Analisi Numerica

Alessandro Sosso, Marco Ambrogio Bergamo

Anno 2023-2024

Indice

Preliminari Meschini

Definizione 0.1 : Spazio vettoriale, Norma

Teo. 1 (Disuguaglianza di Cauchy-Schwarz)**.** $\langle x, y\rangle = \left|x^T \cdot y\right| \leq \|x\|_2 \cdot \|y\|_2$

Definizione 0.2 (Norma matriciale)**:** $\|A \cdot B\| \leq \|A\| \cdot \|B\|$ e Norma matriciale indotta

$$\|A\| \coloneqq \sup_{x \in \mathbb{R}^n} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|$$

Definizione 0.3 : Spettro di una matrice $\Lambda(A)$, Raggio spettrale $\rho(A) = \max_{\lambda \in \Lambda(A)} |\lambda|$

Teo. 2 . Se A simmetrica $\implies \|A\|_2 = \rho(A)$

Prop. 1 . $\|A\|$ norma matriciale indotta $\implies \rho(A) \leq \|A\|$

Teo. 3 . $\rho(A) = \inf \|A\|$ per tutte le possibili norme matriciali

Teo. 4 (Decomposizione in valori singolari)**.** $A \in \mathbb{R}^{m \times n}$, allora $\exists U \in \mathbb{R}^{m \times m}, V \in \mathbb{R}^{n \times n}$ ortogonali, $\exists \Sigma \in \mathbb{R}^{m \times n}$ diagonale (di elementi detti valori singolari) tali che $A = U \Sigma V^T$

Dimostrazione. P. 11. Supponiamo $m \geq n$. Per induzione su m □

Definizione 0.4 (Raggio spettrale)**:** $\rho(A) \coloneqq \max_{\lambda \in \sigma(A)} |\lambda|$ è l'autovalore di modulo massimo

Parte I
Primo semestre

1 Buona posizione e numero di condizionamento di un problema

1.1 Problema numerico

Definizione 1.1 (Problema ben posto/stabile)**:** $d \in D$, $x \in X$ con D, X sottoinsiemi di spazi **normati**. Un problema ben posto è

$$F(x, d) = 0$$

tale che

(i) Esiste una soluzione

(ii) La soluzione è unica

(iii) La soluzione dipende con continuità coi dati

*Esempio (**Problema mal posto**)* Numero di soluzioni di $P_a(x) = x^4 - x^2(2a-1) + a(a-1)$ in funzione di a

Definizione 1.2 (Funzione risolvente)**:** Se il problema è ben posto, ovvero ammette soluzione unica, allora esiste una funzione

$$\begin{array}{rccc} G: & D & \to & X \\ & d & \mapsto & x \end{array} \quad \textbf{continua}$$

e abbiamo $F(G(d), d) = 0$ e $x + \delta x = G(d + \delta d)$.

Definizione 1.3 (Errore)**:** In generale i dati saranno perturbati $d + \delta d$ e $G(d + \delta d) = x + \delta x$ in modo che

$$F(x + \delta x, d + \delta d) = 0$$

$$\textsc{Assoluto}\ \|\delta x\|_X \qquad \textsc{Relativo}\ \frac{\|\delta x\|_X}{\|x\|_X} \text{ se } \|x\|_X \neq 0$$

Proprietà di Lipschitzianità In generale richiediamo oltre alla dipendenza continua dai dati, anche la lip., ovvero che valga:

$$\boxed{\exists K_0 = K_0(d) \mid \forall \delta d : d + \delta d \in D,\ \|\delta x\|_X \leq K_0 \|\delta d\|_D}$$

Questa proprietà è più adatta ad esprimere il concetto di **stabilità numerica**: piccole perturbazioni sui dati danno luogo a perturbazioni **dello stesso ordine di grandezza** sulla soluzione.

Definizione 1.4 (Numero di condizionamento)**:** Data $G\colon D \to X$ risolvente e $\delta x \coloneqq G(d + \delta d) - G(d)$

$$\textsc{Relativo}\quad K(d) = \lim_{\delta \to 0} \left(\sup_{\|\delta d\| \leq \delta} \left\{ \frac{\|\delta x\|/\|x\|}{\|\delta d\|/\|d\|} \right\} \right) = \lim_{\delta \to 0} \left(\sup_{\|\delta d\| \leq \delta} \left\{ \frac{\|\delta x\|}{\|\delta d\|} \frac{\|d\|}{\|x\|} \right\} \right)$$

$$\textsc{Assoluto}\quad \hat{K}(d) = \lim_{\delta \to 0} \left(\sup_{\|\delta d\| \leq \delta} \left\{ \frac{\|\delta x\|}{\|\delta d\|} \right\} \right)$$

L'assoluto è utile per es. quando $d = 0$ o $x = 0$. È una **misura** della bontà di dipendenza continua dai dati.

Osservazione Il libro la definisce

$$\textsc{Relativo}\ K(d) = \sup \left\{ \frac{\|\delta x\|/\|x\|}{\|\delta d\|/\|d\|},\ \delta d \neq 0,\ d + \delta d \in D \right\} \quad \textsc{Assoluto}\ K_{\text{abs}}(d) = \sup \left\{ \frac{\|\delta x\|}{\|\delta d\|},\ \delta d \neq 0,\ d + \delta d \in D \right\}$$

Definizione 1.5 (Buon condizionamento)**:** $K(d)$ piccolo (o $\hat{K}(d)$). Cioè ad un errore relativo sui dati corrisponde un errore piccolo sulla soluzione.

Osservazione $K(d) < +\infty$ vuol dire $\exists k_0 > 0$ tale che $\frac{\|\delta x\|}{\|x\|} \leq k_0 \frac{\|\delta d\|}{\|d\|}$ per $\|\delta d\|$ sufficientemente piccolo.
Infatti, fissato δ "piccolo", ho che

$$\sup_{\|\delta d\| \leq \delta} \left\{ \frac{\|\delta x\|/\|x\|}{\|\delta d\|/\|d\|} \right\} \approx K(d) \geq \frac{\|\delta x\|/\|x\|}{\|\delta d\|/\|d\|} \quad \forall \delta d \in D : \|\delta D\| \leq \delta$$

e quindi $\frac{\|\delta x\|}{\|x\|} \leq K(d) \frac{\|\delta d\|}{\|d\|}$

*Esempio (**Sistema lineare**)* Per semplicità perturbo solo il termine noto:

$$\begin{cases} Ax = b \\ A(x + \delta x) = b + \delta b \end{cases} \implies A\delta x = \delta b \implies \delta x = A^{-1}\delta b$$

allora

$$\sup_{\delta b \neq 0} \frac{\|\delta x\|}{\|\delta b\|} = \sup \frac{\left\|A^{-1}\delta b\right\|}{\|\delta b\|} = \left\|A^{-1}\right\|$$

e quindi

$$K(b) = \left\|A^{-1}\right\| \cdot \frac{\|b\|}{\|x\|} = \frac{\left\|A^{-1}\right\| \|Ax\|}{\|x\|} \overset{\star}{\leq} \left\|A^{-1}\right\| \|A\| =: K(A)$$

$\star$ poiché $\|Ax\| \leq \|A\| \|x\|$. In conclusione abbiamo che

$$\boxed{\frac{\|\delta x\|}{\|x\|} \leq K(A) \frac{\|\delta b\|}{\|b\|}} \quad \text{con } K(A) = \left\|A^{-1}\right\| \|A\|$$

*Osservazione (**Funzione risolvente differenziabile**)* Supponiamo $D \subseteq \mathbb{R}^m$ e $X \subseteq \mathbb{R}^n$, G differenziabile. Allora

$$G(d + \delta d) = G(d) + G'(d) \cdot \delta d + o(\|\delta d\|) \quad \text{Taylor}$$

allora $\delta x = G(\tilde{d}) - G(d) = G'(d) \cdot \delta d + o(\|\delta d\|)$ e

$$K_{\text{abs}}(d) = \lim_{\delta \to 0} \left(\sup_{\|\delta d\| \leq \delta} \left\{ \frac{\|\delta x\|}{\|\delta d\|} \right\} \right) \approx \lim_{\delta \to 0} \left(\sup_{\|\delta d\| \leq \delta} \left\{ \frac{\|G'(d) \cdot \delta d\| + o(\|\delta d\|)}{\|\delta d\|} \right\} \right) = \|G'(d)\|$$

$$K(d) \approx \|G'(d)\| \frac{\|d\|}{\|G(d)\|}$$

ovvero:

Definizione 1.6 (Numero di condizionamento al primo ordine)**:** Dal ragionamento appena fatto per funzioni risolventi derivabili:

$$\boxed{K(d) \approx \|G'(d)\| \frac{\|d\|}{\|G(d)\|} \qquad K_{\text{abs}}(d) \approx \|G'(d)\|}$$

*Esempio (**Equazione non lineare**)*

1.2 Metodo numerico

Definizione 1.7 (Metodo numerico)**:** Successione di problemi $F_n(x_n, d_n) = 0$ (detti problemi approssimati) tali che $x_n \to x$ quando $d_n \to d$, ovvero che la **soluzione numerica** x_n converga alla **soluzione esatta** x. Consideriamo il problema ben posto con risolvente $G : D \to X$. Un metodo numerico è un **insieme di trasformazioni elementari**

$$\phi_i : D_i \to D_{i+1} \quad i = 1, \ldots, r$$

tali che

- $D_0 = D, \quad D_{r+1} = X$
- $x_0 = d, \quad x_{i+1} = \phi_i(x_i), \quad x_{r+1} = x$
- $\phi_r \circ \cdots \circ \phi_0 =: \widetilde{G} \approx G$

NB: ogni operazione è affetta da **roundoff**: $\tilde{x}_{i+1} = fl(\phi_i(\tilde{x}_i))$

Definizione 1.8 (Metodo ben posto)**:** Se i problemi approssimati $F_n(x_n, d_n) = 0$ sono ben posti $\forall n$

Definizione 1.9 (Metodo consistente)**:** Un metodo numerico $\{F_n\}_n$ si dice

$$\text{Consistente } F_n(x,d) \longrightarrow F(x,d) = 0 \qquad \text{Fortemente consistente } F_n(x,d) = 0 \quad \forall n$$

Definizione 1.10 (Metodo stabile/ben condizionato)**:** Se i problemi approssimati $F_n(x_n, d_n) = 0$ sono ben condizionati, ovvero $K_n(d_n)$ piccolo $\forall n$, o anche

$$K_{\phi_i}(x_i) = \frac{\|J\phi_i(x_i)\|\|x_i\|}{\|\phi_i(x_i)\|} \quad \text{è "piccola"}$$

$\{F_n\}_n$ stabile se $\exists\, k_0 > 0, \delta_0 > 0$ tali che $\forall\, n\ \|\delta x_n\| \le k_0 \|\delta d_n\|$ con $\|\delta d_n\| \le \delta_0 \quad \forall\, \delta d_n$, ovvero i problemi sono uniformemente ben condizionati

Definizione 1.11 (Numero di condizionamento asintotico rel/ass)**:**

$$K^{num}(d) = \lim_{k\to\infty} \sup_{n\ge k} K_n(d) \qquad K^{num}_{abs}(d) = \lim_{k\to\infty} \sup_{n\ge k} K_{abs,n}(d)$$

Osservazione $\{F_n\}_n$ stabile $\iff \sup_n K_n(d_n) \le +\infty$

Definizione 1.12 (Convergenza)**:** $\{F_n\}_n$ convergente se

$$\forall\, \varepsilon > 0 \quad \exists \begin{cases} n_0(\varepsilon) \\ \delta_0(n_0, \varepsilon) > 0 \end{cases} \text{ tali che } \forall \begin{cases} n \ge n_0 \\ \delta d_n \text{ con } \|\delta d_n\| \le \delta_0 \end{cases} \text{ vale } \quad \|G(d) - G_n(d + \delta d_n)\| \le \varepsilon$$

Definizione 1.13 (Errore ass/rel di un metodo)**:** Danno misura della convergenza di x_n:

$$E(x_n) = \|x - x_n\| \qquad E_{rel}(x_n) = \frac{\|x - x_n\|}{\|x\|} \quad (x \neq 0)$$

Teo. 5 (Lax-Richtmyer)**.** Un metodo

$$\text{convergente} \underset{\text{+consistente}}{\overset{\Longrightarrow}{\Longleftarrow}} \text{stabile}$$

Dimostrazione (per i sistemi lineari).

$$\Longrightarrow \quad \left\| \overbrace{G_n(d) - G_n(d + \delta d_n)}^{\delta x_n} \right\| \le \left\| \overbrace{G_n(d) - G(d)}^{\longrightarrow 0 \text{ per convergenza}} \right\| + \left\| \overbrace{G(d) - G(d + \delta d_n)}^{\le k_0 \|\delta d\| \longrightarrow 0} \right\| + \left\| \overbrace{G(d + \delta d_n) - G_n(d + \delta d_n)}^{\longrightarrow 0 \text{ per convergenza}} \right\|$$

$$\Longleftarrow \quad \|G(d) - G_n(d + \delta d_n)\| \le \left\| \overbrace{G(d) - G_n(d)}^{\|x - A_n^{-1} d\| \le \|A_n x - d\| \longrightarrow 0} \right\| + \left\| \overbrace{G_n(d) - G_n(d + \delta d_n)}^{\le k_0 \|\delta d\| \longrightarrow 0} \right\|$$

□

Definizione 1.14 (Analisi a priori e a posteriori dell'errore)**:**

- Analisi a priori:
  - *Analisi in avanti*: dato l'errore δd studio l'effetto sull'errore della soluzione δx
  - *Analisi all'indietro*: dato l'errore della soluzione δx cerco di determinare δd
- Analisi a posteriori: Stimo l'errore della soluzione dopo il calcolo (per esempio col residuo $r = F(x_n, d)$)

Definizione 1.15 (Sorgenti di errore)**:**

$$e = \begin{cases} e_{\text{matematico}} \begin{cases} e_{\text{sul PF}} = \text{errore sul modello matematico del problema fisico} \\ e_{\text{sui dati}} = \text{errore sulla misurazione dei dati} \end{cases} \\ + \\ e_{\text{computazionale}} \begin{cases} e_n = \text{discretizzazione numerica (troncamento, numero finito di passi nei } \lim_{\to\infty}) \\ e_a = \text{di arrotondamento (roundoff)} \end{cases} \end{cases}$$

Un metodo numerico è convergente se l'$e_{\text{computazionale}}$ può essere ridotto arbitrariamente aumentando lo sforzo computazionale.

Definizione 1.16 (Altre caratteristiche di un metodo numerico)**:** Abbiamo:

- **Accuratezza**: errori piccoli rispetto a tolleranza fissata. Quantificata con l'ordine di infinitesimo di e_n risp. al parametro della discretizzazione.
- **Affidabilità**: se applicato a tanti test l'errore totale può essere tenuto al di sotto di una tolleranza con probabilità maggiore rispetto a quella prestabilita
- **Efficienza**: colplessità computazionale sia la più bassa possibile
- **Complessità (di un problema)**: minima complessità tra tutti gli algoritmi (che è il loro tempo di esecuzione)

2 Aritmetica finita

Un calcolatore può rappresentare solo un numero finito di numeri $\mathbb{F} \subset \mathbb{R}$. Ci sono due limitazoni:

(i) Non si possono rappresentare numeri troppo grandi o troppo piccoli

(ii) I numeri rappresentati sono in un numero discreto

Numeri floating point $x \in F$ è rappresentato come

$$x = (-1)^s (0.a_1 \dots a_t)\beta^e$$

con

- $s = \{0, 1\}$ segno
- a_i cifre significative
- t numero di cifre significative
- β base
- e ordine di grandezza

Doppia precisione

Definizione 2.1 (Unità di roundoff)**:** Aritmetica finita $F \subseteq \mathbb{R}$, u unità di roundoff per cui $\forall\, x \in \mathbb{R}\ \exists\, x' \in F$ tale che $|x - x'| \leq u\,|x|$

Definizione 2.2 (ε macchina)**:** Sia $x \in \mathbb{R}$, $x_{min} < |x| < x_{max} \implies \exists \varepsilon \in \mathbb{R}$, $|\varepsilon| \leq u \quad | \quad fl(x) = x(1 - \varepsilon)$

$$\varepsilon_{mc} = \min\{x \in F \mid 1 \oplus x > 1\}$$

2.1 Operazioni in $\mathbb{F}$

$$* \rightarrow \text{ operazione in } \mathbb{R}$$
$$\circledast \rightarrow \text{ operazione in } F : \boxed{x \circledast y \coloneqq fl(x * y)}$$

usando l'ε macchina si ha $\exists \varepsilon \in \mathbb{R}$, $|\varepsilon| \leq u$ tale che

$$x \circledast y = (x * y)(1 + \varepsilon)$$

Osservazione Molte proprietà delle operazioni in $\mathbb{R}$ non valgono in aritmetica finita, come la **proprietà associativa**

Stabilità In somma multipla sommare prima le coppie che hanno somma in modulo minore

Cancellazione L'errore relativo di $a - b$ diventa molto grande se sono vicini

3 Metodi diretti per sistemi lineari

Definizione 3.1 (Metodi diretti/iterativi)**:** Abbiamo:

- **Metodi diretti**: calcolano la soluzione esatta in un numero finito di passaggi
- **Metodi iterativi**: calcolano la soluzione esatta in un numero infinito di passaggi, ovvero calcolano una successione $(x_k)_k$ a valori in $\mathbb{R}^n$ t.c. $x_k \to x$

3.1 Sistemi lineari

Definizione 3.2 (Sistema lineare approssimato)**:** $Ax = b \quad \longrightarrow \quad (A + \delta A)(x + \delta x) = (b + \delta b)$

Definizione 3.3 (Numero di condizionamento di una matrice)**:** È

$$K(A) := \|A\| \cdot \left\|A^{-1}\right\|$$

Osservazione $K(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$, ($\sigma$ valori singolari), per A SPD vale $K_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$ (λ autovalori)

Dimostrazione. $\|A\|_2 \overset{\star}{=} \sqrt{\rho(A^T A)} = \sqrt{\rho(A^2)} = \sqrt{\sigma_{max}^2} = \sigma_{max}$ dove non dimostriamo $\star$ e ρ è il raggio spettrale.
Inoltre vale $\sigma(A^{-1}) = 1/\sigma(A)$ e dalla def. di numero di cond. segue la tesi. □

Osservazione $K(A) > 1$ in quanto $1 = \|I\| = \left\|A \cdot A^{-1}\right\| \leq \|A\| \cdot \left\|A^{-1}\right\| = K(A)$

Prop. 2 (Stima a priori dell'errore)**.** Abbiamo che

$$\boxed{\frac{\|\delta x\|}{\|x\|} \leq K(A) \frac{\|\delta b\|}{\|b\|}} \quad \text{con } K(A) = \left\|A^{-1}\right\| \|A\|$$

Dimostrazione. Per semplicità perturbo solo il termine noto:

$$\begin{cases} Ax = b \\ A(x + \delta x) = b + \delta b \end{cases} \implies A\delta x = \delta b \implies \delta x = A^{-1}\delta b$$

allora

$$K_{abs} = \sup_{\delta b \neq 0} \frac{\|\delta x\|}{\|\delta b\|} = \sup \frac{\left\|A^{-1}\delta b\right\|}{\|\delta b\|} = \left\|A^{-1}\right\|$$

e quindi

$$K(b) = K_{abs} \cdot \frac{\|b\|}{\|x\|} = \left\|A^{-1}\right\| \cdot \frac{\|b\|}{\|x\|} = \frac{\left\|A^{-1}\right\| \|Ax\|}{\|x\|} \overset{\star}{\leq} \left\|A^{-1}\right\| \|A\| =: K(A)$$

$\star$ poiché $\|Ax\| \leq \|A\| \|x\|$. □

Lemma 1 . Preso $\|B\| < 1 \implies I + B$ invertibile e $\left\|(I + B)^{-1}\right\| \leq \frac{1}{1 - \|B\|}$

Dimostrazione. Per assurdo, $I + B$ non invertibile $\implies \exists x \in \ker(I + B)$ non nullo $\implies I \cdot x = -B \cdot x \implies \|x\| = \|B \cdot x\| \leq \|B\| \cdot \|x\| \implies \|B\| \geq 1$, ↯

$$I = (I + B)^{-1}(I + B) \quad \implies \quad (I + B)^{-1} = I - B(I + B)^{-1} \quad \implies \quad \left\|(I + B)^{-1}\right\| \leq 1 + \|B\| \cdot \left\|(I + B)^{-1}\right\|$$

□

Cor. 1 . Se A invertibile e $\left\|A^{-1}\right\| \cdot \|\delta A\| < 1 \quad \implies \quad A + \delta A$ invertibile

Dimostrazione. $A + \delta A$ invertibile $\iff \quad A^{-1}(A + \delta A) = I + A^{-1}\delta A$ invertibile

$$\left\|A^{-1}\delta A\right\| \leq \left\|A^{-1}\right\| \|\delta A\| < 1$$

da cui la tesi e $\|\delta A\| \leq \frac{1}{\|A^{-1}\|}$ per il lemma □

Teo. 6 . Preso δA tale che $\left\|A^{-1}\right\| \cdot \|\delta A\| < 1$, allora per il Sistema lineare approssimato vale:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{K(A)}{1 - K(A)\frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right)$$

Dimostrazione.

$$(A + \delta A)(x + \delta x) = (b + \delta b) \implies \cancel{A \cdot x} + \delta A \cdot x + (A + \delta A)\delta x = \cancel{b} + \delta b$$

$$\implies \delta x = \overbrace{(A + \delta A)^{-1}}^{\clubsuit} \overbrace{(\delta b - \delta A \cdot x)}^{\spadesuit}$$

$$\|\clubsuit\| \overset{\star}{=} \left\| A^{-1}\left(I + A^{-1}\delta A\right)^{-1} \right\| \leq \left\|A^{-1}\right\| \left\| \left(I + A^{-1}\delta A\right)^{-1} \right\| \overset{\text{per (1)}}{\leq} \frac{\left\|A^{-1}\right\|}{1 - \left\|A^{-1}\delta A\right\|} \leq \frac{\left\|A^{-1}\right\|}{1 - \left\|A^{-1}\right\| \|\delta A\|}$$

in $\star$: $(A + \delta A)^{-1} = (AA^{-1}(A + \delta A))^{-1} = (A(I_n + A^{-1}\delta A))^{-1}$

$$\|\spadesuit\| \leq \|\delta A\| \|x\| + \|\delta b\| \overset{\star}{\leq} \|A\| \|x\| \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right) \qquad \text{in } \star \text{ moltiplico per } 1 = \frac{\|Ax\|}{\|b\|} \leq \frac{\|A\| \|x\|}{\|b\|}$$

□

Osservazione Se $\frac{\|\delta A\|}{\|A\|}, \frac{\|\delta b\|}{\|b\|} = O(u)$, con u unità di roundoff e $O(\bullet)$ <<ordine di>> allora

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{K(A)}{1 - K(A)O(u)} O(u) = K(A)O(u) \tag{1}$$

3.2 Sistemi triangolari

Definizione 3.4 (Matrice triangolare superiore/inferiore)**:** $A \in \mathbb{R}^{n \times n}$ con $a_{ij} = 0$ per $i > j$ (superiore) o $i < j$ (inferiore)

Prop. 3 (Determinante di matirice triangolare)**.** $\det A = \prod_{i=1}^{n} a_{ii}$

Cor. 2 . Abbiamo

- A triangolare è invertibile $\iff a_{ii} \neq 0 \quad \forall i = 1 \dots n$
- A triangolare $\implies \sigma(A) = \{a_{ii} \mid i = 1 \dots n\}$

Prop. 4 . A, B triangolari superiori (inferiori) $\implies A \cdot B$ triangolare supreiore (inferiore)

Backward substitution Se A trangolare **superiore**

$$Ax = b \implies \begin{cases} a_{11}x_1 + \ldots \cdots + & a_{1n}x_n = b_1 \\ a_{22}x_2 + \cdots + & a_{2n}x_n = b_2 \\ & \vdots \\ & a_{nn}x_n = b_n \end{cases}$$

Per risolvere il sistema si usa backsubstitution:

$$\begin{aligned} x_n &= \frac{b_n}{a_{nn}} \\ x_{n-1} &= \frac{b_{n-1} - a_{n-1,n}x_n}{a_{n-1,n-1}} \\ &\vdots \\ x_i &= \frac{b_i - \sum_{j=i+1}^{n} a_{ij}x_j}{a_{ii}} \end{aligned}$$

Forward substitution Se A trangolare **inferiore** il sistema il sistema si risolve con forwardsubstitution:

$$x_1 = \frac{b_1}{a_{11}}$$
$$\vdots$$
$$x_i = \frac{b_i - \sum_{j=1}^{i-1} a_{ij}x_j}{a_{ii}}$$

Definizione 3.5 (Costo computazionale)**:** numero di operazioni floating point (**flops**) effettuate dall'algoritmo

Osservazione Per i gli algoritmi di backsubstitution/forwardsubstitution il numero di operazioni per calcolare x_i è: $(n-i)$ prodotti, $(n-i)$ somme, 1 divisione. Quindi il costo computazionale è:

$$\sum_{i=1}^{n} 2(n-i)+1 \approx 2\sum_{i=1}^{n}(n-i) = 2\left[\sum_{i=1}^{n} n - \sum_{i=1}^{n} i\right] = 2(n^2 - \frac{n(n-1)}{2}) = 2n^2 - n^2 + n = n^2 + n \approx n^2 = O(n^2)$$

Teo. 7 (Inversa di una matrice triangolare)**.** A triangolare superiore (inferiore) $\implies$ A^{-1} triangolare superiore (inferiore)

Dimostrazione. Sia $A^{-1} = \left[x^{(1)}, \dots, x^{(n)}\right]$, risolvo $AA^{-1} = I \to Ax^{(i)} = e_i$ con backsubstitution ottengo $x^{(i)} = \left(x_1^{(i)}, \dots, x_i^{(i)}, 0, \dots, 0\right)^T$ □

Teo. 8 . $|\delta T| \leq \left(nu + O\left(u^2\right)\right)|T|$

Cor. 3 . $\frac{\|\delta x\|}{\|x\|} \leq \frac{K(T)\cdot n\cdot u}{1-K(T)\cdot n\cdot u} = K(T)\cdot n\cdot u + O\left(u^2\right)$

3.3 Eliminazione di Gauss e Fattorizzazione LU

3.3.1 Eliminazione di Gauss

Voglio trovare un algoritmo che mi porti da un sistema lineare ad uno triangolare:

$$Ax = b \longrightarrow Ux = y \quad U \text{ triangolare}$$

faccio successione di sistemi lineari $A^{(i)}x = b^{(i)}$ t.c. $\begin{cases} A^{(1)} = A \\ b^{(1)} = b \end{cases}$ e $\begin{cases} A^{(n)} = U \\ b^{(n)} = y \end{cases}$

(1) $\begin{cases} A^{(1)} = A \\ b^{(1)} = b \end{cases}$

(2) $\begin{cases} a_{ij}^{(2)} = a_{ij}^{(1)} - l_{i1}a_{1j}^{(1)} \\ b_i^{(2)} = b_i^{(1)} - l_{i1}b_1^{(1)} \end{cases}$ per $i = 2\dots n$, $j = 1\dots n$ con $l_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} \implies$ ho "eliminato" la prima colonna (tranne a_{11})

(k+1) $\begin{cases} a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik}a_{kj}^{(k)} \\ b_i^{(k+1)} = b_i^{(k)} - l_{ik}b_k^{(k)} \end{cases}$ per $i = k+1\dots n$, $j = k+1\dots n$ con $l_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \implies$ ho "reso triangolari" le prime k colonne

ovvero

$$\begin{cases} A^{(k+1)} = L_k A^{(k)} \\ b^{(k+1)} = L_k b^{(k)} \end{cases} \quad \text{con} \quad L_k = \boldsymbol{I}_n - l_k \cdot e_k^T = \begin{bmatrix} 1 & 0 & \dots & & \dots & 0 \\ 0 & \ddots & & & & \\ \vdots & & 1 & & & \vdots \\ & & -l_{k+1,k} & 1 & & \\ & & -l_{k+2,k} & 0 & 1 & \\ & & \vdots & \vdots & \ddots & 0 \\ 0 & \dots & -l_{nk} & 0 & \dots & 0 & 1 \end{bmatrix} \text{ e } l_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ l_{k+1,k} \\ \vdots \\ l_{nk} \end{pmatrix}$$

oppure: colonna della sottomatrice che parte dalla riga $k+1$-esima e colonna k-esima:

$$A_j^{(k+1)} = A_j^{(k)} - a_{kj}^{(k)} \boldsymbol{l}_k \quad \text{con } (\boldsymbol{l}_k)_i = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$$

(n) $\begin{cases} A^{(n)} = U \\ b^{(n)} = y \end{cases}$

Costo computazionale al passo k ho

- $n-k$ divisioni
- $2(n-k)(n-k+1)$ prodotti e sottrazioni
- $2(n-k)$ ...

$= (n-k)(1+2(n-k+1)+2) = (n-k)(2n-2k+5) \implies \sum_{k=1}^{n-1}(n-k)(2n-2k+5) \approx 2\sum_{k=1}^{n-1}(n-k)^2 = 2\sum_{k=1}^{n-1} k^2 = 2\frac{(2n-1)(n-1)}{6} \approx \frac{2}{3}n^3 = O(n^3)$

Osservazione $e_j^T l_k = 0 \ \forall j \leq k$

Osservazione Abbiamo

$$L_k^{-1} = I_n + l_k e_k^T$$

Dim: $(I_n - l_k e_k^T)(I_n + l_k e_k^T) = I_n - \cancel{l_k e_k^T} + \cancel{l_k e_k^T} - l_k \underbrace{(e_k^T l_k)}_{=0} e_k^T = I_n$

3.3.2 Fattorizzazione LU

Vediamo da prima con L_k k-esimo passaggio dell'eliminazione di Gauss che

$$U = A^{(n)} = L_{n-1}A^{(n-1)} = L_{n-1}L_{n-2}A^{(n-2)} = \cdots = L_{n-1}\dots L_1 \underbrace{A^{(1)}}_{=A}$$

Allora

$$A = \underbrace{L_1^{-1}L_2^{-1}\dots L_{n-1}^{-1}}_{:=L} U = LU$$

Vediamo che L è triangolare: dato che $L_k^{-1} = \left(\boldsymbol{I}_n - l_k \cdot e_k^T\right)^{-1} = \boldsymbol{I}_n + l_k \cdot e_k^T$, allora $L = \prod_{k=1}^{n}\left(\boldsymbol{I}_n + l_k \cdot e_k^T\right) = \boldsymbol{I}_n + \sum_{k=1}^{n} l_k \cdot e_k^T$

$$[\dots]$$

$$L = \boldsymbol{I}_n + \sum_{i=1}^{n-1} l_i e_i^T = \begin{bmatrix} 1 & 0 & \dots & 0 \\ l_{21} & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ l_{n1} & l_{n2} & \dots & 1 \end{bmatrix}$$

Le entrate di L sono i coefficienti dell'eliminazione di Gauss

Definizione 3.6 (Fattorizzazione LU)**:** data $A \in \mathbb{R}^{n\times n}$ è la sua scomposizione in $A = LU$ con $L \in \mathbb{R}^{n\times n}$ triangolare inferiore e $U \in \mathbb{R}^{n\times n}$ triangolare superiore

$Ax = b \ \Rightarrow \ LUx = b$ e quindi risolvo $\begin{cases} L\hat{b} = b \\ Ux = \hat{b} \end{cases} \longrightarrow$ è equivalente all'eliminazione di Gauss

Teo. 9 . Sia $A \in \mathbb{R}^{n\times n}$, esiste ed è unica la sua fattorizzazione LU $\iff$ la matrice $A_k = \begin{bmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{1k} & \dots & a_{kk} \end{bmatrix}$ (minore principale k-esimo) è non singolare ($\det \neq 0$) per $k = 1 \cdots n-1$

Dimostrazione. □ P. 25

3.3.3 Per matrici a bande

Definizione 3.7 (Matrice a banda)**:** A matrice a banda di ampiezza p se $a_{ij} = 0$ per $|i-j| > p$

Prop. 5 . Sia A a banda di ampiezza p, se $\exists!$ la sua fattorizzazione $LU \implies L, U$ anch'esse a banda con la stessa ampiezza

Dimostrazione. □ P. 26

Osservazione Posso modificare l'algoritmo per calcolare la fattorizzazione LU se A è a banda per renderlo più efficiente. P. 27

3.3.4 Stabilità dell'eliminazione di Gauss

Abbiamo $A = LU$ e dall'aritmetica finita

$$\begin{cases} \widetilde{L} = L + \delta L \\ \widetilde{U} = U + \delta U \\ \widetilde{A} = A + \delta A = (L + \delta L)(U + \delta U) = LU + \delta LU + L\delta U + \delta L\delta U \end{cases}$$

Supponiamo che

$$\frac{\|\delta L\|}{\|L\|} = O(u), \qquad \frac{\|\delta U\|}{\|U\|} = O(u) \quad \text{per } u \to 0$$

allora

$$\begin{aligned} \|\delta A\| = \left\|\widetilde{A} - A\right\| &\leq \|\delta LU\| + \|L\delta U\| + \overbrace{\cancel{\|\delta L\delta U\|}}^{=O(u^2\|L\|\|U\|)} \\ &\leq \|\delta L\|\|U\| + \|L\|\|\delta U\| \\ &= O(u\|L\|\|U\|) + O(u\|L\|\|U\|) \end{aligned}$$

E quindi

$$\frac{\|\delta A\|}{\|A\|} = O\left(u\frac{\|L\|\|U\|}{\|A\|}\right), \qquad \frac{\|\delta x\|}{\|x\|} = O\left(K(A)\frac{\|\delta A\|}{\|A\|}\right) \quad \text{ricorda 1}$$

Osservazione Se $\|L\|, \|U\| \gg \|A\|$ può esserci instabilità. Vediamo che scambiando le equazioni (ovvero le righe) queste due norme possono passare da essere molto maggiori di quella di A ad essere piccole, per esempio mi **conviene mettere alla prima riga la riga col primo elemento maggiore**. Quindi adottiamo il seguente:

3.3.5 Pivoting (scambiare righe)

Sia $A^{(k)}$ la matrice ottenuta al k-esimo passo del metodo di Gauss.

Definizione 3.8 (Elemento pivotale)**:** Nel MEG l'elemento $a_{kk}^{(k)}$

Definizione 3.9 (Pivoting parziale)**:** Al passo k scambio

$$\text{riga } k\text{-esima} \longleftrightarrow \text{riga } i\text{-esima con } i = \arg\max_{s=k\dots n} \left|a_{sk}^{(k)}\right|$$

ovvero a ogni passaggio dell'eliminazione di Gauss faccio $A^{(k+1)} = L_k P_k A^{(k)}$ (P matrice di permutazione delle righe per il pivoting parziale), quindi $L_{n-1}P_{n-1}\dots L_1P_1A = U$, definisco ora $L'_k = P_{n-1}\dots P_{k+1}L_kP_{k+1}^{-1}\dots P_{n-1}^{-1}$ da cui $\underbrace{L'_{n-1}\dots L'_1}_{L^{-1}}\underbrace{P_{n-1}\dots P_1}_{P}A = U$ e quindi $PA = LU$.

Osservazione L'_k ha la stessa struttura di L_k:

$$L'_k = \boldsymbol{I}_n - P_{n-1}\dots P_{k+1}l_k\underbrace{e_k^TP_{k+1}^{-1}\dots P_{n-1}^{-1}}_{e_k^T} = \boldsymbol{I}_n - (P_{n-1}\dots P_{k+1}l_k)\,e_k^T$$

Definizione 3.10 (Matrice di permutazione)**:** Prendo la matrice identità e ne permuto le righe. Se moltiplicata a sx ad una matrice mi permuta le righe

Prop. 6 . La fattorizzazione con il pivoting esiste sempre se A non singolare:

$$A \in \mathbb{R}^{n\times n} \text{ non singolare } \implies \exists P \in \mathbb{R}^{n\times n} \text{ matrice di perm. : } PA = LU$$

Osservazione Il sistema lineare diventa

$$Ax = b \longrightarrow \underbrace{PA}_{=LU} x = Pb \longrightarrow \begin{cases} Ly = Pb \\ Ux = y \end{cases}$$

Definizione 3.11 (Pivoting totale)**:**

*Esempio (**Calcolo del determinante**)* Normalmente $\det A = \sum_{i=1}^{i+j} a_{ij} \det(A_{ij})$ che ha costo $O(n!)$.

$$PA = LU \implies A = P^{-1}LU \implies \det(A) = \overbrace{\det\left(P^{-1}\right)}^{(-1)^\delta}\overbrace{\det(L)}^{1}\det(U) = (-1)^\delta \prod_{i=1}^{n} u_{ii}$$

Ha il costo della LU ovvero $O(\frac{2}{3}n^3)$

*Esempio (**Calcolo dell'inversa**)* $A \in \mathbb{R}^{n\times n}$ invertibile, $A^{-1} \coloneqq X$, $\begin{cases} X = (x^{(1)}, \dots, x^{(n)}) \\ I_n = (e^{(1)}, \dots, e^{(n)}) \end{cases}$, $AX = I_n \iff$ $Ax^{(j)} = e^{(j)}$ con $j = 1, \dots, n$. Allora

$$PA = LU \implies A = P^{-1}LU \implies L\underbrace{Ux^{(j)}}_{y^{(j)}} = Pe^{(j)} \iff \begin{cases} Ux^{(j)} = y^{(j)} \\ Ly^{(j)} = Pe^{(j)} \end{cases}$$

algoritmo di Thomas [...]

Fattorizzazione di Cholesky Per A SPD $\implies$ si può scrivere come $A = H^T H$ con H triangolare superiore.

$$A = \begin{bmatrix} a & w^T \\ w & \boldsymbol{K} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ w & \boldsymbol{I}_{n-1} \end{bmatrix} \cdot \begin{bmatrix} a & w^T \\ 0 & \boldsymbol{K} - \frac{ww^T}{a} \end{bmatrix} = \begin{bmatrix} \alpha & 0 \\ \frac{w}{\alpha} & \boldsymbol{I}_{n-1} \end{bmatrix} \cdot \begin{bmatrix} \alpha & w^T/\alpha \\ 0 & \boldsymbol{K} - \frac{ww^T}{a} \end{bmatrix} = \begin{bmatrix} \alpha & 0 \\ \frac{w}{\alpha} & \boldsymbol{I}_{n-1} \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & \boldsymbol{K} - \frac{ww^T}{a} \end{bmatrix} \cdot \begin{bmatrix} \alpha & w^T/\alpha \\ 0 & \boldsymbol{I}_{n-1} \end{bmatrix}$$

3.4 Fattorizzazione QR

Definizione 3.12 (fattorizzazione QR)**:** $A \in \mathbb{R}^{m\times n}$ con $m \geq n$. Una fattorizzazione del tipo

$$A = QR$$

con $Q \in \mathbb{R}^{m\times m}$ ortogonale e $R \in \mathbb{R}^{m\times n}$ trapezioidale con le righe dalla $n+1$ in poi nulle

Definizione 3.13 (fattorizzazione QR ridotta)**:** $A = \widetilde{Q}\widetilde{R} \in \mathbb{R}^{m\times n}$ con $\widetilde{Q} \in \mathbb{R}^{m\times n}$ a colonne ortonormali e $\widetilde{R} \in \mathbb{R}^{n\times n}$ triangolare superiore

Osservazione $A = QR = (\widetilde{Q} \mid X)\begin{pmatrix} \widetilde{R} \\ \mathbf{0} \end{pmatrix} = \widetilde{Q}\widetilde{R} + \cancel{X\mathbf{0}}$

Osservazione Metodi per il calcolo:

- Completa: metodo di Householder
- Ridotta: metodo di Gram-Schmidt standard (instabile) e modificato

Osservazione A ha rango massimo $\iff$ $\widetilde{R}$ invertibile

Prop. 7 . $\sigma_i(A) = \sigma_i(R)$ e quindi $K(A) = \frac{\sigma_1(A)}{\sigma_n(A)} = K(R)$

Dimostrazione. Abbiamo: $A^TA = V\Sigma^T U^T U\Sigma V^T = V\Sigma^T\Sigma V^T$ quindi $\boxed{\sigma_i(A) = \sqrt{\lambda(A^TA)}}$.
Allora $\sigma_i(A) = \sqrt{\lambda(A^TA)} = \sqrt{\lambda(R^TQ^TQR)} = \sqrt{\lambda(R^TR)} = \sigma_i(R)$ □

metodo di Gram-Schmidt standard (s.g.s) Presi $A = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix}$ e $Q = \begin{bmatrix} q_1 & \dots & q_n \end{bmatrix}$ abbiamo

$$A = QR \begin{cases} a_1 = r_{11}q_1 \\ a_2 = r_{12}q_1 + r_{22}q_2 \\ \vdots \\ a_n = r_{1n}q_1 + r_{2n}q_2 + \dots + r_{nn}q_n \end{cases} \quad \begin{cases} q_1 = {}^1\!/_{r_{11}}(a_1) \\ q_2 = {}^1\!/_{r_{22}}(a_2 - r_{12}q_1) \\ \vdots \\ q_n = {}^1\!/_{r_{nn}}\left(a_n - \sum_{i=1}^{n-1} r_{in}q_i\right) \end{cases} \quad \begin{cases} r_{ij} = \langle q_i, a_j \rangle \quad i < j \\ r_{jj} = \|\widetilde{q_j}\|_2 \end{cases}$$

$$\widetilde{q_j} = a_j - \sum_{i=1}^{j-1} r_{ij}q_i$$

Costo: $2mn^2 + O(mn)$

metodo di Gram-Schmidt modificato (s.g.m) Più stabile in aritmetica discreta, al posto di fissare a_j e rimuovere le proiezioni di a_j dai vettori q_i, fisso q_i e ne sottraggo le proiezioni a tutti i vettori a_j
Costo: $2mn^2$

metodo di Householder p. 43-44

3.5 Sistemi sovradeterminati

Sia $Ax = b$ con $A \in \mathbb{R}^{m \times n}$, $m \geq n$, $b \in \mathbb{R}^m$, rango$(A) = n$. In generale non ha soluzione, a meno che $b \in$rango(A): al posto di trovare $Ax = b$, ovvero $Ax - b = 0$, trovo il vettore x che ha immagine **più vicina** a b, ovvero il

$$\min_{x \in \mathbb{R}^n} \|b - Ax\|_2^2 = \min_{x \in \mathbb{R}^n} \sum_{i=1}^{n} (b_i - \sum_{j=1}^{m} a_{ij}x_j)^2$$

detto **problema dei minimi quadrati**. Devo trovare il minimo della funzione

$$\begin{aligned} \varphi(x) &= \|b - Ax\|_2^2 \\ &= (b - Ax)^T(b - Ax) \\ &= \|b\|_2^2 + x^T A^T Ax \underbrace{-b^T Ax - x^T A^T b}_{-2x^T A^T b} \quad b^T Ax = \text{al trasposto poiché scalare} \\ &= \|b\|_2^2 + x^T A^T Ax - 2x^T A^T b \quad \text{scegliere } x^T \text{ nel raccogliere} \end{aligned}$$

- $\nabla\varphi(x) = 2A^T Ax - 2A^T b$
- $H_\varphi = 2A^T A > 0$ poiché $x^T(A^T A)x = (Ax)^T(Ax) = \|Ax\|_2^2 > 0$ per $x \neq 0$ poiché A iniettiva

allora esiste punto di minimo (unico) in corrispondenza di $\nabla\varphi(x) = 0 \iff A^T Ax = A^T b$

Osservazione Questo sistema può essere molto mal condizionato

$$K(A^T A) = (K(A))^2$$

Dimostrazione. Dobbiamo usare

- $K(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$

  $\|A\|_2 \overset{\star}{=} \sqrt{\rho(A^T A)} = \sqrt{\rho(A^2)} = \sqrt{\sigma_{max}^2} = \sigma_{max}$ dove non dimostriamo $\star$ e ρ è il raggio spettrale.
  Inoltre vale $\sigma(A^{-1}) = 1/\sigma(A)$ e dalla def. di numero di cond. segue la tesi.

- $\sigma_i(A) = \sqrt{\lambda_i(A^T A)}$
  abbiamo: $A^T A = V\Sigma^T U^T U\Sigma V^T = V\Sigma^T \Sigma V^T$ quindi $\sigma_i(A) = \sqrt{\lambda(A^T A)}$.

Quindi

$$K(A^T A) = \frac{\lambda_{\max}(A^T A)}{\lambda_{\min}(A^T A)} = \frac{\sigma_{\max}^2(A)}{\sigma_{\min}^2(A)} = \left(\frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}\right)^2 = K(A)^2$$

□

*Osservazione (**con** QR)* Se $A = QR$ abbiamo $\cancel{R^T} Q^T QRx = \cancel{R^T} Q^T b \implies \boxed{Rx = Q^T b}$ e so che $K(R) = K(A)$

4 Metodi iterativi per sistemi lineari

Definizione 4.1 (metodo iterativo)**:** $\lim_{k\to+\infty} x^{(k)} = x$ con x soluzione di $Ax = b$

4.1 Metodi iterativi stazionari

Definizione 4.2 (metodo iterativo stazionario)**:** $x^{(k+1)} = Bx^{(k)} + f$

Definizione 4.3 (consistenza)**:** Un metodo iterativo stazionario si dice consistente se vale $x = Bx + f$ per $Ax = b$

Prop. 8 . In un metodo iterativo convergente $\implies$ consistente

Dimostrazione. $x^{(k+1)} = Bx^{(k)} + f \overset{x^{(k)} \text{ converge}}{\underset{k\to+\infty}{\implies}} x = Bx + f$ □

*Osservazione (**Errore**)* Se il metodo è consistente ($\iff x = Bx + f$) risulta

$$\underbrace{x - x^{(k+1)}}_{:=e^{(k+1)}} = (Bx + \cancel{f}) - (Bx^{(k)} + \cancel{f}) = B\underbrace{(x - x^{(k)})}_{:=e^{(k)}} \implies \boxed{e^{(k+1)} = Be^{(k)}}$$

e quindi

$$\begin{aligned} e^{(k+1)} &= Be^{(k)} \\ e^{(k)} &= B^k e^{(0)} \\ \left\|e^{(k)}\right\| \le \|B^k\| \left\|e^{(0)}\right\| &\implies \frac{\left\|e^{(k)}\right\|}{\left\|e^{(0)}\right\|} \le \|B^k\| \end{aligned}$$

Teo. 15 . Un metodo iterativo stazionario consistente converge $\forall x_0 \in \mathbb{R}^n \iff \rho(B) < 1$ ($\rho(B) := \max_{\lambda\in\Lambda(B)} |\lambda|$)

Dimostrazione. Abbiamo:

$\implies$ Se fosse $\rho(B) \ge 1$ prendo autovalore λ con $|\lambda| \ge 1$ e $x^{(0)} = x - v$ con x sol. esatta e v autovettore associato, così

$$e^{(k+1)} = B^{k+1} \underbrace{e^{(0)}}_{=x-x^{(0)}=v} = B^{k+1}v = \lambda^{k+1}v$$

e quindi $\left\|e^{(k+1)}\right\| = |\lambda|^{k+1}\|v\|$ diverge per $k \to \infty$

$\impliedby$ $\rho(B) = \inf_{\|\cdot\|} \|B\|$, quindi $\exists\, \|\cdot\|_*$ per cui $\rho(B) \le \|B\|_* < 1$, dunque $\left\|e^{(k)}\right\|_* \le \|B\|_*^k \left\|e^{(0)}\right\|_* \overset{k\to\infty}{\to} 0$

□

Osservazione Se $\|B\| < 1 \implies$ il metodo converge e la convergenza è monotona (in quella norma), ovvero $\left\|e^{(k+1)}\right\| = \left\|Be^{(k)}\right\| \le \|B\| \cdot \left\|e^{(k)}\right\| < \left\|e^{(k)}\right\|$

metodo di Splitting Pongo

$$A = P - N$$

con P (invertibile) detto *precondizionatore*. Quindi si ha

$$(P - N)x = b \implies Px = Nx + b \implies x^{(k+1)} = P^{-1}\left(Nx^{(k)} + b\right)$$

Vale che

$$x^{(k+1)} = P^{-1}\left(Nx^{(k)} + b\right) = P^{-1}\left((P - A)x^{(k)} + b\right) = x^{(k)} + P^{-1}\left(b - Ax^{(k)}\right) = x^{(k)} + P^{-1}r^{(k)}$$

Quindi posso riscrivere come

$$\boxed{x^{(k+1)} = x^{(k)} + P^{-1}r^{(k)} = x^{(k)} + \delta^{(k)}}$$

dove $\boxed{r^{(k)} = b - Ax^{(k)}}$ è chiamato **residuo al passo** k. Ovvero a ogni passo prendo il vettore precedente e ci aggiungo la preimmagine (tramite P) del residuo.

Osservazione ***(Convergenza)*** Dato che $B = P^{-1}\underbrace{N}_{P-A} = I - P^{-1}A$, allora

$$\boxed{\text{metodo splitting converge} \iff \rho(I_n - P^{-1}A) < 1}$$

Una situazione ideale è quella in cui $I_n - P^{-1}A \approx 0$ ovvero $P \approx A$, ovvero P "simile" ad A, ma più facile da invertire.
Vediamo che $\sigma(I - P^{-1}A) = \{1 - \lambda,\ \lambda \in \sigma(P^{-1}A)\}$ quindi deve essere $|1 - \lambda| < 1 \quad \forall \lambda \in \sigma(P^{-1}A)$ e quindi

$$\text{Metodo converge} \iff \sigma(P^{-1}A) \subseteq \{z \in \mathbb{C} \mid |1 - z| < 1\}$$

Osservazione Il metodo di Splitting è consistente, $Ax = b \ \Rightarrow\ (P - N)\,x = b \ \Rightarrow\ x = P^{-1}\,(Nx + b)$.

Osservazione ***(Criterio d'arresto)*** L'algoritmo si ferma se

$$\frac{\left\|r^{(k)}\right\|_2}{\left\|r^{(0)}\right\|_2} \leq \eta \longrightarrow \text{ tolleranza fissata}$$

Se il criterio d'arresto è soddisfatto vale $r^{(k)} = b - Ax^{(k)} = A(\underbrace{A^{-1}b}_{x} - x^{(k)}) = Ae^{(k)}$

4.1.1 Metodi di Jacobi e di Gauss-Seidel

Considerando ora

$A = D - E - F \quad D$ diagonale $(d_{ii} = a_{ii}) \quad E$ strett. inferiore $(e_{ij} = -a_{ij}) \quad F$ strett. superiore $(f_{ij} = -a_{ij})$

metodo di Jacobi Preso $A = \overbrace{D}^{P} - \overbrace{(E + F)}^{N}$, ho

$$Dx^{(k+1)} = (E + F)\,x^{(k)} + b \qquad x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1, j\neq i}^{n} a_{ij}x_j^{(k)}\right) \qquad B_J = I - D^{-1}A$$

metodo di Gauss-Seidel Preso $A = \overbrace{D - E}^{P} - \overbrace{F}^{N}$, ho

$$(D - E)\,x^{(k+1)} = Fx^{(k)} + b \qquad x_i^{(k+1)} = \underbrace{\frac{1}{a_{ii}}}_{D^{-1}}\left(\underbrace{b_i - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}}_{Fx^{(k)}+b} - \underbrace{\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)}}_{Ex^{(k+1)}}\right) \qquad B_{GS} = (D - E)^{-1}\,F$$

Lemma 2 . $\|A\|_\infty = \sup_{x\in\mathbb{R}^n} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_i \sum_{j=1}^n |a_{ij}|$

Dimostrazione. □ p. 49

Definizione 4.4 (dominanza diagonale stretta per righe)**:** $|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| \qquad \forall\, i = 1, \ldots, n$

Teo. 19 . Se A è a dominanza diagonale stretta per righe per righe, il metodo di Jacobi e il metodo di Gauss-Seidel sono convergenti, cioè $Ax = b$ converge $\forall x^{(0)} \in \mathbb{R}^n$

Dimostrazione.

$$(B_J)_{ij} = D^{-1}\,(E + F) = \begin{cases} -\frac{a_{ij}}{a_{ii}} & \text{se } i \neq j \\ 0 & \text{se } i = j \end{cases} \qquad \|B_J\|_\infty = \max_i \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} = \max_i \frac{1}{a_{ii}} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| < 1$$

quindi $\rho(B_J) \leq \|B_J\|_\infty < 1 \iff$ converge □

Teo. 20 . Se A è SPD $\implies$ il metodo di Gauss-Seidel converge.

Teo. 21 . Se A è tridiagonale $\implies \rho(B_{GS}) = \rho^2(B_J)$

Cor. 4 . Se A è SPD e tridiagonale, il metodo di Jacobi converge.

*Osservazione (**Costo computazionale**)* Sia

$$x^{(k+1)} = x^{(k)} + \underbrace{P^{-1}\overbrace{\left(b - Ax^{(k)}\right)}^{=r^{(k)}}}_{=\delta^{(k)}}$$

quindi

- Calcolo del residuo: $b - Ax^{(k)} = r^{(k)} \implies 2n^2$
- Risoluzione di $P\delta^{(k)} = r^{(k)} \implies n$ (Jacobi) o n^2 (Gauss-Seidel)
- Calcolo di $x^{(k+1)} = x^{(k)} + \delta^{(k)}$

in tutto abbiamo $= O(n^2)$ flops per iterazione, quindi se il numero di iterazioni per avere la convergenza è $\ll n$ questi approcci sono convenienti rispetto alle $O(n^3)$ operazioni dei metodi diretti (come la LU)

Esempio Esempi di criteri di arresto

Tolleranza scelta $\left\|r^{(k)}\right\| \leq \eta$ (vale $\left\|e^{(k)}\right\|_2 \leq K(A)\left\|r^{(k)}\right\|_2$ Numero di iterazioni $k > N_{\max}$

4.1.2 Metodi di Richardson (NO)

Definizione 4.5 (metodo di Richardson)**:** $x^{(k+1)} = x^{(k)} + \alpha P^{-1} r^{(k)}$. P (invertibile) e $R_\alpha = I - \alpha P^{-1}A$

Osservazione Il metodo di Richardson è consistente, $x = x + \alpha P^{-1}\cancel{(b - Ax)}$.

Teo. 22 . Un metodo di Richardson converge $\forall\, x_0 \in \mathbb{R}^n \iff \frac{2\cdot \mathrm{Re}(\lambda)}{\alpha|\lambda|^2} > 1 \quad \forall\, \lambda \in \Lambda\left(P^{-1}A\right)$

Dimostrazione.

$$\rho(R_\alpha) = \max_{\lambda \in \Lambda(P^{-1}A)} |1 - \alpha\lambda| < 1 \qquad |1 - \alpha\lambda|^2 = (1 - \alpha \cdot \mathrm{Re}(\lambda))^2 + \alpha^2 \,\mathrm{Im}^2(\lambda) < 1 \qquad \alpha|\lambda|^2 < 2 \cdot \mathrm{Re}(\lambda)$$

□

Cor. 5 . Se gli autovalori di $P^{-1}A$ sono reali e positivi $\lambda_1 \geq \ldots \geq \lambda_n > 0$, la condizione diventa $0 < \alpha < \frac{2}{\lambda_1}$. Inoltre $\alpha_{\text{opt}} = \frac{2}{\lambda_1 + \lambda_n}$ minimizza il raggio spettrale, ovvero $\rho\left(R_{\alpha_{\text{opt}}}\right) = \min_\alpha \rho(R_\alpha) = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}$

Dimostrazione. Conseguenza diretta del teorema precedente. $\rho(R_\alpha) = \max(1 - \alpha\lambda_n, \alpha\lambda_1 - 1)$, da cui $\alpha_{\text{opt}} = \frac{2}{\lambda_1 + \lambda_n}$ □

Definizione 4.6 (condizionamento spettrale)**:** $\chi(M) = \frac{\lambda_1}{\lambda_n}$, con M di autovalori reali e positivi $\lambda_1 \geq \ldots \geq \lambda_n > 0$

Osservazione $\chi(M) \geq 1$, e se M SPD $\chi(M) = K_2(M) = \|M\|_2 \cdot \left\|M^{-1}\right\|_2$

Osservazione $\rho\left(R_{\alpha_{\text{opt}}}\right) = \frac{\chi\left(P^{-1}A\right) - 1}{\chi\left(P^{-1}A\right) + 1}$

Definizione 4.7 : Data A SPD, pongo $\|M\|_A = \sqrt{v^T A v}$

Teo. 23 . Siano P, A SPD e α che soddisfa la condizione di 5, allora $\left\|e^{(k+1)}\right\|_A \leq \rho(R_\alpha)\left\|e^{(k)}\right\|_A$.

Dimostrazione.

$$\left\|e^{(k+1)}\right\|_A = \left\|R_\alpha e^{(k)}\right\|_A = \left\|A^{1/2}R_\alpha e^{(k)}\right\|_2 = \left\|A^{1/2}R_\alpha A^{-1/2}A^{1/2}e^{(k)}\right\|_2 \leq \left\|A^{1/2}R_\alpha A^{-1/2}\right\|_2 \left\|A^{1/2}e^{(k)}\right\|_2$$
$$= \left\|I-\alpha A^{1/2}P^{-1}A^{-1/2}\right\|_A \left\|e^{(k)}\right\|_A$$
$$\left\|I-\alpha A^{1/2}P^{-1}A^{-1/2}\right\|_A = \rho\left(I-\alpha A^{1/2}P^{-1}A^{-1/2}\right) = \rho\left(I-\alpha P^{-1}A\right) = \rho\left(R_\alpha\right)$$

□

Osservazione Se $P=I$, allora $\rho\left(R_{\alpha_{\text{opt}}}\right) = \frac{\chi(A)-1}{\chi(A)+1} = \frac{K_2(A)-1}{K_2(A)+1}$, se $K_2(A) \gg 1$ la convergenza può essere lenta. Il precondizionatore serve a ridurre $\chi\left(P^{-1}A\right) \approx 1$

4.2 Metodo del gradiente (non stazionario)

Se A SPD, per risolvere $Ax=b$ basta trovare il minimo di

$$\Phi(y) = \frac{1}{2}y^TAy - y^Tb$$

siccome $\nabla\Phi(y) = \frac{1}{2}\left(A^T+A\right)y - b = Ay-b$, e dunque $\nabla\Phi(x) = 0 \Leftrightarrow Ax=b$.
Inoltre se in x il gradiente è nullo, tale punto è di minimo globale, infatti:

$$\Phi(y) = \Phi(x+(y-x)) = \frac{1}{2}x^TAx + \frac{1}{2}(y-x)^TA(y-x) + (y-x)^TAx - x^Tb - (y-x)^Tb =$$
$$= \Phi(x) + \underbrace{\frac{1}{2}(y-x)^T\cancel{(Ax-b)}}_{=0} + \underbrace{\frac{1}{2}(y-x)^TA(y-x)}_{>0} > \Phi(x) \qquad \Longrightarrow \qquad \frac{1}{2}\|y-x\|_A = \Phi(y) - \Phi(x)$$

Impostazione del metodo $x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$ dove $\begin{cases}\alpha_k \textbf{ passo}\\ \boldsymbol{p^{(k)}} \textbf{ direzione di spostamento}\end{cases}$ con

- $p^{(k)}$ da scegliere $\forall k$
- $\alpha_k = \arg\min_\alpha \Phi\underbrace{(\overbrace{x^{(k)}+\alpha_k p^{(k)}}^{x^{(k+1)}})}_{:=F(\alpha)}$

  poiché
  - $F(\alpha) = \frac{1}{2}(x^{(k)}+\alpha p^{(k)})^TA(x^{(k)}+\alpha p^{(k)}) - (x^{(k)}+\alpha p^{(k)})^Tb = \frac{\alpha^2}{2}p^{(k)T}Ap^{(k)} + \alpha p^{(k)^T}Ax^{(k)} - \alpha p^{(k)^T}b + \text{costanti}$ (**parabola**)
  - $F'(\alpha) = \alpha p^{(k)^T}Ap^{(k)} + p^{(k)^T}(Ax^{(k)}-b) = 0 \iff \alpha = \frac{p^{(k)T}r^{(k)}}{p^{(k)T}Ap^{(k)}}$

  Quindi sarà

$$\boxed{\alpha_k = \frac{p^{(k)T}r^{(k)}}{p^{(k)T}Ap^{(k)}} = \frac{(p^{(k)},r^{(k)})}{(p^{(k)},p^{(k)})_A}}$$

  quindi otteniamo l'algoritmo (bisogna scegliere $p^{(k)}$ per ogni k):

$$\alpha_k = \frac{p^{(k)T}r^{(k)}}{p^{(k)T}Ap^{(k)}}$$
$$x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$$
$$r^{(k+1)} \overset{\star}{=} r^{(k)} - \alpha_k Ap^{(k)}$$

  dove $\star = b - Ax^{(k+1)} = b - A(x^{(k)}+\alpha_k p^{(k)}) = b - Ax^{(k)} - \alpha_k Ap^{(k)} = \star$.

Osservazione Questa scelta di α_k implica ed è equivalente a $r^{(k+1)} \perp p^{(k)}$, infatti (scrivendo p per $p^{(k)}$)

$$(r^{(k+1)},p^{(k)}) = (r-\alpha_kAp,p) = (r,p) - \alpha_k(Ap,p) = (r,p) - \frac{(p,r)}{\cancel{(Ap,p)}}\cancel{(Ap,p)} = (r,p) - (r,p) = 0$$

metodo del gradiente Scegliamo

- $p^{(k)} = -\nabla\Phi\left(x^{(k)}\right) = b - Ax^{(k)} = r^{(k)}$ (direzione in cui Φ decresce più velocemente)
- $\alpha_k = \frac{r^{(k)T} r^{(k)}}{r^{(k)T} A r^{(k)}} = \frac{(r^{(k)}, r^{(k)})}{(r^{(k)}, r^{(k)})_A}$

quindi otteniamo l'algoritmo

$$\begin{aligned} \alpha_k &= \frac{r^{(k)T} r^{(k)}}{r^{(k)T} A r^{(k)}} \\ x^{(k+1)} &= x^{(k)} + \alpha_k r^{(k)} \\ r^{(k+1)} &\overset{\star}{=} r^{(k)} - \alpha_k A r^{(k)} \end{aligned}$$

dove $\star = b - Ax^{(k+1)} = b - A(x^{(k)} + \alpha_k r^{(k)}) = b - Ax^{(k)} - \alpha_k A r^{(k)} = \star$.

Osservazione Questa scelta di α_k implica ed è equivalente a $r^{(k+1)} \perp r^{(k)} = p^{(k)}$ è quindi $p^{(k+1)} \perp p^{(k)} = p^{(k)}$, infatti (scrivendo r per $r^{(k)}$)

$$(r^{(k+1)}, r^{(k)}) = (r - \alpha_k Ar, r) = (r,r) - \alpha_k (Ar, r) = (r,r) - \frac{(r,r)}{\cancel{(Ar,r)}} \cancel{(Ar,r)} = (r,r) - (r,r) = 0$$

ovvero ogni direzione è perpendicolare a quella precedente, ovvero con questo passo arrivo a tangere una curva di livello.
In questo modo la convegenza è molto lenta se $\frac{\lambda_1}{\lambda_n}$ è grande (immagina ellissi molto stirate).

Teo. 26 . Sia A SPD, allora per il metodo del gradiente vale $\left\|e^{(k+1)}\right\|_A \leq \frac{K_2(A)-1}{K_2(A)+1} \left\|e^{(k)}\right\|_A$.

Dimostrazione. Direttamente da 23 con $P = I$ e A SPD. □

4.2.1 Metodo del gradiente coniugato

Voglio $x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$ con

$$p^{(k)} = r^{(k)} + \overset{\text{diff. da prima}}{\boxed{\beta_k p^{(k-1)}}}$$

con β_k scelto in modo che $p^{(k-1)} A p^{(k)} = 0$ (A-ortogonale). Allora deve essere

$$p^{(k-1)} A p^{(k)} = p^{(k-1)} A (r^{(k)} + \beta_k p^{(k-1)}) = 0 \implies \beta_k = -\frac{p^{(k-1)T} A r^{(k)}}{p^{(k-1)T} A p^{(k-1)}} = \boxed{-\frac{(p^{(k)-1}, r^{(k)})_A}{(p^{(k-1)}, p^{(k-1)})_A}}$$

Prop. 9 . $\alpha_k = \frac{\left\|r^{(k)}\right\|_2^2}{p^{(k)T} A p^{(k)}}$ e $\beta_k = \frac{\left\|r^{(k+1)}\right\|_2^2}{\left\|r^{(k)}\right\|_2^2}$

Osservazione Ciò, come vedremo con spazi di Krylov, implica che $p^{(j)} A p^{(k)} = 0 \quad \forall j < k$, ovvero ogni nuova direzione è A-ortogonale a **tutte** quelle prima.

Definizione 4.8 (spazio di Krylov)**:** $V_k = \left\langle r^{(0)}, Ar^{(0)}, \ldots, A^{k-1} r^{(0)} \right\rangle$, $V_j \subseteq V_k$ per $j \leq k$. In generale sono definiti

$$K_k(A, \boldsymbol{v}) = \operatorname{span}\{\boldsymbol{v}, A\boldsymbol{v}, \ldots, A^{k-1}\boldsymbol{v}\}$$

Osservazione Abbiamo

- $V_k \subseteq V_{k+1}$
- $v \in V_k \implies Av \in V_{k+1}$
- Se $x^{(0)} = 0 \implies r^{(0)} = b - Ax^{(0)} = b$ e $V_k = \left\langle b, Ab, \ldots, A^{k-1} b \right\rangle$

Lemma 3 . $V_k = \langle r^{(0)}, \dots, r^{(k-1)} \rangle = \langle p^{(0)}, \dots, p^{(k-1)} \rangle$, in particolare

$$x^{(k+1)} = \sum_{i=0}^{k} \alpha_i p^{(i)} + x^{(0)} \in V_k + x^{(0)} \quad \text{(spazio affine)}$$

Dimostrazione. Per induzione su k

- $k = 1$ ovvio
- Dimostriamo che $V_k = \langle r^{(0)}, \dots, r^{(k-1)} \rangle = \langle p^{(0)}, \dots, p^{(k-1)} \rangle \implies V_{k+1} = \langle r^{(0)}, \dots, r^{(k)} \rangle = \langle p^{(0)}, \dots, p^{(k)} \rangle$. Abbiamo

$$\begin{cases} p^{(k-1)} \in \langle r^{(0)}, \dots, r^{(k-1)} \rangle & \text{ip. indutt.} \\ p^{(k)} = r^{(k)} + \beta_k p^{(k-1)} \end{cases} \implies \begin{cases} p^{(k)} = r^{(k)} + \beta_k p^{(k-1)} \in \langle r^{(0)}, \dots, r^{(k)} \rangle \\ r^{(k)} = p^{(k)} - \beta_k p^{(k-1)} \in \langle p^{(0)}, \dots, p^{(k)} \rangle \end{cases} \implies \text{tesi}$$

- $r^{(k)} = b - Ax^{(k)} = b - A(x^{(k-1)} + \alpha_{k-1} p^{(k)}) = \underbrace{r^{(k-1)}}_{\in V_k} - \alpha_k \underbrace{Ap^{(k-1)}}_{\in A_{k+1}} \in A_{k+1}$
- $A^k r^{(0)} = A(A^{k-1} r^{(0)}) \overset{\text{ip.ind.}}{=} A(\sum_{i=0}^{k-1} \alpha_i p^{(i)}) = \sum_{i=0}^{k-1} \alpha_i A p^{(i)} \in \langle r^{(0)}, \dots, r^{(k)} \rangle$
  Se $\alpha_i \neq 0 : Ap^{(i)} = \frac{r^{(i+1)} - r^{(i)}}{\alpha^{(i)}} \in \langle r^{(0)}, \dots, r^{(i+1)} \rangle$

□

Definizione 4.9 (*A*-ortogonalità a uno spazio)**:** $u \in \mathbb{R}^n, V \subseteq \mathbb{R}^\infty$ allora $u \perp_A V$ se $\forall v \in V : u^T A v = 0$ (A SPD)

Teo. 27 . Un metodo della forma $x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$ per cui valga l'A-ortogonalità a uno spazio soddisfa

$$\begin{cases} r^{(k)} \perp V_k \\ p^{(k)} \perp_A V_k \end{cases} \quad \forall k \in \mathbb{N}$$

Dimostrazione. Per induzione su k.

- Per $k = 1$, ho $V_1 = \langle p^{(0)} \rangle$ e $\begin{cases} r^{(1)} \perp p^{(0)} & \text{per def. di } \alpha_0 \\ p^{(1)} \perp_A p^{(0)} & \text{per def. di } \beta_0 \end{cases}$
- Supponendo $\begin{cases} r^{(k)} \perp V_k \\ p^{(k)} \perp_A V_k \end{cases} \iff \begin{cases} p^{(j)T} r^{(k)} = 0 & \forall j = 0, \dots, k-1 \\ p^{(j)T} A p^{(k)} = 0 & \forall j = 0, \dots, k-1 \end{cases}$, siccome $V_k = \langle p^{(0)}, \dots, p^{(k-1)} \rangle$
  - Dimostriamo $r^{(k+1)} \perp V_{k+1}$. Abbiamo

$$r^{(k+1)} \perp V_{k+1} \iff p^{(j)T} r^{(k+1)} = 0 \quad \forall j \leq k$$

    Vediamo che:

$$p^{(j)T} r^{(k+1)} = \begin{cases} p^{(k)T} r^{(k+1)} = 0 & \text{per costr. di } \alpha_k \qquad \text{se } j = k \\ p^{(j)T} r^{(k+1)} = p^{(j)T} \left(r^{(k)} - \alpha_k A p^{(k)} \right) = \cancel{p^{(j)T} r^{(k)}} - \alpha_k \cancel{p^{(j)T} A p^{(k)}} = 0 & \text{se } j < k \end{cases}$$

  - Dimostriamo $p^{(k+1)} \perp_A V_{k+1}$. Abbiamo

$$p^{(k+1)} \perp_A V_{k+1} \iff p^{(j)T} A p^{(k+1)} = 0 \quad \forall j \leq k$$

    Vediamo che:

$$p^{(j)T} A p^{(k+1)} = \begin{cases} p^{(k)T} A p^{(k+1)} = 0 & \text{per costr. di } \beta_k \qquad \text{se } j = k \\ p^{(j)T} A p^{(k+1)} = p^{(j)T} A \left(r^{(k+1)} + \beta_k p^{(k)} \right) = \cancel{p^{(j)T} A r^{(k+1)}} + \beta_k \cancel{p^{(j)T} A p^{(k)}} = 0 & \text{se } j < k \end{cases}$$

    dove il primo addendo è zero poiché per il lemma $p^{(j)} \in V_{j+1} \implies Ap^{(j)} \in V_{j+2} \subseteq V_{k+1}$ (poiché $j \leq k+1$). Quindi $p^{(j)T} A r^{(k+1)} = \overset{\in V_{k+1}}{(Ap^{(j)})}^T r^{(k+1)} = 0$ in quanto $r^{(k+1)} \perp V_k$

□

metodo del gradiente coniugato $x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$ con

$$p^{(k)} = \begin{cases} r^{(0)} & k = 0 \\ r^{(k)} + \beta_{k-1} p^{(k-1)} & k \geq 1 \end{cases} \qquad \beta_k = \frac{p^{(k)T} A r^{(k+1)}}{p^{(k)T} A p^{(k)}} \qquad \alpha_k = \frac{p^{(k)T} r^{(k)}}{p^{(k)T} A p^{(k)}}$$

β_k è tale che $p^{(k)T} A p^{(k+1)} = 0$, siccome $p^{(k)T} A p^{(k+1)} = p^{(k)T} A \left(r^{(k+1)} + \beta_k p^{(k)} \right) = p^{(k)T} A r^{(k+1)} + \beta_k p^{(k)T} A p^{(k)}$

Prop. 10 . Se $\underline{k}$ tale che $\dim\left(V_{\underline{k}}\right) = \dim\left(V_{\underline{k}+1}\right) = \underline{k}$ (equivalentemente $V_{\underline{k}} = V_{\underline{k}+1}$) $\implies x^{(\underline{k})} = x$ soluzione esatta

Dimostrazione. $V_{\underline{k}} = V_{\underline{k}+1} = \langle r_0, \dots, r_{\underline{k}} \rangle \implies r_{\underline{k}} \in V_{\underline{k}}$, ma $r_{\underline{k}} \perp V_{\underline{k}}$ quindi $r_{\underline{k}} = 0$ □

Cor. 6 . Il metodo del gradiente coniugato converge sempre in al più n iterazioni

Teo. 29 . Sia $\|\cdot\|_A$ la norma indotta da A (indotta dal prod. scalare $(x, y)_A = x^T A y$), allora

$$\left\| e^{(k)} \right\|_A = \left\| x - x^{(k)} \right\|_A = \min_{y \in x^{(0)} + V_k} \|x - y\|_A$$

ovvero ogni soluzione approssimata $x^{(k)}$ è il punto più vicino (rispetto alla distanza indotta da A) dello spazio $x^{(0)} + V_k$ alla soluzione esatta x.
In particolare $\left\| e^{(k+1)} \right\|_A \leq \left\| e^{(k)} \right\|_A$

Dimostrazione. $x^{(k)} = x^{(0)} + \sum_{j=0}^{k-1} \alpha_j p^{(j)} \in x^{(0)} + V_k$. Sia $y \in x^{(0)} + V_k$, $y \neq x^{(k)} \implies y - x^{(k)} \in V_k$. Allora

$$\|y - x\|_A^2 = \left\| \left(y - x^{(k)} \right) + \left(x^{(k)} - x \right) \right\|_A^2 = \left\| \underbrace{x^{(k)} - x}_{-e^{(k)}} \right\|_A^2 + \left\| y - x^{(k)} \right\|_A^2 - 2 \cancel{\left(y - x^{(k)} \right) \underbrace{A e^{(k)}}_{r^{(k)}}} \geq \left\| x^{(k)} - x \right\|_A^2$$

□

Prop. 11 . $V_k = \left\{ r^{(0)}, A r^{(0)}, \dots, A^{k-1} r^{(0)} \right\} = \left\{ p(A)\, r^{(0)} \mid p \in \mathbb{P}_{k-1} \right\}$ con $\mathbb{P}_k$ polinomi di deg $\leq k$

Dimostrazione. Per induzione su k □

Teo. 30 . Sia $\Pi_k := \{ p \in \mathbb{P}_k \mid p(0) = 1 \} = \{ a_0 + a_1 X + \cdots + a_k X^k : a_0 = 1 \}$, allora

$$\frac{\left\| e^{(k)} \right\|_A}{\left\| e^{(0)} \right\|_A} = \min_{p \in \Pi_k} \frac{\left\| p(A)\, e^{(0)} \right\|_A}{\left\| e^{(0)} \right\|_A} \leq \min_{p \in \Pi_{k-1}} \left(\max_{\lambda \in \Lambda(A)} |p(\lambda)| \right)$$

Dimostrazione. □ p. 56

Cor. 7 . Se A ha l autovalori distinti, il metodo del gradiente coniugato converge in al più l iterazioni.

Dimostrazione. □ p. 57

Teo. 31 . Vale

$$\min_{p \in \Pi_{k-1}} \left(\max_{\lambda \in \Lambda(A)} |p(\lambda)| \right) \leq 2 \left(\frac{\sqrt{K_2(A)} - 1}{\sqrt{K_2(A)} + 1} \right)^k$$

e dunque in particolare

$$\left\| e^{(k)} \right\|_A \leq 2 \left(\frac{\sqrt{K_2(A)} - 1}{\sqrt{K_2(A)} + 1} \right)^k \left\| e^{(0)} \right\|_A$$

Osservazione Il metodo del gradiente coniugato converge più rapidamente del metodo del gradiente siccome $\sqrt{K_2(A)} \leq K_2(A)$ (26 e 31).
Se $K(A) \approx 1$ la convergenza è rapida, se $K(A) \gg 1$ la convergenza potrebbe essere lenta

4.2.2 Metodo del gradiente coniugato precondizionato

Provo ora a risolvere con il metodo del gradiente coniugato $P^{-1/2}AP^{-1/2}\widetilde{x} = P^{-1/2}b$ sostituendo $\widetilde{x}^{(k)} = P^{-1/2}$, $x^{(k)}\widetilde{r}^{(k)} = P^{-1/2}r^{(k)}$ e $\widetilde{p}^{(k)} = P^{-1/2}p^{(k)}$

metodo del gradiente coniugato precondizionato $x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}$ con

$$p^{(k)} = \begin{cases} r^{(0)} & k = 0 \\ r^{(k)} + \beta_{k-1}p^{(k-1)} & k \geq 1 \end{cases} \qquad z^{(k)} = P^{-1}r^{(k)} \qquad \beta_k = \frac{r^{(k+1)T}z^{(k+1)}}{r^{(k)T}r^{(k)}} \qquad \alpha_k = \frac{r^{(k)T}z^{(k)}}{p^{(k)T}Ap^{(k)}}$$

4.3 Precondizionatori algebrici, Sparsità

Considerando ora

$$A = D - E - F \qquad D \text{ diagonale} \qquad E \text{ strett. inferiore} \qquad F \text{ strett. superiore}$$

Esempio Esempi di precondizionatori algebrici

$$\text{Jacobi } P_J = D \qquad \text{Gauss-Seidel simmetrico } P_{SGS} = (D-E)\,D^{-1}\,(D-E)^T$$
$$\text{Gauss-Seidel } P_{GS} = D - E \qquad \text{Cholesky incompleto} A = \hat{R}^T\hat{R}$$

5 Calcolo di Autovalori e Autovettori

Dato che non esiste una formula per calcolare le radici di un polinomio di grado maggiore o uguale a 5 (teo.), tutti i metodi sono **iterativi**.

5.1 Localizzazione geometrica degli autovalori

Siccome $|\lambda| \leq \|A\| \quad \forall\, \lambda \in \Lambda(A)$, tutti gli autovalori si trovano in un cerchio di raggio $\|A\|$ centrato nell'origine

Teo. 33 (teorema dei cerchi di Gershgorin)**.** Data $A \in \mathbb{C}^{n\times n}$, allora

$$\Lambda(A) \subseteq S_R \coloneqq \bigcup_{i=1}^{n} R_i \qquad \text{con} \quad R_i \coloneqq \left\{ z \in \mathbb{C} \mid |z - a_{ii}| \leq \textstyle\sum_{j=1, j\neq i}^{n} |a_{ij}| \right\}$$

R_i detti cerchi riga. Analogamente $\Lambda(A) \subseteq S_C$ per le colonne (S_R per A^T)

Dimostrazione. Sia $\lambda \in \Lambda(A)$. Se $\lambda = a_{ii}$ per qualche i ho finito. Preso $\lambda \neq a_{ii}$, il suo autovettore v e $D = \operatorname{diag}(A)$ diagonale di A, vale

$$\begin{aligned} Av &= \lambda v \\ Av - Dv &= \lambda v - Dv \\ (A-D)v &= (\lambda I - D)v \quad \text{con } (\lambda I - D) \text{ diag. e invertibile} \\ v &= (\lambda I - D)^{-1}(A-D)v \end{aligned}$$

da cui

$$\|v\|_\infty \leq \left\| (\lambda I - D)^{-1}(A-D) \right\|_\infty \|v\|_\infty \quad \overset{\|v\|\neq 0}{\Longrightarrow} \quad 1 \leq \left\| (\lambda I - D)^{-1}(A-D) \right\|_\infty \overset{\exists i}{=} \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|a_{ij}|}{|\lambda - a_{ii}|} = \frac{1}{|\lambda - a_{ii}|} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|$$

nella sommatoria abbiamo messo $j \neq i$ poiché per $j = i$ il numeratore è 0 essendo elemento della matrice $A - D$ che ha la diagonale vuota. □

Cor. 8 . Una matrice a dominanza diagonale stretta per righe è non singolare

Dimostrazione. Siccome $|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|$, nessun R_i interseca l'origine e dunque 0 non può essere autovalore □

Teo. 34 (primo teorema di Gershgorin)**.** $\Lambda(A) \subseteq S_R \cap S_C$

Dimostrazione. $\Lambda(A) = \Lambda(A^T)$ □

Teo. 35 (secondo teorema di Gershgorin)**.** Se $\exists\, 1 \leq m \leq n$ tale che $S_1 \coloneqq \bigcup_{i=1}^{m} R_i$ e $S_2 \coloneqq \bigcup_{i=m+1}^{n} R_i$ sono disgiunti, allora S_1 contiene esattamente m autovalori (con molteplicità) e S_2 ne contiene $m - n$. Analogamente per le colonne.

Cor. 9 . Un cerchio di Gershgorin disgiunto dagli altri contiene esattamente 1 autovalore, e se $A \in \mathbb{R}^{n\times n}$ tale autovalore è reale (poiché se fosse complesso dovrebbe contenere anche il coniugato)

5.2 Analisi del condizionamento

$\widetilde{A} = A + E$ con E matrice di perturbazione ($\|E\| \ll \|A\|$)

Teo. 36 (teorema di Bauer-Fike)**.** $A \in \mathbb{C}^{n\times n}$ diagonalizzabile e $E \in \mathbb{C}^{n\times n}$, allora $\forall\, \mu \in \Lambda(A + E)$ vale

$$\min_{\lambda \in \Lambda(A)} |\lambda - \mu| \leq K_p(X) \|E\|_p$$

con X matrice con colonne gli autovettori di A (e dunque $K_p(X) = \|X\|_p \cdot \left\|X^{-1}\right\|_p$).

Dimostrazione. Preso $\lambda \notin \Lambda(A)$, autovettore v, e posto $v = X\hat{v}$ e $D = X^{-1}AX$ diagonale

$$(A+E)\,v = \mu v \qquad X^{-1}(A+E)\,X\hat{v} = \mu\hat{v} \qquad \left(D + X^{-1}EX\right)\hat{v} = \mu\hat{v} \qquad (\mu I - D)\,\hat{v} = X^{-1}EX\hat{v}$$

$$\|\hat{v}\|_p = \left\|(\mu I - D)^{-1} X^{-1}EX\hat{v}\right\|_p \leq \left\|(\mu I - D)^{-1}\right\|_p \|E\|_p \underbrace{\left\|X^{-1}\right\|_p \|X\|_p}_{K_p(X)} \|\hat{v}\|_p$$

□

5.3 Metodi delle potenze

Data $A \in \mathbb{C}^{n\times n}$ diagonalizzabile di autovalori $|\lambda_1| \geq \ldots \geq |\lambda_n| \geq 0$ e base di autovettori unitari $\mathscr{B} = \{x_1, \ldots, x_n\}$

metodo delle potenze Per determinare l'autovalore di modulo max: $q^{(k+1)} = \dfrac{Aq^{(k)}}{\left\|Aq^{(k)}\right\|_2} = \dfrac{A^k q^{(0)}}{\left\|A^k q^{(0)}\right\|_2}$ con $\left\|q^{(0)}\right\|_2 = 1$ e $\nu^{(k)} = \left(q^{(k)}\right)^* Aq^{(k)}$

$$q^{(0)} = \sum_{i=1}^{n} \alpha_i x_i \qquad\qquad A^k q^{(0)} = \sum_{i=1}^{n} \alpha_i A^k x_i = \sum_{i=1}^{n} \alpha_i \lambda_i^k x_i$$

Definizione 5.1 (*O*-grande)**:** Se $g(x)$ è definitivamente $\neq 0$ per $x \to x_0$ allora $f(x) = O(g(x))$ significa $\frac{f(x)}{g(x)}$ definitivamente limitato per $x \to x_0$, ovvero

$$\frac{f(x)}{g(x)} \leq c \iff f(x) \leq cg(x) \quad \text{defin. per } x \to x_0$$

Teo. 38 (Ordine di convergenza)**.** Se $|\lambda_1| > |\lambda_2|$ e $\alpha_1 \neq 0$, allora

$$\min_{\substack{x\in\langle x_1\rangle \\ \|x\|_2=1}} \left\|x - q^{(k)}\right\|_2 \leq c\left|\frac{\lambda_2}{\lambda_1}\right|^k = O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) \qquad\qquad \left|\lambda_1 - \nu^{(k)}\right| = O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)$$

Il min nella tesi serve perché so che $q^{(k)}$ tende alla **direzione** dell'autovalore max (ovvero all'autospazio $\langle x_1\rangle$) ma se l'autovalore è negativo continua a oscillare avvicinandosi a $+x_1$ o $-x_1$, quindi l'errore lo calcolo da quello più vicino dei due.

Dimostrazione. **(I)** Scompongo

$$A^k q^{(0)} = \sum_{i=1}^{n} \alpha_i \lambda_i^k x_i = \alpha_1 \lambda_1^k (x_1 + \overbrace{\sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1}\left(\frac{\lambda_i}{\lambda_1}\right)^k x_i}^{W^{(k)}})$$

Ho che

$$\begin{cases} x^{(k)} := \frac{\alpha_1\lambda_1^k}{|\alpha_1\lambda_1^k|} x_1 \in \langle x^{(1)}\rangle, \quad \left\|x^{(k)}\right\|_2 = 1 \\ q^{(k)} = \frac{A^k q^{(0)}}{\|A^k q^{(0)}\|} = \frac{\alpha_1\lambda_1^k}{|\alpha_1\lambda_1^k|} \frac{x^{(1)}+W^{(k)}}{\|x^{(1)}+W^{(k)}\|} \end{cases}$$

$$\begin{aligned} \left\|q^{(k)} - x^{(k)}\right\|_2 &= \left\|\left(\frac{x_1 + W^{(k)}}{\|x_1 + W^{(k)}\|} - x_1\right)\frac{\alpha_1\lambda_1^k}{|\alpha_1\lambda_1^k|}\right\|_2 \\ &= \left\|\frac{x_1 + W^{(k)}}{\|x_1 + W^{(k)}\|} - x_1\right\|_2 \\ &\leq \left|\frac{1}{\|x_1 + W^{(k)}\|} - 1\right| + \frac{\|W^{(k)}\|}{\|x_1 + W^{(k)}\|} \quad \text{ho raccolto } x^{(1)} \text{ che ha norma 1} \\ &\overset{\star}{=} O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) \quad \star \text{ per dim. vedi sotto} \end{aligned}$$

Vediamo che $\left\|W^{(k)}\right\| = O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)$:

$$\left\|W^{(k)}\right\| = \left\|\sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1}\left(\frac{\lambda_i}{\lambda_1}\right)^k x^{(i)}\right\| = \left\|\sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1}\left(\frac{\lambda_i}{\lambda_1}\right)^k\right\| \leq \left|\frac{\lambda_i}{\lambda_1}\right|^k \sum_{i=2}^{n}\left|\frac{\alpha_i}{\alpha_1}\right| = c\left|\frac{\lambda_i}{\lambda_1}\right|^k$$

Quindi valgono (ricordando la def. di O grande):

$$\begin{cases} \left\|x^{(1)} + W^{(k)}\right\| = 1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) \\ o\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) / \left(1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right) = O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) \end{cases}$$

e così otteniamo la tesi.
(II) Abbiamo (* =trasposto coniugato):

$$\begin{cases} \nu^{(k)} = q^{(k)*} A q^{(k)} = q^{(k)*} A(q^{(k)} - x^{(1)}) + q^{(k)*} A x^{(1)} \\ \lambda_1 = x^{(1)*} A x^{(1)} \end{cases} \implies \nu^{(k)} - \lambda_1 = q^{(k)*} A(q^{(k)} - x^{(1)}) + (q^{(k)} - x^{(1)})^* A x^{(1)}$$

Quindi

$$\begin{aligned} \left|\nu^{(k)} - \lambda_1\right| &\leq \left|q^{(k)*} A(q^{(k)} - x^{(1)})\right| + \left|(q^{(k)} - x^{(1)})^* A x^{(1)}\right| \\ &\leq \underbrace{\left\|q^{(k)}\right\|}_{=1} \|A\| \underbrace{\left\|q^{(k)} - x^{(1)}\right\|}_{o\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)} + \underbrace{\left\|q^{(k)} - x^{(1)}\right\|}_{o\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)} \|A\| \underbrace{\left\|x^{(1)}\right\|}_{=1} \end{aligned}$$

□

Osservazione La convergenza è rapida se $\lambda_1 \gg \lambda_2$

Cor. 10 (Velocità di convergenza per mat. reali simm.)**.** Se $|\lambda_1| > |\lambda_2|$ e A è inoltre **reale e simmetrica**, allora

$$\left|\lambda_1 - \nu^{(k)}\right| = O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right)$$

Se $|\lambda_1| = |\lambda_2|$ allora prendere il **primo autovalore in modulo minore di** λ_1.

Dimostrazione. □ p. 62

*Osservazione (**Criterio d'arresto**)* Se (λ, v) autocoppia di A, allora $Av = \lambda v \iff \|Av - \lambda v\| = 0$. Allora se $(\nu^{(k)}, q^{(k)})$ approssimano (λ, v) abbiamo $\left\|Aq^{(k)} - \nu^{(k)} q^{(k)}\right\| \approx 0$ e quindi un criterio d'arresto può essere

$$\left\|r^{(k)}\right\|_2 = \left\|Aq^{(k)} - \nu^{(k)} q^{(k)}\right\|_2 \leq \varepsilon$$

Presa $E^{(k)} := -r^{(k)} \left(q^{(k)}\right)^*$ si ha

$$\left(A + E^{(k)}\right) q^{(k)} = Aq^{(k)} - r^{(k)} = \nu^{(k)} q^{(k)} \qquad \left\|E^{(k)}\right\|_2 = \left\|r^{(k)}\right\|_2 \left\|q^{(k)}\right\|_2 = \left\|r^{(k)}\right\|_2 \leq \varepsilon$$

E quindi per il teorema di Bauer-Fike $\min_{\lambda \in \Lambda(A)} \left|\lambda - \nu^{(k)}\right| \leq K_2(X) \|E\|_2 \leq \varepsilon \cdot K_2(X)$

*Osservazione (**Costo computazionale**)* $z^{(k)} = Aq^{(k-1)}$: $2n^2$ flops se A densa, $2 \cdot \mathrm{nnz}(A)$ flops se A sparsa (con $\mathrm{nnz}(A)$ i non zeri di A)

metodo delle potenze inverse Per determinare l'autovalore più vicino a μ (detto **shift**) applico il metodo delle potenze su $(A-\mu I)^{-1}$ infatti ha gli stessi autovettori:

$$\begin{aligned}(\lambda, v) \text{ autocoppia di } A &\iff Av = \lambda v \\ &\iff Av - \mu v = \lambda v - \mu v \\ &\iff (A-\mu I)v = (\lambda - \mu)v \\ &\iff (A-\mu I)^{-1}v = \frac{1}{(\lambda-\mu)}v \\ &\iff \left(\frac{1}{\lambda-\mu}, v\right) \text{ autocoppia di } (A-\mu I)^{-1}\end{aligned}$$

e quindi l'autovalore di modulo massimo di $(A-\mu I)^{-1}$ corrisponde all'autovalore di A più vicino a μ

*Osservazione (**Ordine di convergenza**)* Il Ordine di convergenza assicura che se $\exists m \in \{1, \dots, n\}$ t.c. $|\mu - \lambda_m| < |\mu - \lambda_i| \quad \forall i \neq m$ allora $\begin{cases} q^{(k)} \to x_m \\ \nu^{(k)} \to \lambda_m \end{cases}$.
L'ordine di convergenza è

$$O\left(\frac{|\lambda_m - \mu|}{|\lambda_t - \mu|}\right)^k \quad \text{dove } |\lambda_t - \mu| \geq |\lambda_i - \mu| \quad \forall i \neq m$$

Osservazione Nel caso in cui $|\lambda_1| = |\lambda_2|$ il metodo delle potenze

- per $\lambda_1 = \lambda_2$ converge ancora, siccome $A^k q^{(0)} = \sum_{i=1}^n \alpha_i \lambda_i^k x_i \approx \lambda_1^k (\alpha_1 x_1 + \alpha_2 x_2)$
- per $\lambda_1 \neq \lambda_2$ la convergenza non è garantita

5.4 Metodi basati sulla fattorizzazione QR (NO)

metodo QR Per approssimare tutti gli autovalori: Fattorizzato $T_{k-1} = Q_k R_k$, scrivo $T_k = R_k Q_k$ ($T_0 = A$).

Osservazione Vale che $T_k = (Q_1 Q_2 \dots Q_k)^T A (Q_1 Q_2 \dots Q_k)$, quindi tutti i T_k sono simili a A attraverso matrici ortogonali. Inoltre la stabilità non cambia, in quanto

$$\widetilde{T_k} = \widehat{Q}_k^T \widetilde{A} \widehat{Q}_k^T = \widehat{Q}_k^T (A+E) \widehat{Q}_k^T = T_k + \widehat{Q}_k^T E \widehat{Q}_k^T \qquad \left\| \widehat{Q}_k^T E \widehat{Q}_k^T \right\|_2 = \|E\|_2$$

Teo. 41 . Presa $A \in \mathbb{R}^{n \times n}$ di autovalori $|\lambda_1| > \dots > |\lambda_n| > 0$, allora

$$\lim_{k \to +\infty} T_k = \begin{bmatrix} \lambda_1 & t_{1\,2} & \cdots & t_{1\,n} \\ 0 & \lambda_2 & \cdots & t_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

e se A è simmetrica, T_k converge a una diagonale

Lemma 4 . $A^k = Q_1 \cdots Q_k R_k \cdots R_1 \quad \forall k \geq 1$

Dimostrazione. Per induzione, dato $A^k = Q_1 \cdots Q_k R_k \cdots R_1$ ho che

$$Q_1 \cdots Q_{k+1} R_{k+1} \cdots R_1 = Q_1 \cdots Q_k T_{k+1} R_k \cdots R_1 = Q_1 \cdots Q_{k-1} T_k Q_k R_k \cdots R_1 = \dots = \overbrace{T_0}^{A} \overbrace{Q_1 \cdots Q_k R_k \cdots R_1}^{A^k} = A^{k+1}$$

□

Dimostrazione del teorema. Prendo $A = X \Lambda Y$, con Λ diagonale con gli autovalori in ordine decrescente, X colonne autovettori, $Y = X^{-1}$, $X = QR$ e $Y = LU$, da cui

$$Q_1 \cdots Q_k R_k \cdots R_1 = A^k = X \Lambda^k Y = QR \Lambda^k LU = QR \left(\overbrace{\Lambda^k L \Lambda^{-k}}^{\longrightarrow I \text{ per } k \to +\infty} \right) \Lambda^k U$$

$$\underbrace{Q_1 \cdots Q_k}_{\hat{Q}_k} \underbrace{R_k \cdots R_1 U^{-1} \Lambda^{-k} R^{-1}}_{\hat{R}_k} D = Q \overbrace{R \Lambda^k L \Lambda^{-k}}^{\longrightarrow I \text{ per } k \to +\infty} R^{-1} D \qquad \text{dunque} \begin{cases} \hat{Q}_k \longrightarrow QD \\ \hat{R}_k D \longrightarrow I \end{cases}$$

$$T_k = \left(\underbrace{Q_1 \cdots Q_k}_{\hat{Q}_k} \right)^T T_0 \left(\underbrace{Q_1 \cdots Q_k}_{\hat{Q}_k} \right) = \hat{Q}_k^T X \Lambda X^{-1} \hat{Q}_k = \underbrace{\hat{Q}_k^T Q}_{\longrightarrow D} R \Lambda R^{-1} \underbrace{Q^T \hat{Q}_k}_{\longrightarrow D} \longrightarrow DR\Lambda R^{-1} D$$

□

Osservazione Il metodo QR può essere visto come una serie di metodi delle potenze in cui parto da una matrice I moltiplico le colonne per A e poi a ogni passaggio ortogonalizzo.

Definizione 5.2 (matrice di Hessenberg)**:** H di Hessenberg se $h_{ij} = 0 \quad \forall\, i > j+1$

metodo di Hessenberg-QR Trasformo A in una matrice di Hessenberg tramite 44 e poi applico il metodo QR (la complessità computazionale per la fattorizzazione QR di una matrice di Hessenberg è di $O\left(n^2\right)$)

Definizione 5.3 (trasformazioni di Householder)**:**

$$P = I - 2\frac{vv^T}{\|v\|_2^2} \qquad v = x \pm e_m \|x\|_2 \qquad Px = \pm e_m \|x\|_2 = \begin{pmatrix} 0 & \dots & 0 & \|x\|_2 & 0 & \dots & 0 \end{pmatrix}^T$$

triangolarizzazione di Householder $Q_{n-1} \dots Q_1 A = R$ con $Q_j = \begin{bmatrix} I_{j-1} & 0 \\ 0 & P_j \end{bmatrix}$ e P_j matrice di Householder opportuna

$$\underset{A}{\begin{bmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{bmatrix}} \longrightarrow \underset{Q_1 A}{\begin{bmatrix} \times & \times & \times \\ \bigcirc & \times & \times \\ \bigcirc & \times & \times \end{bmatrix}} \longrightarrow \underset{Q_1 Q_2 A}{\begin{bmatrix} \times & \times & \times \\ \bigcirc & \times & \times \\ \bigcirc & \bigcirc & \times \end{bmatrix}}$$

riduzione a matrice di Hessenberg Devo trovare $T_0 = Q_0^T A Q_0$, $Q_0 = P_1 \dots P_{n-2}$ con $Q_k = \begin{bmatrix} I_k & 0 \\ 0 & \widehat{P_k} \end{bmatrix}$ e $\widehat{P_k}$ matrice di Householder

$$\underset{A}{\begin{bmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{bmatrix}} \longrightarrow \underset{P_1^T A P_1}{\begin{bmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \bigcirc & \times & \times & \times \\ \bigcirc & \times & \times & \times \end{bmatrix}} \longrightarrow \underset{P_2^T P_1^T A P_1 P_2}{\begin{bmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \bigcirc & \times & \times & \times \\ \bigcirc & \bigcirc & \times & \times \end{bmatrix}}$$

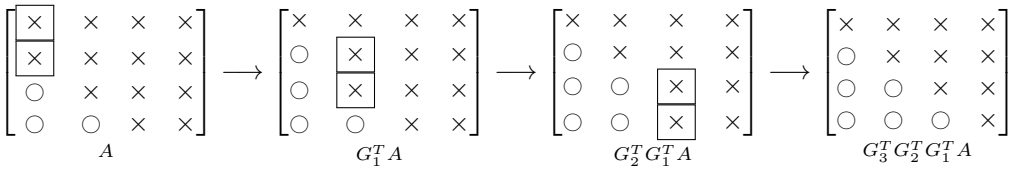
Definizione 5.4 (matrici elementari di Givens)**:** $i, k \in \{1, \dots, n\}$ e $\theta \in [0, 2\pi]$, $G(i,k,\theta)$ ortogonale è la rotazione di θ sul piano $\langle e_i, e_k \rangle$

$$G(i,k,\theta) = \begin{bmatrix} I_{i-1} & & & & \\ & \cos\theta & & \sin\theta & \\ & & I_{k-i-1} & & \\ & -\sin\theta & & \cos\theta & \\ & & & & I_{n-k} \end{bmatrix}$$

Osservazione $G(i,k,\theta)$ permette di annullare 1 componente di un vettore lasciandone inalterate $n-2$

$$\theta = \arctan\left(-\frac{x_k}{x_i}\right) \qquad \begin{bmatrix} y_i \\ y_k \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_i \\ x_k \end{bmatrix} = \begin{bmatrix} \sqrt{x_i^2 + x_k^2} \\ 0 \end{bmatrix}$$

fattorizzazione QR di una matrice di Hessenberg A matrice di Hessenberg, $R = Q^T A$ con $Q^T = G_{n-1}^T \dots G_1^T$ (applico le matrici di Givens per annullare l'ultima componente per ogni colonna)



Teo. 46 (formula di Shermann-Morris). $A \in \mathbb{R}^{n\times n}$ invertibile e $u, v \in \mathbb{R}^n$, $A+uv^T$ invertibile $\iff 1+v^T A^{-1}u \neq 0$, e

$$\left(A + uv^T\right)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1} u}$$

Parte II
Secondo semestre

6 Zeri di funzione e ottimizzazione

Teo. 47 (Condizionamento del problema di ottimizzazione). Radice di molteplicità m,

$$K_{\text{abs}}(d) \simeq \left| \frac{m! \cdot \delta d}{f^{(m)}(\alpha)} \right|^{1/m} \frac{1}{|\delta d|}$$

In particolare per $m = 1$ abbiamo

$$K_{\text{abs}}(d) \simeq \frac{1}{|f'(x)|} \qquad K(d) \simeq \frac{1}{|f'(x)|} \frac{|d|}{|\alpha|}$$

Dimostrazione.

$$\cancel{\alpha} + \delta d = f(\alpha + \delta\alpha) = \cancel{f(\alpha)} + \sum_{k=1}^{m} \frac{f^{(k)}(\alpha)}{k!} (\delta\alpha)^k + o(\|\delta\alpha\|^m) \overset{\star}{\Longrightarrow} \delta d = \frac{f^{(m)}(\alpha)}{m!} (\delta\alpha)^m$$

$$\Longrightarrow \delta\alpha = \frac{m! \cdot \delta d}{f^{(m)}(\alpha)}^{1/m}$$

$\star$ dato che le derivate $k = 1 \dots m-1$ si annullano per ipotesi. Allora

$$K_{\text{abs}}(d) \simeq \frac{|\delta\alpha|}{|\delta d|} \simeq \left| \frac{m! \cdot \delta d}{f^{(m)}(\alpha)} \right|^{1/m} \frac{1}{|\delta d|}$$

□

Osservazione Per $|f'(x)|$ piccola il problema è mal condizionato.

Definizione 6.1 (Ordine di convergenza)**:** $x^{(k)}$ converge ad α con ordine p se:

$$\frac{\left|x^{(k+1)} - \alpha\right|}{\left|x^{(k)} - \alpha\right|^p} \leq C$$

definitivamente in k

Impostazione del metodo $f\colon [a,b] \to \mathbb{R}$ continua, $f(\alpha) = 0$, per il teo. di Lagrange $\exists \xi$ t.c.

$$\frac{f(\alpha) - f(x)}{\alpha - x} = f'(\xi) \quad \xi \in [\alpha, x]$$

da cui otteniamo, essendo $f(\alpha) = 0$

$$f(x) + (\alpha - x) f'(\xi) = 0$$

che suggerisce il metodo iterativo: dato $x^{(k)}$ si calcola $x^{(k+1)}$ risolvendo

$$f(x^{(k)}) + (x^{(k+1)} - x^{(k)}) q_k = 0$$

dove q_k è opportuna approssimazione di $f'(x^{(k)})$. Ciò equivale a trovare il punto di intersezione tra asse x e retta di pendenza q_k passante per $(x^{(k)}, f(x^{(k)}))$, scritta in forma più esplicita (voglio $y = 0$)

$$y - f(x^{(k)}) = q_k(\overset{\text{incognita}}{\boxed{x^{(k+1)}}} - x^{(k)})$$

esplicitando l'incognita:

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

Metodo di Bisezione IPOTESI: $f\colon [a,b] \to \mathbb{R}$ continua, $f(a) \cdot f(b) < 0$

Vari metodi

Metodo delle corde	$q_k = \frac{f(b)-f(a)}{b-a}$	(con $f(a)\cdot f(b)<0$)	ordine 1
Metodo delle secanti	$q_k = \frac{f(x^{(k)})-f(x^{(k-1)})}{x^{(k)}-x^{(k-1)}}$	(assegnati i due dati iniziali $x^{(-1)}, x^{(0)}$)	ordine $\frac{1+\sqrt{5}}{2}$
Regula falsi	$q_k = \frac{f(x^{(k)})-f(x^{(k')})}{x^{(k)}-x^{(k')}}$	(con k' più grande $< k$ tale per cui $f(x^{(k)})\cdot f(x^{(k')})<0$)	
Metodo di Newton	$q_k = f'(x^{(k)})$	(con $f\in C^1(J)$, $f'(x)\neq 0\ \forall x\in J$)	ordine 2 (se rad. sempl.)

Teo. 51 (Teorema di convergenza locale)**.** Sia $f\colon I\to\mathbb{R}$, $x_0\in I$, $f\in C^2$, se $\exists M$ tale che $\left|\frac{f''(x)}{f'(y)}\right|\leq M$ $\forall\,(x,y)\in I^2$, allora $|x_0-\alpha|<\frac{2}{M}$ (con α zero di f) $\Longrightarrow$ Metodo di Newton converge con ordine 2.

Dimostrazione. Sviluppo di Taylor in $x^{(k)}$ al primo ordine con errore di Lagrange:

$$f(x)=f(x^{(k)})+f'(x^{(k)})(x-x^{(k)})+\frac{f''(\xi)}{2}\left(x-x^{(k)}\right)^2$$

In particolare vale per $x=\alpha$:

$$\begin{aligned}0=f(\alpha)&=f(x^{(k)})+f'(x^{(k)})(\alpha-x^{(k)})+\frac{f''(\xi)}{2}\left(\alpha-x^{(k)}\right)^2\\&=\underbrace{\boxed{\frac{f\left(x^{(k)}\right)}{f'\left(x^{(k)}\right)}-x^{(k)}}}_{-x^{(k+1)}}+\alpha+\frac{f''(\xi)}{2f'\left(x^{(k)}\right)}\left(\alpha-x^{(k)}\right)^2\quad\text{ho raccolto } f'(x^{(k)})\end{aligned}$$

Quindi

$$\frac{\left(\alpha-x^{(k+1)}\right)}{\left(\alpha-x^{(k)}\right)^2}\overset{\star}{=}-\frac{1}{2}\frac{f''(\xi)}{f'\left(x^{(k)}\right)}\quad\text{ovvero}\quad|e_{k+1}|\leq\frac{1}{2}\left|\frac{f''(\xi)}{f'\left(x^{(k)}\right)}\right|e_k^2$$

Dimostriamo che converge: abbiamo, con J intorno di α

$$\begin{aligned}f\in C^2(J)&\Longrightarrow\frac{1}{2}\left|\frac{f''(x)}{f'(y)}\right|\leq M\quad\text{per un } M \text{ e } \forall x,y\in J\\&\Longrightarrow|e_{k+1}|\leq Me_k^2\\&\Longrightarrow M|e_{k+1}|\leq(Me_k)^2\leq(Me_0)^{2k+1}\\&\Longrightarrow|e_{k+1}|\leq\frac{1}{M}(Me_0)^{2k+1}\\&\Longrightarrow|e_k|\leq\frac{1}{M}(Me_0)^{2k}\overset{k\to+\infty}{\longrightarrow}0\qquad\text{se } |Me_0|<1\iff|e_0|<\frac{1}{M}\end{aligned}$$

quindi se $|e_k|=|\alpha-x_0|$ è abbastanza piccolo il metodo converge e da $\star$

$$\frac{\left|\alpha-x^{(k+1)}\right|}{\left|\alpha-x^{(k)}\right|^2}=\left|\frac{1}{2}\frac{f''(\xi)}{f'\left(x^{(k)}\right)}\right|\leq M\implies\text{converge di ordine 2}$$

□

6.1 Metodo di Punto Fisso

Data $f:[a,b]\to\mathbb{R}$ la posso scomporre in $f(x)=x-\phi(x)$ e quindi $\phi(x)=x-f(x)$ e ottengo

$$f(x)\iff x-\phi(x)=0\iff\phi(x)=x\quad(x\text{ punto fisso di }\phi)$$

La scelta di ϕ non è unica: mi basta scegliere $\phi(x)=x+F(f(x))$ con F continua e $F(0)=0$ (stessi zeri di f)

Metodo di Picard o di punto fisso Presa $\phi\colon\mathbb{R}\to\mathbb{R}$ cerchiamo $\alpha\in\mathbb{R}$ tale che $\phi(\alpha)=\alpha$. Prendo x_0 fissato e $x_{n+1}=\phi(x_n)$

Teo. 53 (Teorema delle Contrazioni)**.** Presa $\phi\colon[a,b]\to[a,b]$ continua, essa ha un punto fisso. Se è contrattiva ($|\phi(x)-\phi(y)|\leq L|x-y|$ $\forall x,y\in[a,b]$ con $L<1$), tale punto è unico.

Dimostrazione. ESISTENZA Prendo $g(x) = \phi(x) - x$, ho

$$\begin{cases} g(a) \leq 0 & \text{essendo } \phi(x) \geq a \ \forall x \in [a,b] \text{ quindi } \phi(a) \geq a \\ g(b) \geq 0 & \text{essendo } \phi(x) \leq b \ \forall x \in [a,b] \text{ quindi } \phi(b) \leq b \end{cases} \implies \text{teorema degli zeri}$$

UNICITÀ Per assurdo α_1 e α_2 punti fissi distinti, $|\alpha_1 - \alpha_2| = |\phi(\alpha_1) - \phi(\alpha_2)| \leq L|\alpha_1 - \alpha_2|$ con $L < 1$ □

Teo. 54 (Teorema di convergenza globale). Per il metodo di Picard, se ϕ continua e contrattiva allora

$$\lim_{k\to\infty} x_k = \alpha \quad \forall x_0 \in [a,b] \quad \text{convergenza globale}$$

$$|x_k - \alpha| \leq \frac{L^k}{1-L}|x_1 - x_0| \quad \text{stima a priori dell'errore}$$

$$|x_k - \alpha| \leq \frac{L}{1-L}|x_k - x_{k-1}| \quad \text{stima a posteriori dell'errore}$$

Dimostrazione. **Convergenza globale**: per ogni x_k

$$\begin{aligned} |x_k - \alpha| &= |\phi(x_{k-1}) - \phi(\alpha)| \\ &\overset{\spadesuit}{\leq} \boxed{L|x_{k-1} - \alpha|} \\ &\leq [\dots] \text{ ripetere questo procedimento a } |x_{k-1} - \alpha| \\ &\overset{\heartsuit}{\leq} \boxed{L^k|x_0 - \alpha|} \overset{k\to+\infty}{\longrightarrow} 0 \quad (\text{essendo } L < 1) \end{aligned}$$

Stima a priori:

$$\begin{aligned} |x_0 - \alpha| &\leq |x_0 - x_1| + |x_1 - \alpha| \\ &\leq |x_0 - x_1| + L|x_0 - \alpha| \\ &\leq \frac{|x_0 - x_1|}{1-L} \quad \text{ho portato a sx un addendo e poi diviso} \end{aligned}$$

ora sostituire in ♡.
Stima a posteriori:

$$\begin{aligned} |x_{k-1} - \alpha| &\leq |x_{k-1} - x_k| + |x_k - \alpha| \\ &\leq |x_{k-1} - x_k| + L|x_{k-1} - \alpha| \\ &\leq \frac{|x_{k-1} - x_k|}{1-L} \quad \text{ho portato a sx un addendo e poi diviso} \end{aligned}$$

ora sostituire in ♠. □

Cor. 11 (Convergenza globale nei chiusi). Se ϕ derivabile con $|\phi'(x)| \leq L < 1 \ \forall x \in [a,b]$, allora $x_k \longrightarrow \alpha \ \forall x_0 \in [a,b]$. Ovvero

$$|\phi'(x)| < 1 \ \forall x \in [a,b] \quad (\alpha \in [a,b]) \implies \text{convergente in } [a,b]$$

Teo. 55 (Teorema di convergenza locale). Se α punto fisso di $\phi \in C^1(J)$, J intorno di α

$$\text{se } |\phi'(\alpha)| < 1 \implies \exists \delta > 0 : \quad x_k \longrightarrow \alpha \quad \forall |x_0 - \alpha| \leq \delta$$

. Vale inoltre che

$$\lim_{k\to+\infty} \frac{x_{k+1} - \alpha}{x_k - \alpha} = \phi'(\alpha)$$

Dimostrazione. **Convergenza locale**: $\begin{cases} |\phi'(\alpha)| < 1 \\ \phi(x) \quad \text{continua} \end{cases} \implies \exists \delta > 0 \text{ t.c. } |\phi'(\alpha)| < 1 \quad \forall x \in [\alpha - \delta, \alpha + \delta]$,
quindi ϕ è contrazione su $[\alpha - \delta, \alpha + \delta]$. Dobbiamo verificare che ϕ mappa tale intervallo in se stesso: per induzione (no sbatti). Applico teo. convergenza globale a tale int.
Limite:

$$x_{k+1} - \alpha = \phi(x_k) - \phi(\alpha) \overset{\text{teo.Lagr.}}{=} \phi'(\xi_k)(x_k - \alpha) \quad \xi_k \in [\alpha, x_k]$$

Abbiamo $\begin{cases} x_k \to \alpha \implies \xi_k \to \alpha \\ \phi' \quad \text{continua in intorno di } \alpha \end{cases} \implies \phi'(\xi_k) \to \phi'(\alpha)$
Quindi la tesi applicando questo all'ugualianza appena sopra. □

Osservazione Se $|\phi'(\alpha)| = 1$ non si può dire niente a priori sulla convergenza, se $|\phi'(\alpha)| > 1$ non si può avere convergenza

Prop. 12 . Se $|\phi'(\alpha)| > 1$ non si può avere convergenza

Dimostrazione.

$$|x_{k+1} - \alpha| = |\phi(x_k) - \phi(\alpha)| \overset{\text{teo.Lagr.}}{=} |\phi'(\xi_k)(x_k - \alpha)| = |\phi'(\xi_k)||(x_k - \alpha)|$$

essendo ϕ' continua in un intorno di $\alpha \implies$ per x_k suffic. vicino ho $\phi'(\xi_k) > 1 \implies |x_{k+1} - \alpha| > |x_k - \alpha| \implies$ no convergenza □

*Osservazione (**Newton come caso particolare di punto fisso**)* $\phi(x) = x - \frac{f(x)}{f'(x)}$, punti fissi di $\varphi \leftrightarrow$ zeri di f

Teo. 56 (Ordine di convergenza di Picard). Dato $\alpha \in I \subseteq \mathbb{R}$ punto fisso di $\phi \in C^{p+1}(I)$ tale che $\begin{cases} \phi^{(i)}(\alpha) = 0 \quad i = 1, \dots, p \\ \phi^{(p+1)}(\alpha) \neq 0 \end{cases}$, allora il metodo di punto fisso converge localmente con ordine $p+1$ e vale

$$\lim_{k \to +\infty} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^{p+1}} = \frac{\phi^{(p+1)}(\alpha)}{(p+1)!}$$

Dimostrazione. Taylor con errore di Lagrange:

$$x_{k+1} - \alpha = \phi(x_k) - \phi(\alpha) = \cancel{\sum_{i=1}^{p} \frac{\phi^{(i)}(\alpha)}{i!}(x_k - \alpha)^i} + \frac{\phi^{(p+1)}(\xi)}{(p+1)!}(x_k - \alpha)^{p+1}$$

quindi si ha la prima tesi dividendo per l'ultimo fattore e la seconda tesi ricordando che $\xi \overset{k \to \infty}{\to} \alpha$ poiché $x_k \overset{k \to \infty}{\to} \alpha$ □

Osservazione A parità di ordine di convergenza, tanto più piccola è $\phi^{(p+1)}(\alpha)$ tanto più rapida è la convergenza. (devo scegliere bene ϕ)

Prop. 13 (Ordine metodo Newton con punto fisso). Se α radice semplice ha ordine 2, altrimenti 1

Dimostrazione. Se α radice semplice $\phi(x) = x - \frac{f(x)}{f'(x)}$, calcolare $\phi'(x)$ e vedi che $\phi'(\alpha) = 0$, poi $\phi''(x)$ e vedi che $\phi''(\alpha) \neq 0$.
Se α radice multipla scrivo $f(x) = (x - \alpha)^m h(x)$ con $m > 1$ molteplicità e $h(\alpha) \neq 0$, rifaccio il procedimento di prima e vedo $\phi'(\alpha) = 1 - \frac{1}{m} \neq 0$ □

Metodo di Newton modificato Se la radice α ha molteplicità $m > 1$ voglio trovare un modo per ripristinare l'ordine quadratico. Ho visto che in Newton ho $\phi'(\alpha) = 1 - \frac{1}{m}$ quindi se avessi $\phi'(\alpha) = 1 - \frac{1}{m} \cdot m = 0$ sarei a posto. Quindi pongo:

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

*Osservazione (**criteri di arresto**)* Ponendo $f(\alpha + \delta\alpha) = \delta d$ (residuo) e ricordando

$$\frac{|\delta\alpha|}{|\delta d|} \approx K_{abs} = \frac{1}{|f'(x)|} \implies |\delta\alpha| \approx \frac{1}{|f'(\alpha)|}|\delta d|$$

nel caso di metodi iterativi $\alpha + \delta\alpha = x_k$ e $\delta d = f(x_k)$.
Esempi di criteri d'arresto:

- **Controllo del residuo**:

$$\boxed{|f(x_k)|} \leq \varepsilon \qquad |e_k| = |\alpha - x_k| \leq \frac{1}{|f'(\alpha)|}\boxed{|f(x_k)|}$$

Vediamo che se

$$\begin{cases} |f'(\alpha)| \approx 1 \implies |e_k| \approx \varepsilon \\ |f'(\alpha)| \ll 1 \implies \frac{1}{|f'(x)|} \gg 1 \implies \text{test non affidabile} \\ |f'(\alpha)| \gg 1 \implies \frac{1}{|f'(x)|} \ll 1 \implies \text{test troppo restrittivo} \end{cases}$$

- **Controllo dell'incremento**:

$$\text{si arresta se } \boxed{|x_{k+1} - x_k|} \leq \varepsilon \quad x_{k+1} - x_k = e_k - e_{k+1} \quad |e_k| \approx \frac{1}{|1 - f'(\alpha)|} \boxed{|x_{k+1} - x_k|}$$

$$\begin{cases} x_{k+1} - x_k = (x_{k+1} - \alpha) + (\alpha - x_k) = e_k - e_{k+1} \\ e_{k+1} = \alpha - x_{k+1} = \phi(\alpha) - \phi(x_{k+1}) \overset{\text{Lagr.}}{=} \phi'(\xi_k)(\alpha - x_k) = \phi'(\xi_k) e_k \end{cases}$$

Allora

$$x_{k+1} - x_k = e_k - e_{k+1} = e_k - \phi'(\xi_k) e_k = (1 - \phi'(\xi_k)) e_k$$

e quindi

$$e_k = \frac{1}{1 - \phi'(\xi_k)} (x_{k+1} - x_k) \quad \text{dove } \phi'(\xi_k) \approx \phi'(\alpha) \text{ se } \phi' \text{ varia poco nell'intorno}$$

$$|e_k| \approx \frac{1}{|1 - \phi'(\alpha)|} \overbrace{|x_{k+1} - x_k|}^{< \varepsilon}$$

Vediamo che

$$\begin{cases} \phi'(\alpha) \approx 1 \implies \text{test non affidabile} \\ \phi'(\alpha) = 0 \implies \text{test è ottimale} \\ -1 \geq \phi'(\alpha) \geq 0 \implies \text{test affidabile} \end{cases}$$

6.1.1 Stabilità del metodo delle iterazioni di punto fisso

Vogliamo vedere come si propagano gli errori di arrotondamento dovuti al fatto che non operiamo in aritmetica esatta.

Teo. 58 (Stabilità del metodo di punto fisso)**.** $\tilde{x}_{k+1} = \phi(\tilde{x}_k) + \delta_k$, $\phi : \mathbb{R} \to \mathbb{R}$ derivabile con derivata $\leq L < 1$ e $|\delta_k| \leq \delta$, allora

$$|\tilde{x}_k - \alpha| \leq \frac{\delta}{1 - L} + \frac{L^k}{1 - L} |\tilde{x}_1 - \tilde{x}_0|$$

Dimostrazione. □ P. 18

6.2 Sistemi non lineari

Metodo di Newton vettoriale $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - [D\phi(\boldsymbol{x}_k)]^{-1} \phi(\boldsymbol{x}_k)$ (con $\det(D\phi) \neq 0$)

6.3 Radici di Polinomi

Definizione 6.2 (Formula di Horner)**:** $P_n(x) = a_0 + x(a_1 + x(a_2 + \ldots + x(a_{n-q} + a_n x)\ldots))$

Definizione 6.3 (Metodo di deflazione)**:** Trovo uno zero, divido con Ruffini, ripeto.

Definizione 6.4 (Funzione Logistica)**:** $f_\lambda \colon [0,1] \to \mathbb{R} \quad x \mapsto \lambda x(1-x)$, con $1 \leq \lambda \leq 4$ (per avere almeno un punto fisso e $f_\lambda \leq 1$). Il punto fisso $x = \frac{\lambda - 1}{\lambda}$ è attrattivo per $1 < \lambda < 3$

7 Approssimazione di funzioni

Definizione 7.1 (Funzione interpolante)**:** $p\colon \mathbb{R} \to \mathbb{R}$ interpola $\binom{\text{nodi}}{x_i}, \binom{\text{dati}}{y_i}$ se vale $p(x_i) = y_i \; \forall i$

Tipi di interpolazione Abbiamo

- **polinomiale**: con polinomio
- **trigonoetrica**: con polinomio trigonometrico
- **composita (spline)**: è solo localmente un polinomio

A cosa serve :

- sostituire f con funzione più semplice da trattare (integrare/derivare)
- trovare funzione polinomiale che descriva dati sperimentali (se sono molti)

7.1 Interpolazione polinomiale di Lagrange

Date $n+1$ coppie $\{(x_i, y_i)\}_{i=0,\dots,n}$ si cerca polinomio interpolatore di grado m $\Pi_m(x) \in \mathbb{P}_m$

- $n < m$ problema sottodeterminato
- $n = m$ unicità (vedi sotto)
- $n > m$ problema sovradeterminato (minimi quadrati)

Teo. 60 (Lagrange - esistenza e unicità del polinomio interpolatore)**.** Date $n+1$ coppie $\{(x_i, y_i)\}_{i=0,\dots,n}$ esiste un unico polinomio p di grado al più n che le interpola

Dimostrazione. Dimostrando unicità e esistenza

Unicità. Prendo $p, q \in \mathbb{P}_n$ interpolanti, $p - q = \omega$ con $\omega(x_i) = 0 \quad \forall i = 0, \dots, n$, quindi è di grado n con $n+1$ zeri $\overset{\text{th.Fond.Algebra}}{\implies}$ $\omega \equiv 0$ e dunque $p \equiv q$.

Esistenza. Costruzione esplicita di p dalla base di Lagrange $\{l_i\}_{i=0}^n$ e $p(x) = \sum_{i=0}^n l_i(x) \cdot y_i$

$$l_i(x) = \frac{(x-x_0)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)} = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j}$$

Quindi abbiamo

$$\begin{cases} l_i(x_i) = 1 \\ l_i(x_j) = 0 \end{cases} \implies l_i(x_j) = \delta_{ij}$$

Dimostriamo che sono linearmente indipendenti: dobbiamo vedere che

$$q(x) = \sum_{i=0}^{n} a_i l_i(x) \equiv 0 \implies a_i = 0 \quad \forall i = 0, \dots, n$$

in effetti questo dovrebbe valere in particolare per gli x_j:

$$q(x_j) = \sum_{i=0}^{n} a_i \underbrace{l_i(x_j)}_{=\delta_{ij}} = 0 \implies a_j = 0 \quad \forall j = 0, \dots, n$$

Quindi essendo $\{l_i\}_{i=0}^n$ $n+1$ polinomi di grado n linearmente indip. formano una base per $\mathbb{P}_n$. Verificare l'inerpolazione (ovvia). □

Definizione 7.2 (Polinomio interpolatore e nodale)**:** Polinomio interpolatore $\Pi_n f$ di grado al più n, Polinomio nodale

$$\omega_{n+1} := \prod_{i=0}^{n} (x - x_i)$$

Osservazione Possiamo riscrivere gli elementi della base di Lagr. in funzione del polinomio nodale:

$$l_i(x) = \frac{\omega_{n+1}(x)}{(x-x_i)\omega'_{n+1}(x_i)}$$

Teo. 61 (Errore di Lagrange in uno sviluppo di Taylor). Sia $f : (a,b) \to \mathbb{R}$ derivabile n volte in $x_0 \in (a,b)$ allora sia $T_n(x)$ il polinomio di Taylor di grado n generato da f con centro x_0. Se inoltre $f \in C^{n+1}((a,b))$ allora $\forall x \in (a,b)\ \exists \xi \in [x_0, x]$ t.c.

$$f(x) - T_n(x) \coloneqq E_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1} \qquad \xi \in [x_0, x]$$

Dimostrazione. WLOG $x > x_0$ Vogliamo dimostrare

$$\frac{g(x)}{w(x)} \coloneqq \frac{\overbrace{f(x) - T_n(x)}^{\coloneqq g(x)}}{\underbrace{(x-x_0)^{n+1}}_{\coloneqq w(x)}} = \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

Osserviamo che

$$\begin{cases} g(x_0) = g'(x_0) = \cdots = g^{(n)}(x_0) = 0 \\ g^{(n+1)}(x) = f^{(n+1)}(x) \quad (\star) \end{cases} \qquad \begin{cases} w(x_0) = w'(x_0) = \cdots = w^{(n)}(x_0) = 0 \\ w^{(n+1)}(x) = (n+1)! \quad (\star) \end{cases}$$

Vediamo che $g(x), w(x)$ soddisfano le ip. del th. di Cauchy in $[x_0, x]$: $\exists \xi_1 \in (x_0, x)$ t.c

$$\frac{g(x) - \overbrace{g(x_0)}^{=0}}{w(x) - \underbrace{w(x_0)}_{=0}} = \frac{g'(\xi_1)}{w'(\xi_1)} \implies \frac{g(x)}{w(x)} = \frac{g'(\xi_1)}{w'(\xi_1)}$$

Vediamo che $g'(x), w'(x)$ soddisfano le ip. del th. di Cauchy in $[x_0, x]$: $\exists \xi_2 \in (x_0, x)$ t.c

$$\frac{g'(x) - \overbrace{g'(x_0)}^{=0}}{w'(x) - \underbrace{w'(x_0)}_{=0}} = \frac{g''(\xi_2)}{w''(\xi_2)} \implies \frac{g'(x)}{w'(x)} = \frac{g''(\xi_1)}{w''(\xi_1)}$$

Procedere così ottenendo una sequenza di $(n+1)$ numeri $\xi_1, \dots, \xi_{n+1}$ t.c. $x_0 < \xi_{n+1} < \cdots < \xi_1$ e ponendo $\xi \coloneqq \xi_{n+1}$

$$\frac{g(x)}{w(x)} = \frac{g'(\xi_1)}{w'(\xi_1)} = \frac{g''(\xi_2)}{w''(\xi_2)} = \cdots = \frac{g^{(n+1)}(\xi_{n+1})}{w^{(n+1)}(\xi_{n+1})} \overset{(\star)}{=} \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

□

Teo. 62 (Errore di interpolazione di Lagrange). Sia $x \in \text{Dominio}(f)$, supponiamo $f \in C^{n+1}(I_x)$, con I_x il più piccolo intervallo che contiene i nodi $\{x_i\}_{i=0}^n$ e il punto x. Allora l'errore di interpolazione nel punto x è dato da:

$$E_n(x) \coloneqq f(x) - \Pi_n f(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\omega_{n+1}(x) \qquad \xi_x \in I_x$$

Dimostrazione. La tesi è ovvia per $x = x_i$ per un certo $i = 0, \dots, n$. Allora **fissiamo** $x \neq \{x_i\}_i$ e definisco

$$G(t) = E_n(t) - \omega_{n+1}(t)\underbrace{\frac{E_n(x)}{\omega_{n+1}(x)}}_{\text{cost.}} \qquad t \in I_x$$

Vediamo che $\begin{cases} f \in C^{n+1}(I_x) \\ \omega_{n+1} \in \mathbb{P}_{n+1} \end{cases} \implies G \in C^{n+1}(I_x)$ e che G si annulla in $n+2$ punti distinti $\{x_i\}_{i=1}^n$, x:

$$G(x_i) = \underbrace{E_n(x_i)}_{=0} - \underbrace{\omega_{n+1}(x_i)}_{=0}\frac{E_n(x)}{\omega_{n+1}(x)} = 0$$

$$G(x) = E_n(x) - \cancel{\omega_{n+1}(x)}_{=0}\frac{E_n(x)}{\cancel{\omega_{n+1}(x)}} = 0$$

Quindi

$$\begin{cases} G \in C^{n+1}(I_x) \\ n+2 \text{ zeri} \end{cases} \overset{\text{th.Rolle}}{\Longrightarrow} G'(t) \text{ ha } n+1 \text{ zeri} \overset{\text{reiterato}}{\Longrightarrow} G^{(j)} \text{ ha } n+2-j \text{ zeri} \implies G^{(n+1)} \text{ ha zero in } \xi_x$$

Calcoliamo

$$G^{(n+1)}(t) = E_n^{(n+1)}(t) - \omega_{n+1}^{(n+1)}(t)\frac{E_n(x)}{\omega_{n+1}(x)}$$

vediamo che

$$\begin{cases} E_n^{(n+1)}(t) = f^{(n+1)}(t) - \overbrace{(\Pi_n f)^{(n+1)}(t)}^{=0} = f^{(n+1)}(t) \quad \text{poiché } \Pi_n f \in \mathbb{P}_n \\ \omega_{n+1}^{(n+1)}(t) = (n+1)! \quad \text{poiché derivata } (n+1) \text{ di } p \in \mathbb{P}_{n+1} \text{ è la derivata del solo termine di grado } n+1 \end{cases}$$

e quindi

$$G^{(n+1)}(t) = f^{(n+1)}(t) - (n+1)!\frac{E_n(x)}{\omega_{n+1}(x)} \overset{G^{(n+1)}(\xi_x)=0}{\Longrightarrow} E_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\omega_{n+1}(x)$$

□

Definizione 7.3 (matrice di interpolazione)**:** X matrice triangolare inferiore con n-esima riga la n-upla di punti di interpolazione all'n-esimo passaggio (riga $\leftrightarrow$ set di nodi in numero crescente)

Definizione 7.4 (Norma del massimo/infinito)**:** Sia $f \in C^0([a,b])$, allora

$$\|f\|_\infty := \max_{x\in[a,b]} |f(x)|$$

Definizione 7.5 (Polinomio di miglior approssimazione uniforme)**:** Fissata f e la matrice di interpolazione X, indichiamo:

$$E_{n,\infty} := \|f - \Pi_n f\|_\infty \qquad n = 0, 1, \dots \text{ (righe di } X)$$

Allora $p_n^* \in \mathbb{P}_n$ è il polinomio di miglior approssimazione uniforme (ovvero nella norma infinito) se

$$E_n^* := \|f - p_n^*\|_\infty = \min_{q_n \in \mathbb{P}_n} \|f - q_n\|_\infty$$

Non è detto che p_n^* sia un polinomio interpolatore di f

Definizione 7.6 (Operatore di interpolazione)**:** Esso è

$$\begin{array}{rccc} \Pi_n : & C^0([a,b]) & \to & \mathbb{P}_n \subset C^0([a,b]) \\ & f & \mapsto & \Pi_n f \end{array}$$

è lineare e $\Pi_n p = p \quad \forall p \in \mathbb{P}_n$. Inoltre

$$\|\Pi_n\|_\infty := \sup_{\substack{f\in C^0([a,b]) \\ \|f\|_\infty = 1}} \|\Pi_n f\|_\infty$$

Definizione 7.7 (Costante di Lebesgue)**:** Sia X matrice di interpolazione, è

$$\Lambda_n(X) = \left\| \sum_{i=0}^{n} |l_i(x)| \right\|_\infty$$

Teo. 63 (Significato della costante di Lebesgue)**.** Vale

$$\Lambda_n(X) = \|\Pi_n\|_\infty$$

Dimostrazione. Facciamo $\geq$ e poi $\leq$.

- $\underline{\|\Pi_n\|_\infty \leq \left\|\sum_{i=0}^n \left|l_i^{(n)}\right|\right\|_\infty}$:

$$\begin{aligned}\|\Pi_n\|_\infty &= \sup_{\substack{f\in C^0[a,b]\\ \|f\|_\infty=1}} \|\Pi_n f\|_\infty \\ &= \sup_{\substack{f\in C^0[a,b]\\ \|f\|_\infty=1}} \max_{x\in[a,b]} |\Pi_n f(x)| \\ &\leq \sup \max \sum_{i=0}^n \overbrace{|f(x_i)|}^{\leq \|f\|_\infty=1} \left|l_i^{(n)}\right| \\ &\leq \sup \max_{x\in[a,b]} \sum_{i=0}^n \left|l_i^{(n)}\right| = \left\|\sum_{i=0}^n \left|l_i^{(n)}\right|\right\|_\infty\end{aligned}$$

- $\underline{\|\Pi_n\|_\infty \geq \left\|\sum_{i=0}^n \left|l_i^{(n)}\right|\right\|_\infty}$: dimostrando che esiste $\overline{f}\in C^0[a,b]$, $\|\overline{f}\|_\infty=1$, tale che $\left\|\Pi_n\overline{f}\right\|_\infty \geq \left\|\sum_{i=0}^n \left|l_i^{(n)}\right|\right\|_\infty$. Presi

$$\begin{cases}\overline{x} \text{ t.c. } \left\|\sum_{i=0}^n \left|l_i^{(n)}\right|\right\|_\infty = \sum_{i=0}^n \left|l_i^{(n)}(\overline{x})\right|_\infty \\ \overline{f} = \operatorname{sign}\left(l_i^{(n)}(\overline{x})\right)\end{cases}$$

abbiamo

$$\begin{aligned}\left\|\Pi_n\overline{f}\right\|_\infty &= \max_{x\in[a,b]} \left|\Pi_n\overline{f}(x)\right| \\ &\geq \left|\Pi_n\overline{f}(\overline{x})\right| \\ &= \left|\sum_{i=0}^n \overline{f}(x_i)\, l_i^{(n)}(\overline{x})\right| \\ &= \left|\sum_{i=0}^n \operatorname{sign}\left(l_i^{(n)}(\overline{x})\right) l_i^{(n)}(\overline{x})\right| \\ &= \sum_{i=0}^n \operatorname{sign}\left(l_i^{(n)}(\overline{x})\right) l_i^{(n)}(\overline{x}) \\ &= \sum_{i=0}^n \left|l_i^{(n)}(\overline{x})\right| = \max_{x\in[a,b]} \sum_{i=0}^n \left|l_i^{(n)}\right| = \left\|\sum_{i=0}^n \left|l_i^{(n)}\right|\right\|_\infty\end{aligned}$$

□

Cor. 12 . Vale che

$$E_{n,\infty} \leq (1+\Lambda_n(X))\, E_n^*$$

Dimostrazione. (Lunga, p. 37)

$$E_{n,\infty} := \|f-\Pi_n f\|_\infty \leq \|f-p_n^*\|_\infty + \left\|\overbrace{p_n^*-\Pi_n f}^{\Pi_n(p_n^*-f)}\right\|_\infty \leq (1+\|\Pi_n\|_\infty)\,\|f-p_n^*\|_\infty$$

con $\|\Pi_n\|_\infty := \sup_{\substack{f\in C^0[a,b]\\ \|f\|_\infty=1}} \|\Pi_n f\|_\infty$, poi applico Significato della costante di Lebesgue. □

Osservazione Abbiamo

- L'effetto della scelta dei nodi d'interpolazione su $E_{n,\infty}(x)$ dipende dal valore della costante di Lebesgue
- Si ha che $\exists X^*$ matrice d'interpolazione che minimizza la costante di Lebesgue, ma in generale non è possibile determinare esplicitamente gli elementi
- $\forall$ matrice di interpolazione X $\exists C>0$ tale che

$$\Lambda_n(X) \geq \frac{2}{\pi}\log(n+1) - C$$
$$\implies \Lambda_n \overset{n\to+\infty}{\longrightarrow} +\infty$$

- $\forall$ matrice di interpolazione X su $[a,b]$ $\exists f \in C^0([a,b])$ t.c. $\Pi_n f$ **non converge uniformemente** a $f \implies$ non è possibile approssimare con l'interpolazione polinomiale tutte le funzioni continue (v. controesempio di Runge)

*Esempio (**Fenomeno di Runge**)* Interpolando su $[-5,5]$ con nodi equispaziati $f(x) = \frac{1}{1+x^2}$, vale che $\lim_{n\to+\infty} |f - \Pi_n f| = +\infty$ per $|x| > 3.63$

Definizione 7.8 (nodi di Chebyshev)**:** (per far funzionare meglio le cose) $x_i = \cos\left(\frac{2i-1}{2n}\pi\right)$ oppure $x_i = \cos\left(\frac{i}{n}\pi\right)$

Osservazione Per i nodi equispaziati vale $\Lambda_n(X) = \frac{2^{n+1}}{e\cdot n\cdot \log n}$, mentre i nodi di Chebyshev minimizzano $\Lambda_n(X) = O(\log(n))$, e vale che $\lim_{n\to+\infty} |f - \Pi_n f| = 0$

Teo. 64 (Stabilità di Π_n)**.** Per quanto riguarda la stabilità dell'operatore Π_n, vale che

$$\left\|\Pi_n f - \Pi_n \tilde{f}\right\|_\infty \leq \Lambda_n(X) \left\|f - \tilde{f}\right\|_\infty$$

dove

- $\left\|f - \tilde{f}\right\|_\infty$ è la massima perturbazione sui dati
- $\Lambda_n(X)$ è il coefficiente di amplificazione delle perturbazioni sui dati

quindi la perturbazione massima è controllata da $\Lambda_n(X)$

Dimostrazione. Siano $\tilde{f}(x_i)$ approssimazioni dei valori $f(x_i)$ $\forall i = 0,\dots,n$

$$\left\|\Pi_n f - \Pi_n \tilde{f}\right\|_\infty = \max_{x\in[a,b]} \left|\sum_{i=0}^n \left(f(x_i) - \tilde{f}(x_i)\right) l_i^{(n)}(x)\right| \leq \max_{x\in[a,b]} \sum_{i=0}^n \left|f(x_i) - \tilde{f}(x_i)\right| \left|l_i^{(n)}(x)\right| =$$

$$\leq \max_{i=0,\dots,n} \left|f(x_i) - \tilde{f}(x_i)\right| \underbrace{\max_{x\in[a,b]} \sum_{i=0}^n \left|l_i^{(n)}(x)\right|}_{\Lambda_n(X)} \leq \Lambda_n(X) \max_{i=0,\dots,n} \left|f(x_i) - \tilde{f}(x_i)\right|$$

□

Osservazione Abbiamo

- A piccole perturbazioni sui dati corrispondono piccole variazioni sul polinomio interpolatore purché $\Lambda_n(X)$ sia piccola
- $\Lambda_n(X)$ fornisce il **numero di condizionamento** del problema dell'interpolazione polinomiale
- per n grande l'iterpolazione polin. sui nodi equispaziati può essere instabile

Prop. 14 . $\Lambda_n(X) \geq 1$

Dimostrazione. Abbiamo

$$\Pi_n 1 = 1 = \sum_{j=0}^n 1 \cdot l_j(x) \implies \sum_{j=0}^n l_j(x) = 1$$

Quindi

$$1 = \|1\|_\infty = \left\|\sum_{j=0}^n l_j(x)\right\|_\infty = \max_{[a,b]} \left|\sum_{j=0}^n l_j(x)\right| \leq \max_{[a,b]} \sum_{j=0}^n |l_j(x)| = \left\|\sum_{j=0}^n |l_j(x)|\right\|_\infty = \Lambda_n$$

□

7.2 Forma di Newton del polinomio interpolatore

Impostazione Date $\{(x_i, y_i)\}_{i=0}^{n}$ vogliamo scrivere il polinomio interpolatore come

$$\Pi_n f = \Pi_{n-1} f + q_n \quad \text{con} \quad \begin{cases} \Pi_{n-1} f \text{ interpolatore su } \{(x_i, y_i)\}_{i=0}^{n-1} \\ q_n(x_i) = 0 \text{ per } x_i = 0, \dots, n-1 \end{cases}$$

Dovendo essere q_n di grado n con n zeri deve essere per forza

$$q_n(x) = \Pi_n f - \Pi_{n-1} f = a_n \underbrace{(x - x_0)(x - x_1) \dots (x - x_{n-1})}_{=\omega_n} \quad \text{con} \quad \begin{cases} a_n \in \mathbb{R} \\ \omega_n \text{ polinomio nodale } n\text{-esimo} \end{cases}$$

quindi

$$\boxed{q_n(x) = a_n \cdot \omega_n(x)}$$

Un modo comodo per trovare a_n è valutarlo in x_n (impongo $\Pi_n f(x_n)$ interpolatore):

$$a_n = \frac{q_n(x_n)}{w_n(x_n)} = \frac{\Pi_n f(x_n) - \Pi_{n-1} f(x_n)}{w_n(x_n)} = \frac{f(x_n) - \Pi_{n-1} f(x_n)}{\omega_n(x_n)} =: f[x_0, \dots, x_n]$$

Definizione 7.9 (n-esima differenza divisa)**:** Dall'impostazione precedente è

$$f[x_0, \dots, x_n] := \frac{f(x_n) - \Pi_{n-1} f(x_n)}{\omega_n(x_n)} = a_n$$

Definizione 7.10 (Rappresentazione di Newton)**:** Se poniamo $f[x_0] = f(x_0)$ e $\omega_0(x) \equiv 1$ otteniamo per ricorsione

$$\Pi_n f = \sum_{j=0}^{n} f[x_0, \dots, x_j] \omega_j(x) \tag{2}$$

Osservazione Notiamo che

$$\Pi_n f = \sum_{j=0}^{n} f[x_0, \dots, x_j] \omega_j(x) = \sum_{j=0}^{n} f(x_j) l_j^{(n)}(x) = \sum_{j=0}^{n} \overbrace{\boxed{\frac{\omega_{n+1}(x)}{(x - x_j)\omega'_{n+1}(x_j)}}}^{=l_j} f(x_j) = \omega_{n+1}(x) \sum_{j=0}^{n} \frac{f(x_j)}{(x - x_j)\omega'_{n+1}(x_j)}$$

e quindi uguagliando i coefficienti di ω_n (facendo attenzione alla semplificazione a destra $\omega_{n+1}/(x-x_n) = \omega_n$):

$$f[x_0, \dots, x_n] = \sum_{j=0}^{n} \frac{f(x_j)}{\omega'_{n+1}(x_j)}$$

Teo. 65 . Vale che

$$f[x_0, \dots, x_n] = \frac{f[x_1, \dots, x_n] - f[x_0, \dots, x_{n-1}]}{x_n - x_0}$$

Dimostrazione. (LUNGA, p. 47) Uguagliando i coefficienti di x^n di Rappresentazione di Newton si ha $f[x_0, \dots, x_n] = \sum_{j=0}^{n} \frac{f(x_j)}{\omega'_{n+1}(x_j)}$, da cui

$$\begin{aligned} \frac{f[x_1, \dots, x_n] - f[x_0, \dots, x_{n-1}]}{x_n - x_0} &= \frac{\sum_{j=1}^{n} \frac{f(x_j)}{\omega'_{n+}(x_j)} - \sum_{j=0}^{n-1} \frac{f(x_j)}{\omega'_{n-}(x_j)}}{x_n - x_0} \\ &= \frac{\frac{f(x_n)}{\omega'_{n+}(x_n)} + \sum_{j=1}^{n-1} f(x_j) \left(\frac{1}{\omega'_{n+}(x_j)} - \frac{1}{\omega'_{n-}(x_j)} \right) - \frac{f(x_0)}{\omega'_{n-}(x_0)}}{x_n - x_0} \\ &= \frac{\frac{f(x_n)(x_n - x_0)}{\omega'_{n+}(x_n)} + \sum_{j=1}^{n-1} f(x_j) \frac{x_n - x_0}{\omega'_{n+1}(x_j)} - \frac{f(x_0)(x_n - x_0)}{\omega'_{n-}(x_0)}}{x_n - x_0} = f[x_0, \dots, x_n] \end{aligned}$$

□

*Osservazione (**Costo computazionale**)* p. 50

Prop. 15 (Errore di interpolazione con le differenze)**.**

7.3 Interpolazione composita di Lagrange (polinomiale a tratti)

Impostazione Dato che in generale con nodi equispaziati $\Pi_n f$ non converge uniformemente a f per $n \to \infty$ allora, dato intervallo $[a,b]$

- Divido $[a,b]$ in N sottointervalli $I_j = [x_j, x_{j+1}]$ con $j = 0, \dots, N-1$ di ampiezza h_j e sia $h \coloneqq \max h_j$.
- Uso interpolazione di Lagr. su ogni sottointervallo: ognuno lo divido in $k+1$ nodi equispaziati (k grado del polin.)

Definizione 7.11 (Spazio dei polinomi a tratti di grado k)**:** esso è

$$X_h^k = \{v \in C^0([a,b]) : v|_{I_j} \in \mathbb{P}_k(I_j) \quad \forall j = 0, \dots, N-1\}$$

Definizione 7.12 (Polinomio interpolatore composito)**:** è $\Pi_h^k \in X_h^k$ ovvero

$$\left(\Pi_h^k f\right)|_{I_j} = \Pi_h^k \left(f|_{I_j}\right)$$

*Esempio (**Interpolazione lineare a tratti**)*

Prop. 16 (Errore)**.** In generale vale:

$$\boxed{\left\|f - \Pi_h^k f\right\|_\infty \le C h^{k+1} \left\|f^{(k+1)}\right\|_\infty}$$

in particolare, per l'interpolazione lineare a tratti ($k=1$) vale $C = 1/8$, ovvero

$$\left\|f - \Pi_h^1 f\right\|_\infty \le \frac{1}{8} h^2 \|f''\|_\infty$$

Dimostrazione. □ p. 53

Osservazione Se $h \to 0$ allora converge sempre uniformemente.

7.4 Funzioni spline

Siano $x_0, \dots, x_n$ $n+1$ nodi distinti in $[a,b]$ t.c. $\boxed{a = x_0} < x_1 < \cdots < \boxed{x_n = b}$

Definizione 7.13 (Spline)**:** $s_k \colon [a,b] \to \mathbb{R}$ si dice spline di grado $k \ge 1$ se $s_k|_{[x_i, x_{i+1}]} \in \mathbb{P}_k$ e $s_k \in C^{k-1}([a,b])$

*Osservazione (**Gradi di libertà**)* Abbiamo

- $s_k|_{[x_i,x_{i+1}]} \in \mathbb{P}_k \implies$ [coeff. per ogni I_j] $\cdot$ [numero di intervalli] $= (k+1)n$ coefficienti
- $s_k \in C^{k-1}([a,b]) \implies$ [k condizioni di continuità] $\cdot$ [numero di nodi interni] $= k(n-1)$

Quindi [gradi di libertà]-[restrizioni]$= (k+1)n - k(n-1) = \boxed{n+k}$ gradi di libertà

- Interpolatoria: $s_k(x_i) = f(x_i) \quad i = 0, \dots, n \implies$ [condizioni di interpolazione] $= n+1$

In totale $n + k - (n+1) = \boxed{k-1}$ gradi di libertà.

Definizione 7.14 (Spline periodiche e naturali)**:** Dato che le spline interpolatorie hanno $k-1$ gradi di libertà, posso imporre $k-1$ condizioni aggiuntive ai bordi:

$$\text{Spline Periodiche } s_k^{(m)}(a) = s_k^{(m)}(b) \quad \forall m = 1, \dots, k-1$$
$$\text{Spline Naturali } s_k^{(l+j)}(a) = s_k^{(l+j)}(b) = 0 \quad \forall j = 0, \dots, l-2 \quad \text{con } k = 2l-1,\ l \ge 1 \quad k \text{ dispari}$$

*Esempio (**spline cubiche**)* per $k=3$ le condizioni diventano

$$\text{Spline Periodiche } s'(x_0) = f'(x_0), s'(x_n) = f'(x_n)$$
$$\text{Spline Naturali } s''(x_0) = s''(x_n) = 0$$

7.4.1 Spline di grado 1 ($k = 1$)

Stima dell'errore:

$$\boxed{\left\|f - \Pi_h^1 f\right\|_\infty \leq \frac{M}{8} h^2} \qquad \text{con } M : |f''(x)| \leq M \ \forall x \in [a,b] \qquad \text{e } h = \max_{i=1\dots n} h_i$$

7.4.2 Spline di grado 3 ($k = 3$)

Parte grossa da p. 59 a 63

Teo. 66 (Proprietà di minimo dell'energia). s spline cubica naturale o vincolata, vale che

$$\int_{x_0}^{x_n} \left(s''(x)\right)^2 dx \leq \int_{x_0}^{x_n} \left(g''(x)\right)^2 dx$$

$\forall\, g \in C^{k-1}([x_0, x_n])$ che allo stesso modo interpola i punti e ha le stesse condizioni in x_0 e x_n

Dimostrazione. Ricordando $(\alpha-\beta)^2 = \alpha^2+\beta^2-2\alpha\beta = \alpha^2-\beta^2+2\beta^2-2\alpha\beta = \alpha^2-\beta^2-2\beta(\alpha-\beta)$, quindi ponendo $\alpha = g''$ e $\beta = s''$

$$0 \leq \int_{x_0}^{x_n} \left(\left(g(x) - s(x)\right)''\right)^2 dx = \int_{x_0}^{x_n} \left(g''(x)\right)^2 dx - \int_{x_0}^{x_n} \left(g''(x)\right)^2 dx - \boxed{2\int_{x_0}^{x_n} s''(x)\left(g''(x) - s''(x)\right) dx}$$

(voglio dimostrare $=0$)

Vediamo che

(I) $\left[s''(g'-s')\right]' = s^{(3)}(g'-s') + s''(g''-s'') \overset{\text{teo.Fond.Int.}}{\Longrightarrow} \left[s''(g'-s')\right]_{x_k}^{x_{k+1}} = \int s^{(3)}(g'-s') + \int s''(g''-s'')$

(II) $\left[s^{(3)}(g-s)\right]' = s^{(4)}(g-s) + s^{(3)}(g'-s') \overset{\text{teo.Fond.Int.}}{\Longrightarrow} \left[s^{(3)}(g-s)\right]_{x_k}^{x_{k+1}} = \int s^{(4)}(g-s) + \int s^{(3)}(g'-s')$

$$\Longrightarrow \int_{x_k}^{x_{k+1}} s''\left(g''-s''\right) dx = \left[s''(g'-s')\right]_{x_k}^{x_{k+1}} - \left[s^{(3)}(g-s)\right]_{x_k}^{x_{k+1}} + \int s^{(4)}(g-s)$$

$$\begin{aligned}
\int_{x_0}^{x_n} s''(x)\left(g''(x) - s''(x)\right) dx &= \sum_{k=0}^{n-1} \int_{x_k}^{x_{k+1}} s''\left(g''-s''\right) dx \\
&= \sum_{k=0}^{n-1} \left[s''\left(g'-s'\right)\right]_{x_k}^{x_{k+1}} - \sum_{k=0}^{n-1} \left[s^{(3)}(g-s)\right]_{x_k}^{x_{k+1}} + \sum_{k=0}^{n-1} \left[\int_{x_k}^{x_{k+1}} s^{(4)}(g-s)\right] \qquad \text{per (I-II)} \\
&= \overbrace{\left[s''\left(g'-s'\right)\right]_{x_0}^{x_n}}^{\clubsuit} - \overbrace{\left[s^{(3)}(g-s)\right]_{x_0}^{x_n} + \sum_{k=0}^{n-1} \left[\int_{x_k}^{x_{k+1}} s^{(4)}(g-s)\right]}^{\spadesuit} \qquad \text{i nodi interni si canc.} \\
&= 0
\end{aligned}$$

$\clubsuit = s''(x_n)\left(g'(x_n) - s'(x_n)\right) - s''(x_0)\left(g'(x_0) - s'(x_0)\right) = 0$ per le condizioni in x_0 e x_n

$\spadesuit = \sum_{k=0}^{n-1} \left(\left[s^{(3)}(g-s)\right]_{x_k}^{x_{k+1}} - \int_{x_k}^{x_{k+1}} s^{(4)}(g-s)\, dx \right) = 0$ perché $\begin{cases} g(x_i) = s(x_i) \\ s^{(4)}(x) \equiv 0 \end{cases}$

□

Osservazione Quindi tra tutte le funzioni $C^2([a,b])$ che interpolano g nei nodi $a = x_0 < x_1 < \cdots < x_n = b$, la spline cubica naturale è quella che minimizza la curvatura, ovvero che oscilla di meno.

Prop. 17 (Proprietà d'approssimazione). p. 65

7.5 Interpolazione in più variabili

Cerco una base di Lagrange per i nodi su cui voglio interpolare per avere l'unisolvenza. Di solito interpolo sui triangoli per i polinomi generici o sui quadrati per i polinomi omogenei

7.6 Interpolazione astratta

Definizione 7.15 (Interpolazione in senso astratto)**:** Prendo un $\mathbb{R}$-spazio vettoriale V (può essere di funzioni, come $\mathbb{P}$) di base $\{\varphi_1, \dots, \varphi_n\}$ e con una base del duale $\langle \mathscr{L}_1, \dots, \mathscr{L}_n \rangle = V^*$ con

$$\mathscr{L}_i : V \to \mathbb{R}$$

detti gradi di libertà o **funzionali lineari** (funzionali perché possono prendere in pancia funzioni, come la valutazione della funzione di input in un punto). $v \in V$ interpola in senso astratto $(y_1, \dots, y_n) \in \mathbb{R}^n$ se

$$\mathscr{L}_i(v) = y_i \quad \forall i = 1, \dots, n$$

Definizione 7.16 (Unisolvenza)**:** Un sistema di gradi di libertà è unisolvente se $\forall (y_1, \dots, y_n)$ esiste unico $v \in V$ che li interpola.

Definizione 7.17 (Matrice di Haar)**:** $(H)_{ij} = \mathscr{L}_i(\varphi_j)$, $H = \begin{pmatrix} \mathscr{L}_1(\varphi_1) & \cdots & \mathscr{L}_1(\varphi_n) \\ \vdots & \ddots & \vdots \\ \mathscr{L}_n(\varphi_1) & \cdots & \mathscr{L}_n(\varphi_n) \end{pmatrix}$

Teo. 67 . Un sistema di gradi di libertà è unisolvente $\iff \det(H) \neq 0$, ovvero i funzionali sono linearmente indipendenti.

Dimostrazione. p. 67 □

*Esempio (**Interpolazione polinomiale**)*

7.7 Interpolazione nel senso dei minimi quadrati

Definizione 7.18 (Prodotto scalare e norma L^2 (euclidea))**:** Per f, g funzioni $[a, b] \to \mathbb{R}$, ho

$$(f, g) = \int_a^b f(x)\, g(x)\, dx \qquad \|f\|_{L^2} = \sqrt{\int_a^b |f(x)|^2 dx} = \sqrt{(f, f)}$$

Mentre per il caso **discreto**: $f(x_i)$ con $i = 0, \dots, n$

$$(f, g) = \sum_{i=0}^{n} f(x_i)\, g(x_i) \qquad \|f\|_{L^2} = \sqrt{\sum_{i=0}^{n} |f(x_i)|^2} = \sqrt{(f, f)}$$

Definizione 7.19 (Funzioni ortogonali)**:** p. 70

Definizione 7.20 (Sistema ortonogonale/ortonormale)**:**

Prop. 18 .

Dimostrazione. □

Cor. 13 .

Dimostrazione. □

Teo. 68 (di miglior approssimazione nel senso dei minimi quadrati - caso continuo). Siano $f \in C^0([a, b])$, $\{\phi_0, \dots, \phi_n\}$ linearmente indipendenti, $\phi_i \in C^0([a, b])$. Allora

$$\exists! g_n^* \in U \coloneqq \operatorname{span}\{\phi_0, \dots, \phi_n\} \text{ t.c } \|f - g_n^*\|_2 = \min_{g_n \in U} \|f - g_n\|_2$$

Inoltre g_n^* risolve il seguente sistema lineare

$$(f - g_n^*, \phi_k) = 0 \quad \forall k = 0, \dots, n \quad \textbf{equazioni normali}$$

dove $(\cdot, \cdot)$ è il prodotto scalare sopra definito.

Dimostrazione. Abbiamo

$$\begin{cases} g_n^* \in U = \text{span}\{\phi_0, \dots, \phi_n\} \implies g_n^* = \sum_{j=0}^n c_j^* \phi_j \\ (f - g_n^*, \phi_k) = 0 \overset{\text{lin.}}{\implies} (g_n^*, \phi_k) = (f, \phi_k) \end{cases} \implies \sum_{j=0}^n c_j^* \overset{\text{noti}}{(\phi_j, \phi_k)} = \overset{\text{noti}}{(f, \phi_k)} \quad \forall k = 0, \dots, n \quad \text{Sist. lin.}$$

- Esistenza/unicità: i coefficienti c_j^* sono soluzione di tale sistema lineare. Essendo la matrice $(\phi_j, \phi_k)_{jk}$ quadrata mi basta dimostrare che è non singolare (invertibile) e ho sia la suriettività (**esistenza della sol.**) che l'inniettività (**unicità della sol.**).
  Se **per assurdo fosse singolare**, il sistema omogeneo ammetterebbe soluzione non nulla ($\ker \neq 0$), ovvero:
  $$\exists\, c_0, \dots, c_n \text{ non tutti nulli } \mid \sum_{j=0}^n c_j(\phi_j, \phi_k) = 0 \quad \forall k = 0, \dots, n$$
  e quindi
  $$\left\| \sum_{j=0}^n c_j \phi_j \right\|_2^2 = \left(\sum_{k=0}^n c_k \phi_k, \sum_{j=0}^n c_j \phi_j \right) = \sum_{k=0}^n c_k \underbrace{\left(\sum_{j=0}^n c_j (\phi_k, \phi_j) \right)}_{=0} = 0 \overset{\text{norma}}{\implies} \sum_{j=0}^n c_j \phi_j = 0 \quad \lightning \text{ essendo } \{\phi_j\} \text{ l.i.}$$

- Proprietà di minimo: dimostriamo che $\forall g_n = \sum_{j=0}^n c_j \phi_j$ con $c_j \neq c_j^*$ per almeno un indice j è tale che $\|f - g_n\|_2 \geq \|f - g_n^*\|_2$. Abbiamo
  $$\begin{cases} f - g_n = f - g_n^* + \overbrace{\sum_{j=0}^n c_j^* \phi_j}^{g_n^*} - \overbrace{\sum_{j=0}^n c_j \phi_j}^{g_n} = f - g_n^* + \sum_{j=0}^n (c_j^* - c_j)\phi_j \\ (f - g_n^*,\ \sum_{j=0}^n (c_j^* - c_j)\phi_j) = \sum_{j=0}^n (c_j^* - c_j) \underbrace{(f - g_n^*,\ \phi_j)}_{=0} = 0 \end{cases}$$
  e quindi
  $$\|f - g_n\|_2^2 = (f - g_n,\ f - g_n) = \|f - g_n^*\|_2^2 + \underbrace{\left\| \sum_{j=0}^n (c_j + c_j^*)\phi_j \right\|_2^2}_{\geq 0} + 2 \underbrace{\left(f - g_n^*,\ \sum_{j=0}^n (c_j^* - c_j)\phi_j \right)}_{=0}$$
  quindi $\|f - g_n\|_2^2 \geq \|f - g_n^*\|_2^2 \implies \|f - g_n\|_2 \geq \|f - g_n^*\|_2$

□

Definizione 7.21 (minimi quadrati discreti)**:**

$$(f, g) = \sum_{i=0}^n \omega_i f(x_i)\, g(x_i)\, dx \qquad \|f\|_{L^2\text{-discreta}} = \sqrt{\sum_{i=0}^n \omega_i f^2(x_i)}$$

Teo. 69 (di miglior approssimazione nel senso dei minimi quadrati - caso discreto)**.** Dati $f(x_i)$, $\{x_i\} = S$ con $i = 0, \dots, m$ (m grado del polinomio, caso significativo quando $m > n$), siano $\{\phi_0, \dots, \phi_n\}$ linearmente indipendenti, $\phi_i \in C^0([a, b])$. Allora

$$\exists! g_n^* \in U \coloneqq \text{span}\{\phi_0, \dots, \phi_n\} \text{ t.c } \|f - g_n^*\|_2 = \min_{g_n \in U} \|f - g_n\|_2$$

Inoltre g_n^* risolve il seguente sistema lineare

$$(f - g_n^*, \phi_k) = 0 \quad \forall k = 0, \dots, n \quad \textbf{equazioni normali}$$

dove $(\cdot, \cdot)$ è il prodotto scalare sopra definito.

Dimostrazione. Abbiamo

$$\begin{cases} g_n^* \in U \implies g_n^* = \sum_{i=0}^n c_i^* \phi_i \\ g_n \in U \implies g_n = \sum_{i=0}^n c_i \phi_i \end{cases} \qquad \text{e} \qquad \|f - g_n\|_{2,S}^2 = \sum_{i=0}^m |f(x_i) - g_n(x_i)|^2$$

allora cerco il minimo della funzione

$$\begin{array}{rccc} d: & \mathbb{R}^{n+1} & \to & \mathbb{R} \\ & (c_0,\dots,c_n) & \mapsto & \|f-g_n\|_{2,S}^2 \end{array}$$

risolvendo $\boldsymbol{\nabla} d = \mathbf{0}$, ovvero risolvendo

$$\frac{\partial d}{\partial c_k} = \sum_{i=0}^{m} 2\left(f(x_i) - \sum_{j=0}^{n} c_j\phi_j(x_i)\right)\phi_k(x_i) = 0 \quad \forall k = 0,\dots,n$$

che divenda

$$\sum_{i=0}^{n}\sum_{j=0}^{n} c_j\phi_j(x_i)\phi_k(x_i) = \sum_{i=0}^{n} f(x_i)\phi_k(x_i) \quad \forall k = 0,\dots,n$$

Vediamo che definendo

$$\operatorname{tab} f \coloneqq \begin{pmatrix} f(x_0) \\ \vdots \\ f(x_m) \end{pmatrix} \qquad c \coloneqq \begin{pmatrix} c_0 \\ \vdots \\ c_n \end{pmatrix} \qquad A \coloneqq [\operatorname{tab}\phi_0 \mid \cdots \mid \operatorname{tab}\phi_n] \in \mathbb{R}^{(m+1)\times(n+1)}$$

è equivalente a risolvere

$$\sum_{i=0}^{m}(Ac)_i\phi_k(x_i) = (A^T \operatorname{tab} f)_k \implies (A^TAc)_k = (A^T \operatorname{tab} f)_k \implies \boxed{A^TAc = A^T \operatorname{tab} f}$$

dove A^TA è SPD $\implies$ rango massimo $\implies$ soluzione unica. □

Teo. 70 . Se c vettore dei coefficienti di g e A matrice matrice di valutazione della base nei punti $\{x_i\}$, allora

$$\|f-g\|^2 = \sum_{i=0}^{n}\left(\overset{b}{f}(x_i) - \overset{A\cdot c}{g}(x_i)\right)^2 = \|b - A\cdot c\|_2^2 \qquad \text{minimo quando} \quad A^T\cdot A\cdot c = A^T\cdot b$$

Dimostrazione.

$$\begin{aligned} \phi(x) &= \|b-A\cdot c\|_2^2 = (b,b) - (b,A\cdot c) - (A\cdot c,b) + (A\cdot c, A\cdot c) = b^T\cdot b - 2\cdot c^T\cdot A^T\cdot b + c^T\cdot A^T\cdot A\cdot c \\ \nabla\phi(x) &= -2\cdot A^T\cdot b + 2A^T\cdot A\cdot c = 0 \qquad \implies \qquad A^T\cdot A\cdot c = A^T\cdot b \end{aligned}$$

□

Definizione 7.22 (Polinomi ortogonali di Legendre)**:** $\phi_0(x) = 1, \phi_1(x) = x, \phi_{i+1}(x) = \frac{2i+1}{i+1}x\phi_i(x) - \frac{1}{i+1}\phi_{i-1}(x)$

Definizione 7.23 (Polinomi ortogonali di Chebyshev)**:** $T_i(x) = \cos(i\arccos(x))$, oppure $T_{i+1}(x) = 2xT_i(x) - T_{i-1}(x)$

8 Integrazione Numerica

Definizione 8.1 (formula di quadratura)**:** è una formula del tipo

$$I_n(f) = \sum_{i=0}^{n} \alpha_i f(x_i) \quad \text{con} \quad \begin{matrix} \alpha_i & \text{pesi di quadratura} \\ x_i & \text{punti di quadratura} \end{matrix}$$

che serve per approssimare $\int_a^b f(x)\,dx = I(f)$.

Definizione 8.2 (formula interpolatoria)**:** Formula di quadratura in cui i pesi sono definiti come

$$\boxed{\alpha_i \coloneqq \int_a^b l_i(x)dx}$$

con

$$f_n = \Pi_n f = \sum_{i=0}^{n} f(x_i) l_i(x) \qquad l_i(x) = \text{i-esimo polin. base lagr.}$$

Quindi è ottenuta sostituendo il polinomio interpolatore di grado n nei punti di quadratura

$$I_n(f) = \int_a^b \sum_{i=0}^{n} f(x_i)\, l_i(x)\, dx = \sum_{i=0}^{n} f(x_i) \overbrace{\int_a^b l_i(x)\, dx}^{\alpha_i}$$

ovvero è

$$I_n(f) \coloneqq I(f_n)$$

Definizione 8.3 (grado di esattezza/precisione)**:** Massimo $r \geq 0$ per cui $I_n(p) = I(p) \quad \forall p \in \mathbb{P}_r$

Teo. 71 . Una formula di quadratura a $n+1$ punti è interpolatoria $\iff$ ha grado di esattezza/precisione $\geq n$

Dimostrazione. Doppia implicazione

$\implies$ ovvio per la def.: $\Pi_n f = f$

$\impliedby$ Dato che la formula ha grado di precisione n, integra esattamente $(\star)$ i polinomi di grado n, in particolare gli l_i:

$$\int_a^b l_i(x) \overset{\star}{=} \sum_{j=0}^{n} \alpha_j \underbrace{l_i(x_j)}_{=\delta_{ij}} = \alpha_i \quad \forall i = 0, \ldots, n$$

□

8.1 Formule di quadratura interpolatorie - formule di Newton-Cotes

8.1.1 Formula del punto medio ($n = 0$)

formula del punto medio si sostituisce f con $\Pi_0 f$ relativo al nodo $x_0 = \frac{a+b}{2}$

$$I_0(f) = (b-a) f\left(\frac{a+b}{2}\right) \qquad E_0(f) \coloneqq I(f) - I_0(f) = \frac{(b-a)^3}{24} f''(\xi) \quad \xi \in (a,b) \quad (\text{Se } f \in C^2([a,b]))$$

Dimostrazione dell'errore. Preso $c = \frac{a+b}{2}$ e la serie di Taylor $f(x) = f(c) + f'(c)(x-c) + \frac{f''(\xi_x)}{2}(x-c)^2$

$$E_0(f) \coloneqq \int_a^b f(x)\,dx - \int_a^b f(c)\,dx = \cancelto{\text{dispari in } [a,b]}{\int_a^b f'(c)(x-c)dx} + \int_a^b \frac{f''(\xi_x)}{2}(x-c)^2\,dx \overset{\text{media integ.}}{=} \frac{f''(\xi)}{2}\int_a^b (x-c)^2\,dx$$

$$\overset{\star}{=} \frac{f''(\xi_x)}{2}(x-c)^2 \frac{2}{3}\left(\frac{b-a}{2}\right)^3 = \frac{(b-a)^3}{24} f''(\xi)$$

dove $\star$: $\int_a^b (x-c)^2 dx = 2\int_c^b (x-c)^2 = \frac{2}{3}(x-c)^3|_c^b \overset{\spadesuit}{=} \frac{2}{3}\left(\frac{b-a}{2}\right)^3$

dove $\spadesuit$: $\cancel{(c-c)^3} - (b-c)^3 = -(b - \frac{a+b}{2})^3 = -(\frac{a-b}{2})^3$ □

Osservazione (***Grado di precisione***) ha g.d.p 1: $f \in \mathbb{P}_1 \implies f'' = 0 \implies E_0(f) = 0$

formula del punto medio composita m sottointervalli di ampiezza $h = \frac{b-a}{m}$, con nodi $x_k = a + (2k+1)\frac{h}{2}$ (multipli dispari di $h/2$)

$$I_{0,m}(f) = h\sum_{k=0}^{m-1} f(x_k) \qquad E_0(f) = \sum_{k=0}^{m-1} \frac{h^3}{24} f''(\xi_k) = \frac{h^3}{24} m f''(\xi) = \frac{h^2(b-a)}{24} f''(\xi) \quad \xi \in (a,b)$$

Teo. 74 (teorema della media integrale)**.** Presi $f \in C^0([a,b])$ e g integrabile che non cambia segno in $[a,b]$, allora $\exists \xi \in (a,b)$ tale che

$$\int_a^b f(x)\, g(x)\, dx = f(\xi) \int_a^b g(x)\, dx$$

Teo. 75 (teorema del valor medio discreto)**.** Presi $u \in C^0([a,b])$, $x_{i=0,\dots,s} \in [a,b]$ e $\delta_{i=0,\dots,s}$ costanti con lo stesso segno (ad esempio ≥ 0), allora $\exists\, \eta \in (a,b)$ tale che

$$\sum_{i=0}^{s} \delta_i u(x_i) = u(\eta) \sum_{i=0}^{s} \delta_i$$

Dimostrazione. Presi $u_m = \min_{x\in[a,b]} u(x) = u(x_m)$, $u_M = \max_{x\in[a,b]} u(x) = u(x_M)$, ho

$$u_m \sum_{i=0}^{s} \delta_i \leq \sum_{i=0}^{s} \delta_i u(x_i) \leq u_M \sum_{i=0}^{s} \delta_i$$

da cui definendo $U(x) = u(x)\sum_{i=0}^{s}\delta_i$ ho $U(x_m) \leq \sum_{i=0}^{s}\delta_i u(x_i) \leq U(x_M)$, e dunque $U(\eta) = \sum_{i=0}^{s}\delta_i u(x_i)$ □

8.1.2 Formula del trapezio ($n = 1$)

formula del trapezio si sostituisce f con $\Pi_1 f$ relativo ai nodi $x_0 = a$ e $x_1 = b$

$$I_1(f) = \frac{b-a}{2}(f(a) + f(b)) \qquad E_1(f) = -\frac{(b-a)^3}{12} f''(\xi) \quad \xi \in (a,b) \quad (\text{Se } f \in C^2([a,b]))$$

Dimostrazione dell'errore. Dalla formula dell'Errore di interpolazione di Lagrange

$$E_1(f) \coloneqq \int_a^b f(x) - \Pi_1 f(x)\, dx = \int_a^b \frac{f''(\xi_x)}{2}(x-a)(x-b)\, dx \overset{\text{media integ.}}{=} \frac{f''(\xi)}{2}\int_a^b (x-a)(x-b)\, dx$$

□

Osservazione g.d.p.=1

formula del trapezio composita m sottointervalli di ampiezza $h = \frac{b-a}{m}$, $x_k = a + kh$

$$I_{1,m}(f) = \sum_{k=0}^{m} \frac{h}{2}(f(x_i) + f(x_{i+1})) = \frac{h}{2}\left(f(a) + 2\sum_{k=1}^{m-1} f(x_k) + f(b)\right)$$

$$E_{1,m}(f) = -\sum_{k=0}^{m-1} \frac{h^3}{12} f''(\xi_k) = -\frac{h^2(b-a)}{12} f''(\xi) \quad \xi \in (a,b)$$

8.1.3 Formula di Cavalieri-Simpson ($n = 2$)

formula di Cavalieri-Simpson si sostituisce f con $\Pi_2 f$ relativo ai nodi $x_0, x_1, x_2 = a, \frac{a+b}{2}, b$, i pesi sono $\alpha_i = \int_a^b l_i(x)\, dx$

$$I_2(f) = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) \qquad E_2(f) = \frac{1}{90}\left(\frac{b-a}{2}\right)^5 f^{(4)}(\xi) \quad \xi \in (a,b)$$

Osservazione g.d.p.=3

formula di Cavalieri-Simpson composita $h = \frac{b-a}{m}$, $x_k = a + k\frac{h}{2}$

$$I_{2,m}(f) = \frac{h}{6}\left(f(a) + 2\sum_{k=1}^{m-1} f(x_{2k}) + 4\sum_{k=1}^{m-1} f(x_{2k+1}) + f(b)\right) \qquad E_{2,m}(f) = -\frac{b-a}{180}\left(\frac{h}{2}\right)^4 f^{(4)}(\xi) \quad \xi \in (a,b)$$

8.1.4 Formule di Newton-Cotes (generalizzazione)

formule di Newton-Cotes si sostituisce f con $\Pi_n f$ relativo ai nodi equispaziati $x_k = x_0 + k\frac{x_n - x_0}{n} = x_0 + kh$

chiuse se $x_0 = a$ $x_n = b$

, aperte se $x_0 = a + h$ $x_n = b - h$

Attuo cambio di variabile $x = x_0 + th$:

$$l_i(x) = \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j} = \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{\cancel{a} + th - \cancel{a} - jh}{\cancel{a} + ih - \cancel{a} - jh} = h\prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-j}{i-j} \qquad \alpha_i = \int_a^b l_i(x)\,dx = h\int_0^n \overbrace{\prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-j}{i-j}}^{\varphi_i(t)}\,dt$$

$$I_n(f) = \sum_{i=0}^{n} \alpha_i f(x_i) \qquad E_n(f) = \begin{cases} M_n (b-a)^{n+3} f^{(n+2)}(\xi) & n \text{ pari} \quad f \in C^{n+2}([a,b]) \\ K_n (b-a)^{n+2} f^{(n+1)}(\xi) & n \text{ dispari} \quad f \in C^{n+1}([a,b]) \end{cases}$$

*Osservazione (**Grado di precisione**)* $\begin{cases} n \text{ pari} \implies n+1 \\ n \text{ dispari} \implies n \end{cases}$

formule di Newton-Cotes composite [...]

Teo. 82 . $E_{n,m}(f) = O(h^{n+p})$ con $p = 2$ per n pari, $p = 1$ per n dispari

Dimostrazione. Per n pari

$$E_{n,m}(f) = \sum_{k=1}^{m} M_n h^{n+3} f^{(n+2)}(\xi_k) = M_n h^{n+3} \sum_{k=1}^{m} f^{(n+2)}(\xi_k) = M_n h^{n+3} m f^{(n+2)}(\xi) = M_n (b-a) h^{n+2} f^{(n+2)}(\xi)$$

□

*Osservazione (**analisi a posteriori dell'errore**)* Siccome $E_{n,m}(f) = O(h^{n+p})$, vale che $E_{n,2m}(f) \approx \frac{1}{2^{n+p}} E_{n,m}(f)$, da cui

$$E_{n,2m}(f) \approx \frac{1}{2^{n+p} - 1}\left(I_{n,2m}(f) - I_{n,m}(f)\right)$$

Dimostrazione.

$$\left(1 - \frac{1}{2^{n+p}}\right) I(f) \approx I_{n,2m}(f) - \frac{1}{2^{n+p}} I_{n,m}(f) \qquad I(f) \approx \frac{2^{n+p}}{2^{n+p} - 1} I_{n,2m}(f) - \frac{1}{2^{n+p} - 1} I_{n,m}(f)$$

Sottraendo da ambo i lati $I_{n,2m}(f)$ ho la tesi □

8.1.5 Adattività

[...]

8.2 Integrazione Gaussiana

Approccio alternativo a Newton-Cotes. Sia $w\colon [-1,1] \to \mathbb{R}$ non negativa, integrabile e assolutamente continua (**funzione peso**).

$$L^2_w([-1,1]) = \left\{ f\colon [-1,1] \to \mathbb{R}, \int_{-1}^{1} f(x)\,w(x)\,dx < +\infty \right\} \qquad (f,g)_w = \int_{-1}^{1} f(x)\,g(x)\,w(x)\,dx$$

$$I_w(f) = \int_{-1}^{1} f(x)\,w(x)\,dx \approx I_{n,w}(f) \coloneqq \sum_{i=0}^{n} \alpha_i f(x_i)$$

$$I_{n,w}(f) = \int_{-1}^{1} \Pi_n f(x)\, w(x)\, dx \qquad \alpha_i = \int_{-1}^{1} l_i(x)\, w(x)\, dx$$

Teo. 83 (teorema di Jacobi). Preso $m > 0$, $I_{n,w}(f) = \sum_{i=0}^{n} \alpha_i f(x_i)$ ha grado di esattezza/precisione $n+m$ $\iff$ il polinomio nodale ω_{n+1} è ortogonale a $\mathbb{P}_{m-1}$, ovvero

$$\int_{-1}^{1} \omega_{n+1}(x)\, p_{m-1}(x)\, w(x)\, dx = 0 \quad \forall p \in \mathbb{P}_{m-1}$$

Dimostrazione. Doppia implicazione

$\Longleftarrow$ Preso $f \in \mathbb{P}_{n+m}$, posso scriverlo $f = \omega_{n+1} p_{m-1} + q_n$ con $\begin{cases} p_{m-1} \in \mathbb{P}_{m-1} \\ q_n \in \mathbb{P}_n \end{cases}$ e quindi $q_n = f - \omega_{n+1} p_{m-1}$.

Se la formula è interpolatoria ($n+1$ nodi) $\implies$ ha g.d.p. almeno n $\implies$ $q_n \in \mathbb{P}_n$ integrato esattamente:

$$\sum_{i=0}^{n} \alpha_i f(x_i) = \sum_{i=0}^{n} \alpha_i \underbrace{\omega_{n+1}(x_i)}_{=0} p_{m-1}(x_i) + \underbrace{\sum_{i=0}^{n} \alpha_i q_n(x_i)}_{=\int_{-1}^{1} q_n(x) w(x) dx} = \int_{-1}^{1} f(x) w(x) dx - \int_{-1}^{1} \underbrace{\omega_{n+1}(x_i)}_{=0} p_{m-1} w(x)$$

Quindi $\int_{-1}^{1} f(x) w(x) dx = \sum_{i=0}^{n} \alpha_i f(x_i) \overset{f \in \mathbb{P}_{n+m}}{\Longrightarrow}$ la formula di quadratura ha g.d.p. $n+m$

$\implies$ Preso $p_{m-1} \in \mathbb{P}_{m-1}$,

$$\int_{-1}^{1} \underbrace{\omega_{n+1}(x)\, p_{m-1}(x)}_{\in \mathbb{P}_{m+n}} w(x)\, dx \overset{\text{g.d.p}=m+n}{=} \sum_{i=0}^{n} \alpha_i \overbrace{\omega_{n+1}(x_i)}^{0} p_{m-1}(x_i) = 0$$

□

Cor. 14 . Il grado massimo di una formula con $n+1$ punti è $2n+1$

Dimostrazione. Per assurdo $2n+2$, prendo ad esempio $p = \omega_{n+1}$ e ottengo $\int_{-1}^{1} \omega_{n+1}^2(x)\, w(x)\, dx = 0$, ↯ □

Osservazione Scegliendo $m = n+1$, si ha che

$$\begin{aligned} I_{n,\omega}(f) \text{ ha g.d.p } 2n+1 \text{ (massimo)} &\iff \int_{-1}^{1} \omega_{n+1}(x)\, p(x)\, w(x)\, dx = 0 \quad \forall p \in \mathbb{P}_n \\ &\iff \overset{\in \mathbb{P}_{n+1}}{\omega_{n+1}} \perp \mathbb{P}_n \\ &\iff \text{nodi } x_i \text{ sono radici di polin. (grad. } n+1) \perp \mathbb{P}_n \end{aligned}$$

Tale polinomio è il polinomio nodale ω_{n+1} ortogonale a $\mathbb{P}_n$. La dimensione di $\mathbb{P}_n^{\perp} \subseteq \mathbb{P}_{n+1}$ è 1, quindi esisterà solo un polinomio nodale di questo tipo. Voglio dunque trovare una successione di polinomi **monici** $p_n \in \mathbb{P}_n$ tali che

$$p_n \perp \underbrace{p_0, \dots, p_{n-1}}_{\text{base}} \quad \mathbb{P}_{n-1} = \text{span}\{p_0, \dots, p_{n-1}\}$$

i nodi saranno gli zeri di tale p_n.
I pesi si trovano risolvendo:

$$\int_{-1}^{1} p_j(x) \omega(x) dx = \sum_{i=0}^{n} \alpha_i p_j(x_i) \quad \forall j = 0, \dots, n-1 \quad \text{(impongo il g.d.p. max)}$$

oppure

$$\alpha_i = \int_{-1}^{1} l_i(x) \quad \text{(def. di pesi interpolatori)}$$

con $l_i = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}$

Definizione 8.4 (polinomi ortogonali della quadratura gaussiana)**:** Dati $p_{-1}(x) = 0$ e $p_0(x) = 1$

$$p_{k+1}(x) = (x - \gamma_k) p_k(x) - \beta_k p_{k-1}(x) \qquad \gamma_k := \frac{(xp_k, p_k)_w}{(p_k, p_k)_w} \qquad \beta_k := \frac{(p_k, p_k)_w}{(p_{k-1}, p_{k-1})_w}$$

Dimostrazione. Per induzione, $k = 1$ ovvio, suppongo $p_i \perp p_j$ per $j < i \leq k$, voglio dimostrare $p_{k+1} \perp p_j$ per $j \leq k$

$$j \leq k-2 \qquad (p_{k+1}, p_j)_w = (xp_k, p_j)_w - \gamma_k \cancel{(p_k, p_j)_w} - \beta_k \cancel{(p_{k-1}, p_j)_w} = (p_k, \overbrace{xp_j}^{\in \mathbb{P}_{k-1}})_w = 0$$

$$j = k-1 \quad (p_{k+1}, p_{k-1})_w = (xp_k, p_{k-1})_w - \gamma_k \cancel{(p_k, p_{k-1})_w} - \beta_k (p_{k-1}, p_{k-1})_w = (p_k, \overbrace{xp_j}^{p_k})_w - (p_k, p_k)_w = 0$$

$$j = k \qquad (p_{k+1}, p_{k-1})_w = \cancel{(xp_k, p_k)_w} - \gamma_k (p_k, p_k)_w - \beta_k \cancel{(p_{k-1}, p_k)_w} = 0$$

□

Lemma 5 . Se $w(x)$ è pari, allora $p_k(-x) = (-1)^k p_k(x)$

Dimostrazione. [...] □

Prop. 19 . Se $w(x)$ è pari, allora vale $\gamma_k = 0 \quad \forall k$

Dimostrazione. Dal teorema sopra, in γ_k vale $(xp_k, p_k)_w = \int_{-1}^{1} xp_k^2(x) w(x) dx = 0$ perché l'integranda è dispari □

Definizione 8.5 (polinomi ortogonali di Lagrange)**:** Per $w(x) \equiv 1$, allora i polinomi sono detti di Lagrange e valgono

$$p_0(x) = 1 \qquad p_1(x) = x \qquad p_{i+1}(x) = xp_i(x) - \beta_k p_{i-1}(x) \qquad \beta_k := \frac{(p_k, p_k)}{(p_{k-1}, p_{k-1})}$$

e si dimostra che $(p_k, p_k) = \left(k + \frac{1}{2}\right)^{-1}$ e dunque $\beta_k = \frac{k - \frac{1}{2}}{k + \frac{1}{2}} = \frac{2k-1}{2k+1}$

9 Approssimazione di Equazioni Differenziali Ordinarie

Definizione 9.1 (problema di Cauchy)**:** Dato $I = [a,b] \subseteq \mathbb{R}$, $f\colon I \times \mathbb{R} \to \mathbb{R}$ continua, $t \in I, y_0 \in \mathbb{R}$, trovare $y \in C^1(I)$ tale che

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(t_0) = y_0 \end{cases}$$

Definizione 9.2 (equazione integrale di Volterra)**:** $y(t) = y_0 + \int_{t_0}^{t} f(s, y(s))\, ds \quad \forall t \in (a,b)$

Definizione 9.3 (lipschitzianità)**:**

Definizione 9.4 (stabilità secondo Lyapunov)**:** Il problema di Cauchy è detto stabile se $\exists\, C > 0$ tale che $\forall\, \varepsilon > 0$ sufficientemente piccolo, il problema perturbato

$$\begin{cases} z'(t) = f(t, z(t)) + \delta(t) \\ z(t_0) = y_0 + \delta_0 \end{cases} \qquad \text{con } |\delta_0| < \varepsilon \text{ e } |\delta(t)| < \varepsilon\ \forall t \in I \tag{3}$$

soddisfa $|y(t) - z(t)| < C\varepsilon \quad \forall t \in I$

Teo. 84 . Se f Lipschitz $\implies$ Il problema di Cauchy ammette una e una sola soluzione ed è stabile secondo Lyapunov

9.1 Metodi ad un passo

Presi $N+1$ punti equispaziati in $I = [t_0, t_0 + T]$, $t_j = t_0 + jh$ con $h = \frac{T}{N}$, approssimo $y(t_j) =: y_j \approx u_j$ (con $u_0 = y_0$)

Definizione 9.5 (metodi a un passo)**:** Se $\forall\, j \geq 0$ la stima u_{j+1} dipende solo da u_j

metodo di Eulero in avanti $u_{j+1} = u_j + hf(t_j, u_j)$

Definiti ora $e_j := y_j - u_j = \overbrace{(y_j - z_j)}^{\text{discretiz.}} + \overbrace{(z_j - u_j)}^{\text{propagaz.err.}}$ e $z_{j+1} := y_j + hf(t_j, y_j)$

Definizione 9.6 (errore locale/globale di troncamento)**:** abbiamo

$$\text{Errore locale di troncamento:} \qquad \tau_j(h) := \frac{y_j - z_j}{h}$$
$$\text{Errore globale di troncamento:} \qquad \tau(h) := \max_{j=0,\dots,N} |\tau_j(h)|$$

Da Taylor

$$y_{j+1} = y_j + h \underbrace{y'(t)}_{f(t_j, y_j)} + \frac{h^2}{2} y''(\xi_j) \quad \Rightarrow \quad \tau_{j+1}(h) = \frac{h}{2} y''(\xi_j)$$

Si ha che

$$|y_{j+1} - z_{j+1}| \leq h\tau(h) \quad |z_{j+1} - u_{j+1}| \leq |y_j + hf(t_j, y_j) - u_j - hf(t_j, u_j)| \leq |e_j| + h\overbrace{|f(t_j, y_j) - f(t_j, u_j)|}^{L|y_j - u_j|}$$

$$|e_{j+1}| = |y_{j+1} - u_{j+1}| \leq |z_{j+1} - u_{j+1}| + |y_{j+1} - z_{j+1}| \leq (1 + hL)|e_j| + h\tau(h) \leq \frac{e^{LT} - 1}{L} \tau(h)$$

Teo. 86 . Il metodo di Eulero in avanti con $u_0 = y_0$, $f \in C^1(I \times \mathbb{R})$ e $M := \max_I |y''(\xi)|$ soddisfa

$$|e_{j+1}| \leq \frac{e^{L(t_{j+1} - t_0)} - 1}{L} \frac{M}{2} h$$

Con gli errori di arrotondamento $\overline{u_0} = y_0 + \eta_0$, $\overline{u_{j+1}} = \overline{u_j} + hf(t_j, \overline{u_j}) + \eta_{j+1}$ e $\eta = \max |\eta_k|$ diventa

$$|e_{j+1}| \leq e^{L(t_{j+1} - t_0)} \left(|\eta_0| + \frac{1}{L} \left(\frac{M}{2} h + \frac{\eta}{h} \right) \right)$$

non è più $\to 0$ per $h \to 0^+$, ma ci sarà un h_{opt} che minimizzerà l'errore

metodo di Eulero all'indietro/implicito $u_{j+1} = u_j + hf(t_{j+1}, u_{j+1})$

metodo di Crank-Nicolson $u_{j+1} = u_j + \frac{h}{2}(f(t_j, u_j) + f(t_{j+1}, u_{j+1}))$
Proviene da

$$y_{j+1} - y_j = \int_{t_j}^{t_{j+1}} f(s, y(s))ds \overset{\text{metodoTrapezi}}{\approx} \frac{h}{2}(\underbrace{f(t_j, y_j)}_{\text{eul.esplicito}} + \underbrace{f(t_{j+1}, y_{j+1})}_{\text{eul.implicito}})$$

metodo di Heun (versione esplicita di C-N) $u_{j+1} = u_j + \frac{h}{2}(f(t_j, u_j) + f(t_{j+1}, u_j + hf(t_j, u_j)))$

Prop. 20 . Se $hL < 1$, allora il metodo di Eulero all'indietro/implicito e il metodo di Crank-Nicolson hanno una e una sola soluzione.

Dimostrazione (Eulero). La funzione $g(z) = u_j + hf(t_{j+1}, z)$ è una contrazione per $hL < 1$ □

metodo a un passo generalizzato $u_{j+1} = u_j + h\Phi(t_j, u_j, h, f)$, da cui si può scrivere $y(t+h) = y(t) + h\Phi(t, y(t), h, f) + \overset{\text{residuo}}{\sigma(t,h)}$

Definizione 9.7 (errore di troncamento globale)**:** $\tau(h) = \max_{t\in[t_0,t_0+T]} \frac{|\sigma(t,h)|}{h}$

Definizione 9.8 (consistenza di un metodo)**:** Se $\lim_{h\to 0^+} \tau(h) = 0$. Consistenza di ordine p se $\tau(h) = O(h^p)$
errore di troncamento/discretizzazione controllato

Prop. 21 . Se $f \in C^1(I \times \mathbb{R})$ il metodo di Eulero all'indietro/implicito è consistente di ordine 1

Dimostrazione.

$$\Phi(t, y(t), h, f) = f(t+h, z) \quad z = y(t) + hf(t+h, z) \;\to\; z - y(t) = O(h) \;\to\; f(t+h, z) = f(t, y(t)) + O(h)$$

$$\sigma(t,h) = \overbrace{y(t+h) - y(t)}^{hf(t,y(t))+O(h^2)} - hf(t+h, z) = O(h^2)$$

□

Prop. 22 . Se $f \in C^2(I \times \mathbb{R})$ il metodo di Crank-Nicolson è consistente di ordine 2

Prop. 23 . Se $f \in C^2(I \times \mathbb{R})$ il metodo di Heun è consistente di ordine 2

Definizione 9.9 (zero stabilità per metodi a un passo)**:** Se $\exists h_0 > 0, C > 0$ tali che $\forall h \in (0, h_0)$ vale che $|z_j - u_j| \le C\varepsilon$ con $\{z_j\}$ soluzione del problema perturbato

$$z_{j+1} = z_j + h[\Phi(t_j, u_j, h, f) + \delta_{j+1}] \qquad z_0 = y_0 + \delta_0 \qquad \text{con } |\delta_i| \le \varepsilon$$

errore di propagazione controllato

Teo. 91 . Se Φ è Lipschitziana di costante Λ rispetto al secondo argomento uniformemente rispetto a $h \le h_0$ e $t \in [t_0, t_0 + T]$, allora il metodo è zero-stabile.

Dimostrazione.

$$|e_j| = |z_j - u_j| = |z_{j-1} + h\Phi(t_{j-1}, z_{j-1}) + h\delta_j - u_{j-1} - h\Phi(t_{j-1}, u_{j-1})| \le$$

$$\le |e_j| + h\underbrace{|\Phi(t_{j-1}, z_{j-1}) - \Phi(t_{j-1}, u_{j-1})|}_{\Lambda|e_j| \quad \Lambda \text{ indipedete da } h \text{ o } t_{j-1}} + h|\delta_j| \le (1 + h\Lambda)|e_{j-1}| + h\varepsilon \le e^{\Lambda T}\left(1 + \frac{1}{\Lambda}\right)\varepsilon$$

□

Definizione 9.10 (convergenza di un metodo)**:** Se $\lim_{h\to 0^+} |y_j - u_j| = 0$, convergenza di ordine p se $|y_j - u_j| = O(h^p)$

Teo. 92 . Un metodo $\begin{cases}\text{consistente (errore di troncamento controllato)}\\ \text{zero stabile (errore di propagazione controllato)}\end{cases} \implies$ convergente. Inoltre consistenza di ordine $p \implies$ convergenza di ordine p

Dimostrazione.

$$y_{j+1} = y_j + h\left[\Phi(t_j, y_j) + \frac{\sigma(t_j,h)}{h}\right] \qquad u_{j+1} = u_j + h\Phi(t_j, u_j) \qquad \frac{\sigma(t_j,h)}{h} \le \tau(h) \overset{h\to 0^+}{\longrightarrow} 0$$

La tesi da (91) con $z_j = y_j$, $\frac{\sigma(t_j,h)}{h} = \delta_{j+1}$ e $\varepsilon = \tau(h)$ □

Dato $\lambda \in \mathbb{C}^- = \{z \in \mathbb{C} \mid \mathrm{Re}(z) < 0\}$, introduciamo il problema modello

$$\begin{cases} y'(t) = \lambda \cdot y(t) \\ y(t_0) = 1 \end{cases}$$

Definizione 9.11 (assoluta stabilità)**:** Un metodo è assolutamente stabile se per il problema modello $u_j \overset{t_j\to+\infty}{\longrightarrow} 0$. La regione di assoluta stabilità è $A = \left\{z = h\lambda \in \mathbb{C} \mid u_j \overset{t_j\to+\infty}{\longrightarrow} 0\right\}$

Definizione 9.12 (*A*-stabilità)**:** Quanto a regione di assoluta stabilità è $A = \mathbb{C}^-$

9.2 Sistemi di Equazioni Differenziali Ordinarie

Ez

9.3 Metodi a più passi (multistep)

Definizione 9.13 (metodo multistep)**:** [...]

metodo del punto medio $u_{j+1} = u_{j-1} + 2hf(t_j, u_j)$

metodo di Simpson $u_{j+1} = u_{j-1} + \frac{h}{3}\left[f(t_{j-1}, u_{j-1}) + 4f(t_j, u_j) + f(t_{j+1}, u_{j+1})\right]$

metodo multistep generalizzato Dato un metodo multistep a $p+1$ passi ($b_{-1} = 0$ per metodo esplicito):

$$u_{j+1} = \sum_{k=0}^{p} a_k u_{j-k} + h\sum_{k=-1}^{p} b_k f(t_{j-k}, u_{j-k})$$

metodo di Adams $u_{j+1} = u_j + h\sum_{k=-1}^{p} b_k f(t_{j-k}, u_{j-k})$ $\begin{matrix}\text{implicito se} & b_{-1} \neq 0 & \text{(Adams-Moulton)}\\ \text{esplicito se} & b_{-1} = 0 & \text{(Adams-Baskford)}\end{matrix}$ con i coefficienti ottenuti interpolando f con un polinomio nei punti $t_{j-p}, \dots, t_j\,(, t_{j+1})$

Definizione 9.14 (residuo)**:** $\sigma(t,h) = y(t+h) - \sum_{k=0}^{p} a_k y(t-kh) - h\sum_{k=-1}^{p} b_k y'(t-kh)$

Teo. 97 . Un metodo multistep è consistente (9.8) $\iff \sum_{k=0}^{p} a_k = 1$ e $-\sum_{k=0}^{p} k a_k + \sum_{k=-1}^{p} b_k = 1$; è consistente di ordine $q \iff$ vale la condizione precedente e $\sum_{k=0}^{p}(-k)^q a_k + q\sum_{k=-1}^{p}(-k)^{q-1} b_k = 1$

Dimostrazione. Sostituendo $y(t-kh) = y(t) - khy'(t) + o(h)$ e $f(t-kh, y(t-kh)) = f(t, y(t)) + o(1)$ in 9.14

$$\begin{aligned}\sigma(t,h) &= y(t) - h\overbrace{y'(t)}^{f(t,y(t))} - \sum_{k=0}^{p} a_k\left(y(t) - khy'(t)\right) - h\sum_{k=-1}^{p} b_k f(t, y(t)) + o(h)\\ &= \left(1 - \sum_{k=0}^{p} a_k\right) y(t) + hf(t, y(t))\left(1 - \sum_{k=0}^{p} a_k k - \sum_{k=-1}^{p} b_k\right) + o(h)\\ \frac{\sigma(t,h)}{h} &= \frac{\left(1 - \sum_{k=0}^{p} a_k\right)}{h} y(t) + f(t, y(t))\left(1 - \sum_{k=0}^{p} a_k k - \sum_{k=-1}^{p} b_k\right) + o(1) \longrightarrow 0\end{aligned}$$

Per l'ordine q la dimostrazione è analoga sviluppando con Taylor fino al grado q e $q+1$ rispettivamente □

Osservazione Se $y \in C^2$, un metodo consistente lo è di ordine almeno 1

9.3.1 Stabilità per i metodi multistep

Studiando il problema $\begin{cases} y'(t) = 0 \\ y(0) = 1 \end{cases}$ si ottiene il metodo multistep $u_{j+1} = \sum_{k=0}^{p} a_k u_{j-k}$

Definizione 9.15 (1° polinomio caratteristico)**:** $\rho(r) = r^{p+1} - \sum_{k=0}^{p} a_k r^{p-k}$ associato al metodo multistep $u_{j+1} = \sum_{k=0}^{p} a_k u_{j-k}$

Osservazione Se il metodo è consistente $\Longrightarrow 1 - \sum_{k=0}^{p} a_k = 0 \Longrightarrow 1$ è radice del 1° polinomio caratteristico $\rho(r)$

Definizione 9.16 (zero stabilità per metodi multistep)**:** Se $\exists\, h_0 > 0, C > 0$ tali che $\forall\, h \in (0, h_0)$ vale che $|z_j - u_j| \leq C\varepsilon$ con $\{z_j\}$ soluzione del problema perturbato

$$z_{j+1} = \sum_{k=0}^{p} a_k z_{j-k} + h\delta_{j+1} \qquad z_k = y_k + \delta_k \text{ per } k = 0, \ldots, p \qquad \text{con } |\delta_i| \leq \varepsilon$$

Condizione della radice: Un metodo multistep soddisfa la condizione della radice se date $r_0, \ldots, r_p$ radici di $\rho(r)$ vale $|r_i| \leq 1$ e se $|r_i| = 1 \Rightarrow r_i$ è una radice semplice.

Teo. 98 . Un metodo multistep consistente è zero stabile $\Longleftrightarrow$ soddisfa la Stabilità per i metodi multistep

Dimostrazione ($\Longrightarrow$). Supponiamo $|r| > 1$ radice reale di $\rho(r)$ e definiamo $w_j = \varepsilon\frac{r^j}{r^p} =: \gamma r^j$, prendo $\delta_j = \begin{cases} w_j & j = 0, \ldots, p \\ 0 & j \geq p+1 \end{cases}$ e $u_0 = \ldots = u_p = 0 \Longrightarrow u_j = 0\ \forall j \geq 0$, da cui $z_j = w_j$ e $|z_j| = |z_j - u_j| \longrightarrow 0$ □

Esempio

metodo del punto medio	$\rho(r) = r^2 - 1$	$r_0 = 1, r_1 = -1$
metodo di Simpson	$\rho(r) = r^2 - 1$	$r_0 = 1, r_1 = -1$
metodo di Adams	$\rho(r) = r^{p-1} - r^p$	$r_1 = \ldots = r_p = 0$

Teo. 99 (teorema di equivalenza)**.** Un metodo multistep è convergente $\Longleftrightarrow$ è zero stabile e l'errore sui dati iniziali $\to 0$ per $h \to 0^+$. In tal caso il metodo ha ordine $q \Longleftrightarrow$ è consistente di ordine q

Teo. 100 (prima barriera di Dahlquist)**.** Un metodo multistep zero stabile a q passi non può avere ordine maggiore di $q + p$ con $p = 2$ per q pari, $p = 1$ per q dispari.

Parte III
Recap ed esame

10 Recap Barabba

Errore di Lagrange	$E_n(x) \coloneqq f(x) - \Pi_n f(x) = \dfrac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x)$
Minimi quadrati	$A^T \cdot A \cdot c = A^T \cdot b$
Metodi Stazionari	Covergenza $\iff \rho(B) < 1$ (serve consistenza) $\qquad \|B\| < 1 \implies \left\|e^{(k+1)}\right\| \leq \|B\| \cdot \left\|e^{(k)}\right\|$
Metodi di Richardson	Covergenza $\iff \dfrac{2 \cdot \mathrm{Re}(\lambda)}{\alpha \lvert\lambda\rvert^2} > 1 \quad \forall \lambda \in \Lambda(P^{-1}A) \qquad \alpha_{\text{opt}} = \frac{2}{\lambda_1 + \lambda_n} \qquad \rho(R_{\alpha_{\text{opt}}}) = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}$
Gershgorin	$\Lambda(A) \subseteq S_R \coloneqq \bigcup_{i=1}^{n} R_i \qquad \text{con} \quad R_i \coloneqq \left\{z \in \mathbb{C} \mid \lvert z - a_{ii}\rvert \leq \sum_{j=1, j\neq i}^{n} \lvert a_{ij}\rvert\right\}$
Matrice di Householder	$P = I - 2\dfrac{vv^T}{\|v\|_2^2} \qquad v = x \pm e_m \|x\|_2$
Shermann-Morris	$\left(A + uv^T\right)^{-1} = A^{-1} - \dfrac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1} u}$
Errore Newton-Cotes	$E_n(f) = \begin{cases} M_n (b-a)^{n+3} f^{(n+2)}(\xi) & n \text{ pari} \\ K_n (b-a)^{n+2} f^{(n+1)}(\xi) & n \text{ dispari} \end{cases} \qquad E_{n,2m}(f) \approx \frac{1}{2^{n+p-1}}(I_{n,2m}(f) - I_{n,m}(f))$
Errore Metodi ODE	$\lvert e_{j+1}\rvert \leq \dfrac{e^{L(t_{j+1}-t_0)} - 1}{L}\tau(h)$
Multistep Generalizzato	$u_{j+1} = \sum_{k=0}^{p} a_k u_{j-k} + h\sum_{k=0}^{p} b_k f(t_{j-k}, u_{j-k}) + hb_{-1} f(t_{j+1}, u_{j+1})$
Consistenza	consistente di ordine $q \iff \sum_{k=0}^{p} a_k = 1$ e $\sum_{k=0}^{p} (-k)^q a_k + q\sum_{k=-1}^{p} (-k)^{q-1} b_k = 1$
1° Polinomio Caratt.	$\rho(r) = r^{p+1} - \sum_{k=0}^{p} a_k r^{p-k}$ condizione della radice: $\lvert r_i\rvert < 0$ o semplici se $\lvert r_i\rvert = 1$
Condizione Zero Stabilità	zero stabile $\iff$ condizione della radice (serve consistenza)

Sistemi lineari	
Backsubstitution	$O(n^2)$ $=n^2$
Forwardsubstitution	

QR			
Standard	$Q \in \mathbb{R}^{m\times m}$ ortog	$Q \in \mathbb{R}^{m\times n}$ colonne ortonorm	$O(mn^2)$ $=2mn^2$
Ridotta	$R \in \mathbb{R}^{m\times n}$ triang inf	$R \in \mathbb{R}^{n\times n}$ triang sup	

LU			
Standard	L triang inf	U triang sup	$\det(A_{1:k,1:k}) \neq 0$ $\quad O(n^3)$ $=\frac{2}{3}n^3$
Pivoting parziale			$\det(A) \neq 0$
Thomas			A tridiagonale $\quad O(n)$ $=8n$
Cholesky			A SPD $\quad O(n^3)$ $=\frac{1}{3}n^3$

11 Scritto: formule e osservazioni per esercizi (da Epic)

11.1 Primo semestre

- Per ricavare il condizionamento del problema del calcolo di una funzione G nel punto d si usa la seguente formula, ricordando che lo jacobiano per $G : I \to \mathbb{R}$ corrisponde a G':

$$K_{ass}(d) = ||J_G(d)||, \quad K_{rel}(d) = \frac{||J_G(d)||\,||d||}{||G(d)||}$$

- Per stimare l'errore relativo in un metodo iterativo di matrice di iterazione B e soluzione x si può sfruttare che $Bx + f = x$, $Bx^{(k-1)} + f = x^{(k)}$ quindi, fissata una norma $||\cdot||$ vale:

$$\frac{||x - x^{(k)}||}{||x||} \leq ||B||^k$$

  Si ricorda che $||B||_\infty = \max_i \sum_j |B_{ij}|$

- Se un cerchio di Gershgorin è disgiunto dagli altri dello stesso tipo (riga o colonna) allora conterrà un unico autovalore, che dovrà essere reale. Per avere matrice con tutti autovalori reali è sufficiente avere cerchi riga o colonna tutti disgiunti fra loro. Posso sfruttare cerchi di Gershgorin per dare stima dello spettro, del raggio spettrale, degli autovalori minimo e massimo della matrice.

- Per risolvere problema ai minimi quadrati $\min_{x\in\mathbb{R}^2} ||b-Ax||_2$ devo risolvere sistema delle equazioni normali: $A^TAx = A^Tb$. Sfruttando la fattorizzazione QR di A mi riduco al sistema triangolare $Rx = Q^Tb$. Altrimenti posso affrontarlo come visto nel secondo semestre (vedere sotto).

- Formula Sherman-Morrison per il calcolo dell'inversa di una matrice a cui sommo una "modifica di rango uno" ottenendo l'inversa della matrice iniziale sommata ancora a una "modifica di rango uno":

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$$

11.2 Secondo semestre

- Per studiare la convergenza di un metodo iterativo $\phi(x)$ devo trovare i punti fissi e guardare il comportamento delle successioni: limitatezza e monotonia (disuguaglianze). Ricordo infatti che se ho convergenza, essa deve essere a un punto fisso di ϕ. Per calcolare l'ordine di convergenza uso teorema di ordine di convergenza guardando il valore delle derivate nei punti fissi: $\phi^{(k)}(\bar{x}) = 0$ allora ho ordine almeno $k + 1$.

- Per trovare polinomio interpolatore $\Pi_n f(x)$ della funzione $f(x)$ di grado massimo n nei nodi $\{x_i\}_{i=0}^n$, posso usare la scrittura di Lagrange $\Pi_n f(x) = \sum_{i=0}^n f(x_i)l_i(x)$, dove $l_i(x)$ sono i polinomi della base di Lagrange:

$$l_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}$$

  O posso usare la scrittura di Newton $\Pi_n f(x) = \sum_{i=0}^n f[x_0, \ldots, x_i]\omega_i(x)$ dove $\omega_i(x) = \prod_{k=0}^{k=i}(x - x_k)$ è il polinomio nodale riferito ai primi i nodi, mentre $f[x_0, \ldots, x_i]$ sono le differenze divise:

$$f[x_0, \ldots, x_j] = \frac{f[x_0, \ldots, x_{j-1}] - f[x_1, \ldots, x_j]}{x_0 - x_j}, \quad f[x_k] = f(x_k)$$

  Per trovare una stima dell'errore di interpolazione uso:

$$f(x) - \Pi_n f(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\omega_{n+1}(x)$$

- Formula dei trapezi e di Cavalieri-Simpson:

$$I_1(f) = \frac{(b-a)}{2}\left[f(a) + f(b)\right], \quad I_2(f) = \frac{(b-a)}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$

- Per la stima dell'errore di interpolazione con una spline lineare (lineare a tratti), uso la disuguaglianza:

$$||f - \Pi_h^1 f||_\infty \leq ||f''||_\infty \frac{h^2}{8}$$

  S spline cubica su $[a, b]$ è naturale se $S''(a) = 0 = S''(b)$.

- Per trovare g_n^* polinomio di migliore approssimazione di f devo risolvere il sistema delle equazioni normali. Prima trovo $\{\phi_i\}_{i=0}^n$ base dello spazio vettoriale su cui sto cercando g_n^*, non mi interessa sia ortogonale, quindi posso considerare base canonica per i polinomi, importante è che sia base, e scrivo $g_n^* = \sum_{i=0}^n c_i\phi_i$.

  (i) Caso continuo su intervallo $[a,b]$: per trovare i coefficienti c_i pongo le condizioni $(g_n^*, \phi_i) = (f, \phi_i)\ \forall i = 0, \dots, n$ dove $(h,g) = \int_a^b h(x)g(x)dx$

  (ii) Caso discreto su punti $\mathbf{x} = \{x_i\}_{i=0}^m$: risolvo sistema delle equazioni normali $A^TA\mathbf{c} = A^T\mathbf{y}$ dove $\mathbf{c} = (c_i)_{i=0}^n$ è il vettore colonna dei coefficienti di g_n^* mentre indicando con $tab\, g = g(\mathbf{x})$ vettore colonna di valutazione di un funzione g nei punti x_i ho, $y = tab\, f$ mentre

$$A = [tab\,\phi_0, \dots, tab\,\phi_n] = \begin{bmatrix} \phi_0(x_0) & \dots & \phi_n(x_0) \\ \dots & \dots & \dots \\ \phi_0(x_m) & \dots & \phi_n(x_m) \end{bmatrix}$$

- Dato problema generale di interpolazione lineare nello spazio V, data $\{\Phi_i\}_{i=0}^n$ base di V dove sto cercando interpolazione, e siano L_i i funzionali lineari del problema, si ha esistenza e unicità della soluzione se $\left|\left(L_i(\Phi_j)\right)_{i,j}\right| \neq 0$. I funzionali possono essere valutazione in un punto, integrale su intervallo, calcolo della derivata in un punto...

- Si ha in generale che una formula di quadratura a $n+1$ nodi è interpolatoria $\iff$ ha gdp almeno n. Grado di precisione n significa che deve essere esatta $\forall p \in \mathbb{P}_n$ quindi per trovare i pesi della formula posso imporre l'esattezza per $p = 1, x, \dots, x^n$.
  Per trovare nodi in modo che gdp $= k$, impongo che la formula di quadratura sia esatta $\forall p \in \mathbb{P}_k$, ovvero per $p = 1, x, \dots, x^k$.
  Una formula di quadratura interpolatoria a $n+1$ nodi ha gdp al massimo pari a $2n+1$, in questo caso è detta formula di Gauss. I nodi della formula di Gauss a $n+1$ nodi, sono dati dalle radici del polinomio monico di grado $n+1$ ortogonale a tutti i polinomi di grado $\leq n$ rispetto al prodotto $(g,h) = \int_a^b g(x)h(x)w(x)dx$ con $w(x)$ funzione peso: considero una base $\mathcal{B}$ di $\mathbb{P}_n$ (canonica va bene) e cerco il polinomio che annulla l'integrale indicato $\forall p \in \mathcal{B}$.

- Dato un metodo multistep a $p+1$ passi ($b_{-1} = 0$ per metodo esplicito):

$$u_{j+1} = \sum_{k=0}^{p} a_k u_{j-k} + h \sum_{k=-1}^{p} b_k f(t_{j-k}, u_{j-k})$$

  Per avere consistenza devo avere:

$$\sum_{k=0}^{p} a_k = 1, \quad \sum_{k=0}^{p} (-k)a_k + \sum_{k=-1}^{p} b_k = 1$$

  Un metodo ha ordine $q \geq 1$ se e solo se:

$$\sum_{k=0}^{p} (-k)^i a_k + i \sum_{k=-1}^{p} (-k)^{i-1} b_k = 1 \quad \forall i = 1, \dots, q$$

  Diciamo $\rho(r) = r^{p+1} - \sum_{k=0}^{p} a_k r^{p-k}$ il primo polinomio caratteristico associato.

  – Un metodo consistente è 0-stabile $\iff$ tutte le radici $r_0, \dots, r_p$ di $\rho(r)$ hanno modulo ≤ 1 e per $|r_i| = 1$ si ha r_i radice semplice.

  – Un metodo consistente è convergente $\iff$ è 0-stabile.

  Diciamo $\Pi(r) = \rho(r) - h\lambda \sum_{k=-1}^{p} b_k r^{p-k}$ il polinomio caratteristico associato.

  – Un metodo è assolutamente stabile se tutte le radici di $\Pi(r)$ hanno modulo < 1, ciò dipende dal valore di h.