Fondamenti di Calcolo Numerico

Anno Accademico 2014/2015

Chapter 1

Errori e condizionamento

1.1 Generalità

Oss. (analisi numerica): l'analisi numerica fornisce risultati approssimati di problemi complessi. Questa approssimazione avviene a due livelli:

1. i problemi in oggetto sono modellati attraverso equazioni della fisica matematica, che vivono in spazi funzionali di dimensione infinita. L'analisi numerica fornisce metodi che danno soluzioni in spazi vettoriali di dimensione finita e uno dei compiti dell'analisi numerica è quindi quello di stabilire quantitativamente in quale relazione stanno queste soluzioni approssimate con il problema analitico originale;

2. i metodi numerici sono implementati in forma di algoritmi in un calcolatore che usa un'aritmetica finita, ovvero approssima i numeri in gioco utilizzando un numero finito di cifre. Un altro compito dell'analisi numerica è quello di stabilire quantitativamente l'impatto di questa approssimazione a un numero finito di cifre sull'accuratezza della soluzione approssimata calcolata attraverso gli algoritmi.

Oss. (errori): l'analisi numerica tratta in modo rigoroso diversi tipi di errori:

- Errore di approssimazione: l'analisi numerica rappresenta in dimensione finita oggetti che vivono in spazi di dimensione infinita. Ne consegue che i metodi numerici risolvono equazioni associate a un modello

che fornisce una rappresentazione approssimata della realtà. L'errore di approssimazione è l'errore introdotto dal modello di descrizione della realtà.

- Errori di misura e di rappresentazione dei dati: l'analisi numerica scrive algoritmi che prendono dati in input. In matematica applicata, questi dati sono forniti da procedure di misurazione effettuate da strumenti di misura. Questi strumenti introducono un errore sul dato, tipicamente caratterizzato da specifiche proprietà statistiche. Gli algoritmi dell'analisi numerica devono tener conto di come questo errore si propaga dal dato alla soluzione.

- Errore dovuto alla rappresentazione in aritmetica finita: sono gli errori di troncamento e arrotondamento che si introducono rappresentando un numero reale per mezzo di un processo finito.

- Errore algoritmico: è l'errore introdotto dall'algoritmo di calcolo.

1.2 Errori

Oss. (rappresentazione decimale): considero $x \in \mathbb{R}$. Una possibile rappresentazione di x è

$$x = sgn(x)\beta^p \sum_{i=1}^{\infty} d_i \beta^{-i}, \tag{1.1}$$

con $0 \leq d_i < \beta$ e $d_1 \neq 0$. Il numero intero p è detto esponente della rappresentazione, i numeri naturali d_i sono detti cifre della rappresentazione e $\sum_{i=1}^{\infty} d_i \beta^{-i}$ è detta mantissa. In seguito considereremo il caso in cui $sgn(x) = 1$ e $\beta = 10$ ma tutti i risultati sono immediatamente generalizzabili.
Oss. (aritmetica finita): una rappresentazione in aritmetica finita (o floating point) di un numero reale x è data da

$$fl_{t,M,N}(x) = \begin{cases} 10^p \sum_{i=1}^{t} \delta_i 10^{-i} & -N \leq p \leq M \\ \text{underflow} & p < -N \\ \text{overflow} & p > M. \end{cases} \tag{1.2}$$

Quindi per rappresentare in aritmetica finita un numero reale è necessario fornire tre numeri naturali t, N, M e un set di t numeri naturali $\delta_1, \ldots, \delta_t$.

Si ha una rappresentazione floating point differente a seconda del modo differente in cui si scelgono le δ_i in relazione ai d_i. Esempi di rappresentazione floating point sono troncamento e arrotondamento.
Def. (precisione di macchina): il numero η tale che

$$\left|\frac{fl(x)-x}{x}\right| < \eta \tag{1.3}$$

per ogni x, è detto precisione di macchina. L'errore che si commette rappresentando x con $fl(x)$ è detto errore di rappresentazione. L'errore di rappresentazione è minore se si usa l'arrotondamento piuttosto che il troncamento.
Esempio (errore di una somma): siano x_1 e x_2 con rappresentazione floating point $fl(x_1)$ e $fl(x_2)$. Sia $u = x_1 + x_2$. Per definizione di rappresentazione floating point si ha che $fl(u) = fl(x_1) + fl(x_2)$ per cui

$$\Delta u := |u - fl(u)| \leq |x_1 - fl(x_1)| + |x_2 - fl(x_2)| = \Delta x_1 + \Delta x_2. \tag{1.4}$$

Quindi

$$\frac{\Delta u}{|u|} \leq \frac{\Delta x_1 + \Delta x_2}{|x_1 + x_2|}. \tag{1.5}$$

Se x_1 e x_2 hanno segno opposto e sono vicini in valore assoluto l'errore relativo su u può diventare estremamente grande (si parla in questo caso di errore di cancellazione).

1.3 Condizionamento

Oss. (stabilità algoritmica): in analisi numerica lo stesso problema può essere affrontato con algoritmi diversi e non necessariamente il risultato finale è lo stesso. Considero, ad esempio, il calcolo di e^{-9} effettuato utilizzando lo sviluppo in serie di Taylor di e^x:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} \ldots \tag{1.6}$$

Posso calcolare e^{-9} in due modi: ponendo $x = -9$ nella (1.6) o ponendo $x = 9$ in

$$e^{-x} = \frac{1}{1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} \ldots} \tag{1.7}$$

Se uso il primo modo, le prime sette cifre del risultato per $n \geq 38$ si stabilizzano sul valore $-0.9639455 \times 10^{-4}$, che è sicuramente sbagliato. Se uso il secondo modo, le prime sette cifre del risultato per $n \geq 23$ si stabilizzano sul valore 0.1234107×10^{-3}, che e' corretto.
Def. (mal posizione): un problema è detto mal posto nel senso di Hadamard se la sua soluzione non esiste, oppure non è unica oppure dipende in modo non continuo dai dati.
Def. (condizionamento): un problema scritto in dimensione finita è tipicamente ben posto. Tuttavia, la sua soluzione numerica può essere instabile. Questa instabilità può essere conseguenza della scelta sbagliata del tipo di algoritmo. Oppure può essere una instabilità intrinseca, che permane qualunque sia l'algoritmo che si applica per la risoluzione. Tipicamente, un problema numerico (ovvero formulato in dimensione finita) che nasce dalla discretizzazione di un problema mal posto è numericamente instabile. Questa patologia nel discreto, figlia di una patologia nel continuo, è detta cattivo condizionamento. Il condizionamento di un problema discreto è un numero (o un ordine di grandezza) che misura l'instabilità numerica intrinseca. Il modo con cui questo numero viene calcolato dipende dal tipo di problema.

1.4 Costo computazionale

Def. (tempo di calcolo): il costo computazionale di un algoritmo è il numero di operazioni elementari (prodotti, divisioni) che l'algoritmo richiede per la sua esecuzione. Il tempo di calcolo è il prodotto tra il costo computazionale e il tempo con cui viene eseguita una singola operazione.
Esempio (sistema lineare): nell'ipotesi in cui una singola operazione venga svolta in 10^{-6} secondi, se si risolve un sistema lineare con il metodo di Cramer (la componente della soluzione è espressa come rapporto tra determinanti di matrici di ordine pieno), per un ordine uguale a 12, il tempo di calcolo è di 104 minuti. Per lo stesso ordine, usando il metodo di Gauss, si impiegano 7.2×10^{-4} secondi.

Chapter 2

Norme

2.1 Vettori

Def. (spazi lineari): uno spazio lineare è un insieme X di elementi detti vettori, per il quale sono ben definite le operazioni lineari di somma tra elementi e di moltiplicazione per uno scalare (complesso). Un insieme di vettori $\{u_1, \ldots, u_N\}$ è detto linearmente indipendente se

$$a_1 u_1 + \ldots + a_N u_N = 0 \Rightarrow a_1 = \ldots = a_N = 0. \tag{2.1}$$

La dimensione di X, dim$X$, è il più grande numero di vettori linearmente indipendenti di X. Un sottospazio di X è un sottoinsieme di X che è anche spazio lineare. Sia S un sottoinsieme di X. Allora span(S) è il sottospazio di X costituito da tutte le combinazioni lineari di elementi di S. Sia N la dimensione di X: allora un qualunque insieme di N vettori linearmente indipendenti è detto base di X. Se $x \in X$ e $\{u_i\}_{i=1}^N$ è una base, allora

$$x = \sum_{i=1}^{N} a_i u_i. \tag{2.2}$$

Def. (norma di un vettore): si definisce norma di un vettore $x \in X$ la funzione $\| \cdot \| : x \to \mathbb{R}^+$ tale che

1. $\|x\| \geq 0$; $\|x\| = 0 \Rightarrow x = 0$;
2. $\|ax\| = |a|\|x\|$;

3. $\|x+y\| \leq \|x\| + \|y\|$.

Esempi (norme p): considero la famiglia a un parametro di funzioni $\|\cdot\|_p : X \to \mathbb{R}^+$ con $1 \leq p < \infty$ e

$$\|x\|_p = \left(\sum_{i=1}^{N} |x_i|^p \right)^{\frac{1}{p}}, \tag{2.3}$$

con $x = \sum_{i=1}^{N} x_i e_i$ e $\{e_i\}_{i=1}^{N}$ base canonica in X. $\|\cdot\|_p$ definisce una norma. Dimostrare i punti 1. e 2. è semplice. Per quanto riguarda il punto 3., dimostro prima la disuguaglianza di Holder

$$\sum_{k=1}^{N} |x_k y_k| \leq \|x\|_p \|y\|_q \tag{2.4}$$

con $1/p + 1/q = 1$. Poiche' la tangente alla curva $y = t^\alpha$ sta sopra la curva a $t = 1$, allora

$$t^\alpha \leq \alpha t + 1 - \alpha \tag{2.5}$$

da cui, sostituendo $t = t/v$,

$$t^\alpha v^\beta \leq \alpha t + \beta v, \tag{2.6}$$

dove $\alpha = 1/p$ e $\beta = 1/q$. Sostituendo

$$t = \frac{|x_k|^p}{\|x\|_p^p} \qquad v = \frac{|y_k|^q}{\|y\|_q^q} \tag{2.7}$$

si ha

$$\sum_k |x_k y_k| \leq \|x\|_p \|y\|_q. \tag{2.8}$$

Per dimostrare la disuguaglianza triangolare:

$$\sum_k |x_k + y_k|^p \leq \sum |x_k| |x_k + y_k|^{p-1} + \sum |y_k| |x_k + y_k|^{p-1}. \tag{2.9}$$

Applicando la disuguaglianza di Hölder ai due termini al secondo membro e tenendo conto del fatto che $q(p-1) = p$:

$$(\sum_k |x_k + y_k|^p)^{1-1/q} \leq \|x\|_p + \|y\|_p \tag{2.10}$$

e poichè $1 - 1/q = 1/p$, questa è la disuguaglianza da dimostrare.
Esempio (norma infinito): definisco il funzionale $\|\cdot\|_\infty : X \to \mathbb{R}^+$ tale che

$$\|x\|_\infty = \max_i |x_i|. \tag{2.11}$$

Anche questo funzionale definisce una norma. Inoltre vale

$$\|x\|_\infty = \lim_{p\to\infty} \|x\|_p. \tag{2.12}$$

Infatti

$$\|x\|_p^p \leq N\|x\|_\infty^p \tag{2.13}$$

per cui

$$\|x\|_p \leq N^{1/p}\|x\|_\infty \tag{2.14}$$

e quindi

$$\lim_{p\to\infty} \|x\|_p \leq \|x\|_\infty. \tag{2.15}$$

D'altra parte

$$\|x\|_p^p = \|x\|_\infty^p + \sum_{i=2}^{N} |x_i|^p \geq \|x\|_\infty^p. \tag{2.16}$$

Teo. (continuità): Th.: $\|\cdot\|$ è una funzione continua di x
Dim.: dalla disuguaglianza triangolare

$$\|x + \delta\| - \|x\| \leq \|\delta\| \tag{2.17}$$

e

$$\|x + \delta\| - \|x\| \geq -\|\delta\|, \tag{2.18}$$

per cui

$$|\|x + \delta\| - \|x\|| \leq \|\delta\| \leq \sum_{i=1}^{N} |\delta_i| \|e_i\| \leq \|\delta\|_\infty \sum_{i=1}^{N} \|e_i\|. \tag{2.19}$$

Prendo δ tale che

$$\|\delta\|_\infty < \frac{\epsilon}{\sum_{i=1}^{N} \|e_i\|} \tag{2.20}$$

e quindi ho la tesi.
Teo. (equivalenza tra norme): Th.: per ogni coppia di norme $\|\cdot\|$ e $\|\cdot\|'$ esistono due numeri positivi m, M tali che

$$m\|x\|' \leq \|x\| \leq \|M\|x\|'. \tag{2.21}$$

Dim.: scelgo $\|\cdot\|' = \|\cdot\|_\infty$ (se due norme sono ciascuna equivalente a $\|\cdot\|_\infty$ allora sono equivalenti tra loro). Sia $S = \{x \;\; t.c. \;\; \|x\|_\infty = 1\}$. S è compatto in X e $\|\cdot\|$ è continua. Per Weierstrass assume massimo e minimo su S, ovvero esistono x_1 e x_2 tali che

$$0 < \|x_1\| \leq \|x\| \leq \|x_2\| \tag{2.22}$$

per ogni $x \in S$. In particolare, poichè $y/\|y\|_\infty \in S$,

$$\|x_1\| \leq \frac{\|y\|}{\|y\|_\infty} \leq \|x_2\| \tag{2.23}$$

e quindi

$$m\|y\|_\infty \leq \|y\| \leq M\|y\|_\infty \tag{2.24}$$

con $m = \|x_1\|$ e $M = \|x_2\|$.

2.2 Matrici

Def. (norma naturale o norma indotta): se $\|\cdot\|$ è una norma vettoriale ben definita e A è una matrice, si definisce norma naturale di A o norma di A indotta dalla norma vettoriale $\|\cdot\|$, la funzione

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}. \tag{2.25}$$

Esistono due definizioni equivalenti (in modo ovvio) alla precedente. La prima:

$$\|A\| = \max_{\|u\|=1} \|Au\|. \tag{2.26}$$

La seconda parte dall'osservazione che $\|Au\|$ è una funzione continua di u e quindi raggiunge un massimo sul compatto di $\mathbb{R}^n$ rappresentato dalla sfera unitaria. Per cui:

$$\|A\| = \|Ay\| \tag{2.27}$$

per un qualche y tale che $\|y\| = 1$.
Oss. (buona definizione): dimostro che la precedente è una buona definizione di norma. I primi due punti sono ovvi. Per quanto riguarda la disuguaglianza triangolare:

$$\|A + B\| = \|(A + B)y\| \leq \|Ay\| + \|By\| \leq \|A\| + \|B\|. \tag{2.28}$$

Inoltre vale

$$\|AB\| \leq \|A\|\|B\|. \tag{2.29}$$

Infatti

$$\|AB\| = \|(AB)y\| \leq \|A\|\|By\| \leq \|A\|\|B\|\|y\| = \|A\|\|B\|. \tag{2.30}$$

Oss. (norma infinito): considero la norma indotta

$$\|A\|_\infty = \sup_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty}. \tag{2.31}$$

Sia x definito come

$$x_k = \begin{cases} \overline{a}_{Jk}/|a_{Jk}| & a_{Jk} \neq 0 \\ 0 & a_{Jk} = 0 \end{cases} \tag{2.32}$$

dove $\sum_{k=1}^n |a_{Jk}| = \max_j \sum_k |a_{jk}|$. Ovviamente $\|x\|_\infty = 1$ e

$$\|Ax\|_\infty = \sum_k |a_{Jk}| \leq \|A\|_\infty. \tag{2.33}$$

D'altra parte, esiste y $\|y\|_\infty = 1$

$$\|A\|_\infty = \|Ay\|_\infty = max_j |\sum_k a_{jk}y_k| \leq max_j \sum_k |a_{jk}| \tag{2.34}$$

e quindi

$$\|A\|_\infty = max_j \sum_k |a_{jk}|. \tag{2.35}$$

Oss. (norma 1): considero la norma indotta

$$\|A\|_1 = \|Ay\|_1 \tag{2.36}$$

con $\|y\|_1 = 1$. Allora

$$\|A\|_1 = \sum_j |\sum_k a_{jk}y_k| \leq \max_m \sum_j |a_{jm}|. \tag{2.37}$$

Se il massimo è raggiunto per $m = K$ allora

$$\|Ae_K\|_1 = \sum_j |\sum_k a_{jk}\delta_{kK}| = \sum_j |a_{jK}| \leq \|A\|_1 \tag{2.38}$$

e quindi

$$\|A\|_1 = max_k \sum_j |a_{jk}|. \tag{2.39}$$

Def. (raggio spettrale): è data una matrice quadrata A di ordine n. Il numero reale o complesso λ è detto autovalore di A se esiste un vettore x tale che $Ax = \lambda x$. Si definisce raggio spettrale di A il numero reale positivo

$$\rho(A) = \max_s |\lambda_s(A)| \tag{2.40}$$

dove λ_s denota l's-esimo autovalore di A.
Oss. (norma 2): si dimostra che

$$\|A\|_2 = \sqrt{\rho(A^*A)}. \tag{2.41}$$

Infatti, prendo y, $\|y\|_2 = 1$ tale che $\|A\|_2 = \|Ay\|_2$. Quindi

$$\|A\|_2^2 = (Ay)^*(Ay) = y^*A^*Ay. \tag{2.42}$$

A^*A è autoaggiunto e quindi ha un insieme completo di n autovettori ortonormali $u_1, \ldots, u_n$. Segue che

$$\lambda_s = u_s^*A^*Au_s. \tag{2.43}$$

Poichè

$$y = \sum_s \alpha_s u_s \tag{2.44}$$

allora, dalla (2.42)

$$\|A\|_2^2 = \sum_s \lambda_s |\alpha_s|^2 \leq \rho(A^*A). \tag{2.45}$$

Se scelgo $y = u_s$ dove $\lambda_s = \rho(A^*A)$, ho che

$$\|Au_s\|_2 = \rho^{1/2}(A^*A). \tag{2.46}$$

Teo. (equivalenza): per ogni coppia di norme indotte $\|\cdot\|$ e $\|\cdot\|'$, esistono due numeri positivi m e M tali che

$$m\|A\|' \leq \|A\| \leq M\|A\|'. \tag{2.47}$$

La dimostrazione è analoga all'analogo teorema per i vettori.
Oss. (norma 2 e raggio spettrale): valgono i seguenti fatti:

- la norma 2 e il raggio spettrale coincidono per matrici autoaggiunte;
- il raggio spettrale è sempre minore uguale della norma indotta di una matrice;
- per ogni matrice, esiste una norma indotta che può essere approssimata bene quanto si vuole dal raggio spettrale; ovvero: per ogni A e per ogni $\epsilon > 0$ esiste una norma indotta $\|\cdot\|$ tale che

$$\rho(A) \leq \|A\| \leq \rho(A) + \epsilon. \tag{2.48}$$

Chapter 3

Sistemi lineari

3.1 Condizionamento

Def. (numero di condizionamento): sia A una matrice quadrata invertibile e considero il sistema lineare

$$y = Ax. \tag{3.1}$$

Se $\|\delta y\|$ è una perturbazione di y, allora

$$\delta y = A\delta x \tag{3.2}$$

per cui

$$\delta x = A^{-1}\delta y. \tag{3.3}$$

Da cui

$$\|\delta x\| \leq \|A^{-1}\|\|\delta y\|. \tag{3.4}$$

D'altra parte

$$\|Ax\| \leq \|A\|\|x\| \tag{3.5}$$

per cui

$$\|x\| \geq \frac{\|Ax\|}{\|x\|}. \tag{3.6}$$

Concludendo

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|\|\delta y\|}{\|y\|/\|A\|} = \|A\|\|A^{-1}\|\frac{\|\delta y\|}{\|y\|}. \tag{3.7}$$

Il numero reale positivo

$$C(A) = \|A\|\|A^{-1}\| \tag{3.8}$$

è detto numero di condizionamento di A ed è una misura della stabilità numerica intrinseca del sistema lineare. Se il numero di condizionamento è grande, il sistema lineare è intrinsecamente instabile.
N.B. 1: il numero di condizionamento di un sistema è una proprietà della matrice.
N.B. 2: il valore esatto del numero di condizionamento dipende dalla scelta della norma matriciale. Però queste sono tutte norme indotte (lo si vede da come ho definito $C(A)$) e quindi equivalenti, per cui l'ordine di grandezza del numero di condizionamento non cambia. Questa è la ragione per cui $C(A)$ è una buona misura dell'instabilità numerica.
Oss. (condizionamento e norma 2): se A è autoaggiunta, $\|A\|_2 = \max_s |\lambda_s|$. D'altra parte se A è autoaggiunta anche A^{-1} è autoaggiunta e $\|A^{-1}\|_2 = 1/min_s|\lambda_s|$. Ne consegue che in norma 2, per matrici autoaggiunte

$$C(A) = \frac{\max_s |\lambda_s|}{\min_s |\lambda_s|}. \tag{3.9}$$

3.2 Eliminazione Gaussiana

Esempio: considero il sistema lineare

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= y_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= y_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= y_3 \end{aligned} \tag{3.10}$$

Per eliminare la variabile x_1 dalla seconda e terza equazione, moltiplico la prima, rispettivamente, per a_{21}/a_{11} e a_{31}/a_{11}, sottraggo e ottengo il sistema

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= y_1 \\ a^{(2)}_{22}x_2 + a^{(2)}_{23}x_3 &= y^{(2)}_2 \\ a^{(2)}_{32}x_2 + a^{(2)}_{33}x_3 &= y^{(2)}_3 \end{aligned} \tag{3.11}$$

dove

$$a^{(2)}_{22} = a_{22} - a_{12}\frac{a_{21}}{a_{11}} \quad ; \quad a^{(2)}_{23} = a_{23} - a_{13}\frac{a_{21}}{a_{11}} \quad ; \quad y^{(2)}_2 = y_2 - y_1\frac{a_{21}}{a_{11}}; \tag{3.12}$$

$$a_{32}^{(2)} = a_{32} - a_{12}\frac{a_{31}}{a_{11}} \quad ; \quad a_{33}^{(2)} = a_{33} - a_{13}\frac{a_{31}}{a_{11}} \quad ; \quad ; \quad y_3^{(2)} = y_3 - y_1\frac{a_{31}}{a_{11}}. \tag{3.13}$$

Per eliminare la variabile x_2 dalla terza equazione della (3.11), moltiplico la seconda per $a_{32}^{(2)}/a_{22}^{(2)}$, sottraggo e ottengo il sistema

$$\begin{array}{c} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = y_1 \\ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 = y_2^{(2)} \\ a_{33}^{(3)}x_3 = y_3^{(3)} \end{array} \tag{3.14}$$

dove

$$a_{33}^{(3)} = a_{33}^{(2)} - a_{23}^{(2)}\frac{a_{32}^{(2)}}{a_{22}^{(2)}} \quad ; \quad y_3^{(3)} = y_3^{(2)} - y_2^{(2)}\frac{a_{32}^{(2)}}{a_{22}^{(2)}}. \tag{3.15}$$

Dall'ultima equazione di (3.15) si trova x_3 e si risale alla seconda e alla terza equazione per determinare x_2 e x_1 risolvendo equazioni di primo grado.
Oss. (generalizzazione): considero il sistema lineare

$$Ax = y \tag{3.16}$$

con A una matrice non singolare $n \times n$. Indico con $A^{(k)}$ la matrice ridotta che descrive il sistema equivalente prima della eliminazione della componente x_k di A e con $a_{ij}^{(k)}$ l'elemento ij di tale matrice. Il processo di eliminazione di cui al punto precedente può essere generalizzato con il seguente schema:

$$a_{ij}^{(1)} = a_{ij} \quad \forall\ i, j \tag{3.17}$$

$$\begin{array}{ll} a_{ij}^{(k)} = 0 & j \le k-1 \ , \ \ i \ge k \\ a_{ij}^{(k)} = a_{ij}^{(k-1)} & i \le k-1 \\ a_{ij}^{(k)} = a_{ij}^{(k-1)} - a_{k-1,j}^{(k-1)}\frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} & i \ge k \ , \ \ j \ge k. \end{array} \tag{3.18}$$

Per quanto riguarda l'aggiornamento del dato y si ha:

$$\begin{array}{ll} y_1^{(1)} = y_1 & \\ y_i^{(k)} = y_i^{(k-1)} & i \le k-1 \\ y_i^{(k)} = y_i^{(k-1)} - y_{k-1}^{(k-1)}\frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} & i \ge k \end{array} \tag{3.19}$$

Questa triangolarizzazione di A porta alle formule finali di risoluzione del sistema:

$$x_n = \frac{y_n^{(n)}}{a_{nn}^{(n)}} \tag{3.20}$$

$$x_i = \frac{1}{a_{ii}^{(n)}}(y_i^{(n)} - \sum_{j=i+1}^{n} a_{ij}^{(n)} x_j) \quad i = n-1, \ldots, 1 \tag{3.21}$$

N.B.: i conti in questo schema sono effettuati assumendo $a_{kk}^{(k)} \neq 0$ per ogni k.
Oss. (costo computazionale): considero le formule nell'osservazione precedente. Fisso k e per calcolare $a_{i,k-1}^{(k-1)}/a_{k-1,k-1}^{(k-1)}$ devo fare $n-k+1$ divisioni. Poi, a i fissato, faccio il prodotto con $a_{k-1,j}^{(k-1)}$ e poi vario i. Ne consegue che ho $(n-k+1)^2$ prodotti. Quindi per triangolarizzare A devo svolgere

$$N_1 = \sum_{k=2}^{n}[(n-k+1) + (n-k+1)^2] = \frac{n^3}{3} - \frac{n}{3} \tag{3.22}$$

operazioni, dove, per fare questo calcolo, ho usato le due formule

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2} \quad ; \quad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}. \tag{3.23}$$

Per calcolare la soluzione uso le (3.20),(3.21). Fisso i e ho $(n-i)$ prodotti più una divisione; sommo su i da 1 a $n-1$ e aggiungo la divisione in (3.20) e trovo

$$N_2 = \sum_{i=1}^{n-1}(n-i) + n = n^2 - \frac{n(n+1)}{2} + n = \frac{n^2}{2} + \frac{n}{2} \tag{3.24}$$

operazioni. Infine, per aggiornare i vettori y ho solo

$$N_3 = \sum_{k=2}^{n}(n-k+1) = \frac{n^2}{2} - \frac{n}{2} \tag{3.25}$$

prodotti (in quanto le divisioni sono gia' state fatte per aggiornare A). In conclusione il numero di operazioni richieste per l'eliminazione Gaussiana è

$$N_{tot} = \frac{n^3}{3} + n^2 - \frac{n}{3}. \tag{3.26}$$

3.3 Pivoting

Oss. (pivot): in un precedente nota bene ho evidenziato il fatto che il processo di eliminazione Gaussiana si interrompe se si trova un $a_{kk}^{(k)} = 0$. Tuttavia sarebbe opportuno poter portare avanti il processo il più possibile.

L'idea è questa: per $k = 1$ provo (i_1, j_1) tale che $a_{i_1,j_1} \neq 0$ ed elimino x_1 da tutte le equazioni diverse dalla i_1-esima. In $A^{(2)}$ cerco (i_2, j_2) tale che $a^{(2)}_{i_2,j_2} \neq 0$ e così via. Trovo pertanto r elementi diversi da zero $a^{(k)}_{i_k,j_k} \quad k = 1, \ldots, r$ che chiamo pivot. Il processo si interrompe se dopo l'r-esimo step tutte le equazioni sono degeneri.

Oss. (pivotizzazione e sistemi equivalenti): considero le matrici

$$P = (e^{(i_1)}, \ldots, e^{(i_n)}) \tag{3.27}$$

e

$$Q = (e^{(j_1)}, \ldots, e^{(j_n)}). \tag{3.28}$$

$P^T A$ è una modificazione di A in cui le righe sono permutate, mentre AQ è una modificazione di A in cui le colonne sono permutate. D'altra parte, i sistemi lineari

$$Ax = y \tag{3.29}$$

e

$$Bf = g \quad B = P^T AQ \quad f = Q^T x \quad g = P^T y \tag{3.30}$$

sono equivalenti. Suppongo che la matrice A abbia rango r. Allora posso trovare r coppie di indici $(i_1, j_1), \ldots, (i_r, j_r)$ a cui corrispondono i pivot $a^{(1)}_{i_1,j_1}, \ldots, a^{(r)}_{i_r j_r}$. Costruisco le matrici di permutazione

$$P = (e^{(i_1)}, \ldots, e^{(i_r)}, \ldots, e^{(i_n)}) \tag{3.31}$$

e

$$Q = (e^{(j_1)}, \ldots, e^{(j_r)}, \ldots, e^{(j_n)}) \tag{3.32}$$

L'eliminazione Gaussiana (r volte) con pivoting vista nell'osservazione precedente, applicata al sistema (3.29) corrisponde all'eliminazione Gaussiana tradizionale, senza pivoting, applicata al sistema (3.30).

3.4 Fattorizzazione LU

Def. (fattorizzazione LU): è data una matrice A di ordine $n \times n$. Si dice fattorizzazione LU di A la trasformazione di A che permette di scrivere la matrice nella forma

$$A = LU \tag{3.33}$$

con L una matrice $n \times n$ unit lower triangular e U una matrice $n \times n$ upper triangular. Se A è LU-fattorizzabile, il sistema $Ax = y$ è facilmente risolvibile. Infatti $LUx = y$ implica $Ux = L^{-1}y =: g$ e quindi

$$x_n = \frac{1}{u_{nn}} g_n \tag{3.34}$$

e

$$x_i = \frac{1}{u_{ii}} (g_i - \sum_{j=i+1}^{n} u_{ij} x_j) \quad i = n-1, \ldots, 1. \tag{3.35}$$

Lo studio della fattorizzazione LU di una matrice si divide in due punti: (1) determinare le condizioni sufficienti per cui A è fattorizzabile; (2) determinare la forma esplicita degli elementi di L e U. Considero due teoremi cha risolvono questi due problemi. Il primo teorema affronta il problema (2), ovvero assume che la fattorizzazione LU sia possibile e mostra come effettuarla. Il secondo teorema fornisce una condizione sufficiente per la fattorizzazione.

Teo. (fattorizzazione LU ed eliminazione Gaussiana): Hp.: l'eliminazione Gaussiana è possible ovvero $a_{kk}^{(k)} \neq 0$ per ogni $k = 1, \ldots, n$. Th.: (1) $A = LU$ con $U = A^{(n)}$ e

$$L_{ik} = \begin{cases} 0 & i < k \\ 1 & i = k \\ \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} & i > k \end{cases} \tag{3.36}$$

(2) $\det A = \det U$; (3) $Ly^{(n)} = y$, dove y è il vettore dei dati.

Dim.: poichè L è unit lower triangular e U è upper triangular

$$(LU)_{ij} = \sum_{k=1}^{\min(i,j)} l_{ik} u_{kj} = \sum_{k=1}^{\min(i,j)} l_{ik} a_{kj}^{(n)}. \tag{3.37}$$

Per la parte upper triangular $a_{kj}^{(n)} = a_{kj}^{(k)}$. Sia $i \leq j$:

$$(LU)_{ij} = \sum_{k=1}^{i-1} l_{ik} a_{kj}^{(k)} + l_{ii} a_{ij}^{(i)}. \tag{3.38}$$

Nel primo addendo, $k < i$, $i \leq j$ e $k < j$ per cui

$$l_{ik} a_{kj}^{(k)} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)}. \tag{3.39}$$

Dall'ultima equazione in (3.18) trovo che

$$a_{ij}^{(k)} - a_{ij}^{(k+1)} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)} \tag{3.40}$$

per cui $l_{ik}a_{kj}^{(k)} = a_{ij}^{(k)} - a_{ij}^{(k+1)}$. Quindi

$$\sum_{k=1}^{i-1} l_{ik}a_{kj}^{(k)} = \sum_{k=1}^{i-1}[a_{ij}^{(k)} - a_{ij}^{(k+1)}] = a_{ij} - a_{ij}^{(i)}. \tag{3.41}$$

Poichè $l_{ii}a_{ij}^{(i)} = a_{ij}^{(i)}$, e poichè i conti sono analoghi per $i > j$, ho la tesi (1). D'altra parte: $\det A = \det LU = \det A \det U = \det U$. Infine:

$$(Ly^{(n)})_i = \sum_{k=1}^{n} L_{ik}y_k^{(n)} = \sum_{k=1}^{i-1} L_{ik}y_k^{(k)} + y_i^{(i)}. \tag{3.42}$$

Per $k < i$

$$L_{ik}y_k^{(k)} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} y_k^{(k)}. \tag{3.43}$$

Ma

$$y_i^{(k)} - y_i^{(k+1)} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} y_k^{(k)} = L_{ik}y_k^{(k)}, \tag{3.44}$$

per cui

$$(Ly^{(n)})_i = \sum_{k=1}^{i-1}(y_i^{(k)} - y_i^{(k+1)}) + y_i^{(i)} = y_i. \tag{3.45}$$

Poichè $\det L = 1 \neq 0$, $y^{(n)} = L^{-1}y$.

Oss. (trasformazioni di Gauss): sono date le n matrici $A^{(k)} \quad k = 1, \ldots, n$ definite dalla procedura di eliminazione Gaussiana (anche qui assumo che $a_{kk}^{(k)} \neq 0$ per ogni k e quindi l'eliminazione Gaussiana è possibile). Definisco gli $n-1$ vettori di Gauss $\tau^{(k)} \quad k = 1, \ldots, n-1$ tale che

$$(\tau^{(k)})^T = \left(0, \ldots, 0, \frac{a_{k+1,k}^{(k)}}{a_{kk}^{(k)}}, \ldots, \frac{a^{(}k)_{nk}}{a_{kk}^{(k)}}\right). \tag{3.46}$$

Definisco le $n-1$ matrici di Gauss $M_k \quad k = 1, \ldots, n-1$ tali che

$$M_k = I - \tau^{(k)}e_k^T. \tag{3.47}$$

Si vede subito che

$$A^{(k)} = M_{k-1} \dots M_1 A \quad k = 2, \dots, n. \tag{3.48}$$

Infatti, nel caso $i \geq k \quad j \geq k$:

$$(M_{k-1} \dots M_1 A)_{ij} = (M_{k-1} A^{(k-1)})_{ij} = \sum_{p=1}^{n} (\delta_{ip} - \tau_i^{(k-1)} (e_{k-1}^T)_p) a_{pj}^{(k-1)} = \tag{3.49}$$

$$= a_{ij}^{(k-1)} - \tau_i^{(k-1)} a_{k-1,j}^{(k-1)} = a_{ij}^{(k-1)} - \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} a_{k-1,j}^{(k-1)}. \tag{3.50}$$

Quindi

$$A^{(n)} = U = M_{n-1} \dots M_1 A. \tag{3.51}$$

Inoltre

$$M_k^{-1} = I + \tau^{(k)} e_k^T \tag{3.52}$$

è unit lower triangular e quindi invertibile. Ne consegue che

$$A = M_1^{-1} \dots M_{n-1}^{-1} U. \tag{3.53}$$

Poichè il prodotto di matrici unit lower triangular è unit lower triangular

$$L = M_1^{-1} \dots M_{n-1}^{-1}. \tag{3.54}$$

Esempio (matrice 3×3):

$$A = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 10 \end{bmatrix} \tag{3.55}$$

$$\tau^{(1)} = \begin{pmatrix} 0 \\ 2/1 \\ 3/1 \end{pmatrix} \tag{3.56}$$

$$M_1 = I - \tau^{(1)} e_1^T = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix} \tag{3.57}$$

$$A^{(2)} = M_1 A = \begin{pmatrix} 1 & 4 & 7 \\ 0 & -3 & -6 \\ 0 & -6 & -11 \end{pmatrix} \tag{3.58}$$

$$\tau^{(2)} = \begin{pmatrix} 0 \\ 0 \\ (-6)/(-3) \end{pmatrix} \tag{3.59}$$

$$M_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{pmatrix} \tag{3.60}$$

$$A^{(3)} = M_2 M_1 A = \begin{pmatrix} 1 & 4 & 7 \\ 0 & -3 & -6 \\ 0 & 0 & 1 \end{pmatrix} \tag{3.61}$$

Teo. (condizioni sufficienti): Hp.: $\det A(1:k, 1:k) \neq 0$ per $k = 1, \ldots, n$. Th.: esistono L e U tali che $A = LU$; tale fattorizzazione è unica.
Dim.: suppongo di aver effettuato $k-1$ step dell'eliminazione Gaussiana. Quindi sono arrivato alla matrice

$$A^{(k)} = M_{k-1} \ldots M_1 A. \tag{3.62}$$

Il pivot allo step k (quello necessario per costruire la matrice $A^{(k+1)}$) è $a_{kk}^{(k)}$. Ora, guardando alla porzione $(1:k) \times (1:k)$ dell'equazione (3.62) e sfruttando il fatto che le trasformazioni di Gauss sono unit lower triangular ho che

$$\det A(1:k, 1:k) = a_{11}^{(k)} \ldots a_{kk}^{(k)} \tag{3.63}$$

e quindi, per l'ipotesi del teorema, $a_{kk}^{(k)} \neq 0$ e l'eliminazione può proseguire. Per quanto riguarda l'unicità, suppongo per assurdo che esistano L_1, L_2, U_1, U_2 tali che

$$A = L_1 U_1 \qquad A = L_2 U_2. \tag{3.64}$$

Da qui

$$L_2^{-1} L_1 = U_2 U_1^{-1}. \tag{3.65}$$

Poichè $L_2^{-1} L_1$ è unit lower triangular e $U_2 U_1^{-1}$ è upper triangular, ambedue devono essere uguali all'identità.

3.5 Fattorizzazione LU: generalizzazione

Oss. (L non unit): la domanda è se è possibile una fattorizzazione LU con una L semplicemente lower triangular. In tal caso, è necessario determinare L e U tale che $LU = A$, ovvero, esplicitamente:

$$l_{kk}u_{kk} = a_{kk} - \sum_{p=1}^{k-1} l_{kp}u_{pk} \qquad k \geq 2 \tag{3.66}$$

$$u_{kj} = \frac{1}{l_{kk}} \left(a_{kj} - \sum_{p=1}^{k-1} l_{kp}u_{pj} \right) \qquad j > k \geq 2 \tag{3.67}$$

$$l_{ik} = \frac{1}{u_{kk}} \left(a_{ik} - \sum_{p=1}^{k-1} l_{ip}u_{pk} \right) \qquad i > k \geq 2. \tag{3.68}$$

$$u_{1j} = \frac{1}{l_{11}} a_{1j} \tag{3.69}$$

$$l_{i1} = \frac{1}{u_{11}} a_{i1}. \tag{3.70}$$

L'idea è quella di scegliere l_{kk} e u_{kk} in modo da soddisfare (3.66) e poi usare (3.67) e (3.68) per determinare gli elementi di L e U rimanenti nelle colonne e righe k-esime. Ci sono casi particolari per la matrice A per cui la scelta di l_{kk} e u_{kk} è del tutto naturale.

Oss. (matrici simmetriche): sia A una matrice simmetrica e suppongo che sia LU-fattorizzabile. Allora si dimostra per induzione (su k) che la scelta $l_{kk} = u_{kk}$ implica che A è fattorizzabile nella forma $LL^T = A$. Infatti, considero il caso $k = 1$.

$$l_{i1} = \frac{1}{u_{11}} a_{i1} = u_{1i}. \tag{3.71}$$

Suppongo che sia vero per k. Allora

$$l_{i,k+1} = \frac{1}{u_{k+1,k+1}} \left(a_{i,k+1} - \sum_{p=1}^{k-1} l_{ip}u_{pk} - l_{ik}u_{k,k+1} \right). \tag{3.72}$$

Ma $u_{k+1,k+1} = l_{k+1,k+1}$ e $a_{i,k+1} = a_{k+1,i}$ per la simmetria di A. Inoltre, per l'ipotesi induttiva, $l_{ik} = u_{ki}$ per ogni $i > k$. Quindi vale per $i = k+1$ e quindi

$u_{k,k+1} = l_{k+1,k}$. Per cui

$$l_{i,k+1} = \frac{1}{l_{k+1,k+1}} \left(a_{k+1,i} - \sum_{p=1}^{k-1} u_{pi} l_{kp} - u_{ki} l_{k+1,k} \right) \tag{3.73}$$

e quindi $l_{i,k+1} = u_{k+1,i}$.
Oss. (Cholesky): se A è simmetrica e definita positiva, allora è LU fattorizzabile. Quindi è fattorizzabile come $A = LL^T$. Questa fattorizzazione per A simmetrica definita positiva è detta fattorizzazione di Cholesky.

Chapter 4

Calcolo degli autovalori di una matrice

4.1 Generalità

Oss.: (autovalori in matematica applicata): il calcolo degli autovalori è un problema ricorrente in molte applicazioni. Un esempio tra i più clamorosi riguarda la meccanica quantistica. L'equazione di Schrödinger 'time independent' per una particella in un potenziale V data da

$$-\frac{h^2}{(2\pi)^2 2m}\triangle\psi + V\psi = E\psi. \tag{4.1}$$

Questa equazione descrive le soluzioni stazionarie dell'equazione 'time-dependent' e tali soluzioni corrispondono agli stati a energia finita. Si vede subito che può essere interpretata come un'equazione agli autovalori per l'operatore

$$H = -\frac{h^2}{(2\pi)^2 2m}\triangle + V. \tag{4.2}$$

Oss.: (calcolo numerico e autovalori): il calcolo numerico affronta tipicamente tre questioni relative al problema agli autovalori:

- dove sono gli autovalori? ovvero, come è possibile localizzare gli autovalori nel piano complesso? Questa questione è il contenuto di due teoremi di localizzazione che sono il teorema di Gerschgorin e il teorema di Weinstein;

- come si fa a calcolare gli autovalori? Qui il processo tipico consiste nel fattorizzare la matrice nel prodotto di matrici il cui risultato è una matrice caratterizzata dagli stessi autovalori della matrice di partenza ma più semplice da trattare. In senso generale, questo è l'approccio della fattorizzazione QR ma anche il più tradizionale metodo delle potenze risponde a questa logica;

- quale è il grado di stabilità degli autovalori rispetto a perturbazioni degli elementi della matrice? In questo caso vi sono risultati locali, che trattano la teoria delle perturbazioni per il singolo autovalore, oppure risultati non-locali, che coinvolgono il set degli autovalori nel loro insieme.

Def. (autovalori): data la matrice A $n \times n$, il numero complesso λ è detto autovalore di A se esiste un vettore x tale che

$$Ax = \lambda x. \tag{4.3}$$

Il vettore x è detto autovettore di A associato all'autovalore λ. In generale, allo stesso autovalore sono associati più autovettori che generano un sottospazio vettoriale detto autospazio associato all'autovalore. La dimensione di questo autospazio è detto molteplicità geometrica $g(\lambda)$ di λ. Gli autovalori di A sono tutte e solo le radici dell'equazione caratteristica

$$\det(A - \lambda I) = 0, \tag{4.4}$$

dove il polinomio di grado n

$$\varphi(\lambda) := \det(A - \lambda I) \tag{4.5}$$

è detto polinomio caratteristico di A. La molteplicità di λ come soluzione dell'equazione caratteristica è detta molteplicità algebrica di $a(\lambda)$ di λ e si dimostra che $a(\lambda) \geq g(\lambda)$.
Def. (forme simili): si dice trasformazione di similitudine per A la trasformazione che manda A nella matrice

$$B = T^{-1}AT, \tag{4.6}$$

dove T è una matrice invertibile. È chiaro che l'insieme degli autovalori di A e B coincidono. Inoltre, se x è autovalore di A, allora $T^{-1}x$ è autovalore

di B. A e B si dicono forme simili della stessa matrice.
Def. (forma di Jordan): data la matrice A, siano $\lambda_1, \ldots, \lambda_k$ i suoi autovalori distinti. Per ogni λ_i con molteplicità algebrica $a(\lambda_i)$ e molteplicità geometrica $g(\lambda_i)$, esistono $g(\lambda_i)$ numeri naturali $\nu_1^i, \ldots, \nu_{g(\lambda_i)}^i$ tali che

$$a(\lambda_i) = \nu_1^i + \ldots + \nu_{g(\lambda_i)}^i \tag{4.7}$$

e una matrice non singolare T tale che

$$J = TAT^{-1} \tag{4.8}$$

è una matrice diagonale a blocchi, in cui i blocchi sono

$$C_{\nu_1^1}(\lambda_1), \ldots, C_{\nu_{g(\lambda_1)}^1}, \ldots, C_{\nu_1^k}(\lambda_k), \ldots, C_{\nu_{g(\lambda_k)}^k}(\lambda_k) \tag{4.9}$$

e $C_\nu(\lambda)$ è una matrice $\nu \times \nu$ con λ sulla diagonale principale, 1 sulla sopradiagonale e zero altrove. Vi è un caso limite interessante, quando la dimensione di ogni blocco di Jordan è uno, ovvero $\nu_j^i = 1 \quad \forall i, j$. In tal caso $J = diag(\lambda_1, \ldots, \lambda_k)$, gli autovalori vengono presi con la loro molteplicità e A è detta diagonalizzabile.
Def. (forma di Schur): ogni matrice A può essere trasformata nella matrice simile $J = Q^* A Q$ dove J è triangolare superiore con gli autovalori (presi con la loro molteplicità) sulla diagonale principale e Q unitaria. Da questo risultato si ha che se A è autoaggiunta, allora può essere diagonalizzata in modo unitario.

4.2 Localizzazione

Oss. (risultati sul raggio spettrale): sia $\rho(A) = \max_s |\lambda_s(A)|$, dove i $\lambda_s(A)$ sono gli autovalori distinti di A. Allora per ogni norma indotta $\|\cdot\|$ vale $\rho(A) \leq \|A\|$. Inoltre, $\forall \epsilon$ esiste una norma indotta tale che $\rho(A) \leq \|A\| \leq \rho(A) + \epsilon$.
Teo. (Gerschgorin): Hp.: A matrice di ordine n e λ un suo autovalore.

$$R_i = \{z \in \mathbb{C} \quad |z - a_{ii}| \leq \sum_{k \neq i} |a_{ik}|\}; \tag{4.10}$$

$$C_k = \{z \in \mathbb{C} \quad |z - a_{kk}| \leq \sum_{i \neq k} |a_{ik}|\}; \tag{4.11}$$

i C_k e R_i sono detti cerchi di Gerschgorin e $\lambda(A)$ indica l'insieme degli autovalori di A.
Th.: λ è contenuto in uno dei cerchi di Gerschgorin R_i o C_k.
Dim.: sia $\lambda \in \lambda(A)$ e x uno dei suoi autovettori. Se $|x_i| = \|x\|_\infty$, allora

$$|\lambda - a_{ii}| = |\sum_{k \neq i} a_{ik} \frac{x_k}{x_i}| \leq \sum_{k \neq i} |a_{ik}| \tag{4.12}$$

e quindi λ è in R_i. L'analogo vale per i cerchi C_k, perchè gli autovalori di A sono anche autovalori di A^T.
Oss. (componenti connesse): scrivo la matrice A nella forma

$$A = D + R, \tag{4.13}$$

dove $D = diag(a_{11}, \ldots, a_{nn})$ e R è tutto il resto. Allora definisco la matrice

$$A(t) = D + tR \quad t \in [0, 1]. \tag{4.14}$$

Gli zeri di un polinomio sono funzioni continue dei suoi coefficienti. Ne consegue che gli autovalori sono funzioni continue di t. Mentre t passa da zero a 1, ogni $\lambda_i(t)$ si muove in modo continuo da a_{ii} a λ_i (e questo per ogni i). Per il teorema di Gerschgorin, l'autovalore $\lambda_i(t)$ è nel cerchio

$$R_i(t) = \{|z - a_{ii}| \leq t \sum_{k \neq i} |a_{ik}(t)|\}. \tag{4.15}$$

Per t crescente, $\lambda_i(t)$ rimane in $R_i(t)$ e per $t = 1$, λ è in R_i. Ma R_i contiene tutti i cerchi e in particolare contiene $R_i(0)$. Tutto questo vale per ogni i, per cui se a $t = 1$ due cerchi hanno un'intersezione non nulla, questa componente connessa deve contenere esattamente due autovalori. Concludendo, in ogni componente massimale connessa di cerchi di Gerschgorin, vi sono esattamente tanti autovalori quanti cerchi.
Oss. (raffinamento): suppongo che sia possibile effettuare la trasformazione di similitudine

$$B = D^{-1} A D \tag{4.16}$$

dove D è una matrice diagonale. Considero i cerchi di Gerschgorin di B:

$$R_i = \{z \quad |z - b_{ii}| \leq \sum_{k \neq i}^{n} \left| \frac{a_{ik} d_k}{d_i} \right| \} \tag{4.17}$$

in cui $b_{ii} = a_{ii}$ e $\lambda(B) = \lambda(A)$. Si tratta di scegliere D in modo tale che

$$r_i = \sum_{k \neq i}^{n} \frac{a_{ik} d_k}{d_i} \tag{4.18}$$

sia piccolo il più possibile.
Teo. (Weinstein): Hp.: sia A una matrice normale di ordine n; sia μ_{ik} il tensore

$$\mu_{ik} = x^* (A^*)^i A^k x. \tag{4.19}$$

Th.: nel cerchio

$$C = \{\mu \quad |\mu - \frac{\mu_{01}}{\mu_{00}}| \leq \sqrt{\frac{\mu_{11} - \mu_{01}\mu_{11}/\mu_{00}}{\mu_{00}}}\} \tag{4.20}$$

esiste almeno un autovalore di A.

4.3 Metodo delle potenze

Oss. (autovalore principale): considero il caso di A diagonalizzabile e tale che

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|. \tag{4.21}$$

Costruisco la successione $\{x_k\} \qquad k = 1, 2, \ldots$ tale che

$$x_k = A x_{k-1} = A^k x_0 \tag{4.22}$$

dove x_0 è un opportuno versore di inizializzazione. Sia $\{u_k\}_{k=1}^n$ una base ortonormale di autovettori. Allora

$$x_k = A^k (\sum_{i=1}^{n} a_i u_i) = \sum_i \lambda_i^k a_i u_i = \tag{4.23}$$

$$= \lambda_1^k [a_1 u_1 + \sum_{j=2}^{n} \left(\frac{\lambda_j}{\lambda_1}\right)^k a_j u_j] \quad k = 1, 2, \ldots \tag{4.24}$$

Pochè $|\lambda_j/\lambda_1| < 1$ per $j \geq 2$, allora la direzione di x_k converge a quella di u_1.
N.B. 1: per $k \to \infty$, λ_1^k va a zero o cresce. Tuttavia questo problema è risolto con un'opportuna operazione di rescaling.
Oss. (algoritmo di calcolo): sulla base della precedente osservazione possiamo formulare il seguente algoritmo di calcolo dell'autovalore principale:

1. $q_0 = a_1 u_1 + \ldots + a_n u_n \quad \|q_0\| = 1$

2. $k = 1, 2, \ldots$

$$z_k = A q_{k-1} \tag{4.25}$$

$$q_k = \frac{z_k}{\|z_k\|} \tag{4.26}$$

$$\lambda_k = q_k^* A q_k \tag{4.27}$$

3. end

N.B. 1: nel metodo delle potenze si assume implicitamente che $a_1 \neq 0$. Ora, in realtà questa ipotesi non è così stringente in quanto la presenza di errori di troncamento contribuisce ad arricchire la componente lungo u_1.
Oss. (metodo dell'iterazione inversa): il metodo dell'iterazione inversa consiste nell'applicare il metodo delle potenze alla matrice $(A - \mu I)^{-1}$. Una forma generale di questo algoritmo è:

1. $\|q_0\| = 1$

2. $k = 1, 2, \ldots$

$$\text{risolvi} \quad (A - \mu I) z_k = q_{k-1} \tag{4.28}$$

$$q_k = \frac{z_k}{\|z_k\|} \tag{4.29}$$

$$\lambda_k = q_k^* A q_k \tag{4.30}$$

3. end

Se $q_0 = \sum_{i=1}^n \beta_i u_i$, con $\{u_i\}_{i=1}^n$ base di autovettori di A, allora

$$(A - \mu I)^{-k} q_0 = \sum_{i=1}^n \frac{\beta_i}{(\lambda_i - \mu)^k} u_i \tag{4.31}$$

per cui, se $\mu \simeq \lambda_j$, $(A - \mu I)^{-k} q_0$ è ricco nella componente u_j. Ne consegue che il metodo dell'iterazione inversa per calcolare un qualunque autovalore, a condizione che se ne conosca a priori una stima decente (in altri termini, questo metodo può essere utilizzato per raffinare la stima di un autovalore).
N.B. 1: il sistema (4.28) è quasi singolare e quindi assai mal condizionato.

4.4 Fattorizzazione QR

Oss. (approccio generale): l'idea generale per il calcolo degli autovalori di una matrice A (densa) diagonalizzabile è la seguente:

1. effettuo un numero finito di trasformazioni di similitudine

$$A_i = T_i^{-1} A_{i-1} T_i \qquad i = 1, \ldots, m; \tag{4.32}$$

2. arrivo alla matrice

$$B = A_m = T^{-1} A T \qquad T = T_1 T_2 \ldots T_m \tag{4.33}$$

   tale che sia semplice calcolare gli autovalori di B.

Questo approccio è efficace se sussistono due condizioni:

- calcolare gli autovalori di B richiede pochissime operazioni;
- passare da A a B è un'operazione ben condizionata.

Un possibile algoritmo generale è il seguente:

1. riduzione a Hessenberg superiore H (tale che $h_{ij} = 0$ se $i > j + 1$):

$$H = U^* A U \tag{4.34}$$

2. $k = 1, 2, \ldots$

$$H_k = Q_k R_k \tag{4.35}$$

$$H_{k+1} = R_k Q_k \tag{4.36}$$

3. end

dove Q è una matrice unitaria e R è una matrice triangolare superiore. La motivazione dello step 1 (riduzione a Hessenberg superiore) è di tipo computazionale: per una matrice densa, la fattorizzazione QR avviene in $4n^3/3$ operazioni. Se la matrice è Hessenberg superiore, in $O(n^2)$ operazioni (se la matrice A è reale simmetrica, la trasformazione in Hessenberg superiore la rende tridiagonale e la fattorizzazione QR avviene in $O(n)$ operazioni).

Oss. (stabilità): Sia ΔA una perturbazione di A e ΔB la corrispondente perturbazione su B. Quindi:

$$B + \Delta B = T^{-1}(A + \Delta A)T \tag{4.37}$$

da cui

$$\Delta B = T^{-1}\Delta A T. \tag{4.38}$$

Quindi ho le disuguaglianze:

$$\|\Delta B\| \leq C(T)\|\Delta A\| \qquad \|A\| \leq C(T)\|B\| \tag{4.39}$$

da cui

$$\frac{\|\Delta B\|}{\|B\|} \leq C(T)^2 \frac{\|\Delta A\|}{\|A\|}. \tag{4.40}$$

Inoltre

$$C(T) \leq C(T_1 T_2 \dots T_M) \leq C(T_1)C(T_2)\dots C(T_m). \tag{4.41}$$

Quindi per avere stabilità numerica ho bisogno che il condizionamento di T sia buono.

Def. (matrice di Householder): sia $w \in C^n$, $w^*w = 1$. Si definisce matrice di Householder con vettore di Householder w la matrice

$$P = I - 2ww^*. \tag{4.42}$$

È immediato notare che $P^* = P$ e $P^*P = PP^* = I$.

Oss. (annullamento di elementi di un vettore): la matrice di Householder è utile per azzerare gli elementi di un vettore. L'idea è quella di trovare w tale che

$$Px = ke_1 \tag{4.43}$$

Anzitutto

$$(Px)^*(Px) = x^*P^*Px = x^*x = \|x\|^2 \tag{4.44}$$

per cui

$$|k|^2 = \|x\|^2 \tag{4.45}$$

Inoltre

$$(x^*Px)^* = x^*P^*x = x^*Px \tag{4.46}$$

e quindi x^*Px è un numero reale; ovvero, se $x_1 = |x_1|e^{i\alpha}$

$$kx^*e_1 = k|x_1|e^{-i\alpha} \tag{4.47}$$

è un numero reale e quindi

$$k = \pm\|x\|e^{i\alpha}. \tag{4.48}$$

Nel seguito scelgo $k = -\|x\|e^{i\alpha}$. Quindi:

$$Px = x - 2(w^*x)w = ke_1. \tag{4.49}$$

Ne segue che (essendo $w^*w = 1$)

$$w = \frac{x - ke_1}{\|x - ke_1\|}. \tag{4.50}$$

Ora

$$\|x-ke_1\|^2 = \|x+\|x\|e^{i\alpha}e_1\|^2 = |x_1+\|x\|e^{i\alpha}|^2+|x_2|^2+\ldots+|x_n|^2 = 2\|x\|^2+2|x_1|\|x\|. \tag{4.51}$$

E quindi

$$2ww^* = \frac{(x - ke_1)(x - ke_1)^*}{\|x\|^2 + |x_1|\|x\|} \tag{4.52}$$

Quindi la matrice di Householder $P = I - 2ww^*$ che trasforma x in ke_1 è

$$P = I - \beta uu^* \tag{4.53}$$

$$u = x - ke_1 = \begin{pmatrix} e^{i\alpha}(|x_1| + \|x\|) \\ x_2 \\ . \\ . \\ x_n \end{pmatrix} \tag{4.54}$$

$$x_1 = |x_1|e^{i\alpha} \qquad \beta = (\|x\|^2 + \|x\||x_1|)^{-1}. \tag{4.55}$$

Oss. (fattorizzazione QR): considero come esempio una matrice A di dimensioni 5×5:

$$A = \begin{matrix} x & x & x & x & x \\ 0 & x & x & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & x & x & x \end{matrix} \tag{4.56}$$

Considero il vettore pieno (a_{33}, a_{43}, a_{53}) e utilizzo la matrice di Householder 3×3 $\tilde{H}_3$ che azzera le due componenti a_{43} e a_{53}. Definisco la matrice

$$H_3 = diag(I, \tilde{H}_3) \tag{4.57}$$

Allora si trova immediatamente che

$$H_3 A = \begin{matrix} x & x & x & x & x \\ 0 & x & x & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & 0 & x & x \\ 0 & 0 & 0 & x & x \end{matrix} \tag{4.58}$$

Procedendo in questo modo si ottiene

$$H_{n-1} \dots H_1 A = R \tag{4.59}$$

con R triangolare superiore e $H_i \quad i = 1, \dots, n-1$ ortogonali e quindi $H_{n-1} \dots H_1$ ortogonale.
N.B.: la fattorizzazione QR non è unica. Sia S una matrice fase di forma

$$S = diag(e^{i\varphi_1}, e^{i\varphi_2}, \dots, e^{i\varphi_n}). \tag{4.60}$$

Allora, se $A = QR$, QS è ancora unitaria, S^*R è ancora upper triangular e $(QS)(S^*R) = QR = A$.
Def. (algoritmo QR): si definisce algoritmo QR il seguente schema:

1. A Hessenberg superiore (o comunque il più sparsa possibile)
2. $A_i = Q_i R_i \quad Q_i^* Q_i = I \quad R_i$ upper triangular
3. $A_{i+1} = R_i Q_i$

Per calcolare la fattorizzazione $A_i = Q_i R_i$ si utilizzano matrici di Householder, ovvero

$$H_{n-1}^{(i)} \dots H_1^{(i)} A_i = R_i \tag{4.61}$$

per cui

$$Q_i = H_1^{(i)} \dots H_{n-1}^{(i)} \tag{4.62}$$

e quindi

$$A_{i+1} = R_i H_1^{(i)} \dots H_{n-1}^{(i)}. \tag{4.63}$$

Teo. (convergenza algoritmo QR): Hp.: sia $A = A_1$ una matrice $n \times n$ con le seguenti proprietà:

1. gli autovalori sono tali che

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|; \tag{4.64}$$

2. $A = Y^{-1}D \quad D = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$ con Y che ammette la fattorizzazione

$$Y = L_Y R_Y \tag{4.65}$$

L_Y unit lower triangular, R_Y upper triangular;

Th.: esistono le matrici fase $S_k = diag(e^{i\varphi_1^k}, \ldots, e^{i\varphi_n^k})$ tale che

$$\lim_{k\to\infty} S_{k-1}^* Q_k S_k = I \tag{4.66}$$

$$\lim_{k\to\infty} S_k^* R_k S_{k-1} = \lim_{k\to\infty} S_{k-1}^* A_k S_{k-1} = R \tag{4.67}$$

con R una matrice upper triangular con gli autovalori sulla diagonale.

$$\lim_{k\to\infty} a_{jj}^{(k)} = \lambda_j \quad j = 1, \ldots, n. \tag{4.68}$$

N.B.1: l'ipotesi 1. sugli autovalori implica che A sia una matrice complessa con autovalori reali (se ci fosse un autovalore complesso, anche il suo complesso coniugato sarebbe autovalore). Nel caso di matrice reale con autovalori complessi, l'algoritmo QR conduce, alla fine delle iterazioni, a una matrice upper triangular in cui, sulla diagonale vi sono blocchi 1×1 (corrispondenti agli autovalori reali) e blocchi 2×2 (corrispondenti alle coppie di autovalori complessi coniugati).
N.B.2: l'ipotesi 2. non e' essenziale per la convergenza. Se non vale, l'algoritmo converge ancora ma gli autovalori nella matrice sulla diagonale della matrice upper triangular finale non sono necessariamente ordinati in modo decrescente.

4.5 Teoria delle perturbazioni

Oss. (teoria locale e teoria non locale): la teoria non locale fornisce risultati che riguardano lo spettro nel suo complesso. Uno di questi risultati è il teorema di Bauer-Fike. Invece la teoria locale definisce il concetto di condizionamento associato al singolo autovalore.

Teo. (Bauer-Fike): Hp.: A matrice $n \times n$ a valori complessi, diagonalizzabile, con $X^{-1}AX = D$. E perturbazione di A; μ autovalore di $A + E$. Th.:

$$\min_{\lambda \in \lambda(A)} |\mu - \lambda| \leq C(X)\|E\|, \tag{4.69}$$

dove il condizionamento è calcolato con la stessa norma $\|\cdot\|$.
Oss. (condizionamento di un autovalore): siano x e y autovettore destro e sinistro di A associati all'autovalore λ, ovvero

$$Ax = \lambda x \quad y^*A = \lambda y^*. \tag{4.70}$$

Allora per ϵ sufficientemente piccolo, trovo $x(\epsilon)$ e $\lambda(\epsilon)$ tali che

$$(A + \epsilon F - \lambda(\epsilon)I)x(\epsilon) = 0 \tag{4.71}$$

con $\|F\| = 1$, $\lambda(0) = \lambda$ e $x(0) = x$. Si ottiene che

$$\lambda'(0) = \frac{y^*Fx}{y^*x}. \tag{4.72}$$

Denotando con

$$\cos(x, y) = \frac{y^*x}{\|x\|_2\|y\|_2}, \tag{4.73}$$

ho che

$$|\lambda'(0)| = \frac{|y^*Fx|}{\|y\|_2\|x\|_2|\cos(x, y)|} \leq \tag{4.74}$$

$$\leq \frac{\|Fx\|_2}{\|x\|_2|\cos(x, y)|} \leq \frac{\|F\|_2}{|\cos(x, y)|}. \tag{4.75}$$

Quindi la sensitività di λ aumenta con il diminuire di $\cos(x, y)$. Nel caso di matrici autoaggiunte $x = y$ e quindi $\cos(x, y) = 1$. Ciò è in accordo con il fatto che gli autovalori di una matrice autoaggiunta sono sostanzialmente stabili.

Chapter 5

Decomposizione in Valori Singolari (SVD)

5.1 Operatori compatti

Def. (topologia): è dato un insieme X. Si definisce topologia su X un insieme $\mathcal{F}$ di sottoinsiemi di X (detti aperti) tale che

1. l'insieme vuoto e X sono aperti;
2. ogni intersezione finita di aperti è aperta;
3. ogni unione arbitaria di aperti è aperta.

Un insieme con la sua topologia è detto spazio topologico.
Def. (insiemi compatti): dato un insieme X, un ricoprimento di X è un insieme di sottoinsiemi di X la cui unione contiene X; dato un ricoprimento di X, un sottoricoprimento di X è un altro ricoprimento di X in cui ogni insieme è sottoinsieme di un insieme del ricoprimento. Uno spazio topologico X è detto compatto se ogni ricoprimento aperto di X possiede un sottoricoprimento finito. Lo spazio topologico è detto relativamente compatto se la sua chiusura è compatta.
Def. (operatori compatti): sia $A : X \to Y$ un operatore lineare tra spazi normati (non necessariamente completi). A è detto compatto se mappa insiemi limitati di X in insiemi relativamente compatti di Y.
Oss. (alcuni risultati su operatori compatti): un operatore $A : X \to Y$

tra spazi di Banach è compatto se e solo se, per ogni successione $\{x_n\} \subset X$, $\{Ax_n\}$ ha una sottosuccessione convergente. Un operatore compatto tra spazi di Banach mappa successioni debolmente convergenti in successioni fortemente convergenti. Sia A un operatore compatto in uno spazio di Hilbert separabile. Allora A è il limite (in norma operatoriale) di una successione di operatori di rango finito.

Teo. (rappresentazione singolare): sia $A : X \to Y$ un operatore compatto tra spazi di Hilbert separabili. Allora esistono due insiemi ortonormali $\{v_k\}$, $\{u_k\}$ in X e Y rispettivamente e un insieme di numeri non negativi $\{\sigma_k\}$ tali che

$$Av_k = \sigma_k u_k \tag{5.1}$$

$$A^* u_k = \sigma_k v_k \tag{5.2}$$

$$Af = \sum_k \sigma_k (g, u_k) v_k \qquad \forall f \in X. \tag{5.3}$$

L'insieme $\{v_k, u_k, \sigma_k\}$ è detto sistema singolare di A mentre la rapresentazione (5.3) è detta rappresentazione singolare di A.

5.2 SVD in spazi euclidei canonici

Oss. (matrici autoaggiunte): considero una matrice autoaggiunta $N \times N$ di rango p. Allora esistono una matrice reale diagonale D $p \times p$ e una matrice unitaria V $N \times p$ tali che

$$V^* A V = D. \tag{5.4}$$

Le colonne di V sono gli autovettori di A e gli elementi diagonali di D sono i corrispondenti autovalori diversi da zero. A ha poi l'autovalore $\lambda = 0$ con molteplicità $N-p$. Se i vettori $v_k \quad k = 1, \ldots, p$ rappresentano le p colonne di V e i vettori $v'_k \quad k = p+1, \ldots, N-p$ rappresentano gli autovettori associati a $\lambda = 0$ (ortonormalizzati), allora

$$f = \sum_{k=1}^{p} (f, v_k) v_k + \sum_{k=p+1}^{N} (f, v'_k) v'_k, \tag{5.5}$$

dove $f \in \mathbb{C}^N$ e $(\cdot, \cdot)$ è il prodotto scalare canonico. Da qui, immediatamente:

$$Af = \sum_{k=1}^{p} \lambda_k (f, v_k) v_k. \tag{5.6}$$

Questa è la rappresentazione spettrale per A autoaggiunta. Il problema che mi pongo è se è possibile generalizzare il concetto di diagonalizzazione e rappresentazione spettrale da una matrice autoaggiunta quadrata a una matrice generica rettangolare $M \times N$ a valori complessi. Tutto questo viene svolto di seguito, assumendo sempre una topologia canonica (la generalizzazione a topologie non-canoniche è comunque possibile).
Teo. (SVD): Hp.: $A : \mathbb{C}^N \to \mathbb{C}^M$ matrice $M \times N$ di rango p; $(\cdot, \cdot)$ e $\| \cdot \|$ prodotto scalare e norma canonici (in $\mathbb{C}^N$ o $\mathbb{C}^M$). Th.: esistono $V \quad N \times N$ e $U \quad M \times M$ unitarie tali che

$$U^* A V = \Sigma. \tag{5.7}$$

Σ è una matrice reale $M \times N$ i cui primi p elementi diagonali $\sigma_1 \geq \sigma_2 \geq \ldots$ sono le radici quadrate dei p autovalori di A^*A e gli altri elementi sono zero. Dim.: procedo per induzione. (I) Sia $M = 1$. Quindi A è un vettore riga. Prendo $U \quad 1 \times 1$ tale che $U = 1$ e $V \quad N \times N$ unitaria tale che la prima colonna è

$$(V)_{i1} = \frac{\overline{a}_i}{\|A\|}. \tag{5.8}$$

Quindi U^*AV è una matrice $1 \times N$ tale che

$$(U^*AV)_{1i} = \sum_k a_{1k} V_{ki} = \sum_k \|A\| V^*_{1k} V_{ki} = \|A\| \delta_{1i} \tag{5.9}$$

(essendo V unitaria). In modo analogo per $N = 1$. (II) Assumo che la tesi sia vera per A di dimensione $(M-1) \times (N-1)$. Sia $A \quad M \times N$, $\sigma = \|A\|$. Sia $V \quad N \times N$ unitaria e tale che la prima colonna di V è il vettore x, $\|x\| = 1$, $\|Ax\| = \sigma$. Sia $U \quad M \times M$ unitaria e tale che la prima colonna di U sia il vettore y, $\|y\| = 1$, $y = \sigma^{-1} Ax$. Allora

$$(U^*AV)_{i1} = \sum_k (U^*)_{ik} \sum_l A_{kl} V_{l1} = \sum_k (U^*)_{ik} (Ax)_k. \tag{5.10}$$

Ma

$$(Ax)_k = \sigma y_k = \sigma U_{k1}. \tag{5.11}$$

Quindi

$$(U^*AV)_{i1} = \sigma \delta_{i1}, \tag{5.12}$$

per cui

$$U^*AV = \begin{pmatrix} \sigma & w^* \\ 0 & C \end{pmatrix}, \tag{5.13}$$

dove w è $(M-1)\times 1$ e C è $(M-1)\times(N-1)$. Quindi

$$(U^*AV)\begin{pmatrix}\sigma\\ w\end{pmatrix}=\begin{pmatrix}\sigma^2+w^*w\\ Cw\end{pmatrix}. \tag{5.14}$$

Da cui

$$\sigma^2+w^*w\leq\|(U^*AV\|\begin{pmatrix}\sigma\\ w\end{pmatrix}\|\leq\|A\|\sqrt{\sigma^2+w^*w}. \tag{5.15}$$

Quindi

$$\sigma^2+w^*w\leq\|A\|^2=\sigma^2 \tag{5.16}$$

per cui $w=0$ e, cioè,

$$U^*AV=\begin{pmatrix}\sigma & 0\\ 0 & C\end{pmatrix}. \tag{5.17}$$

Poichè C ha dimensione $(M-1)\times(N-1)$, posso invocare l'ipotesi induttiva per cui esistono U_C e V_C unitarie e S_C diagonale (con eventuali zeri) tali che

$$U_C^*CV_C=S_C. \tag{5.18}$$

Costruisco allora le matrici $U'=diag(1,U_C)$ e $V'=diag(1,V_C)$ e ho che

$$U'^*U^*AVV'=diag(\sigma,S_C). \tag{5.19}$$

Ribattezzando $U=UU'$, $V=VV'$ e $S=diag(\sigma,S_C)$ ho $U^*AV=S$. Da cui

$$Av_k=\sigma_k u_k \tag{5.20}$$

e

$$A^*u_k=\sigma_k v_k. \tag{5.21}$$

Def. (SVD): data A di dimensioni $M\times N$, la decomposizione

$$U^*AV=\Sigma \tag{5.22}$$

si dice decomposizione in valori singolari di A. I vettori colonna di U e V sono detti vettori singolari mentre gli elementi diagonali non nulli di Σ sono detti valori singolari. Si ha

$$Av_k=\sigma_k u_k\quad A^*u_k=\sigma_k v_k\qquad k=1,\ldots,p \tag{5.23}$$

dove p è il rango della matrice.
Oss. (basi): poichè U e V sono unitarie, gli insiemi di vettori colonna $\{v_k\}_{k=1}^N$ e $\{u_k\}_{k=1}^M$ sono ortonormali. Considero gli insiemi $\{u_k\}_{k=1}^p$ e $\{v_k\}_{k=1}^p$ associati ai valori singolari (che sono p e diversi da zero). I v_k sono ortogonali al nucleo di A, $N(A)$. Infatti, sia $y \in N(A)$. Allora

$$(y, v_k) = (y, \frac{1}{\sigma_k} A^* u_k) = \frac{1}{\sigma_k}(Ay, u_k) = 0. \tag{5.24}$$

Il teorema nullità + rango implica che la dimensione di $N(A)^\perp$ è p per cui $\{v_k\}_{k=1}^p$ è base ortonormale in $N(A)^\perp = R(A^*)$. Analogamente, $\{u_k\}_{k=1}^p$ è base ortonormale in $N(A^*)^\perp = R(A)$.
Def. (rappresentazione singolare): sia $f \in \mathbb{C}^N$. Allora

$$f = \sum_{k=1}^{p} (f, v_k) v_k + \sum_{k=p+1}^{N} (f, v_k') v_k', \tag{5.25}$$

dove i $\{v_k'\}_{k=p+1}^N$ sono una base ortonormale in $N(A)$. Ne segue che

$$Af = \sum_{k=1}^{p} \sigma_k (f, v_k) u_k. \tag{5.26}$$

La (5.26) è detta rappresentazione singolare di A.
Oss. (norma e SVD): nel teorema SVD di cui sopra, dimostro, tra l'altro, che uno dei valori singolari è uguale alla norma 2 di A. Ora faccio vedere che tale valore singolare è quello massimo. Infatti, dalla rappresentazione singolare:

$$\|Af\|^2 \leq \sum_{k=1}^{p} \sigma_k^2 |(f, v_k)|^2 \leq \sigma_1^2 \|f\|^2, \tag{5.27}$$

per cui $\|A\| \leq \sigma_1$. D'altra parte

$$\|Av_1\|^2 = \sum_{k=1}^{p} \sigma_k^2 |(v_1, v_k)|^2 = \sum_{k=1}^{p} \sigma_k^2 \delta_{1k} = \sigma_1^2, \tag{5.28}$$

per cui $\|A\| = \sigma_1$.
Oss. (SVD come problema agli autovalori): definisco la matrice diagonale a blocchi

$$C = \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix}. \tag{5.29}$$

Allora

$$C \begin{pmatrix} u_k \\ v_k \end{pmatrix} = \sigma_k \begin{pmatrix} u_k \\ v_k \end{pmatrix}. \tag{5.30}$$

Analogamente

$$C \begin{pmatrix} u_k \\ -v_k \end{pmatrix} = -\sigma_k \begin{pmatrix} u_k \\ -v_k \end{pmatrix}. \tag{5.31}$$

Chapter 6

Problema ai minimi quadrati

6.1 Equazione di Eulero

Oss. (esistenza e unicità): considero il problema di determinare $f \in \mathbb{C}^N$ da $g \in \mathbb{C}^M$, quando f e g sono legati dall'equazione

$$Af = g, \tag{6.1}$$

con A una matrice a valori complessi di dimensione $M \times N$ e di rango p. Se $N(A)$ non è banale, la soluzione f di questo problema, se esiste, non è unica. Un modo banale per 'restaurare' l'unicità è il seguente. Data la SVD di A, f può essere rappresentato nella forma

$$f = \sum_{k=1}^{p} (f, v_k) v_k + v \tag{6.2}$$

con $v \in N(A)$. Tra tutte le possibili soluzioni scelgo quella con $v = 0$. Tale soluzione (sempre se esiste) è unica. Infatti, per la rappresentazione singolare e per l'indipendenza lineare dei v_k, $Af = 0$ implica $\sigma_k(f, v_k) = 0 \quad \forall k$ e quindi $(f, v_k) = 0 \quad \forall k$ per cui dalla (7.5) con $v = 0$, $f = 0$. Di tutte le soluzioni, quella con $v = 0$ è quella di norma minima (per il teorema di Pitagora). Suppongo ora che $N(A^*) = R(A)^\perp$ non sia banale (ovvero $R(A) \neq \mathbb{C}^M$). Allora $\{u_k\}_{k=1}^p$ non costituisce una base ortonormale di $\mathbb{C}^M$ e quindi

$$g = \sum_{k=1}^{p} (g, u_k) u_k + u \tag{6.3}$$

con $u \in N(A^*)$. Se $u \neq 0$, un confronto tra la rappresentazione singolare e la (6.3) implica che non esistono soluzioni del problema (7.4). Un modo banale per restaurare l'esistenza è quello di cercare soluzioni ai minimi quadrati.
Def. (equazione di Eulero): f è una soluzione ai minimi quadrati di (7.4) se e solo se

$$\|Af - g\|^2 \leq \|A(f + t\phi) - g\|^2 \quad \forall t \in R, \ \ \forall \phi \in \mathbb{C}^N. \tag{6.4}$$

Questo vale se e solo se

$$\|Af-g\|^2 \leq \|Af-g\|^2+t^2\|A\phi\|^2+t^*(\phi, A^*Af-A^*g)+t(A^*Af-A^*g, \phi) \ \ \forall t, \ \ \forall \phi. \tag{6.5}$$

Questo vale se e solo se

$$(A^*Af + A^*g, \phi) = 0 \quad \forall \phi \tag{6.6}$$

se e solo se

$$A^*Af = A^*g. \tag{6.7}$$

Quest'ultima è detta equazione di Eulero, ed è equivalente al problema ai minimi quadrati.
Def. (soluzione generalizzata): un qualunque vettore $f \in \mathbb{C}^N$ è rappresentabile nella forma

$$f = \sum_{k=1}^{p} (f, v_k)v_k + v \tag{6.8}$$

con $v \in N(A)$, mentre un qualunque vettore $g \in \mathbb{C}^M$ è rappresentabile nella forma

$$g = \sum_{k=1}^{p} (g, u_k)u_k + u \tag{6.9}$$

con $u \in N(A^*)$. Se introduco queste due rappresentazioni nell'equazioni di Eulero ho che f è soluzione ai minimi quadrati se e solo se

$$f = \sum_{k=1}^{p} \frac{1}{\sigma_k}(g, u_k)v_k + v. \tag{6.10}$$

Tra tutte le soluzioni ai minimi quadrati, quella di norma minima

$$f = \sum_{k=1}^{p} \frac{1}{\sigma_k}(g, u_k)v_k \tag{6.11}$$

è detta soluzione generalizzata mentre la matrice $A^\dagger$ che manda g in $f^\dagger$ è detta inversa generalizzata.
Oss. (condizionamento): Il concetto di condizionamento può essere generalizzato al caso del problema ai minimi quadrati. Sia $g^\dagger = Af^\dagger$ e sia $\delta g^\dagger$ una sua perturbazione a cui corrisponde la perturbazione $\delta f^\dagger$ della soluzione generalizzata. Allora

$$\|\delta f^\dagger\|^2 = \sum_{k=1}^{p} \frac{1}{\sigma_k^2} |(\delta g^\dagger, u_k)|^2 \leq \frac{1}{\sigma_p^2} \|\delta g^\dagger\|^2. \tag{6.12}$$

Anzitutto, questa disuguaglianza dimostra la continuità di $A^\dagger$. Inoltre

$$\|g^\dagger\|^2 = \|\sum_{k=1}^{p} \sigma_k (f^\dagger, v_k) u_k\|^2 = \sum_{k=1}^{p} \sigma_k^2 |(f^\dagger, v_k)|^2 \leq \sigma_1^2 \|f^\dagger\|^2. \tag{6.13}$$

Per cui

$$\frac{\|\delta f^\dagger\|}{\|f^\dagger\|} \leq \frac{\sigma_1}{\sigma_p} \frac{\|\delta g^\dagger\|}{\|g^\dagger\|}. \tag{6.14}$$

Il condizionamento del problema generalizzato è pertanto σ_1/σ_p.

6.2 Problemi inversi con dati discreti

Oss. (esperimenti): nelle scienze sperimentali, un modo comune per determinare le caratteristiche di un campione fisico, consiste nell'osservare l'interazione tra il campione e la radiazione (elettromagnetica o acustica) emessa da una sorgente nota. Un insieme di detector misura la radiazione diffusa dall'oggetto e il problema consiste nel ricostruire il campione (o un suo parametro) a partire dalle misure e dalla conoscenza delle proprietà fisiche dello strumento di misura. Il modello matematico che descrive la formazione del dato in questo tipo di esperimenti è:

$$g(x_n) = \int k(x_n, y) f(y) dy \qquad n = 1, \ldots, N, \tag{6.15}$$

dove $\{g_n\}_{n=1}^N$ rappresenta le misure; $k(x, y)$ descrive lo strumento di misura e $f(y)$ è il campione da ricostruire. Raffinando un po` il modello: sia X uno spazio di Hilbert; sia $\{F_n\}_{n=1}^N$ un insieme di funzionali lineari limitati

definiti su X e $\{g_n\}_{n=1}^N$ un insieme di numeri reali. Considero il problema di determinare $f \in X$ tale che

$$F_n(f) = g_n \quad n = 1, \ldots, N. \tag{6.16}$$

Essendo ciascun F_n lineare e limitato, il teorema di rappresentazione di Riesz mi garantisce che esiste in X un insieme $\{\phi_n\}_{n=1}^N$ tale che

$$F_n(f) = (f, \phi_n) \quad n = 1, \ldots, N, \tag{6.17}$$

dove $(\cdot, \cdot)$ indica il prodotto scalare canonico in X. Quindi il problema di ricostruzione che voglio affrontare è quello di determinare f conoscendo $\{g_n\}_{n=1}^N$ e $\{\phi_n\}_{n=1}^N$ tali che

$$(f, \phi_n) = g_n \quad n = 1, \ldots, N. \tag{6.18}$$

Oss. (esistenza e unicità): è chiaro che la soluzione del problema (6.18), se esiste, non può essere unica. L'esistenza vale se le ϕ_n sono linearmente indipendenti. Infatti, assumo per assurdo che per un certo elemento $\mathbf{g} = (g_1, \ldots, g_N) \in \mathbb{R}^N$ la soluzione non esista. Questo significa che l'insieme dei funzionali definisce una mappa la cui immagine è strettamente contenuta in $\mathbb{R}^N$. Quindi esiste in $\mathbb{R}^N$ un vettore $\mathbf{c} = (c_1, \ldots, c_N)$ ortogonale a ogni elemento dell'immagine della mappa, ovvero tale che

$$c_1(f, \phi_1) + \ldots + c_N(f, \phi_N) = 0 \tag{6.19}$$

per ogni $f \in X$. Quindi

$$c_1\phi_1 + \ldots + c_N\phi_N = 0 \tag{6.20}$$

e allora le ϕ_n formano un insieme linearmente dipendente, il che è assurdo.
Oss. (indipendenza lineare): considero il caso in cui le ϕ_n formino un insieme linearmente indipendente. Poichè l'insieme delle soluzioni è chiuso e convesso in X, esiste un'unica soluzione di norma minima del problema. Tale soluzione è in $X_N = span(\{\phi_n\}_{n=1}^N)$ e quindi ha forma

$$f = \sum_{n=1}^{N} a_n \phi_n. \tag{6.21}$$

Sostituendo in (6.18) ho che i coefficienti incogniti a_n sono le soluzioni del sistema lineare

$$\sum_{m=1}^{N} G_{mn} a_m = g_n \qquad n = 1, \ldots, N. \tag{6.22}$$

Oss. (caso generale): definisco l'operatore $L : X \to Y$ dove X è uno spazio di Hilbert, Y è uno spazio Euclideo (canonico) di dimensione N e

$$(Lf)_n = (f, \phi_n) \qquad n = 1, \ldots, N. \tag{6.23}$$

Dato $\mathbf{g} \in Y$, considero il problema di determinare $f \in X$ tale che

$$Lf = \mathbf{g}. \tag{6.24}$$

Si dicono pseudosoluzioni del problema l'insieme delle soluzioni di

$$\|Lf - \mathbf{g}\| = minimo. \tag{6.25}$$

Questo problema ai minimi quadrati è equivalente all'equazione di Eulero

$$L^* L f = L^* \mathbf{g}. \tag{6.26}$$

Poichè l'insieme delle pseudosoluzioni è chiuso e convesso in X, esiste un'unica pseudosoluzione di norma minima $f^\dagger$, che è detta soluzione generalizzata.
Oss. (soluzione generalizzata): se $\{\sigma_k; v_k, \mathbf{u_k}\}_{n=1}^{N}$ è il sistema singolare di L, la forma esplicita della soluzione generalizzata è data da

$$f^\dagger = \sum_{n=1}^{N} \frac{1}{\sigma_n} (\mathbf{g}, \mathbf{u}_n) v_n. \tag{6.27}$$

Il modo migliore per calcolare la SVD di L è osservare che LL^* è una matrice $N \times N$. Usando la definizione di aggiunto trovo subito che $L^* : Y \to X$ è tale che

$$L^* \mathbf{g} = \sum_{n=1}^{N} g_n \phi_n, \tag{6.28}$$

da cui

$$LL^* \mathbf{g} = \sum_{m=1}^{N} g_m (\phi_m, \phi_n) = \sum_{m=1}^{N} G_{mn} g_m. \tag{6.29}$$

Quindi $LL^* = G$ con

$$G_{mn} = (\phi_m, \phi_n) \tag{6.30}$$

e i valori singolari di L sono la radice quadrate degli autovalori di G. I vettori singolari $\mathbf{u}_n$ di L sono gli autovettori di G mentre

$$v_n = \frac{1}{\sigma_n} \sum_{m=1}^{N} (\mathbf{u}_n)_m \phi_m. \tag{6.31}$$

La matrice G è detta matrice di Gram.
Oss. (trasformata di Laplace): considero il caso in cui

$$g_n = \int_0^\infty e^{-x_n y} f(y) dy. \tag{6.32}$$

Allora la matrice di Gram ha entrate

$$G_{mn} = \int_0^\infty e^{-(x_n + x_m)y} dy = \frac{1}{x_n + x_m}. \tag{6.33}$$

Nel caso di un campionamento uniforme $x_n = x_0 + n$, si ha che la matrice di Gram è una matrice di Hilbert.

Chapter 7

Soluzione numerica delle equazioni differenziali ordinarie

7.1 Problema di Cauchy

Def. (problema di Cauchy): sia $f : \mathbb{R}^2 \to \mathbb{R}$ una funzione a valori in $\mathbb{R}$ definita su un aperto di $\mathbb{R}^2$. Si definisce problema di Cauchy il problema di determinare $y = y(x)$ tale che

$$\begin{cases} y' = f(x, y) \\ y(x_0) = y_0. \end{cases} \tag{7.1}$$

Per il teorema fondamentale del calcolo integrale, il problema di Cauchy ha soluzione su $I \subset \mathbb{R}$ se e solo se $\forall x \in \mathbb{R}$,

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))dt. \tag{7.2}$$

Def (Lipschitz continuità): una funzione $f = f(x, y)$ definita su un aperto $D \subset \mathbb{R}^2$ è detta Lipshitz-continua in y se esiste $K > 0$ tale che

$$|f(x, y_1) - f(x, y_2)| \leq K|y_1 - y_2| \tag{7.3}$$

per ogni (x, y_1), (x, y_2) in D.
Teo (esistenza e unicità): Hp.: sia

$$R = \{(x, y) \quad , \quad |x - x_0| \leq a \quad |y - y_0| \leq b\}; \tag{7.4}$$

f sia continua su R e Lipschitz-continua in R; sia

$$\delta = \min(a, \frac{b}{M}) \qquad M = \max_{(x,y)\in R} |f(x,y)|. \tag{7.5}$$

Th.: esiste unica $y = y(x)$ definita su $[x_0 - \delta, x_0 + \delta]$ che soddisfa il problema di Cauchy.

7.2 Metodo di Eulero

Def. (metodo di Eulero): è dato il problema di Cauchy (7.1) e una griglia uniforme I_h tale che

$$x_j = x_0 + jh \quad j = 0, \ldots, N \quad h = \frac{b-a}{N}. \tag{7.6}$$

Lo schema

$$u_{j+1} = u_j + hf(x_j, u_j) \quad j = 0, \ldots, N-1 \tag{7.7}$$

con $u_0 = y_0$ definisce il metodo di Eulero per la risoluzione numerica del problema di Cauchy.
NB: lo schema nasce dall'approssimare la derivata con le differenze finite, ovvero

$$\frac{u_{j+1} - u_j}{h} = f(x_j, u_j) \quad j = 0, 1, \ldots, N-1. \tag{7.8}$$

Def. (errori): si definisce errore dello schema di Eulero la quantità

$$e_j = u_j - y_j \quad j = 0, 1, \ldots, N \tag{7.9}$$

con $y_j = y(x_j)$. Si definisce errore di troncamento locale la quantità

$$\tau_{j+1} = \frac{y_{j+1} - y_j}{h} - f(x_j, y_j) \quad j = 0, 1, \ldots, N-1. \tag{7.10}$$

In pratica, l'errore di troncamento locale è l'errore con cui la soluzione esatta del problema di Cauchy fallisce nel soddisfare la relazione alle differenze (7.8). La relazione (7.10) è spesso scritta nella forma

$$y_{j+1} = y_j + hf(x_j, y_j) + h\tau_{j+1} \qquad j = 0, 1, \ldots, N-1. \tag{7.11}$$

Si dice che le equazioni alle differenze (7.8) sono consistenti con l'equazione differenziale del problema di Cauchy se $\tau_j \to 0$ quando $h \to 0$.
Oss. (esempio di consistenza): sia d^2y/dx^2 continua in $[a, b]$. Allora dal teorema di Taylor

$$y_{j+1} = y_j + h\frac{dy}{dx}(x = x_j) + \frac{h^2}{2}\frac{d^2y}{dx^2}(x_j + \theta_j h) \tag{7.12}$$

e quindi

$$\tau_{j+1} = \frac{h}{2}\frac{d^2}{dx^2}(x_j + \theta_j h). \tag{7.13}$$

Teo. (errore): Hp.: sia $y(x)$ soluzione del problema di Cauchy e $\{u_j\}$ lo schema fornito dal metodo di Eulero. Sia $f(x, y)$ Lipschitz-continua in y. Th.:

$$|u_j - y_j| \leq e^{K(x_j - x_0)}\left[|e_0| + \frac{\tau}{K}\right] \qquad j = 0, 1, \ldots, N, \tag{7.14}$$

con

$$\tau = \max_j |\tau_j|. \tag{7.15}$$

Dim.: faccio la sottrazione

$$y_{j+1} = y_j + hf(x_j, y_j) + h\tau_{j+1} \tag{7.16}$$

meno

$$u_{j+1} = u_j + hf(x_j, u_j) \tag{7.17}$$

e ottengo

$$e_{j+1} = e_j + h[f(x_j, u_j) - f(x_j, y_j)] - h\tau_{j+1}. \tag{7.18}$$

Uso la Lipschitz-continuità e ho

$$|e_{j+1}| \leq (1 + hK)|e_j| + h\tau. \tag{7.19}$$

Uso il risultato

$$1 + x + \ldots + x^n = \frac{1 - x^{n+1}}{1 - x} \tag{7.20}$$

e ho

$$|e_{j+1}| \leq (1 + hK)^{j+1}|e_0| + \left[\frac{(1 + hK)^{j+1} - 1}{K}\right]\tau. \tag{7.21}$$

Uso il risultato

$$(1+z)^n \leq e^{nz} \tag{7.22}$$

e ho

$$|e_{j+1}| \leq e^{K(x_{j+1}-x_0)} \left(|e_0| + \frac{\tau}{K} \right). \tag{7.23}$$

7.3 Metodi one-step

Def. (metodi-one step): è data la griglia di punti (7.6) e il problema di Cauchy (7.1) nell'ipotesi in cui la funzione f sia differenziabile molte volte con continuità. Data la funzione $\Phi = \Phi(x, y; h; f)$, si dice classe dei metodi one-step per la risoluzione numerica del problema di Cauchy la classe degli algoritmi

$$u_{j+1} = u_j + h\Phi(x_j, u_j; h; f) \ , \tag{7.24}$$

dove lo schema è inizializzato con un certo u_0. Ogni metodo in questa classe di algoritmi è definito da una specifica scelta della funzione Φ.
Oss. (rapporti incrementali): dall'equazione (7.24) segue immediatamente che

$$\Phi(x_j, u_j; h; f) = \frac{u_{j+1} - u_j}{h} \ , \tag{7.25}$$

ovvero $\Phi(x_j, u_j; h; f)$ è il rapporto incrementale della soluzione approssimata del problema di Cauchy. D'altra parte posso definire la funzione $\Delta = \Delta(x, y, ; f; h)$ tale che

$$\Delta(x, y, ; f; h) = \begin{cases} \frac{y(x+h)-y(x)}{h} & h \neq 0 \\ f(x, y) & h = 0 \end{cases} \tag{7.26}$$

per cui $\Delta(x_j, y_j; h; f)$ è il rapporto incrementale della soluzione esatta del problema di Cauchy.
Def. (troncamento locale): la quantità

$$\tau(x, y; h; f) = \Delta(x, y; h; f) - \Phi(x, y; h; f) \tag{7.27}$$

è detta errore di troncamento locale del metodo one-step. Il metodo è detto consistente se $\lim_{h\to 0} \tau(h) = 0$ ed è detto di ordine p se $\tau(h) = O(h^p)$.
Oss. (metodo di Eulero): il metodo di Eulero è un metodo one-step in

cui $\Phi(x, y; h; f) = f(x, y)$, ovvero Φ non dipende da h. Inoltre è un metodo del primo ordine (e quindi consistente). Infatti, per $h \neq 0$:

$$\tau(x, y; h; f) = \frac{y(x+h) - y(x)}{h} - f(x, y). \tag{7.28}$$

Utilizzando la serie di Taylor per $y(x)$ si ottiene

$$\tau(x, y; h; f) = \frac{y(x) + hy'(x) + \frac{h^2}{2}y''(x) + \ldots - y(x)}{h} - f(x, y) \ . \tag{7.29}$$

Quindi, sfruttando il fatto che $y'(x) = f(x, y)$ si ha

$$\tau(x, y; h; f) = \frac{h}{2}y''(x) + \ldots = \frac{h}{2}[f_x(x, y) + f_y(x, y)f(x, y)] + \ldots \ . \tag{7.30}$$

Oss.: (metodi di ordine superiore): un modo semplice di costruire metodi one-step di ordine superiore consiste nello scegliere $\Phi(x, y; h; f)$ uguale a una somma parziale della serie di Taylor di $\Delta(x, y; h; f)$ data da

$$\Delta(x, y; h; f) = f(x, y) + \frac{h}{2}y''(x) + \frac{h^2}{3!}y'''(x) + \ldots \quad . \tag{7.31}$$

Per esempio, se scelgo

$$\Phi(x, y; h; f) = f(x, y) + \frac{h}{2}y''(x) = f(x, y) + \frac{h}{2}[f_x(x, y) + f_y(x, y)f(x, y)] \ , \tag{7.32}$$

ottengo

$$\tau(h) = \frac{h^2}{3!}y''' + \ldots \tag{7.33}$$

e quindi un metodo di ordine 2. Sul piano numerico, si tratterebbe, tuttavia, di un metodo computazionalmente oneroso, in quanto richiede valutazioni di $f(x, y)$, $f_x(x, y)$ e $f_y(x, y)$ sulla griglia.

Def. (metodo di Runge-Kutta): il metodo di Runge-Kutta corrisponde alla scelta

$$\Phi(x, y; h; f) = \frac{h}{6}[k_1 + k_2 + k_3 + k_4] \quad , \tag{7.34}$$

con

$$k_1 = f(x, y) \qquad k_2 = f(x + \frac{h}{2}, y + \frac{h}{2}k_1) \quad , \tag{7.35}$$

$$k_3 = f(x + \frac{h}{2}, y + \frac{h}{2}k_2) \qquad k_4 = f(x + h, y + hk_3) \ . \tag{7.36}$$

Si dimostra usando il polinomio di Taylor che in questo caso $\tau(h) = O(h^2)$.