Analisi numerica

A.A. 2023-2024

SIMONE SACCANI

ARITMETICA DI MACCHINA E ERRORI

Teorema di rappresentazione dei numeri reali

Sia $B \in \mathbb{N}, B \geq 2$. Allora $\forall x \in \mathbb{R} \setminus \{0\}$ esistono

(1) $\{d_j\}_{j \geq 1}$, $d_j \in \mathbb{N}$, $0 \leq d_j \leq B-1$, $d_1 \neq 0$ e t.c.
$\forall \bar{j}\ \exists j > \bar{j} : d_j \neq B-1$

(2) $p \in \mathbb{Z}$ t.c. $x = \text{sgn}(x)\, B^p \sum_{j=1}^{+\infty} d_j B^{-j}$

Inoltre tale rappresentazione è unica

NOTA: B si dice base, p esponente, $\sum_{j=1}^{+\infty} d_j B^{-j}$ mantissa

Def. Dati $B \in \mathbb{N}, B \geq 2$, $t \in \mathbb{N}, t \geq 1$, $m, M \in \mathbb{N}$, $m, M \geq 1$, si definisce l'**insieme dei numeri di macchina** (o in virgola mobile - floating point):

$$\mathcal{F}(B, t, m, M) = \{0\} \cup \left\{ \pm B^p \sum_{j=1}^{t} d_j B^{-j},\ 0 \leq d_j \leq B-1,\ d_1 \neq 0,\ -m \leq p \leq M \right\}$$

Dato $x \in \mathbb{R} \setminus \{0\}$, se $p > M$, allora x non è rappresentabile (Overflow)
se $p < -m$, allora x non è rappresentabile (Underflow)
Se invece $-m \leq p \leq M$, allora x è approssimato con $\tilde{x} = fl(x)$

$$fl(x) = \text{tron}(x) = \text{sgn}(x)\, B^p \sum_{j=1}^{t} d_j B^{-j}$$

Def. Si chiama **errore di rappresentazione** l'errore relativo $\left| \frac{\tilde{x} - x}{x} \right|$

$$|\tilde{x} - x| = B^p \sum_{j=t+1}^{+\infty} d_j B^{-j} = B^{p-t-1} \sum_{j=0}^{+\infty} d_{t+1+j} B^{-j} < B^{p-t-1} (B-1) \frac{1}{1 - B^{-1}} = B^{p-t}$$

$$|x| = B^p \sum_{j=1}^{+\infty} d_j B^{-j} \geq B^p d_1 B^{-1} \geq B^{p-1}$$

Dunque $\left| \frac{\tilde{x} - x}{x} \right| < \frac{B^{p-t}}{B^{p-1}} = B^{1-t}$

Oss l'errore relativo è indipendente da x

Def. La **precisione di macchina** è $u = B^{1-t}$

esempio Se $B = 2$, $t = 53$, $u \sim 10^{-16}$ (doppia precisione)

$\left| \frac{\tilde{x} - x}{x} \right| < u$: $\exists \varepsilon$ con $|\varepsilon| < u$ t.c. $\tilde{x} = x(1+\varepsilon)$
$\exists \eta$ con $|\eta| < u$ t.c. $\tilde{x} = \frac{x}{1+\eta}$

Def. Dato $x \in \mathbb{R}$, y approssimazione di x, $c \geq 1$ intero
Se $\left| \frac{y - x}{x} \right| < B^{1-c}$, si dice y ha c **cifre significative**

Def. Dati $x, y \in \mathcal{F}$, $op \in \{+, -, *, /\}$,
definiamo l'**operazione di macchina** $[op]$

$$x\,[op]\,y := fl(x\ op\ y)$$

Se $z = x\ op\ y$, $\tilde{z} = x\,[op]\,y$, allora
$\left| \frac{\tilde{z} - z}{z} \right| < u$, cioè $\tilde{z} = z(1+\delta)$, con $|\delta| < u$
δ si dice errore locale dell'operazione aritmetica

Errori nel calcolo di una funzione

Sia $f: \Omega \subset \mathbb{R}^N \longrightarrow \mathbb{R}$

esempio $f((x_1,\dots,x_N)) = \sum_{i=1}^{N} x_i$

$\tilde{x}_i = x_i(1+\varepsilon_i)$ è la rappresentazione di x_i in $\mathcal{F}$, con $|\varepsilon_i| < u$

Viene calcolato $f(\tilde{x})$ invece di $f(x)$.

Def. Si definisce l'**errore inerente** $\varepsilon_{in} = \dfrac{f(\tilde{x}) - f(x)}{f(x)}$, con $f(x) \neq 0$

Sia $\delta_x = \dfrac{\tilde{x}-x}{x}$ l'errore su x

Supponiamo $N=1$ e $f \in C^2(\mathbb{R})$. Posso scrivere:

$$f(\tilde{x}) = f(x) + (\tilde{x}-x)f'(x) + \frac{1}{2}(\tilde{x}-x)^2 f''(\xi) \qquad \text{con } |\xi - x| < |\tilde{x}-x|$$

da cui $\varepsilon_{in} = \dfrac{f(\tilde{x})-f(x)}{f(x)} = \dfrac{f'(x)}{f(x)} x \dfrac{(\tilde{x}-x)}{x} + \dfrac{1}{2}\dfrac{f''(\xi)}{f(x)} \dfrac{(\tilde{x}-x)^2}{x^2} x^2 = \delta_x \cdot c_x + \underbrace{\delta_x^2 \cdot (*)}_{\text{trascurabile poiché } |\delta_x| \ll 1}$

$$\varepsilon_{in} \doteq c_x \cdot \delta_x = \frac{f'(x)x}{f(x)} \delta_x$$

↳ coefficiente di amplificazione $c_x = \dfrac{x f'(x)}{f(x)}$

esempio Se $f(x) = x^p$, $p \geq 1$ intero, $c_x = p$

Se $f(x) = x^{\frac{1}{p}}$, $c_x = \frac{1}{p}$

Se $f: \mathbb{R}^n \longrightarrow \mathbb{R}$

Sia $\delta_{x_i} = \dfrac{\tilde{x}_i - x_i}{x_i}$ $i = 1,\dots,N$

Allora $\varepsilon_{in} = \sum_{i=1}^{n} c_i(x)\, \delta_{x_i}$ dove $c_i(x) = \dfrac{1}{f(x)} x_i \dfrac{\partial f(x)}{\partial x_i}$ $i=1,\dots,N$

esempio Se $f(x) = \sum_{i=1}^{n} x_i$, $c_i(x) = \dfrac{x_i}{f(x)}$: $\varepsilon_{in} \doteq \dfrac{1}{f(x)} \sum_{i=1}^{n} \delta_{x_i} \cdot x_i$

Se $\tilde{x}_i = fl(x_i)$, allora $|\delta_{x_i}| < u$

$|\varepsilon_{in}| \leq \dfrac{1}{|f(x)|} u \sum_{i=1}^{n} |x_i|$

Se $x_i > 0$ $\forall i$ o $x_i < 0$ $\forall i$: $\sum |x_i| = |\sum x_i|$ dunque $|\varepsilon_{in}| \leq u$

Def. Un problema è **mal condizionato** se una piccola perturbazione sui dati induce una grande perturbazione sul risultato

Errore algoritmico

Sia $f: \mathbb{R} \to \mathbb{R}$ razionale

Dato $x \in \mathbb{R}$, voglio calcolare $f(x)$ mediante una sequenza di operazioni elementari

(1) x è approssimato con $\tilde{x}$

(2) l'algoritmo calcola $\varphi(\tilde{x})$ (generalmente diversa da $f(\tilde{x})$), dovuto all'aritmetica f.p.

Def. Si definisce l'**errore algoritmico** $\varepsilon_{alg} = \dfrac{\varphi(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})}$

Abbiamo quindi: $\varepsilon_{in} = \dfrac{f(\tilde{x}) - f(x)}{f(x)}$ $\qquad \varepsilon_{alg} = \dfrac{\varphi(\tilde{x}) - f(\tilde{x})}{f(\tilde{x})}$, da cui

$$\varepsilon_{tot} = \frac{\varphi(\tilde{x}) - f(x)}{f(x)} = \frac{\varphi(\tilde{x})}{f(x)} - 1 = \frac{\varphi(\tilde{x})\, f(\tilde{x})}{f(x)\, f(\tilde{x})} - 1 = (\varepsilon_{alg}+1)(\varepsilon_{in}+1) - 1 = \varepsilon_{in} + \varepsilon_{alg} + \varepsilon_{alg}\varepsilon_{in} \doteq \varepsilon_{in} + \varepsilon_{alg}$$

Errore analitico

Sia $g: \mathbb{R} \to \mathbb{R}$ non razionale

esempio $g(x) = e^x$

$g(x)$ viene approssimata con f razionale

Def. Si definisce l'**errore analitico** $\varepsilon_{anal} = \dfrac{f(x) - g(x)}{g(x)}$

Come prima, abbiamo $\varepsilon_{tot} = \dfrac{\varphi(\tilde{x}) - g(x)}{g(x)} \doteq \varepsilon_{in} + \varepsilon_{alg} + \varepsilon_{anal}$

esempio Errore algoritmico nel calcolo di $f(x_1, x_2) = x_1 \,\mathrm{op}\, x_2$ con $\mathrm{op} \in \{+, -, *, /\}$

Conosco $\tilde{x}_1, \tilde{x}_2$ t.c. $\dfrac{\tilde{x}_i - x_i}{x_i} = \varepsilon_i$

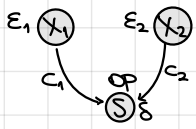
Sia $s = x_1 \,\mathrm{op}\, x_2$, $\tilde{s}$: valore effettivamente calcolato

cioè $\tilde{s} = fl(\tilde{x}_1 \,\mathrm{op}\, \tilde{x}_2) = (\tilde{x}_1 \,\mathrm{op}\, \tilde{x}_2)(1+\delta)$ con $|\delta| < u$ errore locale dell'operazione

Ora $\tilde{x}_1 \,\mathrm{op}\, \tilde{x}_2 = f(\tilde{x}_1, \tilde{x}_2) = f(x_1, x_2)\,(1 + c_1\varepsilon_1 + c_2\varepsilon_2)$ dove $c_i = \dfrac{x_i}{f(x)} \dfrac{\partial f}{\partial x_i}$ $i = 1,2$

Quindi $\tilde{s} = s(1+c_1\varepsilon_1+c_2\varepsilon_2)(1+\delta) \doteq s(1+c_1\varepsilon_1+c_2\varepsilon_2+\delta)$

Dunque $\varepsilon_{alg} = c_1\varepsilon_1 + c_2\varepsilon_2 + \delta$



op	c_1	c_2
$x_1 x_2$	1	1
x_1 / x_2	1	-1
$x_1 + x_2$	$x_1/(x_1+x_2)$	$x_2/(x_1+x_2)$
$x_1 - x_2$	$x_1/(x_1-x_2)$	$-x_2/(x_1-x_2)$

esempio $a = 0.12345678 \qquad b = 0.12345675$

$c = a - b = 0.3\ 10^{-7}$

$\tilde{a} = 0.12345679 \qquad \tilde{b} = 0.12345674$

$\varepsilon_a = \dfrac{\tilde{a}-a}{a} = 0.81 \cdot 10^{-7} \qquad \varepsilon_b = \dfrac{\tilde{b}-b}{b} = -0.81 \cdot 10^{-7}$

$\tilde{c} = \tilde{a} - \tilde{b} = 0.5 \cdot 10^{-7}$

$\dfrac{\tilde{c}-c}{c} = 0.4$

esempio $f(x,y) = x^2 - y^2 = (x+y)(x-y) \qquad x, y \in \mathbb{F}$

A_1: $s_1 = x \cdot x$ $\qquad$ A_2: $s_1 = x + y$

$s_2 = y \cdot y$ $\qquad\qquad$ $s_2 = x - y$

$s_3 = s_1 - s_2$ $\qquad\qquad$ $s_3 = s_1 \cdot s_2$

$$|\varepsilon_3| < u\left(1 + \frac{x^2+y^2}{|x^2-y^2|}\right) \qquad |\varepsilon_3| < 3u$$

Analisi all'indietro dell'errore (algoritmico)

$fl(f(x_1,\dots,x_N)) = \varphi(x_1,\dots,x_N)$

$x_i \in \mathcal{F}$

Cerco $\tilde{x}_i$ approssimazioni di x_i t.c. $\varphi(x_1,\dots,x_N) = f(\tilde{x}_1,\dots,\tilde{x}_N)$

e voluto $\frac{\tilde{x}_i - x_i}{x_i}$; ε_{alg} è ε_{in} dovuto agli errori $\frac{\tilde{x}_i - x_i}{x_i}$

esempio $x_1, x_2 \in \mathcal{F}$

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

$fl(f(x_1,x_2)) = (x_1+x_2)(1+\delta)$ con $|\delta| < u$ errore locale dell'operazione

$= x_1(1+\delta) + x_2(1+\delta) = \tilde{x}_1 + \tilde{x}_2$ dove $\tilde{x}_i = x_i(1+\delta)$

esempio $f(x_1,x_2,x_3) = x_1^2 + x_2 \cdot x_3$ $\quad x_i \in \mathcal{F}$

$(x_1 \cdot x_1)(1+\varepsilon_1)$ errore locale della moltiplicazione $|\varepsilon_i| < u$

$(x_2 \cdot x_3)(1+\varepsilon_2)$

$((x_1 \cdot x_1)(1+\varepsilon_1) + (x_2 \cdot x_3)(1+\varepsilon_2))(1+\varepsilon_3) \doteq \tilde{x}_1^2 + \tilde{x}_2 \cdot \tilde{x}_3$

$\doteq x_1 \cdot x_1 (1+\varepsilon_1+\varepsilon_3) + x_2 \cdot x_3 (1+\varepsilon_2+\varepsilon_3) \doteq \left(x_1\left(1+\frac{\varepsilon_1+\varepsilon_3}{2}\right)\right)^2 + x_2(1+\varepsilon_3)\, x_3(1+\varepsilon_2)$

$\tilde{x}_1 = x_1\left(1+\frac{\varepsilon_1+\varepsilon_3}{2}\right)$ $\quad \tilde{x}_2 = x_2(1+\varepsilon_3)$ $\quad \tilde{x}_3 = x_3(1+\varepsilon_2)$

$$\left|\frac{\tilde{x}_1 - x_1}{x_1}\right| = \left|\frac{\varepsilon_1+\varepsilon_2}{2}\right| < u \qquad \left|\frac{\tilde{x}_2 - x_2}{x_2}\right| = |\varepsilon_3| < u \qquad \left|\frac{\tilde{x}_3 - x_3}{x_3}\right| = |\varepsilon_2| < u$$

def. Un algoritmo è numericamente stabile se l'errore algoritmico è piccolo (in particolare se è limitato superiormente da ku, k costante)

TEOREMI DI GERSCHGORIN

Localizzazione degli autovalori di $A \in \mathbb{C}^{n\times n}$

I teorema di Gerschgorin

Sia $K_i = \{z \in \mathbb{C} \mid |z-a_{ii}| \le \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|\}$ $i=1,\dots,n$ (cerchio di Gerschgorin)

Se λ è autovalore di $A \Rightarrow \lambda \in \bigcup_{i=1}^{n} K_i$

Inoltre, se $Av=\lambda v$, $v \in \mathbb{C}^n \setminus \{0\}$ e se $|v_h| = \max_i |v_i|$

allora $\lambda \in K_h$

DIMOSTRAZIONE

Sia λ autovalore di A, v corrispondente autovettore, $v \in \mathbb{C}^n \setminus \{0\}$ t.c. $Av=\lambda v$

Per componente $\lambda v_i = \sum_{j=1}^{n} a_{ij} v_j \Rightarrow \lambda v_i - a_{ii} v_i = (\lambda - a_{ii}) v_i = \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij} v_j$

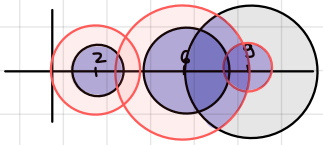
Sia h : $|v_h| = \max_i |v_i|$, da cui $v_h \neq 0$ perché v è autovettore

Per $i=h$. $\lambda - a_{hh} = \sum_{\substack{j=1\\ j\neq h}}^{n} a_{hj} \frac{v_j}{v_h}$

da cui $|\lambda - a_{hh}| = \left| \sum_{\substack{j=1\\ j\neq h}}^{n} a_{hj} \frac{v_j}{v_h} \right| \le \sum_{\substack{j=1\\ j\neq h}}^{n} |a_{hj}| \left|\frac{v_j}{v_h}\right| \le \sum_{\substack{j=1\\ j\neq h}}^{n} |a_{hj}|$, cioè $\lambda \in K_h$ □

esempio $A = \begin{pmatrix} 2 & 1 & 0 \\ -1 & 6 & -1 \\ 1 & -2 & 9 \end{pmatrix}$

A^t



Se H_i sono i cerchi di A^T, λ autovalore

$\lambda \in (\bigcup_{i=1}^{n} K_i) \cap (\bigcup_{i=1}^{n} H_i)$

II teorema di Gerschgorin

Sia $K = \bigcup_{i=1}^{n} K_i$

Supponiamo $K = M_1 \cup M_2$, con $M_1 \cap M_2 = \emptyset$

M_1 formato da h cerchi di Gerschgorin

M_2 formato da $n-h$ cerchi di Gerschgorin

Allora M_1 contiene h autovalori e

M_2 contiene $n-h$ autovalori

DIMOSTRAZIONE

Sia $M_1 = \bigcup_{i=1}^{h} K_i$ $M_2 = \bigcup_{i=h+1}^{n} K_i$

Sia $A(t) = D + t(A-D)$, con $D = \begin{pmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{nn} \end{pmatrix}$ e $t \in [0,1]$

$(A(1)=A,\ A(0)=D)$

Gli autovalori di $A(t)$ sono gli zeri di $p_t(\lambda) = \det(A(t) - \lambda I)$

I coefficienti di $p_t(\lambda)$ dipendono in modo continuo da t

$\Rightarrow$ gli zeri di $p_t(\lambda)$ $\lambda_1(t),\dots,\lambda_n(t)$ dipendono in modo continuo da t

Se $t=0$, $\lambda_i(0) = a_{ii}$

Se $t=1$, $\lambda_i(1) = \lambda_i$ autovalore di A

Per i cerchi di $A(t)$, $K_i(t) \subset K_i$

Per $t=0$: h autovalori $\in M_1$, $n-h$ autovalori $\in M_2$ □

Def. $A \in \mathbb{C}^{n\times n}$ si dice **riducibile** se $\exists P$ di permutazione:

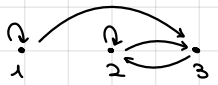
$$PAP^T = \left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline 0 & B_{22} \end{array}\right) \quad \text{con } B_{11}, B_{22} \text{ matrici quadrate}$$

$$P = \begin{pmatrix} 0 & \cdots & 1 & 0 & \cdots \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & 1 \\ \vdots & & & & \end{pmatrix} \quad P^TP = I \quad PAP^T = B \quad b_{ij} = a_{\sigma(i)\sigma(j)}$$

$A \in \mathbb{C}^{n\times n}$ si dice **irriducibile** se non è riducibile

Associamo a $A=(a_{ij})$ un grafo orientato con nodi $\{1,\dots,n\}$ e un arco orientato (i,j) da i a j se $a_{ij} \neq 0$

esempio $A = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 2 & -1 \\ 0 & 3 & 0 \end{pmatrix}$



Def. Un cammino orientato da i a j è una sequenza di archi orientati $(i,i_2)(i_2,i_3)\cdots(i_h,j)$

Un grafo orientato è fortemente connesso se $\forall i,j \in \{1,\dots,n\}$, $i \neq j$, esiste un cammino da i a j

Teorema La matrice A è irriducibile se e solo se il grafo associato è fortemente connesso.

DIMOSTRAZIONE

Oss i grafi di A e P^TAP sono isomorfi (cambiano solo i nomi dei nodi)

$\Leftarrow$ Sia A riducibile

$$P^TAP = \left(\begin{array}{c|c} B_{11} & B_{12} \\ \hline 0 & B_{22} \end{array}\right) \begin{matrix} 1 \\ \vdots \\ h \\ h+1 \\ \vdots \\ n \end{matrix}$$

Il grafo non è fortemente connesso perché non ci sono archi da $i \in \{h+1,\dots,n\}$ a un nodo $j \in \{1,\dots,h\}$

$\Rightarrow$ Il grafo non sia fortemente connesso

Siano $p,q \in \{1,\dots,n\}$ $p \neq q$ t.c. q non è raggiungibile da p (cioè non esiste un cammino da p a q)

Sia $P = \{\text{nodi raggiungibili da } p\}$, $Q = \{\text{nodi non raggiungibili da } p\}$

$q \in Q$, quindi $Q \neq \emptyset$

Applico una permutazione di indici

$$\begin{array}{c} \\ Q \\ P \end{array} \begin{array}{c} \begin{array}{cc} Q & P \end{array} \\ \left(\begin{array}{c|c} * & * \\ \hline 0 & * \end{array}\right) \end{array}$$

0 perché altrimenti avrei un arco da un elemento di P in Q

□

III teorema di Gerschgorin

Sia $A \in \mathbb{C}^{n\times n}$ irriducibile, λ autovalore di A t.c.
se $\lambda \in K_i \Rightarrow \lambda \in \partial K_i$ per i opportuno
allora $\lambda \in K_i \ \forall i$
In particolare $\lambda \in \partial K_i \ \forall i$, dunque $\lambda \in \bigcap_{i=1}^{n} \partial K_i$

DIMOSTRAZIONE

$Ax = \lambda x \quad , \quad x \in \mathbb{C}^n \setminus \{0\}$

$(\lambda - a_{hh}) x_h = \sum_{\substack{j=1 \\ j\neq h}}^{n} a_{hj} x_j$ (h-esima riga del prodotto)

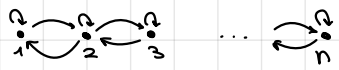
Sia $|x_h| = \max_j |x_j|$

$\lambda - a_{hh} = \sum_{\substack{j=1 \\ j\neq h}}^{n} a_{hj} \frac{x_j}{x_h} \Rightarrow |\lambda - a_{hh}| \leq \sum_{\substack{j=1 \\ j\neq h}}^{n} |a_{hj}| \frac{|x_j|}{|x_h|} \leq \sum_{\substack{j=1 \\ j\neq h}}^{n} |a_{hj}|$

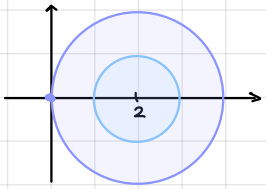
$\Rightarrow \lambda \in K_h$ e per ipotesi $\lambda \in \partial K_h$, quindi sono delle uguaglianze
Dunque $|x_j| = |x_h|$ quando $a_{hj} \neq 0$
A irriducibile $\Rightarrow \exists$ sequenza di archi $(h,k_2)(k_2,k_3)\cdots(k_{\ell-1},k_\ell)$
$k_1 = h$, cioè $a_{k_i,k_{i+1}} \neq 0$, $\ell \geq n$
tali che $\{k_1, k_2, \dots, k_\ell\} = \{1,\dots,n\}$
$a_{h,k_2} \neq 0 \Rightarrow |x_{k_2}| = |x_h| \Rightarrow \lambda \in K_{k_2}$
Ripeto il ragionamento, dove h è sostituito da k_2
$a_{k_2,k_3} \neq 0 \Rightarrow \lambda \in K_{k_3}$
Poiché i nodi sono tutto $\{1,\dots,n\}$, concludo che
$\lambda \in K_h \quad \forall h$ □

esempio

$$A = \begin{pmatrix} 2 & 1 & & 0 \\ 1 & 2 & 1 & \\ & \ddots & \ddots & \ddots \\ & & 1 & 2 \\ 0 & & 1 & 2 \end{pmatrix}$$



quindi A è irriducibile



0 non può essere autovalore
per il III teorema di Gerschgorin

Def $A \in \mathbb{C}^{n\times n}$ si dice **irriducibilmente dominante diagonale** (IDD) se
(1) è irriducibile
(2) $|a_{hh}| \geq \sum_{j\neq h} |a_{hj}|$
(3) $\exists k$ per cui $|a_{kk}| > \sum_{j\neq k} |a_{kj}|$

corollario A IDD $\Rightarrow \det A \neq 0$

Def $A \in \mathbb{C}^{n\times n}$ si dice **strettamente dominante diagonale** (SDD)
se $|a_{hh}| > \sum_{\substack{j=1 \\ j\neq h}}^{n} |a_{kj}|$

Oss A SDD $\Rightarrow 0 \notin \bigcup_{i=1}^{n} K_i \Rightarrow \det A \neq 0$

FORMA NORMALE DI SCHUR

Def. $U \in \mathbb{C}^{n\times n}$ si dice unitaria se $U^H U = I$
$Q \in \mathbb{R}^{n\times n}$ si dice ortogonale se $Q^T Q = I$

esempio $A = \begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix}$

$A_\varepsilon = \begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ \varepsilon & & & 0 \end{pmatrix}$ ha n autovalori distinti $\varepsilon^{\frac{1}{n}} \omega_n^j$, $j = 0, \dots, n-1$

ω_n radice n-esima primitiva dell'unità

Teorema: forma normale di Schur Sia $A \in M(n, \mathbb{C})$. Allora $\exists U$ unitaria e $\exists T$ triangolare t.c.
$$U^H A U = T$$

DIMOSTRAZIONE
Per induzione su n:
- $n = 1$. $A = 1 \cdot A \cdot 1$
- La tesi sia vera per $n-1$

Sia $x \in \mathbb{C}^n \setminus \{0\}$: $Ax = \lambda x$ e $x^H x = 1$
Sia $\{x = y_1, y_2, \dots, y_n\}$ base ortonormale di $\mathbb{C}^n$, cioè $y_i^H y_j = \delta_{ij}$
Sia $Q = (x | y_2 | \dots | y_n)$: $Q^H Q = I$ per costruzione.
$$Q^H A Q\, e_1 = Q^H A x = \lambda Q^H x = \lambda e_1$$
$$Q e_1 = x \Rightarrow e_1 = Q^H x$$
$$Q^H A Q = \left(\begin{array}{c|c} \lambda & u^T \\ \hline 0 \\ \vdots & A_1 \\ 0 \end{array}\right) \qquad \text{con } A_1 \in M(n-1, \mathbb{C})$$

Si applica l'ipotesi induttiva a A_1
Per ipotesi induttiva $\exists Q_1 \in M(n-1, \mathbb{C})$ unitaria t.c. $Q_1^H A_1 Q = T_1$ triangolare superiore
Pongo $\tilde{Q}_1 = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & Q_1 \end{array}\right)$: $\tilde{Q}_1^H Q_1 = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & Q_1^H Q \end{array}\right) = I$ quindi $\tilde{Q}_1$ è unitaria.

Ora $\tilde{Q}_1^H Q^H A Q \tilde{Q}_1 = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & Q_1^H \end{array}\right)\left(\begin{array}{c|c} \lambda & u^T \\ \hline 0 & A_1 \end{array}\right)\left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & Q_1 \end{array}\right) = \left(\begin{array}{c|c} \lambda & u^T Q_1 \\ \hline 0 & Q_1^H A_1 Q_1 \end{array}\right) = \left(\begin{array}{c|c} \lambda & u^T Q_1 \\ \hline 0 & T_1 \end{array}\right) = T$

con T triangolare superiore, $T = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$ con λ_i autovalori di A
Quindi $U = Q\tilde{Q}_1$: U è unitaria $U^H U = (Q\tilde{Q}_1)^H Q\tilde{Q}_1 = \tilde{Q}_1^H Q^H Q \tilde{Q}_1 = I$ □

Oss La forma normale di Schur non è unica:
- ordine degli autovalori
- scelta di una base
- $D = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{pmatrix}$ con $\bar{d}_i d_i = 1$: $\tilde{U} = UD$ è unitaria
  e $\tilde{U}^H A \tilde{U} = D^H U^H A U D = D^H T D = \tilde{T} = \begin{pmatrix} t_{11} & & * \\ & \ddots & \\ 0 & & t_{nn} \end{pmatrix}$

esempio Sia A hermitiana: $A^H = A$

$A = A^H \iff U^H A U = T \quad T = T^H$

Infatti $A = A^H \iff T = U^H A U = U^H A^H U = (U^H A U)^H = T^H$

ma T è triangolare superiore, dunque T è diagonale: $T = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$

Inoltre $\lambda_i = \overline{\lambda_i}$, cioè $\lambda_i \in \mathbb{R}$

esempio Sia A anti-hermitiana: $A^H = -A$

$A = -A^H \iff T = -T^H$ e inoltre $\lambda_i = -\overline{\lambda_i}$, cioè λ_i è immaginario puro.

def. Una matrice $A \in M(n, \mathbb{C})$ si dice normale se $A^H A = A A^H$

teorema A è normale $\iff T$ è diagonale
dove $T = U^H A U$ è la forma normale di Schur

DIMOSTRAZIONE

Oss A è normale $\iff TT^H = T^H T$

$TT^H = U^H A U U^H A^H U = U^H A^H U U^H A U = T^H T$

($\Leftarrow$) Sia T diagonale $\implies TT^H = T^H T$

($\Rightarrow$) Sia $TT^H = T^H T$

Per induzione su n

- $n = 1$ ok
- Sia vero per $n-1$

$$TT^H = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ & \ddots & & \vdots \\ 0 & & & t_{nn} \end{pmatrix} \begin{pmatrix} \overline{t_{11}} & & & 0 \\ \overline{t_{12}} & \ddots & & \\ \vdots & & & \\ \overline{t_{1n}} & \cdots & & \overline{t_{nn}} \end{pmatrix} \qquad T^H T = \begin{pmatrix} \overline{t_{11}} & & & 0 \\ \overline{t_{12}} & \ddots & & \\ \vdots & & & \\ \overline{t_{1n}} & \cdots & & \overline{t_{nn}} \end{pmatrix} \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ & \ddots & & \vdots \\ 0 & & & t_{nn} \end{pmatrix}$$

$$(TT^H)_{1,1} = \sum_{j=1}^{n} t_{1j} \overline{t_{1j}} = \sum_{j=1}^{n} |t_{1j}|^2$$

$$(T^H T)_{1,1} = |t_{11}|^2 \implies t_{1j} = 0 \text{ per } j \geq 2$$

Dunque $T = \left(\begin{array}{c|c} t_{11} & 0 \\ \hline 0 & T_1 \end{array}\right)$

$T^H T = TT^H \implies T_1^H T_1 = T_1 T_1^H$, cioè T_1 è normale

$\implies T_1$ è diagonale per hp. induttiva $\square$

NORME

Def. Una norma su $\mathbb{C}^n$ è un'applicazione
$\|\cdot\| : \mathbb{C}^n \longrightarrow \mathbb{R}$ tale che
(1) $\|x\| \geq 0 \ \forall x \in \mathbb{C}^n$ e $\|x\| = 0 \iff x = 0$
(2) $\forall \alpha \in \mathbb{C}, \forall x \in \mathbb{C}^n \quad \|\alpha x\| = |\alpha| \, \|x\|$
(3) $\forall x, y \in \mathbb{C}^n \quad \|x+y\| \leq \|x\| + \|y\|$

esempio Dato $x = (x_i)_{i=1,\dots,n}$
$\|x\|_1 = \sum_{i=1}^{n} |x_i|$ (norma 1)
$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}$ (norma 2 o euclidea)
$\|x\|_\infty = \max_i |x_i|$ (norma infinito)

Se $\|\cdot\|$ è una norma su $\mathbb{C}^n$ e se $S \in M(n, \mathbb{C})$, $\det S \neq 0$
$\| \ \|_S : x \longmapsto \|Sx\|$ è una norma

Se $\langle \, , \cdot \rangle : \mathbb{C}^n \times \mathbb{C}^n \longrightarrow \mathbb{C}^n$ è un prodotto hermitiano, definito positivo
allora $\langle x, x \rangle^{\frac{1}{2}}$ è una norma su $\mathbb{C}^n$

Teorema: equivalenza delle norme Per ogni coppia di norme $\|\cdot\|', \|\cdot\|''$ su $\mathbb{C}^n$
$\exists \alpha, \beta \in \mathbb{R}, \alpha, \beta > 0$ tali che
$\alpha \|x\|' \leq \|x\|'' \leq \beta \|x\|' \quad \forall x \in \mathbb{C}^n$

DIMOSTRAZIONE
Se $x = 0$: ok
Sia $x \in \mathbb{C}^n \setminus \{0\}$
Basta dimostrare il teorema con $\|x\|' = \|x\|_\infty$
(il caso generale si deduce concatenando le disuguaglianze)
$\| \ \|''$ norma su $\mathbb{C}^n$
$S_\infty = \{ x \in \mathbb{C}^n : \|x\|_\infty = 1 \}$, $\| \ \| : \mathbb{C}^n \longrightarrow \mathbb{R}$
S_∞ è controimmagine di $\{1\}$ limitato e chiuso $\Rightarrow$ S_∞ è compatto
$\|\cdot\|'' : S_\infty \longrightarrow \mathbb{R}$
Poiché S_∞ è compatto $\Rightarrow \exists \ \beta = \max_{x \in S_\infty} \|x\|''$ e $\alpha = \min_{x \in S_\infty} \|x\|''$
Se $x \in \mathbb{C}^n \setminus \{0\}$, sia $y = \frac{x}{\|x\|_\infty}$, quindi $y \in S_\infty$
Perciò $\alpha \leq \|y\|'' \leq \beta$, cioè $\alpha \leq \frac{\|x\|''}{\|x\|_\infty} \leq \beta$ □

Teorema $\forall x \in \mathbb{C}^n$

$$\|x\|_\infty \le \|x\|_1 \le n\|x\|_\infty$$
$$\|x\|_2 \le \|x\|_1 \le \sqrt{n}\|x\|_2$$
$$\|x\|_2 \le \|x\|_1 \le \sqrt{n}\|x\|_2$$

Oss Se $y = Ux$, $x \in \mathbb{C}^n$, U unitaria

$\Rightarrow \|y\|_2 = \|x\|_2$

Infatti $\|y\|_2^2 = y^H y = x^H U^H U x = x^H x = \|x\|_2^2$

Def. Una norma di matrice è un applicazione

$\|\cdot\| : M(n,\mathbb{C}) \longrightarrow \mathbb{R}$ tale che

(1) $\|A\| \ge 0 \quad \forall A \in M(n,\mathbb{C})$ e $\|A\| = 0 \iff A = 0$

(2) $\|\lambda A\| = |\lambda| \|A\| \quad \forall \lambda \in \mathbb{C} \ \forall A \in M(n,\mathbb{C})$

(3) $\|A+B\| \le \|A\| + \|B\| \quad \forall A, B \in M(n,\mathbb{C})$

(4) $\|A \cdot B\| \le \|A\| \cdot \|B\| \quad \forall A, B \in M(n,\mathbb{C})$

esempio Norma di Frobenius

$$\|A\|_F = \left(\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2} = (\mathrm{tr}(A^H A))^{1/2}$$

Data $\|\cdot\| : \mathbb{C}^n \longrightarrow \mathbb{R}$ norma vettoriale,

si definisce la norma matriciale indotta $\|A\| = \max_{\|x\|=1} \|Ax\| = \max_{x \in \mathbb{C}^n - \{0\}} \frac{\|Ax\|}{\|x\|}$

La definizione è ben posta perché $\{x \in \mathbb{C}^n \mid \|x\| = 1\}$ è chiuso e limitato,

perciò $\exists \max_{\|x\|=1} \|Ax\|$

Inoltre verifica le proprietà di una norma

Proprietà

- $\|Ax\| \le \|A\| \|x\| \quad \forall A \in M(n,\mathbb{C}), \forall x \in \mathbb{C}^n$
- $\|I\| = 1 \quad \forall \|\cdot\|$ n.m.i.

esempio $\|I\|_F = \sqrt{n} \Rightarrow$ la norma di Frobenius non è una n.m.i.

Def. Sia $\|A\|_\infty = \max_{\|x\|_\infty = 1} \|Ax\|_\infty$, $\|A\|_1 = \max_{\|x\|_1 = 1} \|Ax\|_1$, $\|A\|_2 = \max_{\|x\|_2 = 1} \|Ax\|_2$

Teorema

$$\|A\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}|$$

$$\|A\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|$$

$$\|A\|_2 = (\rho(A^H A))^{1/2}$$

DIMOSTRAZIONE

Sia $x \in \mathbb{C}$ con $\|x\|_1 = 1$

$$\|Ax\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \le \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}||x_j| = \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{ij}| \le \sum_{j=1}^{n} |x_j| \max_\ell \sum_{i=1}^{n} |a_{i\ell}| =$$

$$= \max_\ell \sum_{i=1}^{n} |a_{i\ell}| \cdot \sum_{j=1}^{n} |x_j| = \max_\ell \sum_{i=1}^{n} |a_{i\ell}|$$

Se il massimo si realizza per $j=h$ $\left(\max_\ell \sum_{i=1}^{n} |a_{i\ell}| = \sum_{i=1}^{n} |a_{ih}|\right)$, scelgo $x = e_h$ e vale $\|Ae_h\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}|$

Sia $x \in \mathbb{C}^n$ con $\|x\|_\infty = 1$

$$\|Ax\|_\infty = \max_{i=1,\dots,n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \le \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}| \underbrace{|x_j|}_{\le 1} \le \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|$$

Cerco $x \in \mathbb{C}^n : \|Ax\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|$. Sia $\max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}| = \sum_{j=1}^{n} |a_{rj}|$

Definisco $x = (x_j)$: $x_j = \begin{cases} \frac{\overline{a_{rj}}}{|a_{rj}|} & \text{se } a_{rj} \neq 0 \\ 1 & \text{altrimenti} \end{cases}$

Sia $x \in \mathbb{C}^n$ con $\|x\|_2 = 1$

$\|Ax\|_2^2 = x^H A^H A x$

$A^H A$ è hermitiana e semidefinita positiva

$\exists U$ unitaria t.c. $U^H A^H A U = D$ con $D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$ con $\lambda_i \ge 0 \ \forall i$

$x^H A^H A x = x^H U D U^H x = y^H D y = \sum_{i=1}^{n} |y_i|^2 \lambda_i \le \max_j \lambda_j \sum_{i=1}^{n} |y_i|^2 = \rho(A^H A) \cdot \|y\|_2^2 = \rho(A^H A)$

$y = U^H x : \|y\|_2 = \|x\|_2 = 1$

Vale con x: $A^H A x = \rho(A^H A) x$ e $\|x\|_2 = 1$ □

U, V unitarie, sia $B = UAV$

$B^H B = V^H A^H U^H U A V = V^H (A^H A) V$

$\Rightarrow \|A\|_2 = \|B\|_2 \quad \|A\|_F = \|B\|_F$ perché $A^H A \sim B^H B$

Def. $\rho(A) = \max\{|\lambda| : \lambda \text{ è autovalore di } A\}$

Sia $Ax = \lambda x$, $x \in \mathbb{C}^n \setminus \{0\}$

$\|\cdot\|$ n.m.i

$|\lambda| \|x\| = \|\lambda x\| = \|Ax\| \le \|A\| \cdot \|x\| \quad \Rightarrow \quad |\lambda| \le \|A\|$

Dunque $\rho(A) \le \|A\| \ \forall$ n m.i.

TEOREMA $\forall A \in M(n,\mathbb{C})$ e $\forall \varepsilon > 0$ sufficientemente piccolo
$\exists$ n.m.i. : $\rho(A) \le \|A\| \le \rho(A) + \varepsilon$
Inoltre se gli autovalori di modulo massimo appartengono a blocchi di Jordan di dim 1 allora $\exists$ n.m.i $\|\cdot\| : \rho(A) = \|A\|$

DIMOSTRAZIONE

Oss Data $\|\cdot\| : \mathbb{C}^n \longrightarrow \mathbb{R}$ norma e $S \in M(n,\mathbb{C})$, $\det S \neq 0$
$\Rightarrow \|\cdot\|_S : x \longmapsto \|Sx\|$ è norma vettoriale

$$\|A\|_S = \max_{\|x\|_S = 1} \|Ax\|_S = \max_{\|Sx\|=1} \|SAx\| = \max_{\|y\|=1} \|SAS^{-1}y\|$$

Dunque $\|A\|_S = \|SAS^{-1}\|$

Data $\|\cdot\|$ n.m.i. e data $S \in M(n,\mathbb{C})$, $\det S \neq 0$
allora $\|\cdot\|_S : A \longmapsto \|SAS^{-1}\|$ è n.m.i

Sia $A = WJW^{-1}$ dove $J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_k \end{pmatrix}$, J_i blocchi di Jordan

$$D_\varepsilon^{-1} W^{-1} A W D_\varepsilon = D_\varepsilon^{-1} J D_\varepsilon = \begin{pmatrix} J_1^{(\varepsilon)} & & \\ & \ddots & \\ & & J_k^{(\varepsilon)} \end{pmatrix} \quad \text{con } D_\varepsilon = \begin{pmatrix} 1 & & 0 \\ & \varepsilon & \\ 0 & & \varepsilon^m \end{pmatrix}$$

$$\text{e } J_i^{(\varepsilon)} = \begin{pmatrix} \lambda_i & \varepsilon & & \\ & \lambda_i & \ddots & \\ & & \ddots & \varepsilon \\ & & & \lambda_i \end{pmatrix}$$

$\|A\|_S = \|SAS^{-1}\|_\infty$ dove $S = D_\varepsilon^{-1} W^{-1}$

$$\Rightarrow \|A\|_S = \begin{cases} \rho(A) + \varepsilon & \text{se esiste un autovalore di modulo massimo in blocchi di } J \text{ di dim almeno } 2 \\ \rho(A) & \text{altrimenti, e se } \varepsilon \text{ è abbastanza piccolo} \end{cases}$$

se $|\lambda| < \rho(A)$ e ε non abbastanza piccolo, potrei avere $|\lambda| + \varepsilon > \rho(A)$ □

TEOREMA $\forall \|\cdot\|$ norma di matrice e $\forall A \in M(n,\mathbb{C})$, vale
$$\lim_{k \to +\infty} \|A^k\|^{\frac{1}{k}} = \rho(A)$$

DIMOSTRAZIONE

Se $\exists \lim_{k\to+\infty} \|A^k\|'^{\frac{1}{k}} = \ell$ per un'opportuna $\|\cdot\|'$,
allora $\forall \|\cdot\|$, $\exists \lim_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \ell$

Infatti, date $\|\cdot\|$, $\|\cdot\|'$, $\exists \alpha, \beta > 0$ t.c. $\alpha\|A\|' \le \|A\| \le \beta\|A\|'$ $\forall A \in M(n,\mathbb{C})$
dunque $(\alpha\|A^k\|')^{\frac{1}{k}} \le \|A^k\|^{\frac{1}{k}} \le (\beta\|A^k\|')^{\frac{1}{k}}$, cioè

$$\alpha^{\frac{1}{k}} \|A^k\|'^{\frac{1}{k}} \le \|A^k\|^{\frac{1}{k}} \le \beta^{\frac{1}{k}} \|A^k\|'^{\frac{1}{k}}$$

per $k \to +\infty$: $\ell \le \lim_{k\to+\infty} \|A^k\|^{\frac{1}{k}} \le \ell \Rightarrow \lim_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \ell$

Sia $A = SJS^{-1}$ con J forma normale di Jordan
Data $B \in M(n,\mathbb{C})$, definisco $\|B\|_S = \|S^{-1}BS\|_\infty$
(abbiamo visto che $\|\cdot\|_S$ è n.m.i dalla norma vettoriale $x \longmapsto \|Sx\|_\infty$)

$$\|A^k\|_S = \|J^k\|_\infty \quad (A^k = SJ^kS^{-1})$$

$$J = \begin{pmatrix} J_1 & & 0 \\ & \ddots & \\ 0 & & J_r \end{pmatrix} \Rightarrow J^k = \begin{pmatrix} J_1^k & & 0 \\ & \ddots & \\ 0 & & J_r^k \end{pmatrix}$$

$$\|J^k\|_\infty = \max_{i=1,\dots,r} \|J_i^k\|_\infty$$

Se $\dim J_i = 1 \Rightarrow \|J_i^k\| = |\lambda_i|^k$

Sia $\hat{J} = \underbrace{\begin{pmatrix} \lambda & 1 & & 0 \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{pmatrix}}_{m} = \lambda I + Z$ dove $Z = \begin{pmatrix} 0 & 1 & & 0 \\ & 0 & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix}$

$$\hat{J}^k = (\lambda I + Z)^k = \sum_{i=0}^{k} \binom{k}{i} \lambda^{k-i} Z^i = \sum_{i=0}^{m-1} \binom{k}{i} \lambda^{k-i} Z^i$$

$$\Rightarrow \hat{J}^k = \begin{pmatrix} \lambda^k & \binom{k}{1}\lambda^{k-1} & \cdots & \binom{k}{m-1}\lambda^{k-m+1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \binom{k}{1}\lambda^{k-1} \\ 0 & & & \lambda^k \end{pmatrix}$$

Quindi $\|\hat{J}^k\|_\infty = \sum_{i=0}^{m-1} \binom{k}{i} |\lambda|^{k-i} = |\lambda|^k \sum_{i=0}^{m-1} \binom{k}{i} |\lambda|^{-i}$

$$\|\hat{J}^k\|_\infty^{\frac{1}{k}} = |\lambda| \Big(\underbrace{\sum_{i=0}^{m-1} \binom{k}{i} |\lambda|^{-i}}_{\text{polinomio in } k \text{ di grado costante } (m-1)} \Big)^{\frac{1}{k}} \xrightarrow[k \to +\infty]{} |\lambda|$$

$$\Rightarrow \lim_{k\to+\infty} \|J^k\|_\infty^{\frac{1}{k}} = \max_{i=1,\dots,r} \|J_i^k\|_\infty^{\frac{1}{k}} = \max_i \{|\lambda_i| \mid \lambda_i \text{ autovalore di } J_i\} = \rho(A) \qquad \square$$

Sia $A \in M(n,\mathbb{C})$: $\|A\| < 1$ dove $\|\cdot\|$ è n.m.i.

Dunque $\rho(A) \le \|A\| < 1 \Rightarrow \det(I-A) \neq 0$

$(I-A)(I-A)^{-1} = I \Rightarrow (I-A)^{-1} - A(I-A)^{-1} = I$

$\Rightarrow (I-A)^{-1} = A(I-A)^{-1} + I$

$\Rightarrow \|(I-A)^{-1}\| = \|A(I-A)^{-1} + I\| \le \|A(I-A)^{-1}\| + \underbrace{\|I\|}_{1} \le \|A\| \cdot \|(I-A)^{-1}\| + 1$

$\Rightarrow \|(I-A)^{-1}\| \le \dfrac{1}{1-\|A\|}$

Data $A \in M(n,\mathbb{C})$, $\det A \neq 0$ e $b \in \mathbb{C}^n \setminus \{0\}$

Voglio studiare il condizionamento della soluzione del problema $Ax = b$

Considereiamo il problema $(A+\delta_A)y = b+\delta_b$, conoscendo $\dfrac{\|\delta_B\|}{\|b\|}$ e $\dfrac{\|\delta_A\|}{\|A\|}$

Scrivendo $y = x + \delta x$, $\dfrac{\|\delta x\|}{\|x\|} \le ?$

Caso particolare: $\delta_A = 0$

$Ax = b \qquad A(x+\delta_x) = b+\delta_b$

$\Rightarrow A\delta_x = \delta_b \Rightarrow \delta_x = A^{-1}\delta_b \Rightarrow \|\delta_x\| \le \|A^{-1}\|\,\|\delta_b\|$

$Ax = b \Rightarrow \|b\| = \|Ax\| \le \|A\|\,\|x\| \Rightarrow \|x\| \ge \dfrac{\|b\|}{\|A\|}$

$\Rightarrow \dfrac{\|\delta_x\|}{\|x\|} \le \|A\|\,\|A^{-1}\| \dfrac{\|\delta_b\|}{\|b\|}$

$\|A\|\,\|A^{-1}\|$ numero di condizionamento della matrice A nella norma $\|\cdot\|$

Il sistema $Ax=b$ è ben condizionato se $\mu(A) = \|A\|\,\|A^{-1}\|$ è piccolo

Vale $\mu(A) \ge 1$, infatti $1 = \|I\| = \|AA^{-1}\| \le \|A\|\,\|A^{-1}\| = \mu(A)$

Se $A^H A = AA^H = I$ (A unitaria), allora

$\|A\|_2 = \sqrt{\rho(A^H A)} = 1$

$\|A^{-1}\|_2 = \|A^H\|_2 = \sqrt{\rho(AA^H)} = 1 \Rightarrow \mu_2(A) = 1$

A normale $(A^H A = AA^H)$

allora $\mu_2(A) = \dfrac{\max_{\lambda \in \Lambda(A)} |\lambda|}{\min_{\lambda \in \Lambda(A)} |\lambda|}$

Teorema Data $A \in M(n,\mathbb{C})$, $\det A \neq 0$ e $b \in \mathbb{C}^n \setminus \{0\}$ (problema $Ax = b$)
Considereiamo il problema $(A+\delta_A)y = b+\delta_b$, conoscendo $\dfrac{\|\delta_B\|}{\|b\|} = \varepsilon_b$ e $\dfrac{\|\delta_A\|}{\|A\|} = \varepsilon_A$
Sia $y = x + \delta x$. Inoltre $\|A^{-1}\|\,\|\delta_A\| < 1$
Allora $\det(A+\delta_A) \neq 0$ e

$$\frac{\|\delta_x\|}{\|x\|} \le \frac{\|A\| \cdot \|A^{-1}\|}{1 - \varepsilon_A \|A\| \|A^{-1}\|} (\varepsilon_A + \varepsilon_b)$$

FATTORIZZAZIONI DI MATRICI

$A \in M(n,\mathbb{C})$, $\det A \neq 0$, $b \in \mathbb{C}^n$

Risolvere $Ax = b$

Idea: scrivo $A = BC$, facilmente calcolabili e t.c.

$Ax = b$

$BCx = b$: $\begin{cases} By = b \\ Cx = y \end{cases}$ facilmente risolvibili

esempio (1) $A = LU$ (Lower Upper) L triangolare inf, U triangolare sup

(2) $A = PLU$, con P di permutazione ; $A = P_1 L U P_2$

(3) $A = QR$ Q unitaria, R triangolare sup

Sia A triangolare inferiore, $a_{ii} \neq 0$ $\forall i$

$Ax = b$

$$\begin{pmatrix} a_{11} & & 0 \\ a_{21} & a_{22} & \\ \vdots & \ddots & \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

Metodo di sostituzione in avanti

$x_1 = \frac{b_1}{a_{11}}$

$x_i = (b_i - \sum_{j=1}^{i-1} a_{ij} x_j) \cdot \frac{1}{a_{ii}}$, $i = 2, \dots, n$

Il costo computazionale : $2i-1$ operazioni al passo i-esimo

$\Rightarrow$ #OP totale $= \sum_{i=1}^{n} (2i-1) = n^2$

Se la matrice fosse inferiore, si fa il metodo di sostituzione all'indietro

Sia A unitaria

$Ax = b$: $x = A^H b$

Pose $i = 1, \dots, n$ $x_i = \sum_{j=1}^{n} \overline{a}_{ji} \cdot b_j$

#OP totale $= n(2n-1) = 2n^2 - n$

Fattorizzazione LU

Def. Data $A \in M(n,\mathbb{C})$, una fattorizzazione LU è una fattorizzazione $A = LU$, dove $L = \begin{pmatrix} 1 & & 0 \\ * & \ddots & \\ & & 1 \end{pmatrix}$ è triangolare inferiore e U è triangolare superiore

Data $A = LU$, $\det A \neq 0$, $Ax = b$ si risolve

$LUx = b \iff \begin{cases} Ly = b \\ Ux = y \end{cases}$ con $2n^2$ ops

Def. Una sottomatrice principale di $A \in M(n,\mathbb{C})$

dato $\Omega \subsetneq \{1, \dots, n\}$, è la matrice $\hat{A} = (a_{ij})_{i,j \in \Omega}$

Si dice sottomatrice principale di testa se $\Omega = \{1, \dots, k\}$, $1 \leq k < n$

Teorema Sia $A \in M(n,\mathbb{C})$. Se tutte le sottomatrici principali di testa (non banali) sono non singolari, allora esiste unica la fattorizzazione LU di A

DIMOSTRAZIONE

Per induzione sulla dimensione n

- $n=1$: $A = 1 \cdot A$
- Supponiamo la tesi vera per $n-1$.

$$A = \left(\begin{array}{c|c} A_1 & b \\ \hline {}^tc & a_{nn} \end{array}\right) = \left(\begin{array}{c|c} L_1 & 0 \\ \hline {}^tx & 1 \end{array}\right)\left(\begin{array}{c|c} U_1 & y \\ \hline 0 & u_{nn} \end{array}\right)$$

dove A_1 è la sottomatrice principale di testa di dim $n-1$ verifica le ipotesi del teorema, quindi $\exists!\ A_1 = L_1U_1$

Quindi L_1 e U_1 sono unici.

Per l'ipotesi del teorema, $\det A_1 \neq 0 \Rightarrow \det U_1 \neq 0$

$$\begin{cases} {}^tc = {}^tx U_1 & \Rightarrow {}^tx = {}^tc U_1^{-1} \\ b = L_1 y & \det L_1 \neq 0 \Rightarrow y = L_1^{-1} b \\ a_{nn} = {}^tx y + u_{nn} & \Rightarrow u_{nn} = a_{nn} - {}^tx y \end{cases}$$

$\Rightarrow$ La fattorizzazione è unica. □

Esempio Se non sono verificate le ipotesi, può esistere comunque la fattorizzazione LU

$$\begin{pmatrix} 0 & 1 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$$

$\exists P$ di permutazione t.c. $PA = LU$

Matrici elementari

Def. Una **matrice elementare** è una matrice del tipo

$M = I - \sigma u v^H$ dove $\sigma \in \mathbb{C}$, $u, v \in \mathbb{C}^n$, $(uv^H)_{ij} = u_i \overline{v_j}$

$$v^H u = \sum_{i=1}^{n} u_i \overline{v_i} \in \mathbb{C}$$

$x \in \mathbb{C}^n$, $Mx = x - \sigma u v^H x = x - \sigma (v^H x) u$

Se $x = u$: $Mu = (1 - \sigma(v^H u))u$

Se $v^H x = 0$: $Mx = x$

$\operatorname{tr}(M) = \operatorname{tr}(I) - \sigma \operatorname{tr}(uv^H) = n - \sigma v^H u = n - 1 + (1 - \sigma v^H u)$

Teorema M è non singolare $\iff \sigma(v^H u) \neq 1$

e in tal caso $M^{-1} = I - \tau u v^H$, $\tau = \frac{-\sigma}{1 - \sigma v^H u}$

DIMOSTRAZIONE

Se $\sigma v^H u = 1$, allora $Mu = 0$, $u \neq 0$, quindi M è singolare

Se $\exists x \neq 0$ t.c. $Mx = 0$, cioè $x - \sigma(v^H x)u = 0$,

ossia $x = \sigma(v^H x)u$. Scelgo $x = u$ e $\sigma(v^H u) = 1$

Cerco τ t.c. $(I - \sigma u v^H)(I - \tau u v^H) = I$

$\Rightarrow I - \sigma u v^H - \tau u v^H + \sigma\tau u v^H u v^H = I$

$\Rightarrow (-\sigma - \tau + \sigma\tau(v^H u)) u v^H = 0 \Rightarrow -\sigma - \tau + \sigma\tau(v^H u) = 0$

$\underset{M \neq I}{\Rightarrow} \tau = \frac{-\sigma}{1 - \sigma v^H u}$ □

Teorema $\forall x, y \in \mathbb{C}^n \setminus \{0\}$ $\exists M$ matrice elementare non singolare t.c. $Mx = y$

DIMOSTRAZIONE

Cerco u, v, σ : $(I - \sigma u v^H)x = y \Rightarrow x - \sigma u v^H x = y$

Scelgo v: $v^H x \neq 0$, $\sigma u = \frac{x-y}{v^H x}$

Deve essere $\sigma v^H u \neq 1$ per la non singolarità di M:

$$\sigma v^H u = \frac{v^H x - v^H y}{v^H x} = 1 - \frac{v^H y}{v^H x} \neq 1$$

quindi scelgo v t.c. $v^H y \neq 0$ □

Matrici elementari di Gauss

Def. Dato $x \in \mathbb{C}^n$, $x = (x_i)_{i=1,\dots,n}$ con $x_1 \neq 0$

la **matrice elementare di Gauss** è la matrice elementare

$$M = I - u e_1^t \quad : Mx = x_1 e_1 \qquad \text{con } u = \begin{pmatrix} 0 \\ x_2/x_1 \\ \vdots \\ x_n/x_1 \end{pmatrix}$$

$$M = \begin{pmatrix} 1 & & \\ -u_2 & 1 & \\ \vdots & & \ddots \\ -u_n & 0 & & 1 \end{pmatrix} \qquad M^{-1} = \begin{pmatrix} 1 & & \\ u_2 & 1 & \\ \vdots & & \ddots \\ u_n & 0 & & 1 \end{pmatrix}$$

Sia $A \in M(n, \mathbb{C})$ che verifica le ipotesi di esistenza/unicità della fattorizzazione LU

- M_1 elementare di Gauss: $M_1 A^1 = \begin{pmatrix} a_{11}^{(1)} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ $(\Rightarrow a_{11} \neq 0)$

$$M_1 A = \begin{pmatrix} 1 & & \\ -u_2 & 1 & \\ \vdots & & \ddots \\ -u_n & 0 & & 1 \end{pmatrix} \left(\begin{array}{c|c} a_{11} & b \\ \hline c & A_1 \end{array}\right) = \left(\begin{array}{c|c} a_{11} & b \\ \hline 0 \\ \vdots & B_1 \\ 0 & \end{array}\right)$$

dove $B_1 = A_1 - \begin{pmatrix} u_2 \\ \vdots \\ u_n \end{pmatrix} b$ e $\begin{pmatrix} u_2 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} a_{21}/a_{11} \\ \vdots \\ a_{n1}/a_{11} \end{pmatrix}$

Quindi $B_1 = A_1 - \begin{pmatrix} a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} a_{11}^{-1} (a_{12} \cdots a_{1n}) = \begin{pmatrix} a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & & \vdots \\ a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix}$

con $a_{ij}^{(1)} = a_{ij} - \frac{a_{i1} a_{1j}}{a_{11}}$ $i, j = 2, \dots, n$

- $\hat{M}_2$ elementare di Gauss : $\hat{M}_2 \begin{pmatrix} a_{22}^{(1)} \\ a_{32}^{(1)} \\ \vdots \\ a_{n2}^{(1)} \end{pmatrix} = \begin{pmatrix} a_{22}^{(1)} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

e definisco $M_2 = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & \hat{M}_2 \end{array}\right)$

$$M_2 M_1 A = \left(\begin{array}{c|ccc} a_{11} & a_{12} & \cdots & a_{1n} \\ \hline 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ 0 & & & \\ \vdots & & B_2 & \\ 0 & & & \end{array}\right)$$

B_2 : $(n-2)$

da matrice $M_{n-1} M_{n-2} \cdots M_1 A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ & & \ddots & \vdots \\ 0 & & & a_{nn}^{(n-1)} \end{pmatrix} = U$

da cui $M_{n-1} M_{n-2} \cdots M_1 = L^{-1} \Rightarrow L = M_1^{-1} \cdots M_{n-2}^{-1} M_{n-1}^{-1} = \begin{pmatrix} 1 & & & 0 \\ a_{21}/a_{11} & 1 & & \\ \vdots & a_{32}^{(1)}/a_{22}^{(1)} & 1 & \\ a_{n1}/a_{11} & a_{n2}^{(1)}/a_{22}^{(1)} & \cdots & 1 \end{pmatrix}$

costo computazionale

$A=(a_{ij})_{i,j=1,\dots,n}$ t.c. $\det(A(1:k,1:k))\neq 0$ per $k=1,\dots,n-1$

$\Rightarrow \exists!$ fattorizzazione $A=LU$

$A_{k+1} = M_k A_k$ (M_k: matrice elementare di Gauss)

$$= \begin{pmatrix} I_{k-1} & 0 \\ 0 & \begin{matrix} 1 & 0 \\ m_{k+1,k}^{(k)} & \\ \vdots & \\ m_{n,k}^{(k)} & 0 \quad 1 \end{matrix} \end{pmatrix} \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ & \ddots & \vdots \\ 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & a_{k+1,k}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ & & \vdots & & \vdots \\ & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ & \ddots & \vdots \\ 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & 0 & a_{k+1,k+1}^{(k+1)} & \cdots & a_{k+1,n}^{(k+1)} \\ & & \vdots & \vdots & & \vdots \\ & & 0 & a_{n,k+1}^{(k+1)} & \cdots & a_{nn}^{(k+1)} \end{pmatrix}$$

dove $a_{ij}^{(k+1)} = a_{ij}^{(k)} + m_{ik}^{(k)} a_{kj}^{(k)}$ $\quad i,j=k+1,\dots,n$ $\quad 2(n-k)^2$ flops

$m_{ik}^{(k)} = -\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$ $\quad i=k+1,\dots,n$ $\quad (n-k)$ flops

Costo totale: $\sum_{k=1}^{n-1} 2(n-k)^2+(n-k) = \sum_{k=1}^{n-1} 2k^2+k = 2\left(\frac{n^3}{3}+O(n^2)+O(n^2)\right) \doteq \frac{2}{3}n^3$

$$L = \begin{pmatrix} 1 & & & \\ -m_{21}^{(1)} & 1 & & \\ \vdots & -m_{32}^{(2)} & \ddots & \\ -m_{n1}^{(1)} & \cdots & \cdots & 1 \end{pmatrix}$$

$$A = \left(\begin{array}{c|c} L_k & 0 \\ \hline * & \begin{matrix} 1 & \\ & \ddots \\ & & 1 \end{matrix} \end{array}\right)\left(\begin{array}{c|c} U_k & * \\ \hline 0 & \end{array}\right) = \left(\begin{array}{c|c} A(1:k,1:k) & * \\ \hline * & * \end{array}\right)$$

stabilità numerica

Siano $\tilde{L},\tilde{U}$ i valori effettivamente calcolati di L e U

Sia $\Delta_A = A-\tilde{L}\tilde{U}$

Teorema $|\Delta_A| \leq 2nu\,(|A|+|\tilde{L}|\cdot|\tilde{U}|)+O(u^2)$

(elemento per elemento)

(dove, data $B=(b_{ij})$, $|B|=(|b_{ij}|)$)

strategia di pivoting: pivot parziale

$$A_k = \begin{pmatrix} & & & \\ & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & \vdots & & \vdots \\ & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}$$

Sia $h\geq k$: $|a_{hk}^{(k)}| \geq |a_{ik}^{(k)}|$ per $i=k,\dots,n$

Scambio la riga h-esima con la riga k-esima

$$P_k A_k = \begin{pmatrix} & & * \\ & & a_{hk}^{(k)} \\ & & * \end{pmatrix}$$

$A_{k+1} = M_k P_k A_k$ con $M_k = I - v_k e_k^T$ con $v_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ * \neq 0 \\ * \end{pmatrix}$, $|v_k| \leq 1$

Applicando la strategia a ogni passo, ottengo

$PA=LU$ dove P è un'opportuna matrice di permutazione

Se $\det A \neq 0$, $A = LU \Rightarrow A^{-1} = U^{-1}L^{-1}$

$$\begin{pmatrix} l_{11} & & 0 \\ l_{21} & \ddots & \\ \vdots & \ddots & \\ l_{n1} & \cdots & l_{nn} \end{pmatrix} \begin{pmatrix} b_{11} & & \\ b_{21} & b_{22} & \\ \vdots & \ddots & \ddots \end{pmatrix} = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}$$

(n^2 e $(n-1)^2$)

Costo $2n^3$

Altro metodo per sistemi lineari:

$M_1 A x = M_1 b$ dove M_1 el. di Gauss t.c. $M_1 A e_1 = a_{11} e_1$

$M_2 M_1 A x = M_2 M_1 b$

$\dots$

$U x = M_{n-1} \cdots M_1 A x = M_{n-1} \cdots M_1 b = \tilde{b}$ sistema triangolare

Data $A \in M(n, \mathbb{C})$, $\exists P$ di permutazione t.c.

$PA = LU$

(si realizza con la strategia di pivot parziale)

Matrici elementari di Householder

Def. Una matrice elementare di Householder è una matrice elementare hermitiana e unitaria, cioè

$M = I - \beta u u^H$, $u \in \mathbb{C}^n$, $\beta \in \mathbb{R}$

$M^2 = M M^H = I \longrightarrow (I - \beta u u^H)^2 = I \longrightarrow I - 2\beta u u^H + \beta^2 u u^H u u^H = I$

$\longrightarrow \beta(\beta \cdot u^H u - 2) u u^H = 0$

$\Rightarrow \beta \cdot (\beta u^H u - 2) = 0 \Rightarrow \beta = 0 \ \vee \ \beta = \frac{2}{u^H u}$ se $u \neq 0$

Proprietà: $\forall x \in \mathbb{C}^n \setminus \{0\}$ $\exists M$ elementare di Householder: $Mx = \alpha e_1$

Sia $x \in \mathbb{C}^n$: $Mx = \alpha e_1$

$\|x\|_2 = \|Mx\|_2 = \|\alpha e_1\|_2 = |\alpha|$

(M unitaria)

$M = M^H \Rightarrow x^H M x = x^H \alpha e_1 = \alpha \bar{x}_1 \in \mathbb{R}$ ($x^H M x \in \mathbb{R}$)

Scelgo $\alpha = \theta \|x\|_2$ dove $|\theta| = 1$

e tale che $\alpha \bar{x}_1 = \theta \bar{x}_1 \|x\|_2 \in \mathbb{R} \Rightarrow \theta = \begin{cases} \pm 1 & \text{se } x_1 = 0 \\ \pm \frac{x_1}{|x_1|} & \text{se } x_1 \neq 0 \end{cases}$

$(I - \beta u u^H) x = \alpha e_1$

$x - \beta u (u^H x) = \alpha e_1$

$\beta (u^H x) u = x - \alpha e_1 = \begin{pmatrix} x_1 - \alpha \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$

Scelgo $u = \begin{pmatrix} x_1 - \alpha \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$, $\beta = \frac{2}{u^H u}$

Nell'implementazione, si sceglie $\theta = -\frac{x_1}{|x_1|} \Longrightarrow \alpha = -\frac{x_1}{|x_1|} \|x\|_2$

Il costo computazionale per la costruzione di H è:

- calcolo di $u_1 = x_1 - (\theta\|x\|_2)$ se $x_1 \neq 0$
- calcolo di $\beta = \frac{2}{u^H u} = \frac{2}{(u_1^2 + \sum_{j=2}^{n}|x_j|^2)}$ e $\|x\|_2 = \left(\sum_{j=1}^{n}|x_j|^2\right)^{1/2}$

(i) Calcolo $s = \sum_{j=2}^{n}|x_j|^2$ $\quad n-2+1$ molt., $n-2$ addiz.

(ii) $\beta = \frac{2}{(|u_1|^2 + s)}$ $\quad 3$ op

(iii) $\|x\|_2 = (s + |x_1|^2)^{1/2}$ $\quad 3$ op

(iv) operazioni in numero costante rispetto a n cost

Costo totale: $2n$ a meno di costanti additive

Moltiplicazione Hv:

$$Hv = (I - \beta u u^H)v = v - \beta u u^H v = v - (\beta(u^H v))u$$

(n; n; $2n+c$; 1 op)

Costo totale: $4n$ a meno di costanti additive

Fattorizzazione QR

Una fattorizzazione QR di $A \in M(n, \mathbb{C})$ è

$A = QR$ con Q unitaria, R triangolare superiore

Applicazione: $Ax = b \longleftrightarrow QRx = b \longleftrightarrow \begin{cases} Qy = b \longrightarrow y = Q^H b \\ Rx = y \end{cases}$

Metodo di Householder

Dato $A \in M(n, \mathbb{C})$

Sia H_1 elementare di Householder t.c. $H_1 A e_1 = \begin{pmatrix} \gamma \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

Quindi $A_1 = H_1 A = \left(\begin{array}{c|c} a_{11}^{(2)} & u^T \\ \hline 0 & \\ \vdots & B_1 \\ 0 & \end{array}\right)$

Sia $\tilde{H}_2 \in M(n-1, \mathbb{C})$ di Householder t.c. $\tilde{H}_2 B_1 e_1 = \begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ e $H_2 = \left(\begin{array}{c|c} 1 & 0 \cdots 0 \\ \hline 0 & \\ \vdots & \tilde{H}_2 \\ 0 & \end{array}\right)$

Vale $A_2 = H_2 A_1 = \left(\begin{array}{c|cc} a_{11}^{(2)} & * & \cdots * \\ \hline 0 & a_{22}^{(2)} & * \cdots * \\ & 0 & \\ \vdots & \vdots & B_2 \\ 0 & 0 & \end{array}\right)$

Al passo k-esimo:

$$A_k = H_k \cdots H_1 A = \left(\begin{array}{cc} \text{0 (triangolo)} & * \\ \hline 0 & B_k \end{array}\right) \qquad \text{per } k = 1, \dots, n-1$$

Dunque $A_{n-1} = R$

$R = A_{n-1} = H_{n-1} \cdots H_1 A =$

$\Rightarrow A = H_1^H H_2^H \cdots H_{n-1}^H A = QA$

Oss: la fattorizzazione QR esiste sempre

costo computazionale

Al k-esimo passo:

(1) costruzione di una matrice di Householder $\tilde{H}_k$ $(n-k)\times(n-k)$

(2) calcolo $\tilde{H}_k B_k$ $\quad\overbrace{\qquad}^{(n-k)^2}$

(1): $2(n-k)$ op

(2): $\tilde{H}_k B_k = (I - \beta_k u_k u_k^H) B_k = B_k - \overbrace{\underbrace{\beta_k u_k}_{n}\,\underbrace{(u_k^H B_k)}_{2(n-k)^2}}^{(n-k)^2}$

Costo totale: $\sum_{k=1}^{n-1} 4(n-k)^2 + O(n) \doteq 4\frac{n^3}{3}$

(costo doppio rispetto a LU)

$$Ax=b \longleftrightarrow H_{n-1}\cdots H_2 H_1 Ax = H_{n-1}\cdots H_2 H_1 b \longleftrightarrow Rx=\tilde{b}$$

Teorema $Ax=b$, $\tilde{R}=fl(R)$ calcolata con il metodo di Householder
Se $\tilde{b}_1 = fl(b_1)$ e se $\tilde{x}$ risolve $(A+\Delta_A)\tilde{x}=\tilde{b}_1$, allora
$\|\Delta_A\|_F \le \gamma u (n^2+n)\|A\|_F + O(u^2)$

METODI ITERATIVI

Problema: risolvere $Ax=b$, $A\in M(n,\mathbb{C})$, $\det A\neq 0$

Idea: generare una successione $x^{(k)}\in\mathbb{C}^n$ t.c. $x^{(k)}\xrightarrow[k\to+\infty]{} x$ soluzione del sistema

Metodi stazionari: Sia $A=M-N$ dove $\det M\neq 0$

$Ax=b \longleftrightarrow (M-N)x=b \longleftrightarrow Mx=Nx+b \longleftrightarrow x=M^{-1}Nx+M^{-1}b$ ($P=M^{-1}N$ matrice di iterazione)

Definisco $x^{(k+1)}=M^{-1}Nx^{(k)}+M^{-1}b$, $k=0,1,\dots$, dove $x^{(0)}\in\mathbb{C}^n$ è fissato (definisce un metodo iterativo)

Se $\exists \lim_{k\to+\infty} x^{(k)}=x^*$, allora per continuità $x^*=M^{-1}Nx^*+M^{-1}b$, cioè $Ax^*=b$

def. Un metodo iterativo si dice **convergente** se la successione $x^{(k)}$ converge per ogni scelta di $x^{(0)}$

Teorema Sia $A=M-N$, $\det M\neq 0$ e sia $P=M^{-1}N$.
Se $\exists\|\cdot\|$ n.m.i. tale che $\|P\|<1$ allora $\det A\neq 0$,
il metodo è convergente e $\lim_{k\to+\infty} x^{(k)}=x$,
dove x è l'unica soluzione di $Ax=b$.

DIMOSTRAZIONE

Per assurdo, sia A singolare $\Rightarrow \exists u\in\mathbb{C}^n\setminus\{0\}$ t.c. $Au=0$

$\Rightarrow Mu=Nu \Rightarrow u=M^{-1}Nu \Rightarrow 1$ è autovalore di P

$\Rightarrow \rho(P)\geq 1$ ma $\|P\|\geq\rho(P)\geq 1$ $\forall\|\cdot\|$ n.m.i. ↯

Dunque $\det A\neq 0 \Rightarrow \exists! x$ t.c. $Ax=b$

Sia $e^{(k)}=x^{(k)}-x$, $q=M^{-1}b$

$e^{(k+1)}=x^{(k+1)}-x=Px^{(k)}+q-(Px+q)=P(x^{(k)}-x)=Pe^{(k)}$ $k=0,1,\dots$

Induttivamente $e^{(k)}=P^k e^{(0)}$ per $k=0,1,\dots$

$\Rightarrow \|e^{(k)}\|=\|P^k e^{(0)}\|\leq\|P^k\|\|e^{(0)}\|\leq\|P\|^k\|e^{(0)}\|$

($\|\cdot\|$: norma vettoriale che induce la n.m.i.; $\|P\|^k\to 0$)

dunque $\lim_{k\to+\infty}\|e^{(k)}\|=0$ $\forall x^{(0)}$ □

Teorema Il metodo iterativo è convergente e $\det A\neq 0 \iff \rho(P)<1$

DIMOSTRAZIONE

($\Leftarrow$) Sia $\rho(P)<1$, $\varepsilon>0$ t.c. $\rho(P)+\varepsilon<1$

So che esiste $\|\cdot\|$ n.m.i t.c. $\|P\|\leq\rho(P)+\varepsilon<1$

Quindi applico il teorema precedente

($\Rightarrow$) Sia $\det A\neq 0$, x soluzione del sistema $Ax=b$

So che $e^{(k)}=x^{(k)}-x\xrightarrow[k\to+\infty]{}0$ $\forall x^{(0)}$

So che $e^{(k)}=P^k e^{(0)}$. Scelgo $e^{(0)}$ autovalore di P relativo a λ, $e^{(0)}=u$:

$Pu=\lambda u$, $u\in\mathbb{C}^n\setminus\{0\}$, quindi $e^{(k)}=P^k u=\lambda^k u\xrightarrow[k\to+\infty]{}0$

quindi $|\lambda|<1$, e poiché $\rho(P)<1$ □

Sia $\|\cdot\|$ norma vettoriale

$r_k = \frac{\|e^{(k)}\|}{\|e^{(k-1)}\|}$ riduzione dell'errore al passo k-esimo (per $k \geq 1$)

$\theta_k = \sqrt[k]{r_1 r_2 \cdots r_k}$ media geometrica della riduzione dell'errore nei primi k passi

$$\theta_k = \left(\frac{\|e^{(1)}\|}{\|e^{(0)}\|} \cdot \frac{\|e^{(2)}\|}{\|e^{(1)}\|} \cdot \ldots \cdot \frac{\|e^{(k)}\|}{\|e^{(k-1)}\|} \right)^{1/k} = \left(\frac{\|e^{(k)}\|}{\|e^{(0)}\|} \right)^{1/k}$$

$$= \left(\frac{\|P^k e^{(0)}\|}{\|e^{(0)}\|} \right)^{1/k} \leq \left(\frac{\|P^k\| \cdot \|e^{(0)}\|}{\|e^{(0)}\|} \right)^{1/k} = \|P^k\|^{1/k} \xrightarrow[k \to +\infty]{} \rho(P)$$

Quindi $\theta = \lim_{k \to +\infty} \theta_k \leq \rho(P)$ (riduzione asintotica media dell'errore a ogni passo)

Oss se $e^{(0)} = a$, $Pa = \lambda a$, $|\lambda| = \rho(P)$, allora $\theta = \rho(P)$

$$\|e^{(k)}\| \sim C \cdot \rho(P)^k$$

Se $\rho(P_1) = \rho(P_2)^2$ con $\rho(P_i) < 1$

$\rho(P_1)^k < \varepsilon$

$k \log \rho(P_1) < \log \varepsilon$

Se $\rho(P) = 0 \Rightarrow \exists k \leq n$ t.c. $P^k = 0$

$\Rightarrow e^{(k)} = P^k e^{(0)} = 0$ cioè $x^{(k)} = x$

Metodi di estrapolazione:

$x^{(k+1)} = x^{(k)} - \alpha \cdot (\underbrace{Ax^{(k)} - b}_{\text{errore residuo di } x^{(k)}})$ $\quad \alpha \neq 0$

Se $\exists \lim_{k \to +\infty} x^{(k)} = x$, allora $x = x - \alpha(Ax - b) \Rightarrow Ax = b$

$x^{(k+1)} = (I - \alpha A) x^{(k)} + \alpha b = P_\alpha x^{(k)} + q$

$x = P_\alpha x + q$

La successione converge $\iff \rho(P_\alpha) < 1$

Idea: scegliere α t.c. $\rho(P_\alpha)$ è minima per gli α tali che $\rho(P_\alpha) < 1$

Condizioni di arresto fissata una tolleranza ε e k_{max}

Calcolo i vettori $x^{(k)}$ per $k = 1, \ldots, m$

dove m è il primo intero tale che

$\|x^{(m)} - x^{(m-1)}\| < \varepsilon$

o $\|Ax^{(m)} - b\| < \varepsilon$

oppure $m = k_{max}$

Metodi di Jacobi e Gauß-Seidel

$A = D - L - U$ $\quad$ L strettamente triangolare inf, U strettamente triangolare sup.
D diagonale con $a_{ii} \neq 0$ per $i = 1, \dots, n$

Jacobi: $M = D$, $N = L + U$ $\quad$ $J = D^{-1}(L+U)$
Gauss-Seidel: $M = D - L$, $N = U$ $\quad$ $G = (D-L)^{-1}U$

Metodo di Jacobi per $Ax = b$, $\det A \neq 0$

$$x^{(k+1)} = D^{-1}(L+U)x^{(k)} + D^{-1}b \qquad k = 0, 1, \dots$$
$$x_i^{(k+1)} = a_{ii}^{-1}\left(b_i - \sum_{j \neq i} a_{ij} x_j^{(k)}\right) \qquad \text{per } i = 1, \dots, n$$

Metodo di Gauss - Seidel

$$x^{(k+1)} = (D-L)^{-1}Ux^{(k)} + (D-L)^{-1}b \qquad k = 0, 1, \dots$$
$$(D-L)x^{(k+1)} = Ux^{(k)} + b$$
$$Dx^{(k+1)} = Lx^{(k+1)} + Ux^{(k)} + b$$
$$x^{(k+1)} = D^{-1}\left(Lx^{(k+1)} + Ux^{(k)} + b\right)$$
$$x_i^{(k+1)} = a_{ii}^{-1}\left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)}\right) \qquad \text{per } i = 1, \dots, n$$

condizioni sufficienti di convergenza

Teorema Data $A \in M(n, \mathbb{C})$. Se vale una delle ipotesi:
(i) A o A^H è strettamente dominante diagonale;
(ii) A o A^H è irriducibilmente dominante diagonale;
allora $\rho(J) < 1$ e $\rho(G) < 1$.

DIMOSTRAZIONE

Jacobi:

$a_{ii} \neq 0$ $\forall i$, infatti per (i) è banale
per (ii), avrei una riga nulla, assurdo per l'irriducibilità

$J = D^{-1}(L+U)$

λ autovalore di J: $\det(D^{-1}(L+U) - \lambda I) = 0$
$\Leftrightarrow \det(\lambda D - (L+U)) = 0$

$$A(\lambda) = \lambda D - (L+U) = \begin{pmatrix} \lambda a_{11} & a_{ij} \\ a_{ij} & \lambda a_{nn} \end{pmatrix}$$

Se per assurdo $|\lambda| \geq 1$, allora $A(\lambda)$ è c.d.d se A è idd
$A(\lambda)$ è f.d.d. se A è fdd
$\Rightarrow \det A(\lambda) \neq 0$ ↯ perché λ è autovalore

Gauss - Seidel:

$G = (D-L)^{-1}U$

λ autovalore di G: $\det(\lambda I - (D-L)^{-1}U) = 0$
$\Leftrightarrow \det(\lambda(D-L) - U) = 0$

Se per assurdo $|\lambda| \geq 1$, $\det(D - L - \lambda^{-1}U) = 0$

$$A(\lambda) = D - L - \lambda^{-1}U = \begin{pmatrix} a_{11} & \lambda^{-1}a_{ij} \\ a_{ij} & a_{nn} \end{pmatrix} \qquad \text{e } |\lambda^{-1}a_{ij}| \leq |a_{ij}|$$

dunque $A(\lambda)$ è c.d.d se A è idd
$A(\lambda)$ è f.d.d. se A è fdd
$\Rightarrow \det A(\lambda) \neq 0$ ↯ perché λ è autovalore

□

ZERI DI FUNZIONI

Data $f:[a,b]\longrightarrow\mathbb{R}$, $f\in C([a,b])$
Problema: calcolare $\alpha\in(a,b)$ t.c. $f(\alpha)=0$

Metodo di bisezione

$f\in C([a,b])$, $f(a)\cdot f(b)<0 \Rightarrow \exists\,\alpha\in(a,b): f(\alpha)=0$
Genera $[a_i,b_i]$: $\alpha\in(a_i,b_i)$
e $b_i-a_i=\frac{b-a}{2^i}$ $\quad(a_0=a, b_0=b)$
Per $k=0,1,\dots$. • $c_k=\frac{a_{k-1}+b_{k-1}}{2}$
• $f(c_k)$: se $f(c_k)=0$ stop $\alpha=c_k$
Altrimenti: $f(a_{k-1})f(c_k)<0 \Rightarrow a_k=a_{k-1}, b_k=c_k$
altrimenti $\Rightarrow a_k=c_k, b_k=b_{k-1}$
Condizioni di arresto: • $|f(c_k)|<\varepsilon$
• $|b_k-a_k|<\varepsilon\cdot\min\{|a_k|,|b_k|\}$

$fl(f(x))=f(x)(1+\sigma)$ dove σ dipende da f e dall'algoritmo usato per calcolare f
Si può dimostrare che l'intervallo di incertezza su α è:
$$\left[\alpha-\frac{\sigma}{f'(\alpha)},\ \alpha+\frac{\sigma}{f'(\alpha)}\right]$$
dove σ è l'incertezza su $f(x)$

Metodi del punto fisso

Definisco $g:[a,b]\longrightarrow\mathbb{R}$ t.c. $f(\alpha)=0\iff g(\alpha)=\alpha$

esempio $g(x)=x-\frac{f(x)}{h(x)}$ con $h(x)\neq0$

Problema: risolvere $g(x)=x$ con $g\in C([a,b])$
Idea: definire $\begin{cases}x_{k+1}=g(x_k) & k=0,1,\dots\\ x_0\in[a,b]\end{cases}$
Se $\exists\lim_{k\to+\infty}x_k=\alpha$, per continuità $\alpha=g(\alpha)$

Teorema del punto fisso

Sia $I=[\alpha-\rho,\alpha+\rho]$, $g\in C^1(I)$, $g(\alpha)=\alpha$
$\lambda=\max_{x\in I}|g'(x)|$ con $\lambda<1$.
Allora $\forall x_0\in I$, $x_{k+1}=g(x_k)$ per $k=0,1,\dots$, è tale che
$|x_k-\alpha|\le\lambda^k\rho$
Inoltre α è l'unico punto fisso e
in particolare $\lim_{k\to+\infty}x_k=\alpha$ $\forall x_0$ e $x_k\in I$ $\forall k$

DIMOSTRAZIONE
Per induzione su k: $k=0$, $|x_0-\alpha|\le\rho$ ok
Suppongo la tesi vera per k
$x_{k+1}-\alpha=g(x_k)-g(\alpha)=g'(\xi_k)(x_k-\alpha)$
$x_k\in I$ per hp. ind, $\xi_k\in I$
$|x_{k+1}-\alpha|=|g'(\xi_k)|\,|x_k-\alpha|\le\lambda\lambda^k\rho=\lambda^{k+1}\rho$.
Se $\exists\beta\in I$, $\beta\neq\alpha$ t.c. $g(\beta)=\beta$
$\alpha-\beta=g(\alpha)-g(\beta)=g'(\xi)(\alpha-\beta)$, $\xi\in I$
$\Rightarrow g'(\xi)=1$ ↯ □

Teorema del punto fisso con errore

Sia $g \in C^1(I)$, $I = [\alpha-\rho, \alpha+\rho]$, $\rho > 0$
$g(\alpha) = \alpha$, $\lambda = \max_{x \in I} |g'(x)| < 1$

Sia $\tilde{x}_{k+1} = g(\tilde{x}_k) + \delta_k$ per $k \geq 0$, $\tilde{x}_0 \in I$
Sia $|\delta_k| \leq \delta$, $\sigma = \frac{\delta}{1-\lambda}$
Se $\sigma < \rho$, vale $|\tilde{x}_k - \alpha| \leq \lambda^k(\rho - \sigma) + \sigma$ per $k \geq 0$

DIMOSTRAZIONE
Per induzione su k:
- $k=0$: $|\tilde{x}_0 - \alpha| \leq (\rho - \sigma) + \sigma = \rho$ è verificata
- Supponiamo la tesi sia vera per k

$\tilde{x}_{k+1} - \alpha = g(\tilde{x}_k) + \delta_k - g(\alpha) = g'(\xi_k)(\tilde{x}_k - \alpha) + \delta_k \quad |\xi_k - \alpha| < |\tilde{x}_k - \alpha|$
cioè $\xi_k \in I \Rightarrow |g'(\xi_k)| \leq \lambda$
$|\tilde{x}_{k+1} - \alpha| \leq |g'(\xi_k)||\tilde{x}_k - \alpha| + |\delta_k| \leq \lambda(\lambda^k(\rho-\sigma) + \sigma) + \delta =$
$= \lambda^{k+1}(\rho - \sigma) + \lambda\sigma + \delta$ e $\lambda\sigma + \delta = \frac{\lambda\delta}{1-\lambda} + \delta = \sigma$ □

Sia $g \in C^1([a,b])$, $\max_{x \in I} |g'(x)| < 1$
$\alpha \in (a,b) : g(\alpha) = \alpha$
Scelgo $x_0 = a$ (o b)
se non c'è convergenza scelgo $x_0 = b$ (o a)

Sia $0 < g'(x) < 1 \ \forall x \in I$
$x_{k+1} - \alpha = g(x_k) - g(\alpha) = g'(\xi_k)(x_k - \alpha)$ con $|\xi_k - \alpha| < |x_k - \alpha|$
$\Rightarrow x_{k+1} - \alpha$ e $x_k - \alpha$ hanno lo stesso segno:
$x_0 > \alpha \Rightarrow x_k > \alpha \ \forall k \geq 0$
$x_0 < \alpha \Rightarrow x_0 < \alpha \ \forall k \geq 0$
$|x_{k+1} - \alpha| \leq \lambda |x_k - \alpha| < |x_k - \alpha|$
Se $x_0 > \alpha \Rightarrow \alpha < x_{k+1} < x_k \ \forall k$: x_k monotona ↓
Se $x_0 < \alpha \Rightarrow x_k < x_{k+1} < \alpha \ \forall k$: x_k monotona ↑

Oss Se $g \in C^1([\alpha, \underbrace{\alpha+\rho}_{\text{"}I\text{"}}])$, $0 < g'(x) < 1 \ \forall x \in I$
allora $\forall x_0 \in I$, $x_0 > \alpha$, vale $\alpha < x_{k+1} < x_k \ \forall k$

Sia $-1 < g'(x) < 0 \ \forall x \in I$
$x_{k+1} - \alpha = g(x_k) - g(\alpha) = g'(\xi_k)(x_k - \alpha)$ con $|\xi_k - \alpha| < |x_k - \alpha|$
$\Rightarrow x_{k+1} - \alpha$ e $x_k - \alpha$ hanno segno opposto
$x_0 > \alpha \Rightarrow x_{2k+1} < \alpha, \ x_{2k} > \alpha$
$x_0 < \alpha \Rightarrow x_{2k+1} > \alpha, \ x_{2k} < \alpha$
$|x_{k+1} - \alpha| \leq \lambda |x_k - \alpha| < |x_k - \alpha|$
La convergenza di x_k a α è alternata
$x_0 > \alpha \Rightarrow x_{2k} \downarrow \alpha, \ x_{2k+1} \uparrow \alpha$

condizioni di arresto

$x_{k+1} = g(x_k) \quad k = 0, 1, \dots, h$

dove h è tale che $|x_{h+1} - x_h| < \varepsilon$, dove ε è una tolleranza fissata

x_{h+1} è l'approssimazione di α

$|x_{h+1} - \alpha|$

$x_{h+1} - x_h = g(x_h) - \alpha + \alpha - x_h = g(x_h) - g(\alpha) + \alpha - x_h =$

$= g'(\xi_h)(x_h - \alpha) - (x_h - \alpha) = (g'(\xi_h) - 1)(x_h - \alpha) \qquad$ con $|\xi_h - \alpha| < |x_h - \alpha|$

$$|x_h - \alpha| = \left| \frac{x_h - x_{h+1}}{1 - g'(\xi_h)} \right| \leq \frac{\varepsilon}{|1 - g'(\xi_h)|} \leq \frac{\varepsilon}{1 - \lambda}$$

Se $-1 < g'(x) < 0$, $|x_{h+1} - x_h| < \varepsilon \implies |x_h - \alpha| < \varepsilon$

velocità di convergenza

Def. Sia $\{x_n\}_{n \in \mathbb{N}}$ successione tale che $\lim_{k \to +\infty} x_k = \alpha \in \mathbb{R}$

Supponiamo esista $\lim_{k \to +\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|} = \gamma$

la successione converge in modo:

- **lineare** se $0 < \gamma < 1$
- **sublineare** se $\gamma = 1$
- **superlineare** se $\gamma = 0$

Nel caso superlineare, se $\exists p > 1$ tale che

$$\sigma = \lim_{k \to +\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|^p} \quad \text{è} \quad 0 < \sigma < +\infty$$

si dice che la convergenza è **superlineare di ordine p**.

Se $p = 2$, si parla di convergenza **quadratica**

Se $p = 3$, si parla di convergenza **cubica**

Oss $x_k = \begin{cases} 1/k & \text{se } k \text{ pari} \\ 1/2^k & \text{se } k \text{ dispari} \end{cases}$, $x_k \longrightarrow 0$, $\nexists \gamma$

esempio $x_k = \lambda^k$, $0 < \lambda < 1$

$x_k \longrightarrow 0$ in modo lineare: $\frac{x_{k+1}}{x_k} = \lambda$

esempio $x_k = \frac{1}{k}$

$x_k \longrightarrow 0$ in modo sublineare: $\frac{x_{k+1}}{x_k} = \frac{k}{k+1} \longrightarrow 1$

esempio $x_k = \gamma^{p^k}$, $0 < \gamma < 1$, $p > 1$

$x_k \longrightarrow 0$ in modo superlineare con ordine p

$\frac{x_{k+1}}{(x_k)^p} = \frac{(\gamma^{p^{k+1}})}{(\gamma^{p^k})^p} = 1$

Teorema Sia $g \in C^1([a,b])$, $\alpha \in (a,b)$, $g(\alpha)=\alpha$

(1) Se $\exists x_0 \in [a,b]$: $x_{k+1}=g(x_k)$ è ben definita e converge linearmente con fattore γ allora $\gamma = |g'(\alpha)|$

(2) Se $0<|g'(\alpha)|<1$, allora $\exists I=[\alpha-\rho, \alpha+\rho]$: $\forall x_0 \in I$ $x_{k+1}=g(x_k)$ converge a α linearmente con fattore $\gamma = |g'(\alpha)|$

DIMOSTRAZIONE

(1) Convergenza lineare $\iff \exists \lim_{k\to+\infty} \frac{|x_{k+1}-\alpha|}{|x_k-\alpha|} = \lim_{k\to+\infty} \frac{|g(x_k)-g(\alpha)|}{|x_k-\alpha|} =$

th. Lagrange dove $|\xi_k-\alpha|<|x_k-\alpha|$

$$= \lim_{k\to+\infty} \frac{|g'(\xi_k)(x_k-\alpha)|}{|x_k-\alpha|} = \lim_{k\to+\infty} |g'(\xi_k)| = |g'(\alpha)| = \gamma$$

(2) Sia $0<|g'(\alpha)|<1$, $g \in C^1([a,b]) \implies \exists I=[\alpha-\rho, \alpha+\rho]$: $|g'(x)|<1$ $\forall x \in I$

Quindi per il teorema del punto fisso $\forall x_0 \in I$, $\exists \lim_{k\to+\infty} x_k = \alpha$

Ora

th. Lagrange dove $|\xi_k-\alpha|<|x_k-\alpha|$

$$\lim_{k\to+\infty} \frac{|x_{k+1}-\alpha|}{|x_k-\alpha|} = \lim_{k\to+\infty} \frac{|g(x_k)-g(\alpha)|}{|x_k-\alpha|} \overset{\downarrow}{=} \lim_{k\to+\infty} \frac{|g'(\xi_k)(x_k-\alpha)|}{|x_k-\alpha|} = \lim_{k\to+\infty} |g'(\xi_k)| = |g'(\alpha)| = \gamma$$

dove $0<\gamma<1$ □

Teorema Sia $g \in C^1([a,b])$, $\alpha \in (a,b)$, $g(\alpha)=\alpha$

(1) Se $\exists x_0 \in [a,b]$: $x_{k+1}=g(x_k)$ è ben definita e converge in modo sublineare allora $|g'(\alpha)|=1$

(2) Se $|g'(\alpha)|=1$ e esiste $I \subset [a,b]$ intorno di α : $|g'(x)|<1$ $\forall x \in I \setminus \{\alpha\}$ e $g'(x)$ non cambia segno in I, allora $\forall x_0 \in I$ $\lim_{k\to+\infty} x_k = \alpha$ con convergenza sublineare

DIMOSTRAZIONE

(1) Convergenza sublineare $\iff$ $1 = \lim_{k\to+\infty} \frac{|x_{k+1}-\alpha|}{|x_k-\alpha|} = \lim_{k\to+\infty} \frac{|g(x_k)-g(\alpha)|}{|x_k-\alpha|} =$

th. Lagrange dove $|\xi_k-\alpha|<|x_k-\alpha|$

$$\overset{\downarrow}{=} \lim_{k\to+\infty} \frac{|g'(\xi_k)(x_k-\alpha)|}{|x_k-\alpha|} = \lim_{k\to+\infty} |g'(\xi_k)| = |g'(\alpha)|$$

(2) Osserviamo che, se $x_k \in I$, allora $|x_{k+1}-\alpha| = |g(x_k)-g(\alpha)| = |g'(\xi_k)|\,|x_k-\alpha| < |x_k-\alpha|$ poiché $|g'(\xi_k)|<1$, quindi $x_{k+1} \in I$ e $\{|x_k-\alpha|\}$ è decrescente

- Se $g'(x) \geq 0$ su I, $\{x_k\}$ è monotona e limitata, quindi ha limite β t.c. $\beta = g(\beta)$ se fosse $\beta \neq \alpha$, avrei $|\beta-\alpha| = |g(\beta)-g(\alpha)| = |g'(\xi)|\,|\beta-\alpha| \implies |g'(\xi)|=1$ ↯

Se $g'(x) \leq 0$ su I, definisco $G(x)=g(g(x))$. Osserviamo $x_{2k+2}=G(x_{2k})$, $x_{2k+1}=g(x_{2k})$.

G è t.c. $G'(x)=g'(x)g'(g(x)) \geq 0$

Applicando il ragionamento precedente, si conclude $\lim_{k\to+\infty} x_{2k} = \alpha$

e quindi $\lim_{k\to+\infty} x_{2k+1} = \lim_{k\to+\infty} g(x_{2k}) = g(\lim_{k\to+\infty} x_{2k}) = g(\alpha) = \alpha$ $\quad \forall x_0 \in I$

Come prima:

th. Lagrange dove $|\xi_k-\alpha|<|x_k-\alpha|$

$$\lim_{k\to+\infty} \frac{|x_{k+1}-\alpha|}{|x_k-\alpha|} = \lim_{k\to+\infty} \frac{|g(x_k)-g(\alpha)|}{|x_k-\alpha|} \overset{\downarrow}{=} \lim_{k\to+\infty} \frac{|g'(\xi_k)(x_k-\alpha)|}{|x_k-\alpha|} = \lim_{k\to+\infty} |g'(\xi_k)| = |g'(\alpha)| = 1$$

□

Teorema Sia $g \in C^p([a,b])$, $p>1$, $\alpha \in (a,b)$, $g(\alpha)=\alpha$

(1) Se $\exists x_0 \in [a,b]$: $x_{k+1}=g(x_k)$ converge a α in modo superlineare con ordine p, allora $g'(\alpha)=\dots=g^{(p-1)}(\alpha)=0$, $g^{(p)}(\alpha)\neq 0$

(2) Se $g^{(q)}(\alpha)=0$ per $q=1,\dots,p-1$, $g^{(p)}(\alpha)\neq 0$, allora $\exists I=[\alpha-\rho,\alpha+\rho]\subseteq[a,b]$: $\forall x_0 \in I$ $x_{k+1}=g(x_k)$ converge a α in modo superlineare di ordine p.

DIMOSTRAZIONE

(1) Provo che $g'(\alpha)=0$

Osserviamo che, per $1\le q<p$: ($\le$ sopra, $\to 0$ sopra)

$$\lim_{k\to+\infty}\frac{|x_{k+1}-\alpha|}{|x_k-\alpha|^q}=\lim_{k\to+\infty}\frac{|x_{k+1}-\alpha|}{|x_k-\alpha|^p}\,|x_k-\alpha|^{p-q}=0$$

- $0=\lim_{k\to+\infty}\frac{|x_{k+1}-\alpha|}{|x_k-\alpha|}=|g'(\alpha)|$
- Per induzione : sia $g^{(i)}(\alpha)=0$ $i=1,\dots,h-1$

  dimostro che $g^{(h)}(\alpha)=0$ con $1\le h\le p-1$

$$0\xleftarrow{k\to+\infty}\frac{|x_{k+1}-\alpha|}{|x_k-\alpha|^h}=\frac{|g(x_k)-g(\alpha)|}{|x_k-\alpha|^h}=\frac{|g^{(h)}(\xi_k)(x_k-\alpha)^h|}{|x_k-\alpha|^h\,h!}=\frac{|g^{(h)}(\xi_k)|}{h!}$$

$$g(x_k)=g(\alpha)+g'(\alpha)(x_k-\alpha)+\dots+g^{(h-1)}(\alpha)\frac{(x_k-\alpha)^{h-1}}{(h-1)!}+g^{(h)}(\xi_k)\frac{(x_k-\alpha)^h}{h!}\quad\text{con } |\xi_k-\alpha|<|x_k-\alpha|$$

Dunque $\lim_{k\to+\infty}\left|\frac{g^{(h)}(\xi_k)}{h!}\right|=0$, cioè $g^{(h)}(\alpha)=0$

Ora dimostro che $g^{(p)}(\alpha)\neq 0$

$$0\xleftarrow{}\frac{|x_{k+1}-\alpha|}{|x_k-\alpha|^p}=\frac{|g(x_k)-g(\alpha)|}{|x_k-\alpha|^p}=\left|\frac{g^{(p)}(\xi_k)}{p!}\right|\longrightarrow\left|\frac{g^{(p)}(\alpha)}{p!}\right|$$

(2) $g'(\alpha)=0 \Rightarrow \exists I=[\alpha-\rho,\alpha+\rho]\subset[a,b]$: $|g'(x)|<1$ $\forall x\in I$

Per il teorema del punto fisso, $\forall x_0\in I$, $\exists \lim_{k\to+\infty} x_k=\alpha$

Come sopra:

$$\frac{|x_{k+1}-\alpha|}{|x_k-\alpha|^p}=\frac{|g(x_k)-g(\alpha)|}{|x_k-\alpha|^p}=\left|\frac{g^{(p)}(\xi_k)}{p!}\right|\longrightarrow\left|\frac{g^{(p)}(\alpha)}{p!}\right|$$

□

esempio $g(x)=x^{\frac{4}{3}}$, $g(0)=0$, $x_{k+1}=g(x_k)$

$g\in C^1$: $g'(0)=0$ $\Rightarrow$ convergenza locale (per il th. del punto fisso)

$\lim_{k\to+\infty}\frac{|x_{k+1}-\alpha|}{|x_k-\alpha|}=g'(\alpha)$ $\Rightarrow$ convergenza superlineare (perché $g'(0)=0$)

Per l'ordine di convergenza: cerco $p>1$ t.c.

$$\lim_{k\to+\infty}\frac{x_{k+1}}{x_k^p}=s,\quad 0<s<+\infty$$

$x_{k+1}=x_k^{4/3}$ se $p=4/3$

Def. $\{x_k\}_k$ converge a α con ordine almeno $p>1$ se $\exists \beta>0$: $|x_{k+1}-\alpha|\le\beta|x_k-\alpha|^p$

Oss Se $x_k\to\alpha$ con ordine $q\ge p$ allora $x_k\to\alpha$ con ordine almeno p

confronto tra i metodi

- 2 metodi con convergenza lineare con fattore di convergenza r_1, r_2, $0<r_i<1$

  Riduzione dell'errore al passo k-esimo: $\beta_1 r_1^k$, $\beta_2 r_2^k$ con $\beta_1, \beta_2 > 0$

  $$\beta_1 r_1^{k_1} = \beta_2 r_2^{k_2}$$
  $$\log\beta_1 + k_1 \log r_1 = \log\beta_2 + k_2 \log r_2$$

  In analisi asintotica $k_1 = k_2 \dfrac{\log r_2}{\log r_1}$

  Supponiamo che il costo per passo di ciascun metodo sia c_i

  Il primo metodo è più conveniente del secondo se

  $$\underbrace{k_1 c_1}_{\substack{\text{costo totale di}\\ k_1 \text{ passi per il}\\ \text{primo metodo}}} < \underbrace{k_2 c_2}_{\substack{\text{costo totale di}\\ k_2 \text{ passi per il}\\ \text{secondo metodo}}}$$

  $$k_2 \frac{\log r_2}{\log r_1} c_1 < k_2 c_2 \longrightarrow \frac{\log r_2}{\log r_1} < \frac{c_2}{c_1}$$

- 2 metodi con convergenza superlineare con ordine di convergenza p_1, p_2

  Riduzione dell'errore al passo k-esimo: $\eta_i r_i^{p_i^k}$ con $\eta_i > 0$, $0<r_i<1$

  (Infatti $|x_k - \alpha| = e_k < \beta e_{k-1}^p \le \beta(\beta e_{k-2}^p)^p \le \dots \le \beta\beta^p\beta^{p^2}\cdots\beta^{p^{k-1}} e_0^{p^k} = \eta r^{p^k}$ dove $r = e_0 \beta^{\frac{1}{p-1}}$ e $r<1$ se e_0 è abbastanza piccolo)

  $$\eta_1 r_1^{p_1^{k_1}} = \eta_2 r_2^{p_2^{k_2}}$$
  $$\log\eta_1 + p_1^{k_1}\log r_1 = \log\eta_2 + p_2^{k_2}\log r_2$$

  In analisi asintotica: $k_1 \log p_1 + \log\log r_1 = k_2 \log p_2 + \log\log_2 r_2$

  $$k_1 = k_2 \frac{\log p_2}{\log p_1}$$

  Il primo metodo è più conveniente del secondo se

  $$k_1 c_1 < k_2 c_2$$
  $$k_2 \frac{\log p_2}{\log p_1} c_1 < k_2 c_2 \longrightarrow \frac{\log p_2}{\log p_1} < \frac{c_2}{c_1}$$

$f \in C^1([a,b])$, $\alpha \in (a,b)$, $f(\alpha)=0$

$g(x) = x - \frac{f(x)}{h(x)}$ dove $h(x) \neq 0$ in un intorno di α

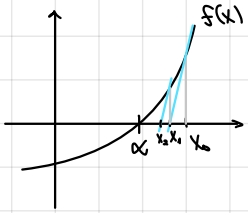
$f(\alpha)=0 \iff g(\alpha)=\alpha$

Metodo delle secanti

$h(x)=m$, $m \in \mathbb{R}\setminus\{0\}$

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

$x_{k+1} = x_k - \frac{f(x_k)}{m}$



Il metodo converge localmente ($\exists I$ intorno di α: $\forall x_0 \in I$ x_k converge)

se $|g'(\alpha)|<1$ in un intorno di α, cioè $\left|1 - \frac{f'(x)}{m}\right| < 1$

$-1 < 1 - \frac{f'(x)}{m} < 1 \longrightarrow 0 < \frac{f'(x)}{m} < 2$

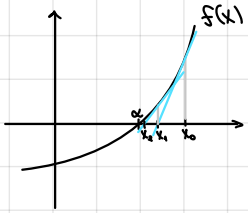
Se $m = f'(\alpha)$, $g'(\alpha)=0$, la convergenza locale è superlineare

Metodo delle tangenti (o di Newton)

$g(x) = x - \frac{f(x)}{f'(x)}$ con $f'(x) \neq 0$ (o $f'(x)\neq 0$ in un intorno di α)

Suppongo $f \in C^2([a,b])$

$\Rightarrow g(x) \in C^1([a,b])$



$g'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2}$ $\quad \because g'(\alpha)=0$

$\Rightarrow$ convergenza locale superlineare

Teorema Sia $f \in C^2([a,b])$, $\alpha \in (a,b)$, $f(\alpha)=0$ e $f'(\alpha)\neq 0$

Allora $\exists I = [\alpha-\rho, \alpha+\rho] \subset [a,b]$ tale che

$\forall x_0 \in I$ $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$ converge a α.

Inoltre, se $f''(\alpha)\neq 0$, la convergenza è superlineare di ordine 2,

altrimenti superlineare di ordine almeno 2

DIMOSTRAZIONE

La convergenza locale superlineare è già stata dimostrata.

Considero l'ordine di convergenza:

$$\lim_{k\to+\infty} \frac{|x_{k+1}-\alpha|}{|x_k-\alpha|^2}$$

$$x_{k+1} - \alpha = x_k - \frac{f(x_k)}{f'(x_k)} - \alpha$$

$$0 = f(\alpha) = f(x_k) + f'(x_k)(\alpha - x_k) + f''(\xi_k)\frac{(\alpha-x_k)^2}{2} \quad \text{con } |\xi_k - x_k| < |\alpha - x_k|$$

$$-\frac{f(x_k)}{f'(x_k)} = (\alpha - x_k) + \frac{f''(\xi_k)}{f'(x_k)}\frac{(\alpha-x_k)^2}{2}$$

$$\Rightarrow x_{k+1} - \alpha = x_k + \alpha - x_k + \frac{f''(\xi_k)}{f'(x_k)}\frac{(\alpha-x_k)^2}{2} - \alpha$$

$$\Rightarrow \frac{|x_{k+1}-\alpha|}{|x_k-\alpha|^2} = \left|\frac{f''(\xi_k)}{f'(x_k)} \cdot \frac{1}{2}\right| \xrightarrow[k\to+\infty]{} \left|\frac{f''(\alpha)}{2f'(\alpha)}\right|$$

Quindi $\lim_{k\to+\infty} \frac{|x_{k+1}-\alpha|}{|x_k-\alpha|^2} = \begin{cases} \neq 0 & \text{se } f''(\alpha)\neq 0 \quad \text{(convergenza quadratica)} \\ 0 & \text{se } f''(\alpha)=0 \quad \text{(convergenza almeno di ordine 2)} \end{cases}$ □

Teorema Sia $f \in C^p([a,b])$, $p \geq 2$, $\alpha \in (a,b)$, $f(\alpha)=0$,
$f'(\alpha)=f''(\alpha)=\dots=f^{(p-1)}(\alpha)=0$ e $f^{(p)}(\alpha)\neq 0$
Allora $\exists I=[\alpha-\rho,\alpha+\rho]\subset[a,b]$ tale che $\forall x\in I\setminus\{\alpha\}$ $f'(x)\neq 0$
e $\forall x_0\in I$ $x_{k+1}=x_k-\frac{f(x_k)}{f'(x_k)}$ converge a α
Inoltre la convergenza è lineare con fattore $\gamma=1-\frac{1}{p}$

DIMOSTRAZIONE
$f'(x)\neq 0$ in un intorno di α, $x\neq\alpha$
$f'(x)=f'(\alpha)+f''(\alpha)(x-\alpha)+\dots+f^{(p-1)}(\alpha)\frac{(x-\alpha)^{p-2}}{(p-2)!}+f^{(p)}(\xi)\frac{(x-\alpha)^{p-1}}{(p-1)!}$ con $|\xi-\alpha|<|x-\alpha|$
$f^{(p)}(\alpha)\neq 0 \Rightarrow \exists U$ intorno di α : $f^{(p)}(x)\neq 0\ \forall x\in U$
$\Rightarrow f'(x)\neq 0\ \forall x\in U\setminus\{\alpha\}$

Ora definisco $g(x)=\begin{cases}\alpha & \text{se } x=\alpha\\ x-\frac{f(x)}{f'(x)} & \text{se } x\in U\setminus\{\alpha\}\end{cases}$

Dimostro che $g\in C^1(U)$ e $g'(\alpha)=1-\frac{1}{p}$, per cui posso applicare il teorema del punto fisso: $\exists I: \forall x_0\in I$ $x_{k+1}=g(x_k)\longrightarrow\alpha$ convergenza lineare con fattore $1-\frac{1}{p}$

(1) $g\in C(U)$:
$\lim_{x\to\alpha} x-\frac{f(x)}{f'(x)} \overset{?}{=} \alpha \longleftarrow \lim_{x\to\alpha}\frac{f(x)}{f'(x)}\overset{H}{=}\dots\overset{H}{=}\lim_{x\to\alpha}\frac{f^{(p-1)}(x)}{f^{(p)}(x)}=0$

(2) $g\in C^1(U)$
Mostro che esiste $\lim_{h\to 0}\frac{g(\alpha+h)-g(\alpha)}{h}=\lim_{h\to 0}\frac{\alpha+h-\frac{f(\alpha+h)}{f'(\alpha+h)}-\alpha}{h}=1-\lim_{h\to 0}\frac{f(\alpha+h)}{hf'(\alpha+h)}$
Ora:
$f(\alpha+h)=f(\alpha)+hf'(\alpha)+\dots+\frac{h^{p-1}}{(p-1)!}f^{(p-1)}(\alpha)+\frac{h^p}{p!}f^{(p)}(\xi)=\frac{h^p}{p!}f^{(p)}(\xi)$ con $|\xi-\alpha|<|h|$
$f'(\alpha+h)=f'(\alpha)+hf''(\alpha)+\dots+\frac{h^{p-2}}{(p-2)!}f^{(p-1)}(\alpha)+\frac{h^{p-1}}{(p-1)!}f^{(p)}(\eta)=\frac{h^{p-1}}{(p-1)!}f^{(p)}(\eta)$ con $|\eta-\alpha|<|h|$
Quindi:
$\lim_{h\to 0}\frac{f(\alpha+h)}{hf'(\alpha+h)}=\lim_{h\to 0}\frac{\frac{h^p}{p!}f^{(p)}(\xi)}{h\cdot\frac{h^{p-1}}{(p-1)!}f^{(p)}(\eta)}=\frac{1}{p}\lim_{h\to 0}\frac{f^{(p)}(\xi)}{f^{(p)}(\eta)}=\frac{1}{p}\frac{f^{(p)}(\alpha)}{f^{(p)}(\alpha)}=\frac{1}{p}$

Quindi g è derivabile in α e vale $g'(\alpha)=1-\frac{1}{p}$
Infine, $g'(x)$ è continua in α.
Infatti, per $x\neq\alpha$, vale $g'(x)=\frac{f''(x)f(x)}{(f'(x))^2}$ e
$f''(\alpha+h)=f''(\alpha)+hf'''(\alpha)+\dots+\frac{h^{p-3}}{(p-3)!}f^{(p-1)}(\alpha)+\frac{h^{p-2}}{(p-2)!}f^{(p)}(\mu)$ con $|\mu-\alpha|<|h|$
Quindi:
$\lim_{x\to\alpha}g'(x)=\lim_{h\to 0}\frac{f''(\alpha+h)f(\alpha+h)}{(f'(\alpha+h))^2}=\lim_{h\to 0}\frac{\frac{h^{p-2}}{(p-2)!}f^{(p)}(\mu)\frac{h^p}{p!}f^{(p)}(\xi)}{\left(\frac{h^{p-1}}{(p-1)!}f^{(p)}(\eta)\right)^2}=\frac{p-1}{p}=1-\frac{1}{p}$ □

Teorema Sia $f \in C^2([\alpha, \alpha+\rho])$, $f(\alpha)=0$, $\rho>0$
t.c. $f'(x)\cdot f''(x)>0 \quad \forall x \in I=[\alpha, \alpha+\rho]$
Allora $\forall x_0 \in I$, $x_{k+1}=x_k-\frac{f(x_k)}{f'(x_k)}$ converge ad α in modo monotono e superlineare di ordine 2.

DIMOSTRAZIONE

OSS $f'(\alpha)\neq 0$, $f''(\alpha)\neq 0 \Rightarrow$ se c'è convergenza, la convergenza è superlineare di ordine (almeno) 2

Supponiamo sia $f'(x)>0$, dunque $f''(x)>0 \ \forall x \in I$

Sia $x_0>\alpha$

Se $x_k>\alpha$, dimostro che $x_{k+1}<x_k$

$x_{k+1}=x_k-\frac{f(x_k)}{f'(x_k)}$ ($f(x_k)>0$, $f'(x_k)>0$) — perché f è crescente: $x_k>\alpha \Rightarrow f(x_k)>f(\alpha)=0$

$\Rightarrow x_{k+1}<x_k$

Dimostro che $x_{k+1}>\alpha$

$x_{k+1}-\alpha = g(x_k)-g(\alpha) = g'(\xi_k)(x_k-\alpha)$ con $|\xi_k-\alpha|<|x_k-\alpha|$ ($x_k-\alpha \to 0$)

con $g(x)=x-\frac{f(x)}{f'(x)} \Rightarrow g'(x)=\frac{f(x)f''(x)}{f'(x)^2}>0$

$\Rightarrow x_{k+1}-\alpha>0$

□

Applicazioni

Dato $a \in \mathbb{R}$: calcolare a^{-1}

a^{-1} è zero di $f(x)=a-\frac{1}{x}$

Metodo di Newton applicato a f: $x_{k+1}=2x_k-ax_k^2$

Dato $a \in \mathbb{R}$, $a \in [1/2, 1]$: calcolare $a^{\frac{1}{2}}$

$a^{1/2}$ è zero di $f(x)=x^2-a$

Metodo di Newton applicato a f: $x_{k+1}=\frac{x_k^2+a}{2x_k}$

Invece se $f(x)=x^{-2}-a^{-1}$: $x_{k+1}=\frac{1}{3}(3x_k-bx_k^3)$ dove $b=a^{-1}$

Zeri di polinomi

$p(x)=\prod_{i=1}^{n}(x-\lambda_i)$

Scelgo x_0. $\{x_k\}$ con Newton. Sia $\alpha=\lim_{k\to+\infty} x_k$ e $\tilde{\alpha}$ la sua approssimazione.

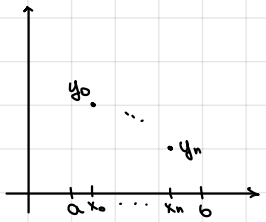
Applico Newton a $\frac{p(x)}{x-\tilde{\alpha}}$

Metodo di Aberth (Ehrlich Aberth)

Date approssimazioni $x_1, \dots, x_{n-1}$ di n zeri di $p(x)$

Il metodo di Aberth è il metodo di Newton applicato a $\frac{p(x)}{\prod_{i=1}^{n-1}(x-x_i)}$

INTERPOLAZIONE



$x_i \in [a,b]$, $i=0,..,n$ $\quad x_i \neq x_j$ per $i \neq j$ (Nodi)
Siano y_i, $i=0,\dots,n$

Problema: determinare $g(x). [a,b] \longrightarrow \mathbb{R}$ t.c. $g(x_i)=y_i$, $i=0,..,n$

Interpolazione lineare

Date $\{\varphi_i(x)\}_{i=0,\dots,n}$, $\varphi_i : [a,b] \longrightarrow \mathbb{R}$ linearmente indipendenti,
determinare $a_0,\dots,a_n$ tali che $g(x) = \sum_{i=0}^{n} a_i \varphi_i(x)$ soddisfa $g(x_i)=y_i$, $i=0,\dots,n$

$$\begin{pmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \cdots & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \cdots & \varphi_n(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_0(x_n) & \varphi_1(x_n) & \cdots & \varphi_n(x_n) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Interpolazione polinomiale

Se $\varphi_i(x) = x^i$, si parla di interpolazione polinomiale

$$V_n = \begin{pmatrix} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^n \end{pmatrix}$$ matrice di Vandermonde

$\det V_n = \prod_{0 \le j < i \le n} (x_i - x_j)$ (si dimostra per induzione)

$x_i \neq x_j \Longrightarrow \det V_n \neq 0 \Longrightarrow$ esistono unici $a_0,\dots,a_n$ t.c.
$p(x) = \sum_{j=0}^{n} a_j x^j$ verifica $p(x_i)=y_i$ $i=0,\dots,n$

Polinomi di Lagrange

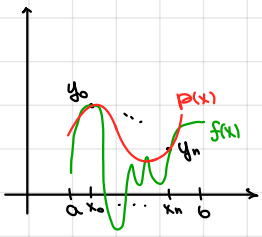
Dati $x_0,\dots,x_n$, $x_i \neq x_j$

$$L_i(x) = \frac{\prod_{j \neq i} (x-x_j)}{\prod_{j \neq i} (x_i - x_j)} = \frac{\pi_n(x)}{x-x_i} \cdot \alpha_i \qquad \text{dove } \pi_n(x) = \prod_{j=0}^{n} (x-x_j),\ \alpha_i = \frac{1}{\prod_{j \neq i} (x_i - x_j)}$$

Osserviamo che $L_i(x_k) = \begin{cases} 0 & \text{se } i \neq k \\ 1 & \text{se } i = k \end{cases}$

Scelgo $\varphi_i(x) = L_i(x)$: la matrice è l'identità
Dunque $p(x) = \sum_{i=0}^{n} y_i L_i(x) = \pi_n(x) \sum_{i=0}^{n} y_i \alpha_i \frac{1}{x-x_i}$

Resto dell'interpolazione polinomiale



$x_i \in [a,b]$, $i=0,\dots,n$ $\quad x_i \neq x_j$ per $i \neq j$ (Nodi)
Siano y_i, $i=0,\dots,n$
$y_i = f(x_i)$ $\forall i$, $f:[a,b] \longrightarrow \mathbb{R}$

Sia $p(x)$, con $\deg p \leq n$, il polinomio di interpolazione: $p(x_i) = y_i$ $\forall i$

Problema: stimare $r(x) = f(x) - p(x)$, $x \in [a,b]$ (resto dell'interpolazione)
$r(x_i) = 0$ $\forall i = 0,\dots,n$

Teorema Siano $x_0,\dots,x_n$ i nodi, $x_i \neq x_j$ se $i \neq j$, $x_i \in [a,b]$,
$f \in C^{n+1}([a,b])$, $p(x)$ che interpola $f(x)$ in x_i
Allora $\forall x \in [a,b]$ $\exists\, \xi(x) \in (a,b)$:

$$r(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \cdot \prod_{i=0}^{n} (x - x_i)$$

DIMOSTRAZIONE
Se $x = x_k$, $r(x) = 0$
Sia $x \in [a,b]$, $x \neq x_i$, $i = 0,\dots,n$

$$g(y) = r(y) - \prod_{i=0}^{n} (y - x_i) \frac{r(x)}{\prod_{i=0}^{n} (x - x_i)} \quad : \quad g:[a,b] \longrightarrow \mathbb{R}$$

$$g(x) = r(x) - \prod_{i=0}^{n} (x - x_i) \frac{r(x)}{\prod_{i=0}^{n} (x - x_i)} = 0$$

$$g(x_k) = r(x_k) - \prod_{i=0}^{n} (x_k - x_i) \frac{r(x)}{\prod_{i=0}^{n} (x - x_i)} = 0$$

Dunque $g(y)$ ha almeno $n+2$ zeri in $[a,b]$
$g \in C^{n+1}([a,b])$ perché $r \in C^{n+1}([a,b])$
g' ha almeno $n+1$ zeri in (a,b), g'' ha almeno n zeri in (a,b), …,
$g^{(n+1)}$ ha almeno 1 zero in (a,b) : $\xi(x) \in (a,b)$

$$g^{(n+1)}(y) = r^{(n+1)}(y) - \frac{r(x)}{\prod_{i=0}^{n} (x - x_i)} D^{(n+1)}\left(\prod_{i=0}^{n} (y - x_i)\right) = f^{(n+1)}(y) - \underbrace{p^{(n+1)}(y)}_{\text{perché } \deg p \leq n} - \frac{r(x)}{\pi(x - x_i)} (n+1)!$$

Dunque

$$0 = g^{(n+1)}(\xi(x)) = f^{(n+1)}(\xi(x)) - \frac{r(x)}{\prod_{i=0}^{n} (x - x_i)} (n+1)!$$

□

Oss Se $|f^{(n+1)}(x)| \leq k$ $\forall x \in [a,b]$, allora

$$|r(x)| \leq \frac{k}{(n+1)!} \prod_{i=0}^{n} (x - x_i)$$

esempio Runge function

Interpolazione delle radici n-esime dell'unità

$n \geq 1$, $\omega_n = \cos\frac{2\pi}{n} + i \sin\frac{2\pi}{n}$ (radice n-esima primitiva dell'unità)

$\{\omega_n^j, j = 0, \dots, n-1\}$ sono le n radici n-esime dell'unità

Dato $y = \begin{pmatrix} y_0 \\ \vdots \\ y_{n-1} \end{pmatrix} \in \mathbb{C}^n$, cerco $z = \begin{pmatrix} z_0 \\ \vdots \\ z_{n-1} \end{pmatrix} \in \mathbb{C}^n$:

$p(x) = \sum_{j=0}^{n-1} z_j x^j$, $p(\omega_n^k) = y_k$, $k = 0, \dots, n-1$

z risolve il sistema $Vz = y$, dove V è la matrice di Vandermonde definita dai nodi

$\Omega_n = (\omega_n^{hk})_{h,k=0,\dots,n-1}$ matrice di Fourier

Lemma $\sum_{j=0}^{n-1} \omega_n^{kj} = \begin{cases} n & \text{se } k \equiv 0 \bmod n \\ 0 & \text{altrimenti} \end{cases}$

DIMOSTRAZIONE

$1 - X^n = (1-X)(1 + X + X^2 + \dots + X^{n-1})$

Se $X = \omega_n^k$ con $k \not\equiv 0 \bmod n$:

$0 = (1 - \underbrace{\omega_n^k}_{\neq 0}) \sum_{j=0}^{n-1} \omega_n^{kj} \implies \sum_{j=0}^{n-1} \omega_n^{kj} = 0$ □

Teorema (1) Ω_n è simmetrica

(2) $\Omega_n^H \Omega_n = n I_n$

(3) $\Omega_n^2 = nP$ dove $P = \left(\begin{array}{c|cc} 1 & 0 & \\ \hline 0 & 0 & 1 \\ & 1 & 0 \end{array}\right)$

DIMOSTRAZIONE

(1) $\Omega_n = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & \omega_n & \dots & \omega_n^{n-1} \\ 1 & \omega_n^2 & \dots & \omega_n^{n-2} \\ \vdots & \vdots & \dots & \vdots \\ 1 & \omega_n^{n-1} & \dots & \omega_n \end{pmatrix}$

$(\Omega_n)_{hk} = \omega_n^{hk} = (\Omega_n)_{kh} \implies \Omega_n^T = \Omega_n$

(2) $(\Omega_n^H \Omega_n)_{rs} = (\overline{\Omega}_n \Omega_n)_{rs} = \sum_{k=0}^{n-1} \overline{\omega_n^{rk}}\, \omega_n^{ks} = \sum_{k=0}^{n-1} \omega_n^{k(s-r)} = \begin{cases} n & \text{se } s-r \equiv 0 \bmod n \\ 0 & \text{altrimenti} \end{cases}$

$\implies \Omega_n^H \Omega_n = nI$

(3) $(\Omega_n^2)_{rs} = \sum_{k=0}^{n-1} \omega_n^{rk} \omega_n^{ks} = \sum_{k=0}^{n-1} \omega_n^{k(r+s)} = \begin{cases} n & \text{se } r+s \equiv 0 \bmod n \\ 0 & \text{altrimenti} \end{cases}$

□

Problema (interpolazione): risolvere $\Omega_n z = y$

$z = \frac{1}{n}\Omega_n^H y$: $z = DFT(y)$ trasformata discreta di Fourier

Problema (valutazione): dato $p(x) = \sum_{j=0}^{n-1} z_j x^j$,
calcolare $y = \Omega_n z$, cioè $y_k = p(\omega_n^k)$
$y = IDFT(z)$ trasformata inversa discreta di Fourier

Condizionamento del problema
Il condizionamento del problema della soluzione di $\Omega_n z = y$
si misura con $K(\Omega_n) = \|\Omega_n\| \, \|\Omega_n^{-1}\|$

$\Omega_n^H \Omega_n = n I \longrightarrow \left(\frac{\Omega_n}{\sqrt{n}}\right)^H \cdot \frac{\Omega_n}{\sqrt{n}} = I$: $F_n = \frac{\Omega_n}{\sqrt{n}}$ è unitaria

$\|F_n\|_2 = (\rho(F_n^H F))^{1/2} = 1$ e $\|F_n^{-1}\|_2 = 1$
$\Rightarrow K_2(F_n) = 1$
Quindi, poiché $K(\Omega_n) = K(F_n)$, $K_2(\Omega_n) = 1$

Algoritmo FFT (Fast Fourier Transform)
per calcolare $y = \Omega_n z$, dove $n = 2^q$

$$y_h = \sum_{k=0}^{n-1} \omega_n^{hk} z_k \qquad h = 0, \dots, n-1$$

$$= \sum_{k=0}^{\frac{n}{2}-1} \omega_n^{2hk} z_{2k} + \sum_{k=0}^{\frac{n}{2}-1} \omega_n^{(2k+1)h} z_{2k+1} = \sum_{k=0}^{\frac{n}{2}-1} \omega_{n/2}^{hk} z_{2k} + \omega_n^h \sum_{k=0}^{\frac{n}{2}-1} \omega_{n/2}^{hk} z_{2k+1}$$

$$z_{pari} = \begin{pmatrix} z_0 \\ z_2 \\ \vdots \\ z_{n-2} \end{pmatrix}, \quad z_{disp} = \begin{pmatrix} z_1 \\ z_3 \\ \vdots \\ z_{n-1} \end{pmatrix}$$

Se $h = 0, \dots, \frac{n}{2}-1$, $y_h = (\Omega_{n/2} z_{pari})_h + \omega_n^h (\Omega_{n/2} z_{disp})_h$

Se $l = 0, \dots, \frac{n}{2}-1$, $y_{l+\frac{n}{2}} = \sum_{k=0}^{\frac{n}{2}-1} \omega_{n/2}^{(l+\frac{n}{2})k} z_{2k} + \omega_n^{l+\frac{n}{2}} \sum_{k=0}^{\frac{n}{2}-1} \omega_{n/2}^{(l+\frac{n}{2})k} z_{2k+1}$

$\omega_n^{l+\frac{n}{2}} = -\omega_n^l$

$$\begin{pmatrix} y_0 \\ \vdots \\ y_{\frac{n}{2}-1} \\ y_{\frac{n}{2}} \\ \vdots \\ y_{n-1} \end{pmatrix} = \begin{pmatrix} \Omega_{n/2} z_{pari} \\ \Omega_{n/2} z_{pari} \end{pmatrix} + \begin{pmatrix} \begin{pmatrix} 1 & & \\ & \ddots & \\ & & \omega_n^{\frac{n}{2}-1} \end{pmatrix} \Omega_{n/2} z_{disp} \\ -\begin{pmatrix} 1 & & \\ & \ddots & \\ & & \omega_n^{\frac{n}{2}-1} \end{pmatrix} \Omega_{n/2} z_{disp} \end{pmatrix}$$

Sia C_n: costo di DFT di ordine n
$C_n = 2C_{n/2} + \frac{n}{2}$ Molt $+ n$ Add $= 2C_{n/2} + \frac{3}{2}n$ op
$C_1 = 0$
Per induzione si dimostra $C_n = \frac{3}{2} n \log_2 n$ $(n = 2^q)$

Applicazione di FFT

Siano $a(t), b(t)$ polinomi a coefficienti in $\mathbb{C}$.

Calcolare $c(t) = a(t)\,b(t)$

Se $a(t) = \sum_{i=0}^{m} a_i t^i$ e $b(t) = \sum_{i=0}^{n} b_i t^i$

allora $c(t) = \sum_{i=0}^{n+m} c_i t^i$ con $c_0 = a_0 b_0$, $c_1 = a_0 b_1 + a_1 b_0, \dots$

Il calcolo dei coefficienti ha un costo $O((m+n)^2)$

Idea: sia $n = 2^q$, $n > \deg c(t)$

(1) Calcolo $a(t)$ e $b(t)$ in ω_n^j per $j = 0, \dots, n-1$: $y_j = a(\omega_n^j)$, $z_j = b(\omega_n^j)$ (2 DFT$_n$ $\cdot$ $2 \cdot \frac{3}{2} n \log_2 n$)

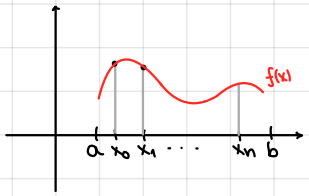
(2) Calcolo $t_j = y_j z_j$ per $j = 0, \dots, n-1$ (n op)

(3) Calcolo i coefficienti del polinomio $c(t)$ t.c. $c(\omega_n^j) = t_j$ per $j = 0, \dots, n-1$ (DFT: $\frac{3}{2} n \log_2 n$)

Costo totale: $O(n \log_2 n)$

INTEGRAZIONE APPROSSIMATA

Sia $f \in C([a,b])$
$S[f] = \int_a^b f(x)\,dx$



Dati $x_0, \dots, x_n \in [a,b]$, $x_i \neq x_j$ per $i \neq j$
$S[f] \simeq S_{n+1}[f] = \sum_{i=0}^{n} w_i f(x_i)$, dove $w_0, \dots, w_n$ sono opportuni pesi
Il resto della formula è
$r_{n+1} = S[f] - S_{n+1}[f]$

def La formula di integrazione approssimata ha grado di precisione $k \geq 0$ se
$r_{n+1} = 0$ se $f(x) = x^j$ per $j = 0, \dots, k$ e $r_{n+1} \neq 0$ se $f(x) = x^{k+1}$

Formule di integrazione interpolatorie
Dati $x_0, \dots, x_n \in [a,b]$, $x_i \neq x_j$ se $i \neq j$, e $f \in C([a,b])$,
sia $p(x)$ il polinomio di grado $\leq n$: $p(x_i) = f(x_i)$ per $i = 0, \dots, n$
$\int_a^b f(x)\,dx$ è approssimato con $S_{n+1}[f] = \int_a^b p(x)\,dx$
$p(x) = \sum_{i=0}^{n} L_i(x) f(x_i)$ dove $L_i(x)$ è l'i-esimo polinomio di lagrange
Perciò
$S_{n+1}[f] = \int_a^b p(x)\,dx = \sum_{i=0}^{n} f(x_i) \underbrace{\int_a^b L_i(x)\,dx}_{w_i}$

In una formula interpolatoria, $w_i = \int_a^b L_i(x)\,dx$

Grado di precisione di formula interpolatoria.
Sia $f(x) = x^j$, per quali j $S_{n+1}[x^j] = S[x^j]$?
Se $j \leq n$, $p(x)$ (che interpola $f(x) = x^j$ in $(x_0, \dots, x_n)$ coincide con f
Quindi le formule interpolatorie hanno grado di precisione almeno n
Vale anche il viceversa: se $S_{n+1}[f]$ ha grado di precisione almeno n,
allora è interpolatoria.
$S_{n+1}[f] = \sum_{i=0}^{n} w_i f(x_i)$ abbia grado di precisione almeno n
Sia $f(x) = L_k(x)$, k-esimo polinomio di lagrange (grado n)
$S_{n+1}[L_k(x)] = S[L_k]$
$\sum_{i=0}^{n} w_i L_k(x_i) = \int_a^b L_k(x)\,dx$
$w_k = \int_a^b L_k(x)\,dx$ dunque la formula è interpolatoria

esempio $\int_{-1}^{1} f(x)\,dx$
Formula interpolatoria con $n=1$, con massimo grado di precisione
Cercare i nodi x_0, x_1 e i pesi w_0, w_1 che danno il massimo grado di precisione
Imponendo le condizioni $\int_{-1}^{1} x^i\,dx = w_0 x_0^i + w_1 x_1^i$ per $i = 0, \dots, 3$
$x_0 = -x_1 = \frac{1}{\sqrt{3}}$, $w_0 = w_1 = 1$
ha grado di precisione 3 $(2n+1)$

Se $f \in C^{n+1}([a,b])$: $f(x) - p(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \prod_{j=0}^{n} (x - x_j)$

$$\int_a^b f(x)\,dx - \int_a^b p(x)\,dx = \frac{1}{(n+1)!} \int_a^b f^{(n+1)}(\xi(x)) \prod_{j=0}^{n} (x - x_j)\,dx$$

$$|S[f] - S_{n+1}[f]| = \frac{1}{(n+1)!} \left| \int_a^b f^{(n+1)}(\xi(x)) \prod_{j=0}^{n} (x - x_j)\,dx \right| \leq \frac{M}{(n+1)!} \int_a^b \left| \prod_{j=0}^{n} (x - x_j) \right| dx$$

se $|f^{(n+1)}(x)| \leq M$

Formule di Newton-Cotes

I nodi sono scelti equispaziati: $h = \frac{b-a}{n}$

$x_i = a + ih$ per $i = 0, \dots, n$

w_i : espressioni razionali

Se $n = 1$: $x_0 = a$, $x_1 = b$

$w_1 = w_0 = \int_a^b L_0(x)\,dx = \frac{h}{2}$

$S_2[f] = \frac{h}{2}(f(x_0) + f(x_1))$

$\int_a^b f(x)\,dx - S_2[f] = -\frac{1}{12} h^3 f''(\xi)$ con $\xi \in (a,b)$

Se $n = 2$:

$w_0 = w_2 = \frac{h}{3}$, $w_1 = \frac{4}{3} h$

$S[f] - S_3[f] = -\frac{1}{90} h^5 f^{(4)}(\eta)$

In generale, per $n \to +\infty$, l'approssimazione non converge.

Formule di Newton-Cotes composte

Siano $z_0, \dots, z_N$, dove $z_j = a + jh$, $h = \frac{b-a}{N}$

$$\int_a^b f(x)\,dx = \sum_{i=0}^{N-1} \int_{z_i}^{z_{i+1}} f(x)\,dx$$

e ciascun $\int_{z_i}^{z_{i+1}} f(x)\,dx$ è approssimato con le formule NC con $n = 1$ o 2

N arbitrario, n = 1

Formula dei trapezi

$$J_2^{(N)}[f] = \sum_{i=0}^{N-1} S_2^{(i)}[f] = \frac{b-a}{2N} \sum_{i=0}^{N-1} (f(z_i) + f(z_{i+1})) = \frac{b-a}{2N} \left(f(z_0) + f(z_N) + 2 \sum_{i=1}^{N-1} f(z_i) \right)$$

$$E_n = S[f] - J_2^{(N)}[f] = \sum_{i=0}^{N-1} -\frac{1}{12} h^3 f''(\xi_i) = -\frac{(b-a)h^2}{12} f''(\xi)$$

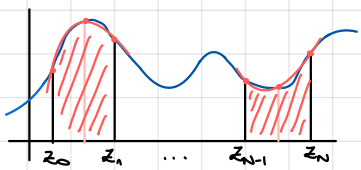
$\Rightarrow S[f] - J_2^{(N)}[f] \xrightarrow[N \to +\infty]{} 0$ come h^2, dove $h = \frac{b-a}{N}$

Se $|f''| < K$, $|E_n| < \varepsilon$ se $\left| \frac{b-a}{12} h^2 K \right| < \varepsilon$ (errore analitico)

N arbitrario, n = 2

Formula di Cavalieri-Simpson

$$S_3^{(i)}[f] = \frac{1}{2} \left(\frac{b-a}{2N} \right) \left(f(z_i) + 4 f\left(\frac{z_i + z_{i+1}}{2} \right) + f(z_{i+1}) \right)$$



Dalla formula del resto dell'interpolazione polinomiale ($h = \frac{b-a}{N}$)

$$\int_{z_i}^{z_{i+1}} f(x)\,dx - S_3^{(i)}[f] = \int_{z_i}^{z_{i+1}} (f(x) - p(x))\,dx = -\frac{1}{90} \left(\frac{h}{2} \right)^5 f^{(4)}(\xi_i) \quad \text{con } \xi_i \in (z_i, z_{i+1})$$

Errore totale:

$$S[f] - \sum_{i=0}^{N-1} S_3^{(i)}[f] = -\frac{1}{90} \left(\frac{h}{2} \right)^5 \underbrace{\sum_{i=0}^{N-1} f^{(4)}(\xi_i)}_{N f^{(4)}(\eta)} \overset{\eta \in (a,b)}{=} -\frac{(b-a)h^4}{2280} f^{(4)}(\eta)$$

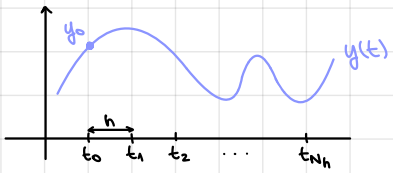
PROBLEMA DI CAUCHY

Sia $I \subset \mathbb{R}$, $f(t,y) : I \times \mathbb{R} \longrightarrow \mathbb{R}$ continua,
$t_0 \in I$, $y_0 \in \mathbb{R}$

Problema: trovare $y \in C^1(I)$ tale che $\begin{cases} y'(t) = f(t, y(t)) & \forall t \in I \\ y(t_0) = y_0 \end{cases}$

Ipotesi: $f(t,y)$ lipschitziana rispetto a y
$\exists L : |f(t,y_1) - f(t,y_2)| \leq L|y_1 - y_2| \quad \forall t \in I, \forall y_1, y_2 \in \mathbb{R}$
$\Longrightarrow \exists\, y(t)$ unica soluzione di (PC)

Caso $I = [t_0, t_0 + T]$, $T > 0$



$t_n = t_0 + nh$, $h > 0$, $n = 0, \dots, N_h$, dove
N_h è il più grande intero t.c. $t_{N_h} \leq t_0 + T$

Notazione: $y_n = y(t_n)$
u_n : approssimazione di $y(t_n)$
$f_n = f(t_n, u_n)$

Metodi a un passo

$$\begin{cases} u_{n+1} = u_n + h\,\Phi(t_n, u_n, f_n, h) & n = 0, 1, \dots, N_h - 1 \\ u_0 = y_0 \end{cases}$$

Definiamo

$$\varepsilon_{n+1} = y_{n+1} - (y_n + h\,\Phi(t_n, y_n, f(t_n, y_n), h)$$

Errore locale di troncamento (LTE): $\tau_{n+1}(h) = \dfrac{\varepsilon_{n+1}}{h}$ nel nodo $n+1$-esimo
Errore globale di troncamento: $\tau(h) = \max\limits_{n=0,\dots,N_h-1} |\tau_{n+1}(h)|$

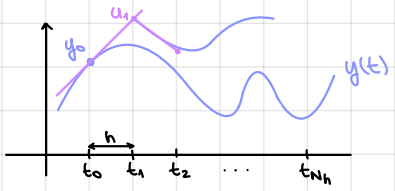
Def. Il metodo è consistente se $\lim\limits_{h \to 0} \tau(h) = 0$
Il metodo è consistente di ordine p, $p \geq 1$, se $\tau(h) = O(h^p)$

Errore globale: $e_n = y_n - u_n$

Def. Il metodo è convergente se $\exists C(h) : \lim\limits_{h \to 0} C(h) = 0$
e $|y_n - u_n| \leq C(h)$ per $n = 0, \dots, N_h$
Il metodo è convergente di ordine p se $C(h) = O(h^p)$, ovvero
$\exists k > 0 : C(h) \leq k h^p \quad \forall h > 0$

Metodo di Eulero (esplicito)

$$(EE)\quad \begin{cases} u_{n+1} = u_n + h f_n \\ u_0 = y_0 \end{cases} \qquad n = 0, 1, \dots, N_h - 1$$



Per il metodo di Eulero (esplicito) ($y \in C^2(I)$):

$$\varepsilon_{n+1} = y_{n+1} - \underbrace{(y_n + h f(t_n, y_n))}_{\text{un passo del metodo di E. a partire da } u_n = y_n}$$

$$\varepsilon_{n+1} = y_{n+1} - y_n - h f(t_n, y_n) = h y'(t_n) + O(h^2) - h f(t_n, y_n) = O(h^2)$$

$$y_{n+1} = y_n + h y'(t_n) + \frac{h^2}{2} y''(\xi_n)$$

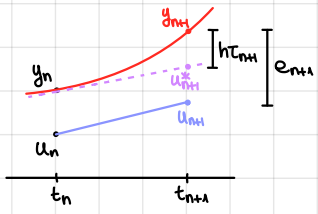
$$\Rightarrow \varepsilon_{n+1} = \frac{h^2}{2} f''(\xi_n)$$

$$\tau_{n+1}(h) = \frac{h}{2} y''(\xi_n) \quad \Rightarrow \quad \tau(h) = O(h)$$

Quindi il metodo di Eulero (esplicito) è **consistente di ordine 1**.

$$y(t_{n+1}) = y(t_n) + h f(t_n, y_n) + h\tau_{n+1} = y(t_n) + h(f(t_n, y_n) + \tau_{n+1})$$

Abbiamo $e_n = y_n - u_n$ e $\varepsilon_{n+1} = h\tau_{n+1}$.
Detto $u^*_{n+1} = y_n + h f(t_n, y_n)$, vale



$$e_{n+1} = y_{n+1} - u_{n+1} = (y_{n+1} - u^*_{n+1}) + (u^*_{n+1} - u_{n+1})$$

dove $\quad y_{n+1} - u^*_{n+1} = \varepsilon_{n+1} = h\tau_{n+1}$

$$u^*_{n+1} - u_{n+1} = y_n - u_n + h f(t_n, y_n) - h f(t_n, u_n) =$$
$$= e_n + h(f(t_n, y_n) - f(t_n, u_n))$$

Poiché $f(\cdot, y)$ è L-Lipschitz, abbiamo:

$$|e_{n+1}| \le |h\tau_{n+1}| + |e_n| + |h| \cdot |f(t_n, y_n) - f(t_n, u_n)| \le Ch^2 + |e_n| + hL|e_n| = (1+hL)|e_n| + Ch^2$$

Procedendo per induzione su $\begin{cases} |e_0| = 0 \\ |e_{n+1}| \le (1+hL)|e_n| + Ch^2 \end{cases}$, otteniamo

$$|e_{n+1}| \le (1+hL)|e_n| + Ch^2 \le (1+hL)^2 |e_{n-1}| + (1+hL)Ch^2 + Ch^2 \le \dots$$

$$\Rightarrow |e_{n+1}| \le Ch^2 \sum_{i=0}^{n} (1+hL)^i = Ch^2 \frac{(1+hL)^{n+1} - 1}{1+hL-1} = \frac{Ch}{L}((1+hL)^{n+1} - 1) \le \frac{Ch}{L}((e^{hL})^{n+1} - 1) = \frac{Ch}{L}(e^{L(t_{n+1}-t_0)} - 1)$$

($1+hL \le e^{hL}$; $h(n+1) = t_{n+1} - t_0$)

Quindi $|e_{n+1}| \le \frac{Ch}{L}(e^{L(t_{n+1}-t_0)} - 1) \sim O(h)$
ossia (EE) è **convergente di ordine 1**.

Metodo di Eulero (implicito)

$$(EI)\quad \begin{cases} u_{n+1} = u_n + h f(t_{n+1}, u_{n+1}) \\ u_0 = y_0 \end{cases}$$

(EI) è **consistente di ordine 1**
e **convergente di ordine 1**

In floating point, $\tau_n \sim Ch + \xi_n$, $|\xi_n| \sim u$ (precisione di macchina)
$\longrightarrow |y_n - u_n| \le Ch + \frac{1}{h}u$

Metodo del trapezio (Crank-Nicolson)

$$(TR)\begin{cases} u_{n+1} = u_n + \frac{h}{2}\left(f(t_n,u_n) + f(t_{n+1},u_{n+1})\right) \\ u_0 = y_0 \end{cases}$$

Il metodo è **consistente di ordine 2**
e **convergente di ordine 2**

Consideriamo il seguente problema (**Problema test di Dahlquist**)

$$\begin{cases} y' = \lambda y \qquad t\in[0,+\infty) \\ y(0)=1 \end{cases}$$

Oss molti problemi si riconducono a questo

(i) $\begin{cases} y' = f(t,y) = f_0 + \frac{\partial f}{\partial y}(t,y)\cdot y + \dots \\ y(t_0) = y_0 \end{cases}$

(ii) $\begin{cases} y' = Ay \qquad y: I \longrightarrow \mathbb{R}^n \\ y(0) = y_0 \end{cases}$

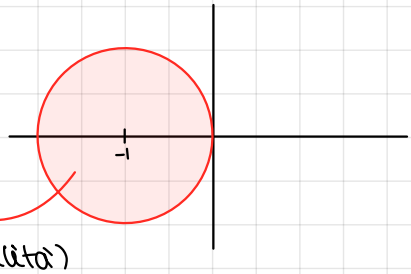
$$\Longrightarrow \begin{cases} (V^{-1}y)' = V^{-1}AVV^{-1}y \\ V^{-1}y = V^{-1}y_0 \end{cases} \quad V^{-1}AV = D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

$$z = V^{-1}y \iff \begin{cases} z' = Dz \\ z = z_0 \end{cases} \iff \begin{cases} (z)_1' = \lambda_1 (z)_1 \\ \vdots \\ (z)_n' = \lambda_n (z)_n \end{cases}$$

$|y(t)| = e^{Re(\lambda)t}$: $y(t)$ limitata per $t\longrightarrow+\infty \iff Re(\lambda)\le 0$

$$(EE)\begin{cases} u_{n+1} = u_n + hf(t_n,u_n) = u_n + h\lambda u_n = (1+h\lambda)u_n \\ u_0 = y_0 \end{cases} \Longrightarrow u_n = (1+h\lambda)^n$$

La soluzione è limitata se $|(1+h\lambda)| < 1 \iff h\lambda \in B(-1,1)$

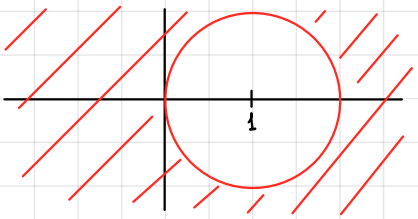


Se $Re(\lambda)\le 0$
Se $\lambda\in\mathbb{R}$, $\lambda\le 0$, $h\lambda\in[-2,0]$
cioè $h\le \frac{2}{|\lambda|}$

$A = \{z\in\mathbb{C} \mid \text{per } z=h\lambda, (u_n) \text{ è lim}\}$
(regione di assoluta stabilità)

$$(EI)\begin{cases} u_{n+1} = u_n + h\lambda u_{n+1} \longrightarrow u_{n+1} = \frac{u_n}{1-h\lambda} = \frac{1}{1-z}u_n \Longrightarrow u_n = \frac{1}{(1-z)^n} \\ u_0 = 1 \end{cases}$$

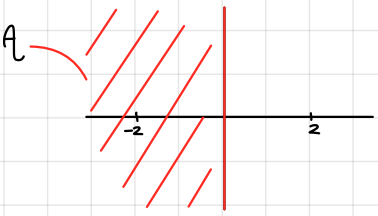
u_n limitata $\iff \left|\frac{1}{1-z}\right| \le 1 \iff |1-z| \ge 1$



$z = h\lambda \overset{?}{\in} A \longrightarrow$ Nessuna condizione
Il metodo è incondizionatamente stabile

$$(TR)\begin{cases} u_{n+1} = u_n + \frac{h}{2}(\lambda u_n + \lambda u_{n+1}) \longrightarrow u_{n+1} = \frac{1+\frac{h\lambda}{2}}{1-\frac{h\lambda}{2}}u_n = \frac{1+\frac{z}{2}}{1-\frac{z}{2}}u_n \Longrightarrow u_n = \left(\frac{1+\frac{z}{2}}{1-\frac{z}{2}}\right)^n \\ u_0 = y_0 \end{cases}$$

$$|u_n| = \left|\frac{1+\frac{z}{2}}{1-\frac{z}{2}}\right|^n = \left|\frac{2+z}{2-z}\right|^n$$



Per $Re(\lambda)\le 0 \Rightarrow u_n$ lim.
$Re(\lambda) > 0 \Rightarrow |u_n| \xrightarrow[n\to+\infty]{} +\infty$

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(\tau)\,y(\tau)\,d\tau$$

Oss (EE) $1+z$
(EI) $(1-z)^{-1}$
(TR) $\frac{1+\frac{z}{2}}{1-\frac{z}{2}}$

e^z è approssimata con lo stesso ordine del metodo
$u_{n+1} = R(z)\,u_n$