503  -0.4082483
  -0.5773503  -0.4082483

 T = [2x2 double]

  -1.7320508  -1.1547005
   0.         -0.8164966
```

I fattori U,T sono ottenuti dai fattori Q,R eliminando, rispettivamente, la terza colonna di Q e la terza riga di R. Infatti, se si esegue il prodotto Q R per colonne, si osserva che, dette q_1,q_2,q_3 le colonne di Q e r_{ij} gli elementi di R, si ha:

$$Q\,R = (r_{11}\,q_1 + 0\,q_2 + 0\,q_3\ ,\ r_{12}\,q_1 + r_{22}\,q_2 + 0\,q_3) = U\,T$$

(4) METODI NUMERICI PER EQUAZIONI DIFFERENZIALI ORDINARIE

(4.01) Esempio (oscillatore armonico smorzato).

I moti di un oscillatore armonico smorzato sono descritti dall'*equazione differenziale*:

$$(*) \qquad x''(t) + a\,x'(t) + b\,x(t) = 0$$

in cui l'incognita è la *funzione* a valori reali x(t). Questa è un'equazione differenziale del *secondo ordine* (lineare, a coefficienti costanti, omogenea). Una *soluzione* dell'equazione è una funzione y(t) a valori reali *con derivata seconda* che soddisfa l'uguaglianza y"(t) + a y'(t) + b y(t) = 0 per *ogni* t in R. L'equazione differenziale determina *tutti* i possibili moti dell'oscillatore (l'equazione (*) ha *infinite* soluzioni). Ciascuno dei moti è individuato dalle *condizioni iniziali*:

$$(CI) \qquad x(t_0) = x_0 \quad , \quad x'(t_0) = v_0$$

Si chiama *Problema di Cauchy* quello di *determinare le soluzioni dell'equazione differenziale che soddisfano le condizioni iniziali.*

L'equazione differenziale del secondo ordine (*) è *equivalente* ad un *sistema di* due *equazioni del primo ordine.* L'equivalenza significa, in questo caso, che: se y(t) è soluzione dell'equazione (*) allora, posto:

$$u_1(t) = y(t) \quad , \quad u_2(t) = y'(t)$$

si ha:

$$u_1'(t) = u_2(t) \quad , \quad u_2'(t) = - a\, u_2(t) - b\, u_1(t)$$

dunque la colonna $(u_1(t), u_2(t))^t$ è soluzione del sistema

(**) $$x_1'(t) = x_2(t) \quad , \quad x_2'(t) = - a\, x_2(t) - b\, x_1(t)$$

Viceversa: se $(y_1(t), y_2(t))^t$ è una soluzione del sistema (**), allora, posto $y(t) = y_1(t)$ si ha: $y'(t) = y_1'(t) = y_2(t)$ e $y''(t) = y_1''(t) = y_2'(t) = - a\, y_2(t) - b\, y_1(t)$ ovvero:

$$y''(t) + a\, y'(t) + b\, y(t) = 0$$

cioè $y(t)$ è soluzione dell'equazione (*). Inoltre, $y(t)$ è soluzione del Problema di Cauchy:

$$x''(t) + a\, x'(t) + b\, x(t) = 0 \quad ; \quad x(t_0) = x_0 \quad , \quad x'(t_0) = v_0$$

se e solo se $(y(t), y'(t))^t$ è soluzione del Problema di Cauchy:

$$x_1'(t) = x_2(t) \quad , \quad x_2'(t) = - a\, x_2(t) - b\, x_1(t) \quad ; \quad x_1(t_0) = x_0 \quad , \quad x_2(t_0) = v_0$$

(4.02) Osservazione.

Le procedure che descriveremo sono pensate per approssimare *la* soluzione del Problema di Cauchy:

(§) $$x'(t) = F(\, t, x(t)\,) \quad , \quad x(t_0) = x_0$$

per t in un intervallo *limitato* $[t_0,t_f]$. L'*incognita* del problema è la funzione x(t) a valori in R^n; i *dati* sono: *la funzione* F definita in R x R^n a valori in R^n, *gli istanti* t_0 *e* $t_f > t_0$ e *la colonna* x_0 in R^n.

L'asserto precedente presuppone che per il problema (§) si abbia *esistenza ed unicità* della soluzione. Vedremo poi che anche per *descrivere le procedure* sarà necessario fare un'ipotesi ulteriore.

(4.03) Ipotesi (di esistenza ed unicità).

Per ogni $\underline{t}$ in R e $\underline{x}$ in R^n esiste una sola soluzione dell'equazione differenziale:

$$x'(t) = F(\, t, x(t)\,)$$

che verifica la condizione iniziale:

$$x(\underline{t}) = \underline{x}$$

Indicheremo tale soluzione con $y(t; \underline{x}, \underline{t})$.

(4.04) Definizione (metodo numerico).

Un *metodo numerico per l'approssimazione della soluzione del Problema di Cauchy* (§) su $[t_0,t_f]$ è una procedura che costruisce, in base al valore di un parametro E controllato dall'utilizzatore, numeri reali $t(0) = t_0, \ldots, t(N)$ in $[t_0,t_f]$, colonne $x(0) = x_0, \ldots, x(N)$ in

R^n e, per k = 0,...,N, suggerisce di utilizzare x(k) come approssimazione di $y(t(k); x_0,t_0)$.

I numeri t(0),...,t(N) si chiamano *istanti di integrazione* e, per k = 0,...,N-1, il numero h(k) = t(k+1) - t(k) si chiama *passo di integrazione all'istante* t(k).

Una realizzazione in *Scilab* di un metodo numerico ha la struttura seguente:

```
function [T,X] = MetodoNumerico(x0,t0,tf,F,E)

k = 0; t(0) = t0; x(0) = x0;
while t(k) < tf,
   SCEGLI h(k) in base al valore di E;
   CALCOLA x(k+1);
   t(k+1) = t(k) + h(k);
   k = k+1;
end;

endfunction
```

Le variabili di uscita sono, rispettivamente, la riga T e la matrice X tali che:

$$T = (t(0),...,t(N)) \quad , \quad X = (x(0),...,x(N))$$

Un metodo numerico è specificato dalle procedure di *scelta* di h(k) e *calcolo* di x(k+1).

Lezione 29 (ore 58,59) - 2 dicembre 2025, 16:30 - 18:30 F3

(4.05) Definizione (errore totale).

Siano t(k) un istante di integrazione e x(k) la corrispondente approssimazione generati da un metodo numerico per l'approssimazione della soluzione del problema

(§) $\quad x'(t) = F(t, x(t)) \quad , \quad x(t_0) = x_0 \quad , \quad t \in [t_0, t_f]$

La colonna:

$$et(k) = x(k) - y(t(k); x_0, t_0) \in R^n$$

si chiama *errore totale all'istante* t(k). La norma di et(k), che si indica con ET(k), è *una misura di quanto il metodo sbaglia, all'istante* t(k), *nel seguire la soluzione del problema* (§).

(4.06) Definizione (metodo convergente per $E \to 0$).

Un metodo numerico per l'approssimazione della soluzione del problema (§) è *convergente per* $E \to 0$ se: *per ogni* $\Delta > 0$ esiste E_* tale che *se* $E < E_*$ *allora* per gli istanti $t(0) = t_0, \ldots, t(N)$ e le colonne $x(0) = x_0, \ldots, x(N)$ determinati dal metodo si ha:

$$t(N) = t_f \qquad e \qquad \max \{ ET(0), \ldots, ET(N) \} < \Delta$$

(4.07) Definizione (errore locale).

Siano t(k-1) e t(k) due istanti di integrazione consecutivi e x(k-1), x(k) le corrispondenti approssimazioni generati da un metodo numerico per l'approssimazione della soluzione del problema (§). La colonna:

$$el(k) = x(k) - y(t(k); x(k-1), t(k-1)) \in R^n$$

si chiama *errore locale all'istante* t(k). La norma di el(k), che si indica con EL(k), è *una misura di quanto il metodo sbaglia, all'istante* t(k), *nel seguire la soluzione dell'equazione differenziale* x'(t) = F(t,x(t)) *che all'istante* t(k-1) *passa per* x(k-1).

(4.08) Osservazione (relazione tra errore locale e totale).

Si ha:

$$et(k) = x(k) - y(t(k); x_0, t_0) = (x(k) - y(t(k); x(k-1), t(k-1))) + (y(t(k); x(k-1), t(k-1)) - y(t(k); x_0, t_0))$$

da cui:

$$et(k) = el(k) + (y(t(k); x(k-1), t(k-1)) - y(t(k); x_0, t_0))$$

Introducendo la notazione:

$$\Delta y(t''; s, t') = y(t''; y(t'; x_0, t_0) + s, t') - y(t''; y(t'; x_0, t_0), t')$$

si riscrive, infine:

$$et(k) = el(k) + \Delta y(t(k);\ et(k-1),\ t(k-1))$$

La quantità $\Delta y(t";\ s,\ t')$ descrive come *l'equazione differenziale* tramanda all'istante t" lo scostamento, s, all'istante t', dalla soluzione $y(t;\ x_0,t_0)$ del problema (§).

(4.A) METODO TS(1) - EULERO ESPLICITO

(4.09) Ipotesi (regolarità delle soluzioni).

Supponiamo che *tutte* le soluzioni dell'equazione differenziale $x'(t) = F(\ t,x(t)\ )$ abbiano *derivata seconda continua*.

La richiesta è certamente soddisfatta se *tutte* le derivate parziali prime della funzione $F(\ t,x\ )$ *esistono* e sono funzioni *continue* di t ed x.

(Infatti: se y(t) è soluzione dell'equazione differenziale si ha:

$$y"(t) = (y'(t))' = (\ F(\ t,y(t)\ ))' = \partial_t F(\ t,y(t)\ ) + \partial_x F(\ t,y(t)\ ) \cdot y'(t)$$

che risulta continua perché lo sono $\partial_t F(\ t,x\ )$, $\partial_x F(\ t,x\ )$, y(t) e y'(t).)

(4.10) Definizione (metodo TS(1) - Eulero esplicito).

Il *metodo TS(1)* (o *metodo di Eulero esplicito*) è definito dalle procedure seguenti.

- SCELTA di h(k). Dati $E > 0$ e $\lambda > 0$, per ogni k si pone:

  $$d(k) = \max \{\ \lambda\ ,\ ||\ y"(t(k);\ x(k),t(k))\ ||\ \}$$

  e poi:

  $$h(k) = \min \{\ \sqrt{\frac{2E}{d(k)}}\ ,\ t_f - t(k)\ \}$$

- CALCOLO di x(k+1). Dopo aver scelto h(k) si pone:

  $$x(k+1) = x(k) + F(\ t(k),x(k)\ )\ h(k)$$

Il nome del metodo è conseguenza del fatto che la funzione $x(k) + F(\ t(k),x(k)\ )\ h$ si ottiene *troncando al termine di ordine uno* la *serie di Taylor* di $y(t(k) + h;\ x(k),t(k))$ in $h = 0$.

(4.11) Osservazione (sulla scelta di h(k)).

Indicando con y(t) la soluzione $y(t;\ x(k),t(k))$ dell'equazione differenziale, sia s la funzione da R in R^n definita da:

$$s(h) = x(k) + F(\ t(k),x(k)\ )\ h - y(t(k) + h)$$

Detto G il grafico di y(t), il valore s(h) rappresenta lo *scostamento* tra G e la retta tangente a G in (t(k),x(k)), misurato all'istante t(k) + h. *Per* h > 0 *la quantità* s(h) *è l'errore locale all'istante* t(k) + h.

Poiché y(t) ha derivata seconda continua, anche s(h) *ha derivata seconda continua*. Per la Formula di Taylor in h = 0 con resto di Lagrange, esiste una funzione z da R in R^n tale che:

$$s(h) = s(0) + s'(0)\ h + \frac{1}{2}\ s"(0)\ h^2 + z(h)\ h^2 \qquad \text{e} \qquad z(h) \to 0 \text{ per } h \to 0$$

e quindi, essendo s(0) = x(k) - y(t(k)) = 0, s'(0) = F(t(k),x(k)) - y'(t(k)) = 0 e s"(0) = - y"(t(k)):

$$s(h) = -\ \frac{1}{2}\ y"(t(k))\ h^2 + z(h)\ h^2 \qquad \text{con} \qquad z(h) \to 0 \text{ per } h \to 0$$

Se y"(t(k)) non è zero *allora*:

- *Per* h *piccolo*: $-\ \frac{1}{2}\ y"(t(k))\ h^2$ *è una buona stima di* s(h)

  (nel senso che l'errore relativo tende a zero per h → 0)

- Si ha:

$$||\ -\ \frac{1}{2}\ y"(t(k))\ h^2\ || = E \quad \Leftrightarrow \quad h = \sqrt{\frac{2E}{||\ y"(t(k))\ ||}}$$

La scelta di h(k) garantisce che, in ogni caso e per ogni $\lambda > 0$, si ha:

$$||\ -\ \frac{1}{2}\ y"(t(k))\ h(k)^2\ || \leqslant E$$

Il parametro λ ha lo scopo di evitare che possa essere d(k) = 0 e garantisce, inoltre, che:

$$\textit{per ogni } k\text{:} \quad d(k) \geqslant \lambda \quad \textit{e quindi} \quad h(k) \leqslant \sqrt{\frac{2E}{\lambda}}$$

Lezione 30 (ore 60,61) - 3 dicembre 2025, 11:30 - 13:30 A13

(4.12) Teorema (convergenza del metodo TS(1)).

Siano t_0 un numero reale, F una funzione definita in R x R^n a valori in R^n e x_0 in R^n e si consideri il Problema di Cauchy:

(§) $\quad x'(t) = F(\, t,x(t)\,) \quad , \quad x(t_0) = x_0 \quad , \quad t \in [t_0,t_f]$

Se tutte le derivate parziali prime di F(t,x) sono funzioni continue di t ed x e il Problema (§) ha una sola soluzione, *allora* per ogni $\lambda > 0$ il metodo TS(1) applicato al Problema (§) è *convergente per* $E \to 0$ e:

- N *tende a infinito come* $1 / \sqrt{E}$;
- *Per ogni* k: ET(k) *tende a zero come* $\sqrt{E}$.

(4.13) Realizzazione in *Scilab* (TS_1_pv).

```
function [T, X, PASSO] = TS_1_pv(x0, t0, tf, F, G2, E, LAMBDA, HMIN)
//
// Integra numericamente, sull'intervallo [t0,tf], il Problema
// di Cauchy in R(n):
//
// x' = F(t,x)
// x(t0) = x0
//
// con il metodo TS(1) - Eulero esplicito - a passo variabile.
//
// x0: condizione iniziale (colonna di n elementi)
// t0: istante iniziale (numero reale)
// tf: istante finale (numero reale)
// F: function che definisce l'equazione differenziale; F(t,x) deve
//    essere una colonna di n numeri reali
// G2: function che restituisce la derivata seconda in t della soluzione
//     dell'equazione differenziale che all'istante t assume valore x;
//     G2(t,x) deve essere una colonna di n numeri reali
// E: valore massimo della stima dell'errore locale (numero reale)
// LAMBDA: numero reale che stabilisce il valore massimo del passo
//         (OPZIONALE - valore predefinito: 1d-5)
// HMIN: valore minimo consentito del passo
//       (OPZIONALE - valore predefinito: (tf - t0) / 1d6)
//
// T = [t(0),...,t(N)]: riga contenente gli istanti di integrazione
// X = [x(0),...,x(N)]: matrice n x (N+1) contenente le approssimazioni
// PASSO = [h(0),...,h(N-1)]: riga contenente i passi di integrazione
//
// Valore degli argomenti opzionali
//
if ~exists('LAMBDA','l') then LAMBDA = 1d-5; end;
```

```
if ~exists('HMIN','l') then HMIN = (tf - t0) / 1d6; end;
//
// Inizializzazione delle variabili di uscita
//
T(1,1) = t0;
X(:,1) = x0;
PASSO = [];
//
// ciclo principale
//
while (T(1,$) < tf), // arresta la costruzione se ha raggiunto tf
  //
  // scelta del passo
  //
  Nd2x = norm(G2(T(1,$),X(:,$)));
  d = max(LAMBDA, Nd2x);
  PASSO(1,$+1) = min(sqrt(2*E/d), tf - T(1,$));
  //
  // calcolo approssimazione e nuovo istante di integrazione
  //
  X(:,$+1) = X(:,$) + F(T(1,$),X(:,$)) * PASSO(1,$);
  T(1,$+1) = T(1,$) + PASSO(1,$);
  //
  // arresta la costruzione se il passo calcolato risulta troppo
  // piccolo e non ha raggiunto tf
  //
  if (PASSO(1,$) < HMIN) & (T(1,$) < tf) then break; end;
  //
end;
//
// Verifica se l'integrazione ha raggiunto tf
//
if T(1,$) < tf then
   printf("\n\nIntegrazione interrotta a T = %3.2e", T(1,$));
end;
//
endfunction
```

(4.14) Esempio (svolto in classe il 4 dicembre).

Si consideri un pendolo realizzato da un punto pesante di massa m collegato da un filo inestensibile di lunghezza L ad un punto fisso. Supposto piano il moto del punto ed adottato l'angolo x tra la verticale discendente ed il filo, misurato in senso antiorario, come coordinata lagrangiana, l'equazione del moto risulta:

(ED) $$x''(t) = -\frac{g}{L}\,\mathrm{sen}\,x(t)$$

Per approssimare nell'intervallo $[t_0,t_f] = [0,3]$ s la soluzione del Problema di Cauchy che si ottiene considerando le condizioni iniziali:

(CI) $$x(0) = x_0 = \pi/4 \text{ rad} \quad , \quad x'(0) = 0$$

si utilizza, in *Scilab*, la procedura TS_1_pv. L'uso della procedura richiede:

- *La determinazione di un sistema di due equazioni differenziali di ordine uno equivalente all'equazione* (ED). Introdotte le variabili $u_1(t) = x(t)$, $u_2(t) = x'(t)$ si ottiene:

  $$(ED') \qquad u_1'(t) = u_2(t) \quad , \quad u_2'(t) = -\frac{g}{L}\,\mathrm{sen}\,u_1(t)$$

  che si completa con le condizioni iniziali:

  $$(CI') \qquad u_1(0) = x_0 \quad , \quad u_2(0) = 0$$

- *La scrittura della* function *che definisce il sistema* (ED'):

```
function y = F(t,u)

   y = [                  u(2)  ;
         - (g/L) * sin( u(1) ) ];

endfunction
```

- *La determinazione della funzione che*, dati t ed u, *restituisce il valore della derivata seconda*, calcolata in t, *della soluzione del sistema* (ED') che passa per u all'istante t:

  $$u''(t) = \begin{bmatrix} u_2'(t) \\ -(g/L)\,u_1'(t)\cos(u_1(t)) \end{bmatrix} = \begin{bmatrix} -(g/L)\,\mathrm{sen}(u_1(t)) \\ -(g/L)\,u_2(t)\cos(u_1(t)) \end{bmatrix}$$

  e la scrittura della relativa function:

```
function y = G2(t,u)

   y = [        - (g/L) * sin( u(1) );
         - (g/L) * u(2) * cos( u(1) ) ];

endfunction
```

- *L'assegnamento dell'istante finale* t_f (s):

```
tf = 3;
```

- *L'assegnamento della colonna delle condizioni iniziali* (CI'):

```
u0 = [x0;0];
```

- *L'assegnamento del valore ai parametri*:

```
g = 9.82; // m/s^2
L = 1; // m
m = 1; // kg
```

- *La scelta del valore massimo consentito per la stima dell'errore locale*, E.

  Per ottenere un valore di E adeguato, occorre un criterio per giudicare l'accuratezza dell'approssimazione ottenuta dalla procedura. Per il sistema fisico in esame possiamo procedere come segue.

  (A) Considerato che durante il moto l'*energia meccanica*:

$$EN(x(t)) = mgL(1-\cos x_1(t)) + \frac{1}{2}mL^2(x_2(t))^2$$

  assume valore costante e pari al valore $EN(t_0)$ assunto all'istante t_0, come misura *relativa* dell'accuratezza dell'approssimazione possiamo scegliere *la variazione relativa dell'energia durante il moto approssimato*:

$$Var\_EN = \frac{\max EN(u(t_k)) - \min EN(u(t_k))}{EN(u(t_0))}$$

  (B) Considerato che il moto del pendolo è periodico e che si ha:

$$\min x_1(t) = -\max x_1(t) \quad \Rightarrow \quad \max x_1(t) + \min x_1(t) = 0$$

  come misura *relativa* dell'accuratezza dell'approssimazione possiamo scegliere *la variazione relativa dell'ampiezza dell'oscillazione durante il moto approssimato*:

$$Var\_A = \frac{\max u_1(t_k) + \min u_1(t_k)}{u_1(t_0)}$$

  Questa scelta è ragionevole se l'intervallo $[t_0,t_f]$ include almeno una oscillazione della funzione $u_1(t_k)$.

  (C) Si ottiene la tabella che segue:

E	N	Var_EN (%)	Var_A (%)
10^{-3}	267	35.89	6.3
10^{-5}	2587	3.25	$5.99\ 10^{-1}$
10^{-7}	25779	0.32	$5.97\ 10^{-2}$

  Quale sia un valore di E adeguato dipende da quello che l'utilizzatore vuole ottenere. La tabelle suggerisce che al diminuire di E l'accuratezza dell'approssimazione aumenta.

(4.15) Osservazione (variazione di N e ET con E).

Siano N e M, rispettivamente, il numero di istanti di integrazione e il massimo valore di ET(k) ottenuto utilizzando la procedura TS_1_pv con E = $\underline{E}$ e N' e M' i corrispondenti valori ottenuti con E = $\alpha\,\underline{E}$. Per quanto detto nel Teorema (4.12) ci si aspetta che:

$$N'/N \approx 1/\alpha^{1/2} \quad \text{e} \quad M'/M \approx \alpha^{1/2}$$

Nella tabella finale dell'esempio precedente si ha $\alpha = 10^{-2}$, dunque ci si aspetta:

$$N'/N \approx 10 \quad \text{e} \quad M'/M \approx 1/10$$

La relazione riguardante l'aumento del numero di istanti di integrazione è evidentemente verificata:

$$2587/267 = 9.69 \quad \text{e} \quad 25779/2587 = 9.96$$

Non avendo possibilità di accedere all'errore totale, ci limitiamo a constatare che per la variazione relativa dell'energia si ha:

$$\text{Var\_EN}'/\text{Var\_EN} = 3.25/35.89 \approx 0.90\ 10^{-1} \quad \text{e} \quad 0.32/3.25 \approx 0.98\ 10^{-1}$$

e per la variazione relativa dell'ampiezza:

$$\text{Var\_A}'/\text{Var\_A} = 5.99\ 10^{-1}/6.3 \approx 0.95\ 10^{-1} \quad \text{e} \quad 5.97\ 10^{-2}/5.99\ 10^{-1} \approx 0.99\ 10^{-1}$$

(4.B) METODO TS(2)

(4.16) Ipotesi (regolarità delle soluzioni).

Supponiamo che *tutte* le soluzioni dell'equazione differenziale $x'(t) = F(t,x(t))$ abbiano *derivata terza continua*.

La richiesta è certamente soddisfatta se *tutte* le derivate parziali *seconde* della funzione $F(t,x)$ *esistono* e sono funzioni *continue* di t ed x.

(Infatti:

$$G_2(t,x) = \partial_t F(t,x) + \partial_x F(t,x) \cdot F(t,x)$$

ha derivate parziali prime continue e quindi:

$$G_3(t,x) = \partial_t G_2(t,x) + \partial_x G_2(t,x) \cdot F(t,x)$$

è continua. Allora, se $y(t)$ è soluzione dell'equazione differenziale:

$$y^{(3)}(t) = ((y'(t))')' = ((F(t,y(t)))')' = (G_2(t,y(t)))' = G_3(t,y(t))$$

è continua perché lo sono $G_3(t,x)$ ed $y(t)$.)

(4.17) Definizione (metodo TS(2)).

Il *metodo* TS(2) è definito dalle procedure seguenti.

- SCELTA di $h(k)$. Dati $E > 0$ e $\lambda > 0$, per ogni k si pone:

$$d(k) = \max \{ \ \lambda \ , \ \| y^{(3)}(t(k); x(k),t(k)) \| \ \}$$

e poi:

$$h(k) = \min \{ \ \sqrt[3]{\frac{6E}{d(k)}} \ , \ tf - t(k) \ \}$$

- CALCOLO di x(k+1). Dopo aver scelto h(k) si pone:

$$x(k+1) = x(k) + F(t(k),x(k)) \, h(k) + \frac{1}{2} \, G_2(t(k),x(k)) \, h(k)^2$$

Il nome del metodo è conseguenza del fatto che la funzione di h utilizzata per il calcolo di x(k+1) si ottiene troncando al termine di ordine *due* la *serie di Taylor* di y(t(k) + h; x(k),t(k)) in h = 0.

(4.18) Osservazione (sulla scelta di h(k)).

Indicando con y(t) la soluzione y(t; x(k),t(k)) dell'equazione differenziale, per lo scostamento s(h) tra y(t(k) + h) e l'approssimazione calcolata dal metodo con un passo di lunghezza h a partire da (t(k),x(k)) si ha, utilizzando la Formula di Taylor in h = 0 con resto di Lagrange:

$$s(h) = - \frac{1}{6} \, y^{(3)}(t(k)) \, h^3 + z(h) \, h^3 \qquad \text{con: } z(h) \to 0 \text{ per } h \to 0$$

Se $y^{(3)}(t(k))$ non è zero *allora*:

- *per* h *piccolo*: $- \frac{1}{6} \, y^{(3)}(t(k)) \, h^3$ *è una buona stima di* s(h)
- si ha:

$$\| -\frac{1}{6} \, y^{(3)}(t(k)) \, h^3 \| = E \quad \Leftrightarrow \quad h = \sqrt[3]{\frac{6E}{\| y^{(3)}(t(k)) \|}}$$

Il parametro λ garantisce che:

$$\text{per ogni k:} \quad d(k) \geqslant \lambda \quad \text{e quindi} \quad h(k) \leqslant \sqrt[3]{\frac{6E}{\lambda}}$$

(4.19) Teorema (convergenza del metodo TS(2)).

Siano t_0 e $t_f > t_0$ numeri reali, F una funzione definita in R x R^n a valori in R^n, x_0 in R^n e si consideri il Problema di Cauchy:

$$(\S) \qquad x'(t) = F(t,x(t)) \ , \ x(t_0) = x_0 \ , \ t \in [t_0,t_f]$$

Se tutte le derivate parziali seconde di F(t,x) sono funzioni continue di t ed x e il Problema (§) ha una sola soluzione, *allora* per ogni $\lambda > 0$ il metodo TS(2) applicato al Problema (§) è *convergente per* E → 0 e:

- N *tende a infinito come* $1 / \sqrt[3]{E}$;

- *Per ogni* k: ET(k) *tende a zero come* $\sqrt[3]{E^2} = E^{2/3}$

(4.20) Osservazione.

Si consideri il Problema di Cauchy (§). Per ogni E > 0, indichiamo con $N_1(E)$ e $ET_1(E)$ il numero di istanti di integrazione e l'errore totale massimo generati dal metodo TS(1) e con $N_2(E)$ e $ET_2(E)$ il numero di istanti di integrazione e l'errore totale massimo generati dal metodo TS(2). Per quanto detto nel Teorema (4.12) e nel Teorema (4.19), per $E \to 0$ si ha:

- $N_1(E) / N_2(E) \to +\infty$ come $1 / \sqrt[6]{E}$, dunque $N_1(E)$ *tende ad* ∞ *più rapidamente di* $N_2(E)$
- $ET_1(E) / ET_2(E) \to +\infty$ come $1 / \sqrt[6]{E}$, dunque $ET_2(E)$ *tende a* 0 *più rapidamente di* $ET_1(E)$

Ci si aspetta allora che, con lo stesso valore di E:

- TS(2) generi un *errore totale massimo più piccolo* di quello generato con TS(1)
- TS(2) raggiunga t_f con un *numero di passi inferiore* rispetto a TS(1)

Lezione 32 (ore 64,65) - 9 dicembre 2025, 16:30 - 18:30 F3

(4.C) METODI RUNGE-KUTTA

(4.21) Esempio.

Nel metodo TS(2) è richiesta all'utilizzatore la determinazione e realizzazione delle funzioni:

$$G_2(t,x) \quad \text{per il calcolo di } x(k+1)$$

e:

$$G_3(t,x) \quad \text{per la scelta di } h(k)$$

In generale il compito è tanto più gravoso quanto più alto è l'ordine del metodo: nel metodo TS(p) l'utilizzatore deve determinare e realizzare le funzioni:

$$G_2(t,x),\ldots,G_p(t,x) \quad \text{per il calcolo di } x(k+1)$$

e:

$$G_{p+1}(t,x) \quad \text{per la scelta di } h(k)$$

I *metodi Runge-Kutta* sono pensati per eliminare questo onere.

Per introdurre la *struttura* dei metodi, vediamo come si trasforma il calcolo di x(k+1) nel metodo TS(2) utilizzando una stima numerica del valore $G_2(t,x)$.

(4.22) Osservazione (stima numerica di G_2)

Il valore $G_2(t(k),x(k)) = y''(t(k))$ può essere *stimato* con le considerazioni seguenti:

(a) Per definizione:

$$\frac{y'(t(k)+\tau) - y'(t(k))}{\tau} \to y''(t(k)) \quad \text{per} \quad \tau \to 0$$

dunque:

per τ piccolo $\dfrac{y'(t(k)+\tau) - y'(t(k))}{\tau}$ *è una buona approssimazione di* $y''(t(k))$

(b) Poiché y(t) è la soluzione dell'equazione differenziale che vale x(k) all'istante t(k) si ha:

$$y'(t(k)) = F(\, t(k), y(t(k))\,) = F(\, t(k), x(k)\,)$$

e:

$$y'(t(k)+\tau) = F(\, t(k)+\tau\ ,\ y(t(k)+\tau)\,)$$

Quest'ultimo valore *non è calcolabile* perché, assegnato τ, la procedura non conosce

$y(t(k) + \tau)$. Allora:

si approssima $y(t(k) + \tau)$ *con* $y(t(k)) + y'(t(k))\ \tau = x(k) + F(\ t(k),x(k)\ )\ \tau$

Complessivamente:

scelto τ *piccolo, si stima* $G_2(t(k),x(k)) = y''(t(k))$ *con*

$$\frac{F(\ t(k) + \tau\ ,\ x(k) + F(\ t(k),x(k)\ )\ \tau\ ) - F(\ t(k),x(k)\ )}{\tau}$$

Questa quantità, dato τ, *è calcolabile senza usare* G_2.

La stima è *ragionevole*. Infatti, indicando con F(k) il valore $F(\ t(k),x(k)\ )$, si consideri la funzione di τ:

$$H(\tau) = F(\ t(k) + \tau\ ,\ x(k) + F(k)\ \tau\ )$$

Poiché si suppone che F(t,x) abbia *derivate parziali prime* continue, anche H ha derivata prima continua. Allora:

$$H(\tau) = H(0) + H'(0)\ \tau + z(\tau)\ \tau \quad \text{con } z(\tau) \to 0 \text{ per } \tau \to 0$$

Ma: $H(0) = F(k)$ e

$$H'(0) = \partial_t F(\ t(k),x(k)\ ) + \partial_x F(\ t(k),x(k)\ ) \cdot F(\ t(k),x(k)\ ) = G_2(\ t(k),x(k)\ ) = y''(t(k))$$

dunque:

$$H(\tau) = F(k) + y''(t(k))\ \tau + z(\tau)\ \tau$$

e:

$$\frac{H(\tau) - F(k)}{\tau} - y''(t(k)) = z(\tau) \to 0 \text{ per } \tau \to 0$$

(4.23) Osservazione (uso della stima numerica).

In TS(2):

$$x(k+1) = x(k) + F(\ t(k),x(k)\ )\ h(k) + \frac{1}{2}\ G_2(\ t(k),x(k)\ )\ h(k)^2$$

Scegliendo $\tau = h(k)$ nella stima dell'Osservazione (4.22) si ottiene:

$$G_2(\ t(k),x(k)\ ) = \frac{F(\ t(k) + h(k)\ ,\ x(k) + F(\ t(k),x(k)\ )\ h(k)\ ) - F(\ t(k),x(k)\ )}{h(k)}$$

da cui (posto $F(k) = F(\ t(k),x(k)\ )$):

$$x(k+1) = x(k) + F(k)\ h(k) + \frac{1}{2}\ [\ F(\ t(k) + h(k)\ ,\ x(k) + F(k)\ h(k)\ ) - F(k)\ ]\ h(k)$$

$$= x(k) + \frac{1}{2}\ [\ F(k) + F(\ t(k) + h(k)\ ,\ x(k) + F(k)\ h(k)\ )\ ]\ h(k)$$

Questa procedura di calcolo di x(k+1) può essere riscritta, in modo più semplicemente generalizzabile, come segue: *il valore* x(k+1) *si ottiene*, dopo aver scelto h(k), *ponendo*:

- $ST_1 = F(t(k),x(k))$
- $ST_2 = F(t(k) + h(k) , x(k) + ST_1\ h(k))$

e poi

$$x(k+1) = x(k) + \frac{1}{2}\ (ST_1 + ST_2)\ h(k)$$

(4.24) Definizione (metodi RK a due e tre stadi).

Si chiamano *metodi Runge-Kutta* (RK) *a due stadi* quelli nei quali, scelti opportunamente numeri reali c_2, a_{21}, b_1 e b_2, il valore x(k+1) si ottiene, dopo aver scelto h(k), ponendo:

- $ST_1 = F(t(k),x(k))$
- $ST_2 = F(t(k) + c_2\ h(k) , x(k) + a_{21}\ ST_1\ h(k))$

e poi

- $x(k+1) = x(k) + (b_1\ ST_1 + b_2\ ST_2)\ h(k)$

Si chiamano *metodi Runge-Kutta* (RK) *a tre stadi* quelli nei quali, scelti opportunamente numeri reali c_2, c_3, a_{21}, a_{31}, a_{32}, b_1, b_2 e b_3, il valore x(k+1) si ottiene, dopo aver scelto h(k), ponendo:

- $ST_1 = F(t(k),x(k))$
- $ST_2 = F(t(k) + c_2\ h(k) , x(k) + a_{21}\ ST_1\ h(k))$
- $ST_3 = F(t(k) + c_3\ h(k) , x(k) + [\ a_{31}\ ST_1 + a_{32}\ ST_2\ ]\ h(k))$

e poi

- $x(k+1) = x(k) + (b_1\ ST_1 + b_2\ ST_2 + b_3\ ST_3)\ h(k)$

(4.25) Definizione (ordine di un metodo per $h \to 0$)

Sia s(h) la funzione scostamento per il metodo in esame. Il numero intero p si dice *ordine del metodo per* $h \to 0$ se:

$$s^{(m)}(0) = 0 \quad \text{per } m = 0,\dots,p \quad \text{e} \quad s^{(p+1)}(0) \neq 0$$

ovvero se il primo termine dello sviluppo di Taylor di s(h) per h = 0 è quello di ordine p+1:

$$s(h) = \frac{1}{(p+1)!}\ s^{(p+1)}(0)\ h^{p+1} + \dots$$

(4.26) Osservazione (determinazione dei parametri in un metodo Runge-Kutta).

In un metodo Runge-Kutta a più stadi i valori dei *parametri* c_i, a_{ij}, b_i sono determinati (*non univocamente*) dalla condizione che: per *ogni* funzione F che definisce il Problema di

Cauchy, l'ordine del metodo per $h \to 0$, sia il *più elevato possibile*.

(4.27) Esempio.

Si consideri il metodo Runge-Kutta a due stadi. Per ogni k, posto $y(t) = y(t;x(k),t(k))$, si ha:

$$s(h) = x(k) + [b_1\, ST_1 + b_2\, ST_2(h)]\, h - y(t(k) + h)$$

Allora:

$$s^{(1)}(h) = b_1\, ST_1 + b_2\, ST_2'(h)\, h + b_2\, ST_2(h) - F[t(k) + h, y(t(k) + h)]$$

$$s^{(2)}(h) = b_2\, ST_2''(h)\, h + 2\, b_2\, ST_2'(h) - \partial_t F[t(k) + h, y(t(k) + h)] - \\ - F[t(k) + h, y(t(k) + h)] \cdot \partial_t F[t(k) + h, y(t(k) + h)]$$

da cui, essendo $ST_2(0) = ST_1 = F[t(k),x(k)]$:

$$s^{(0)}(0) = s(0) = x(k) - y(t(k)) = 0$$

$$s^{(1)}(0) = (b_1 + b_2 - 1)\, F[t(k),x(k)]$$

$$s^{(2)}(0) = 2\, b_2\, ST_2'(0) - \partial_t F[t(k),x(k)] - F[t(k),x(k)] \cdot \partial_t F[t(k),x(k)]$$

Poi, posto $F[t(k) + c_2\, h\ ,\ x(k) + a_{21}\, ST_1\, h] = F_k(h)$ e quindi $ST_1 = F[t(k),x(k)] = F_k(0)$:

$$ST_2'(h) = c_2\, \partial_t F_k(h) + a_{21}\, F_k(0)\, \partial_x F_k(h)$$

da cui:

$$s^{(2)}(0) = (2\, b_2\, c_2 - 1)\, \partial_t F_k(0) + (2\, b_2\, a_{21} - 1)\, F_k(0)\, \partial_x F_k(0)]$$

Infine:

$$s^{(1)}(0) = 0 \text{ per ogni } F \quad \Leftrightarrow \quad b_1 + b_2 - 1 = 0$$
$$s^{(2)}(0) = 0 \text{ per ogni } F \quad \Leftrightarrow \quad 2\, b_2\, c_2 - 1 = 0 \ \text{ e } \ 2\, b_2\, a_{21} - 1 = 0$$

e il metodo risulta di ordine *almeno* due per $h \to 0$ se e solo se:

$$b_1 + b_2 = 1 \quad , \quad 2\, b_2\, c_2 = 1 \quad , \quad 2\, b_2\, a_{21} = 1$$

Ad esempio:

$$b_1 = b_2 = 1/2 \quad , \quad c_2 = a_{21} = 1 \quad \text{(metodo di Heun)}^1$$

$$b_1 = 0 \quad , \quad b_2 = 1 \quad , \quad c_2 = a_{21} = 1/2 \quad \text{(metodo di Eulero modificato o del punto medio)}$$

(4.28) Osservazione.

Per un metodo di ordine p si ha:[2]

- N *tende a infinito come* $1 / \sqrt[p+1]{E}$;
- *Per ogni* k: ET(k) *tende a zero come* $\sqrt[p+1]{E^p} = E^{\frac{p}{p+1}}$

---

1 É il metodo dell'Esempio (4.23), detto anche 'metodo di Eulero migliorato'.

2 Dimostrazione omessa.

(4.29) Osservazione.

La Definizione (4.24) si estende a metodi con un numero qualsiasi di stadi. Inoltre: l'ordine massimo di un metodo ad uno stadio è *uno* (esiste un solo metodo ad uno stadio di ordine uno: il metodo TS(1)), di un metodo a due stadi è *due* e di un metodo a tre stadi è *tre*. In generale, l'ordine massimo di un metodo è *minore o uguale al numero di stadi*.

Lezione 33 (ore 66,67) - 10 dicembre 2025, 11:30 - 13:30 A13

(4.30) Osservazione (scelta di h(k) nei metodi Runge-Kutta).

Coerentemente con l'intento di eliminare l'onere della determinazione e realizzazione delle funzioni $G_j(t,x)$, la scelta di h(k) nei metodi RK avviene, usualmente, come segue.

Siano: RK il metodo di Runge-Kutta, di ordine p per $h \to 0$, scelto per il calcolo di x(k+1) e RK' un altro metodo di Runge-Kutta, di ordine p' = p+1. Allora:

- SCELTA di h(k). Dati $E > 0$ e $\lambda > 0$, per ogni k si sceglie τ piccolo, si calcolano:

  (1) XX = un passo di RK a partire da (x(k),t(k)), di lunghezza τ
  (2) XX' = un passo di RK' a partire da (x(k),t(k)), di lunghezza τ

  si pone:

$$d(k) = \max \{ \lambda , \| XX - XX' \| \}$$

  e poi:

$$h(k) = \min \{ \sqrt[p+1]{\frac{E}{d(k)}}\ \tau \ , tf - t(k) \}$$

Questa procedura di scelta si spiega considerando che:

(a) Il metodo RK ha ordine p per $h \to 0$ dunque, posto $C = \frac{s^{(p+1)}(0)}{(p+1)!}$:

$$\textit{si stima } s(h) \textit{ con } C\, h^{p+1}$$

(b) Poiché:

$$XX - y(t(k) + \tau) = C\, \tau^{p+1} + z(\tau)\, \tau^{p+1} \ , \quad \text{con } z(\tau) \to 0 \text{ per } \tau \to 0 \quad (\text{RK ha ordine p})$$

$$XX' - y(t(k) + \tau) = C'\, \tau^{p+2} + w(\tau)\tau^{p+2} \ , \quad \text{con } w(\tau) \to 0 \text{ per } \tau \to 0 \quad (\text{RK' ha ordine p+1})$$

allora:

$$\frac{XX - XX'}{\tau^{p+1}} = C + [\, z(\tau) - (C' + w(\tau))\ \tau \,] \to C \quad \text{per } \tau \to 0$$

e:

$$\textit{scelto } \tau \textit{ piccolo, si stima } C \textit{ con } \frac{XX - XX'}{\tau^{p+1}}$$

(c) Complessivamente:

$$\textit{scelto } \tau \textit{ piccolo, si stima } s(h) \textit{ con } \frac{XX - XX'}{\tau^{p+1}}\ h^{p+1}$$

dunque:

$$\left\| \frac{XX - XX'}{\tau^{p+1}} \, h^{p+1} \right\| = E \qquad \Leftrightarrow \qquad h = \sqrt[p+1]{\frac{E}{\| XX - XX' \|}} \, \tau$$

(4.31) Realizzazione in *Scilab* (RK12_pv).

Come esempio di realizzazione, consideriamo il metodo RK che utilizza Eulero esplicito, di ordine 1 per h → 0, per il calcolo di x(k+1) e che sceglie h(k) affiancandolo con il metodo dell'Osservazione (4.23), metodo di Heun di ordine 2 per h → 0. Ne risulta un metodo di ordine 1 per h → 0 e quindi convergente di ordine 1/2 per E → 0.

```
01  function [T, X, PASSO] = RK12_pv(x0, t0, tf, F, E, LAMBDA, HMIN, TAU)
02  //
03  // Integra numericamente, sull'intervallo [t0,tf], il Problema
04  // di Cauchy in R(n):
05  //
06  // x' = F(t,x)
07  // x(t0) = x0
08  //
09  // con il metodo di Eulero esplicito (RK di ordine 1) - a passo
10  // variabile - affiancato, per la scelta del passo, dal metodo
11  // RK di ordine 2 definito da c(2) = 1, a(21) = 1 e b(1) = b(2) = 1/2.
12  //
13  // x0: condizione iniziale (colonna di n elementi)
14  // t0: istante iniziale (numero reale)
15  // tf: istante finale (numero reale)
16  // F: function che definisce l'equazione differenziale - F(t,x) deve
17  //    essere una colonna di n numeri reali
18  // E: valore massimo della stima dell'errore locale (numero reale)
19  // LAMBDA: numero reale che stabilisce il valore massimo del passo
20  //         (OPZIONALE - valore predefinito: 1d-5)
21  // HMIN: valore minimo consentito del passo
22  //       (OPZIONALE - valore predefinito: (tf - t0) / 1d6)
23  // TAU: valore del passo per il calcolo delle stime utilizzate
24  //      nella scelta di h(k) (OPZIONALE - valore predefinito: (tf - t0) / 1d3)
25  //
26  // T = [t(0),...,t(N)]: riga contenente gli istanti di integrazione
27  // X = [x(0),...,x(N)]: matrice n x (N+1) contenente le approssimazioni
28  // PASSO = [h(0),...,h(N-1)]: riga contenente i passi di integrazione
29  //
30  // Valore degli argomenti opzionali
31  //
32  if ~exists('LAMBDA','l') then LAMBDA = 1d-5; end;
33  if ~exists('HMIN','l') then HMIN = (tf - t0) / 1d6; end;
34  if ~exists('TAU','l') then TAU = (tf - t0) / 1d3; end;
35  //
36  // Inizializzazione delle variabili di uscita
37  //
38  T(1,1) = t0;
39  X(:,1) = x0;
40  PASSO = [];
41  //
42  // ciclo principale
```

```
43  //
44  while (T(1,$) < tf), // arresta la costruzione se ha raggiunto tf
45    //
46    // scelta del passo
47    //
48    // XX1 = X(:,$) + F(T(1,$),X(:,$)) * TAU;
49    ST1 = F(T(1,$),X(:,$));
50    ST2 = F( T(1,$) + TAU,X(:,$) + ST1 * TAU );
51    // XX2 = X(:,$) + ( (ST1 + ST2)/2 ) * TAU;
52    //
53    //     XX1 - XX2 = (ST1 - ST2)/2 * TAU
54    //
55    d = max(LAMBDA, norm( ((ST1 - ST2)/2) * TAU ));
56    PASSO(1,$+1) = min(sqrt(E/d) * TAU, tf - T(1,$));
57    //
58    // calcolo approssimazione e nuovo istante di integrazione
59    //
60    X(:,$+1) = X(:,$) + F(T(1,$),X(:,$)) * PASSO(1,$);
61    T(1,$+1) = T(1,$) + PASSO(1,$);
62    //
63    // arresta la costruzione se il passo calcolato risulta troppo
64    // piccolo e non ha raggiunto tf
65    //
66    if (PASSO(1,$) < HMIN) & (T(1,$) < tf) then break; end;
67    //
68  end;
69  //
70  // Verifica se l'integrazione ha raggiunto tf
71  //
72  if T(1,$) < tf then
73     printf("\n\nIntegrazione interrotta a T = %3.2e", T(1,$));
74  end;
75  //
76  endfunction
77  //
78  // Esempio per assegnare valori ai parametri opzionali:
79  //
80  //    [T,X,PASSO] = RK12_pv(x0,t0,tf,F,G,E,HMIN = y);
81  //
82  //         => LAMBDA = valore predefinito, HMIN = y, TAU = valore predefinito
83  //
```

Si osservi che:

- Nella scelta del passo la differenza XX1 - XX2 può essere determinata *senza calcolare* XX1 ed XX2 (righe 48-55). Risulta infatti:

$$\text{XX1 - XX2} = \frac{\text{ST1 - ST2}}{2}\ \text{TAU}$$

- Per la scelta del passo si è utilizzato *lo stesso valore di τ ad ogni iterazione.*

Il file che contiene la procedura, insieme ad un esempio di applicazione all'equazione del pendolo (la stessa dell'Esempio (4.14) della Lezione 30), si può trovare nella pagina web del corso, sezione "altro materiale didattico".

Lezione 34 (ore 68,69) - 11 dicembre 2025, 8:30 - 10:30 F3

In questa lezione svolgiamo alcuni esercizi.

Esercizio 1

Sia:

$$F(x) = \begin{bmatrix} x_1 - x_2 - 1 \\ x_1^2 + x_2^2 - 1 \end{bmatrix} : R^2 \to R^2$$

(1) Determinare graficamente gli zeri di F;
(2) Posto G(x) = x - F(x), verificare che gli zeri di F sono tutti e soli i punti uniti di G;
(3) Decidere se il metodo ad un punto definito da G sia utilizzabile per approssimare gli zeri di F;
(4) Dato $x(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, determinare l'elemento x(1) ottenuto utilizzando un passo del metodo di Newton applicato ad F;
(5) Decidere se il metodo di Newton applicato ad F sia utilizzabile per approssimare gli zeri di F.

Soluzione.

(1) Posto:

$$F_1(x) = x_1 - x_2 - 1 \quad \text{e} \quad F_2(x) = x_1^2 + x_2^2 - 1$$

l'equazione F(x) = 0 è equivalente al sistema:

$$F_1(x) = 0 \quad \text{e} \quad F_2(x) = 0$$

L'insieme degli zeri di F_1 è la retta di equazione $x_2 = x_1 - 1$; l'insieme degli zeri di F_2 è la circonferenza di equazione $x_1^2 + x_2^2 = 1$, di centro l'origine e raggio 1. Rappresentando graficamente i due insiemi in un piano cartesiano si determinano i *due* zeri di F:

$$\alpha_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{e} \quad \alpha_2 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

(2) L'equazione x = G(x) si riscrive: x = x + F(x), e quest'ultima è equivalente all'equazione F(x) = 0. Dunque Le equazioni x = G(x) e F(x) = 0 sono equivalenti ossia *hanno le stesse soluzioni*. Le soluzioni della prima sono i *punti uniti di* G, quelle della seconda sono gli *zeri di* F.

(3) Per quanto detto nella Lezione 14, il metodo definito da G è utilizzabile per approssimare il punto unito α_k se e solo se il *raggio spettrale*[1] della matrice jacobiana di G calcolata in α_k, $J_G(\alpha_k)$, è minore di 1. La matrice jacobiana di G è:

1 Si veda la Definizione (2.65) nella Lezione 22.

$$J_G(x) = \begin{bmatrix} 2 & -1 \\ 2x_1 & 2x_2+1 \end{bmatrix}$$

Per α_1 si ha:

$$J_G(\alpha_1) = \begin{bmatrix} 2 & -1 \\ 2 & 1 \end{bmatrix}$$

Il polinomio caratteristico è:

$$\det(J_G(\alpha_1) - \lambda I) = (2 - \lambda)(1 - \lambda) + 2 = \lambda^2 - 3\lambda + 4$$

e gli autovalori sono:

$$\lambda_1 = \frac{3 + i\sqrt{7}}{2} \qquad \text{e} \qquad \lambda_2 = \frac{3 - i\sqrt{7}}{2}$$

Allora: $\rho(J_G(\alpha_1)) > 1$ e il metodo definito da G *non è utilizzabile* per approssimare α_1.

Per α_2 si ha:

$$J_G(\alpha_1) = \begin{bmatrix} 2 & -1 \\ 0 & -1 \end{bmatrix}$$

Gli autovalori sono:

$$\lambda_1 = 2 \qquad \text{e} \qquad \lambda_2 = -1$$

Di nuovo: $\rho(J_G(\alpha_2)) > 1$ e il metodo definito da G *non è utilizzabile* neppure per approssimare α_2.

(4) Il metodo di Newton applicato ad F è il metodo ad un punto definito dalla funzione:

$$N(x) = x - J_F(x)^{-1} F(x) : R^2 \to R^2$$

Si ha:

$$J_F(x) = \begin{bmatrix} 1 & -1 \\ 2x_1 & 2x_2 \end{bmatrix} \qquad \text{e} \qquad J_F(x(0)) = \begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix}$$

La matrice $J_F(x(0))$ è invertibile, dunque x(1) è definito e si ha:

$$x(1) = N(x(0)) \qquad \text{ovvero} \qquad x(1) = x(0) - J_F(x(0))^{-1} F(x(0))$$

Detta v la soluzione del sistema $J_F(x(0))\, z = F(x(0))$, si riscrive:

$$x(1) = x(0) - v$$

Si ha:

$$v = \begin{bmatrix} -2 \\ 0 \end{bmatrix} \qquad \text{e infine:} \qquad x(1) = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

(5) Per quanto detto nell'Osservazione (1.90) della Lezione 14, condizione *sufficiente* per l'utilizzabilità del metodo di Newton per approssimare lo zero α_k di F è che: F abbia derivare (parziali) seconde continue in un intorno di α_k e $J_F(\alpha_k)$ sia invertibile. Nel caso in esame le funzioni F_1 ed F_2 hanno derivate parziali di ogni ordine su R^2 e sia $J_F(\alpha_1)$ che $J_F(\alpha_2)$ sono invertibili. Il metodo di Newton risulta quindi utilizzabile per approssimare entrambi gli zeri di F.

Esercizio 2

Siano:

$$A = \begin{bmatrix} 4 & 1 & & 1 \\ 1 & 4 & & \\ & & 2 & \\ 1 & & & 4 \end{bmatrix} \quad \text{e} \quad b = \begin{bmatrix} 5 \\ 1 \\ 0 \\ 5 \end{bmatrix}$$

(1) Decidere se la matrice A è a predominanza diagonale forte per righe;
(2) Determinare la matrice H_J e la colonna c_J che definiscono il metodo di Jacobi applicato al sistema $A\,x = b$;
(3) Determinare lo spettro ed il raggio spettrale di H_J;
(4) Determinare $\| H_J \|_\infty$;
(5) Decidere se il metodo di Jacobi è convergente;
(6) Dato $x(0) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$, determinare l'elemento x(1) ottenuto utilizzando un passo del metodo di Jacobi.

Soluzione.

(1) Per tutte le righe di A il valore assoluto dell'elemento sulla diagonale è maggiore della somma dei valori assoluti dei restanti elementi della riga. Quindi la matrice *è* a predominanza diagonale forte per righe.
(2) Posto: $A = D + M$ con:

$$D = \text{diag}(A) = \begin{bmatrix} 4 & & & \\ & 4 & & \\ & & 2 & \\ & & & 4 \end{bmatrix} \quad \text{e} \quad M = A - D = \begin{bmatrix} 0 & 1 & & 1 \\ 1 & 0 & & \\ & & 0 & \\ 1 & & & 0 \end{bmatrix}$$

si ha:

$$H_J = -D^{-1} M = -\begin{bmatrix} 0 & 1/4 & & 1/4 \\ 1/4 & 0 & & \\ & & 0 & \\ 1/4 & & & 0 \end{bmatrix} \quad \text{e} \quad c_J = D^{-1} b = \begin{bmatrix} 1/4 \\ 0 \\ 0 \\ 1/4 \end{bmatrix}$$

(3) Il polinomio caratteristico di H_J è:

$$\det(H_J - \lambda I) = \lambda^2 (\lambda^2 - 1/8)$$

dunque:

$$\sigma(H_J) = \{ 0, 0, 1/\sqrt{8}, -1/\sqrt{8} \} \quad \text{e} \quad \rho(H_J) = 1/\sqrt{8}$$

(4) La norma infinito di H_J è, usando la formula di calcolo riportata nell'Osservazione (2.32) della Lezione 18:

$$\| H_J \|_\infty = \max\{ 1/2, 1/4, 0, 1/4 \} = 1/2$$

(5) Per decidere se in questo caso il metodo di Jacobi è convergente si può usare il Teorema di caratterizzazione dei metodi convergenti (Teorema (2.66) della Lezione 22). Dal risultato del punto (3) si ha: $\rho(H_J) = 1/\sqrt{8} < 1$, dunque *il metodo è*

convergente.

Allo stesso risultato si poteva arrivare utilizzando il Teorema (2.72) della Lezione 23: la predominanza diagonale forte per righe di A (stabilita al punto (1)) è una *condizione sufficiente* per la convergenza del metodo di Jacobi. Alternativamente, per il Teorema (2.73) della Lezione 23, $\| H_J \|_\infty < 1$ è una *condizione sufficiente* per avere $\rho(H_J) < 1$ e quindi la convergenza del metodo di Jacobi. Il calcolo di $\rho(H_J)$, che è in generale difficile da fare, non solo consente di decidere *con certezza* della convergenza del metodo (le due condizioni richiamate sopra sono *solo sufficienti*: se non sono verificate...) ma, nel caso in cui il metodo risulti convergente, fornisce anche informazioni sulla *rapidità di convergenza* (Teorema (2.81) della Lezione 23).

(6) Si ha:

$$x(1) = H_J\, x(0) + c_J = \begin{bmatrix} 0 \\ -1/4 \\ 0 \\ 0 \end{bmatrix}$$

Esercizio 3

Si consideri l'equazione differenziale:

$$y''(t) = y(t) + (y'(t))^2 + \operatorname{sen} t$$

(1) Determinare un sistema di equazioni di ordine uno equivalente all'equazione data;
(2) Determinare la funzione $G_2(t,x)$ che restituisce il valore della derivata seconda della soluzione del sistema che all'istante t passa per x;
(3) Dati x(k), t(k) ed h(k), determinare x(k+1) con il metodo TS(1).

Soluzione.

(1) Posto $x_1(t) = y(t)$ e $x_2(t) = y'(t)$, un sistema di equazioni di ordine uno equivalente all'equazione data è:

$$x_1'(t) = x_2(t) \quad , \quad x_2'(t) = x_1(t) + (x_2(t))^2 + \operatorname{sen} t \qquad (\#)$$

(2) Se $x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$ è una soluzione del sistema (#) allora:

$$x_1''(t) = x_2'(t) = x_1(t) + (x_2(t))^2 + \operatorname{sen} t$$

e:

$$x_2''(t) = x_1'(t) + 2\, x_2(t)\, x_2'(t) + \cos t =$$

$$= x_2(t) + 2\, x_2(t)\, [x_1(t) + (x_2(t))^2 + \operatorname{sen} t] + \cos t$$

quindi:

$$G_2(t,x) = \begin{bmatrix} x_1 + x_2^2 + \operatorname{sen} t \\ x_2 + 2\, x_1\, x_2 + 2\, x_2^3 + 2\, x_2 \operatorname{sen} t + \cos t \end{bmatrix}$$

(3) L'approssimazione x(k+1) con TS(1) è:

$$x(k+1) = x(k) + F(t(k),x(k))\,h(k) = \begin{bmatrix} x_1(k) + x_2(k)\,h(k) \\ x_2(k) + [x_1(k) + (x_2(k))^2 + \operatorname{sen} t(k)]\,h(k) \end{bmatrix}$$

Esercizio 4

Per approssimare il grafico della funzione:

$$f(x) = \operatorname{sen} 3x$$

sull'intervallo [a,b] = [0,5], in *Scilab* si utilizzano i seguenti comandi:

```
> x = linspace(0,5,n + 1)';
> plot(x,f(x));
```

L'effetto è quello di disegnare, in un piano cartesiano, il grafico della funzione $\sigma_n(x)$ continua e lineare a tratti sugli intervalli determinati dai punti x(1),...,x(n + 1) che interpola i valori di f in x(1),...,x(n + 1).

Determinare un valore di n in modo che:

$$e_n(f) = \max_{x\in[0,5]} |\sigma_n(x) - f(x)| \leqslant 10^{-2}$$

Soluzione.

La funzione f ha derivata seconda continua: f"(x) = - 9 sen 3x. Per ogni $x \in [x(k),x(k+1)]$ si ha allora (usando il Teorema (3.11) della Lezione 25):

$$|\sigma_n(x) - f(x)| \leqslant \frac{M_2}{2}|x - x(k)||x - x(k+1)| \quad \text{con} \quad M_2 = \max_{x\in[0,5]} |f"(x)| = 9$$

e quindi:

$$\max_{x\in[x(k),x(k+1)]} |\sigma_n(x) - f(x)| \leqslant \frac{M_2}{2} \max_{x\in[x(k),x(k+1)]} |x - x(k)||x - x(k+1)|$$

Inoltre:

$$\max_{x\in[x(k),x(k+1)]} |x - x(k)||x - x(k+1)| = \left(\frac{x(k+1) - x(k)}{2}\right)^2$$

perciò:

$$\max_{x\in[x(k),x(k+1)]} |\sigma_n(x) - f(x)| \leqslant \frac{M_2}{8}\,[\,x(k+1) - x(k)\,]^2 = \frac{M_2}{8}\left(\frac{b - a}{n}\right)^2$$

Si ottiene infine:

$$e_n(f) = \max_{x\in[0,5]} |\sigma_n(x) - f(x)| \leqslant \frac{M_2}{8}\left(\frac{b - a}{n}\right)^2$$

Per ottenere $e_n(f) \leqslant 10^{-2}$ basta che sia:

$$\frac{M_2}{8}\left(\frac{b - a}{n}\right)^2 \leqslant 10^{-2} \quad \text{ovvero} \quad n \geqslant 10\sqrt{\frac{M_2}{8}}\,(b - a) = 53.03\cdots$$

Dunque $n \geqslant 54$.