Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Matematica

Il problema agli autovalori del laplaciano e la sua approssimazione con il metodo degli elementi finiti

Tesi di Laurea in Analisi numerica

Relatore:
Chiar.mo Prof.
Michele Ruggeri

Presentata da:
Jacopo Fantini

Anno Accademico 2024/2025

Introduzione

I problemi agli autovalori occupano un ruolo di fondamentale importanza in numerosi ambiti della matematica applicata, dell'ingegneria e della fisica. Essi emergono naturalmente nello studio di fenomeni descritti da equazioni alle derivate parziali, come la propagazione del calore, le vibrazioni di corpi elastici o la diffusione di onde elettromagnetiche. In termini generali, un problema agli autovalori consiste nel determinare coppie (λ, u), dove λ è uno scalare e u una funzione non nulla, tali che un certo operatore differenziale agisca su u restituendone un multiplo scalare. Tali problemi consentono di individuare le "modalità naturali" di un sistema, fornendo informazioni sulla stabilità, la frequenza e l'energia delle sue soluzioni.

Il caso del laplaciano è di particolare rilievo, poiché questo operatore differenziale autoaggiunto è intimamente legato alla geometria del dominio su cui agisce. Gli autovalori del laplaciano rappresentano, per esempio, le frequenze di vibrazione di una membrana fissata al bordo, mentre le autofunzioni descrivono le corrispondenti forme modali. Lo studio del suo spettro permette quindi di connettere proprietà analitiche, geometriche e fisiche di un sistema, e costituisce un punto di incontro fra analisi funzionale, teoria degli operatori e calcolo numerico.

Dal punto di vista matematico, i problemi agli autovalori vengono analizzati a partire dalla loro formulazione variazionale (o debole), che permette di interpretare il problema in termini di operatori lineari su spazi di Hilbert. In tale contesto, e in particolare nell'ambito della teoria spettrale degli operatori compatti, è possibile dimostrare che, sotto opportune ipotesi, gli autovalori sono reali, possono essere ordinati in una successione crescente e ammettono autofunzioni ortogonali rispetto al prodotto scalare del dominio. Tuttavia, quando tali problemi devono essere risolti in pratica, si rende necessario ricorrere a metodi numerici capaci di fornire approssimazioni affidabili e convergenti.

Tra i metodi più diffusi, il metodo degli elementi finiti (FEM) riveste un ruolo di primaria importanza per la sua flessibilità e per la solidità della sua analisi teorica. Esso consente di discretizzare in modo sistematico equazioni differenziali definite su domini generali, anche di forma complessa, preservando al contempo le principali proprietà strutturali del

problema continuo. Nel caso dei problemi agli autovalori, il FEM permette di ottenere una sequenza di autovalori discreti che approssimano quelli continui, e per i quali è possibile stabilire precise stime di errore.
In questo contesto, un contributo fondamentale è rappresentato dai lavori di Babuška e Osborn, che hanno fornito un quadro astratto per lo studio della convergenza spettrale dei metodi variazionali. Tale teoria, oggi considerata un punto di riferimento nel campo, consente di analizzare la stabilità e la precisione delle approssimazioni numeriche degli autovalori e delle autofunzioni di operatori compatti autoaggiunti. Una sintesi ampia e organica di questi risultati è presentata, tra gli altri, nella review di Daniele Boffi [5], alla quale si farà riferimento nel seguito per completezza bibliografica.
La presente tesi ha come obiettivo lo studio del problema agli autovalori del laplaciano e della sua approssimazione mediante il metodo degli elementi finiti. Il lavoro è organizzato in quattro capitoli.
Nel capitolo 1 vengono introdotti i concetti fondamentali dell'analisi funzionale, che costituiscono la base teorica dello studio successivo: spazi vettoriali, spazi normati e di Hilbert, operatori compatti e le principali proprietà spettrali. Questa parte è dedicata a costruire il linguaggio e gli strumenti matematici necessari per comprendere la formulazione variazionale dei problemi agli autovalori.
Il capitolo 2 sviluppa il quadro astratto dei problemi agli autovalori associati a operatori compatti. Dopo una breve introduzione alla teoria spettrale, si presenta la formulazione variazionale del problema e si enunciano i principali risultati di convergenza per le approssimazioni numeriche, seguendo la linea della teoria di Babuška–Osborn. La trattazione culmina nell'applicazione al caso del laplaciano, di cui si discutono le proprietà matematiche e il significato fisico.
Nel capitolo 3 si analizza nel dettaglio il problema agli autovalori di Laplace in una dimensione. Dopo aver introdotto la formulazione debole classica, vengono presentate diverse discretizzazioni con elementi finiti lineari e quadratici, insieme ad alcune varianti miste. Le soluzioni numeriche ottenute vengono confrontate con quelle esatte, mettendo in evidenza l'ordine di convergenza e il comportamento degli errori al variare della discretizzazione. Questa parte della tesi ha lo scopo di illustrare concretamente l'efficacia e la precisione del metodo degli elementi finiti applicato a un caso di studio semplice ma rappresentativo.
Infine, il capitolo 4 affronta il problema agli autovalori generalizzato, introducendo i principali algoritmi iterativi utilizzati per la sua risoluzione, come il metodo delle potenze e il metodo delle potenze shiftate. Vengono poi proposte alcune applicazioni numeriche, che confermano i risultati teorici e permettono di osservare sperimentalmente la velocità di

convergenza dell'errore relativo e del residuo al variare dello shift, sia per la formulazione standard lineare che la formulazione variazionale mista $P1 - P0$.
Nel complesso, questa tesi intende offrire una panoramica coerente e completa dello studio dei problemi agli autovalori del laplaciano e della loro approssimazione numerica mediante elementi finiti. L'obiettivo è quello di unire rigore teorico e verifica sperimentale, mostrando come l'interazione tra analisi funzionale, teoria spettrale e metodi numerici consenta di affrontare in modo efficace problemi di grande rilievo scientifico e applicativo.

Indice

Capitolo 1

Strutture funzionali e operatori compatti

1.1 Spazi vettoriali

Definizione 1.1 (Spazio vettoriale)**.** *Uno spazio vettoriale su* $\mathbb{K}$ *è un insieme* V *i cui elementi, chiamati vettori, possono essere sommati tra loro e moltiplicati per elementi di* $\mathbb{K}$ *in modo che i risultati di tali operazioni siano ancora elementi di* V *:*

$$v, w \in V, \quad t \in \mathbb{K} \quad \rightsquigarrow \quad v + w, tv \in V.$$

Le operazioni di somma $V \times V \xrightarrow{+} V$ *e di prodotto per scalare* $\mathbb{K} \times V \to V$ *devono soddisfare le seguenti condizioni assiomatiche*

1. *(Proprietà associativa della somma) Comunque si prendano* $v, w, u \in V$ *vale*

$$v + (w + u) = (v + w) + u.$$

2. *(Proprietà commutativa della somma) Comunque si prendano* $v, w \in V$ *vale*

$$v + w = w + v.$$

3. *Esiste un elemento* $0_V \in V$*, detto vettore nullo di* V*, tale che per ogni* $v \in V$ *si ha*

$$v + 0_V = 0_V + v = v.$$

4. *Per ogni vettore* $v \in V$ *esiste un vettore* $-v \in V$*, detto opposto di* v*, tale che*

$$v + (-v) = -v + v = 0_V.$$

5. *Per ogni $v \in V$ vale $1v = v$, dove $1 \in \mathbb{K}$ è l'unità.*

6. *(Proprietà distributive). Comunque si prendano $v, w \in V$ e $a, b \in \mathbb{K}$ si ha che*

$$a(v + w) = av + aw, \quad (a + b)v = av + bv.$$

7. *Comunque si prendano $v \in V$ e $a, b \in \mathbb{K}$ si ha che*

$$a(bv) = (ab)v.$$

Definizione 1.2 (Sottospazio vettoriale)**.** *Sia V uno spazio vettoriale sul campo $\mathbb{K}$. Diremo che un sottoinsieme $U \subseteq V$ è un sottospazio vettoriale se soddisfa le seguenti condizioni:*

1. *$0_V \in U$;*

2. *U è chiuso per l'operazione di somma, ossia se $u_1, u_2 \in U$, allora $u_1 + u_2 \in U$;*

3. *U è chiuso per l'operazione di prodotto per scalare, ossia se $u \in U$ e $a \in \mathbb{K}$, allora $au \in U$.*
   Notiamo che se U è un sottospazio vettoriale di V, allora per ogni vettore $u \in U$ si ha $-u = (-1)u \in U$. Ne segue che U è a sua volta uno spazio vettoriale, con le operazioni di somma e prodotto per scalare indotte da quelle di V.

Definizione 1.3. *Lo spazio vettoriale V si dice di dimensione finita su $\mathbb{K}$, o anche finitamente generato, se esistono vettori $v_1, \ldots, v_n$ in V tali che $V = \mathrm{Span}\,(v_1, \ldots, v_n)$. In questo caso diremo che $\{v_1, \ldots, v_n\}$ è un insieme di generatori di V.*

Definizione 1.4. *La base di uno spazio vettoriale è un insieme di vettori che ha due proprietà fondamentali*

- *Linearmente indipendenti: nessuno dei vettori può essere scritto come combinazione lineare degli altri.*

- *Generano tutto lo spazio: qualsiasi vettore dello spazio può essere scritto come combinazione lineare dei vettori della base.*

Definizione 1.5 (Applicazione lineare)**.** *Siano V, W due spazi vettoriali sullo stesso campo $\mathbb{K}$. Un'applicazione $f : V \to W$ si dice lineare (su $\mathbb{K}$) se commuta con le somme ed i prodotti per scalare, ossia se $f(u + v) = f(u) + f(v), \quad f(tv) = tf(v), \quad$ per ogni $u, v \in V, t \in \mathbb{K}$.*

Ora definiamo due importanti sottospazi vettoriali, rispettivamente uno del dominio e uno del codominio : il nucleo e l'immagine.

Definizione 1.6 (Nucleo)**.** *Il nucleo di un'applicazione lineare $f : V \to W$ è l'insieme*

$$\mathrm{Ker}(f) = \{v \in V \mid f(v) = 0\}.$$

Osservazione 1.7. L'applicazione f è iniettiva se solo se $\ker(f) = 0$

Definizione 1.8 (Immagine)**.** *Il nucleo di un'applicazione lineare $f : V \to W$ è l'insieme*

$$\mathrm{Im}(f) = \{f(v) \mid v \in V\}.$$

Definizione 1.9 (Rango)**.** *Sia $f : V \to W$ applicazione lineare con $\mathrm{Im}(f)$ di dimensione finita, allora pongo*

$$\mathrm{rg}(f) = \dim\big(\mathrm{Im}(f)\big).$$

Se V ha dimensione finita il prossimo teorema ha una gran rilevanza:

Teorema 1.10 (Teorema del Rango)**.** *Sia $f : V \to W$ un'applicazione lineare. Allora V ha dimensione finita se e solo se $Im(V)$ e $\mathrm{Ker}(f)$ hanno entrambi dimensione finita; in tal caso vale la formula*

$$\dim V = \dim \mathrm{Ker}(f) + \mathrm{rg}(f).$$

1.1.1 Autovalori e autovettori nel caso finito-dimensionale

Sia V uno spazio vettoriale di dimensione finita su $\mathbb{K}$.

Definizione 1.11. *Un endomorfismo di V è una qualsiasi applicazione lineare $f : V \to V$*

In dimensione finita si può dimostrare che ogni endomorfismo lineare può essere rappresentato con una matrice.

Definizione 1.12 (Autovalore)**.** *Uno scalare $\lambda \in \mathbb{K}$ si dice un autovalore per f se l'endomorfismo $f - \lambda I : V \to V$ non è invertibile.*

Dunque, se la matrice A rappresenta $f : V \to V$ in una base fissata, allora $A - \lambda I$ rappresenta $f - \lambda I$ e quindi λ è un autovalore per f se e solo se $\det(f - \lambda I) = 0$, se e solo se λ è una radice del polinomio caratteristico di f.
Ad ogni autovalore di un endomorfismo f si possono associare alcuni invarianti numerici, ciascuno dotato di significato algebrico e/o geometrico. I due principali sono riassunti nella prossima definizione.

Definizione 1.13. *Sia λ è un autovalore di un endomorfsmo $f : V \to V$:*

- *la molteplicità di λ come radice del polinomio caratteristico viene detta molteplicità algebrica dell'autovalore;*
- *la dimensione del nucleo di $f - \lambda I : V \to V$ viene detta molteplicità geometrica dell'autovalore.*

Definizione 1.14 (Autovettore)**.** *Un autovettore per f è un vettore non nullo $v \in V$ tale che*

$$f(v) = \lambda v$$

per qualche $\lambda \in \mathbb{K}$. In questo caso si dice che v è un autovettore relativo all'autovalore λ di f.

Ora parliamo di un concetto fondamentale che è un punto di partenza fondamentale per l'analisi spettrale nel caso infinito-dimensionale.

Definizione 1.15 (Diagonalizzazione)**.** *Sia $A \in \mathbb{R}^{n\times n}$ una matrice quadrata. Diciamo che A è diagonalizzabile se esiste una base di $\mathbb{R}^n$ formata da autovettori di A, ovvero se esiste una matrice invertibile P tale che:*

$$P^{-1}AP = D$$

dove D è una matrice diagonale contenente gli autovalori di A.

Un caso particolarmente importante è quello delle matrici simmetriche($A = A^T$). In questo caso vale il seguente teorema:

Teorema 1.16 (Teorema spettrale)**.** *Ogni matrice reale simmetrica è diagonalizzabile mediante una matrice ortogonale. Cioè, esiste una matrice ortogonale Q (cioè $Q^T = Q^{-1}$) tale che:*

$$Q^T AQ = D$$

con D matrice diagonale contenente gli autovalori reali di A.

Questo risultato implica che:

- gli autovalori di una matrice simmetrica sono reali;
- esiste una base ortonormale di autovettori;

Una matrice simmetrica è autoaggiunta rispetto al prodotto scalare standard cioè vale:

$$\langle Ax, y\rangle = \langle x, Ay\rangle \quad \text{per ogni } x, y \in \mathbb{R}^n.$$

1.1.2 Lo spazio duale di uno spazio vettoriale

Sia V uno spazio vettoriale su un campo $\mathbb{K}$. Si definisce spazio duale di V l'insieme V^* formato da tutte le applicazioni lineari da V in $\mathbb{K}$:

$$V^* := \{f : V \to \mathbb{K} \mid f \text{ è lineare}\}.$$

Ogni elemento $f \in V^*$ è detto funzionale lineare.

Osservazione 1.17. Lo spazio V^* è anch'esso uno spazio vettoriale su $\mathbb{K}$, dove la somma e il prodotto per scalari sono definiti punto per punto:

$$(f+g)(v) := f(v) + g(v), \quad (\lambda f)(v) := \lambda f(v).$$

Esempio 1.18. Se $V = \mathbb{R}^n$, ogni funzionale lineare $f \in V^*$ può essere scritto come:

$$f(x) = a_1 x_1 + \cdots + a_n x_n,$$

per univoci coefficienti $a_i \in \mathbb{R}$. In questo caso, $V^* \cong \mathbb{R}^n$.

Osservazione 1.19. Se $\dim V = n < \infty$, allora $\dim V^* = n$, e lo spazio duale è isomorfo a V.

Gli spazi vettoriali forniscono il contesto algebrico per lo studio delle applicazioni lineari. Tuttavia, per introdurre concetti analitici come la distanza tra elementi o la convergenza di successioni, è necessario considerare strutture diverse: gli spazi metrici. Questi ultimi non richiedono una struttura lineare, ma introducono una metrica che consente di affrontare problemi di natura topologica e analitica, preparando così il terreno per spazi più strutturati come quelli normati e di Hilbert.

1.2 Spazi metrici

Definizione 1.20. *Uno spazio metrico è una coppia (X, d), dove X è un insieme non vuoto e $d : X \times X \to \mathbb{R}$ è una funzione, detta metrica, che soddisfa, per ogni $x, y, z \in X$*

- *(Positività) $d(x, y) \geq 0$, e $d(x, y) = 0 \iff x = y$;*
- *(Simmetria) $d(x, y) = d(y, x)$;*
- *(Disuguaglianza triangolare) $d(x, z) \leq d(x, y) + d(y, z)$.*

Esempio 1.21 (Spazio euclideo $\mathbb{R}^n$). La metrica classica è $d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$.

Esempio 1.22. Spazio euclideo $C([0,1])$ con metrica uniforme:

$$d(f,g) = \sup_{x\in[0,1]} |f(x) - g(x)|.$$

Esempio 1.23 (Spazio discreto). Dato un insieme X si definisce $d(x,y) = \begin{cases} 0 & \text{se } x = y \\ 1 & \text{se } x \neq y \end{cases}$

Un concetto centrale in uno spazio metrico è la convergenza di successioni.

Definizione 1.24. *Una successione $(x_n) \subset X$ si dice convergente a un punto $x \in X$ se:*

$$\lim_{n\to\infty} d(x_n, x) = 0.$$

Definizione 1.25. *Una successione $(x_n) \subset X$ si dice di Cauchy se:*

$$\forall \varepsilon > 0,\ \exists N \in \mathbb{N} \text{ tale che } d(x_n, x_m) < \varepsilon \quad \text{per ogni } n, m \geq N.$$

Definizione 1.26. *Uno spazio metrico è detto completo se ogni successione di Cauchy converge a un punto dello spazio.*

La metrica d induce su X una topologia naturale, in cui le palle aperte:

$$B(x,r) := \{y \in X \mid d(x,y) < r\}$$

formano una base per gli insiemi aperti.

Definizione 1.27 (Continuità tra spazi metrici)**.** *Siano (X, d_X) e (Y, d_Y) due spazi metrici. Una funzione $f : X \to Y$ si dice continua in un punto $x_0 \in X$ se:*

$$\forall\, \varepsilon > 0,\ \exists\, \delta > 0 \text{ tale che } d_X(x, x_0) < \delta \Rightarrow d_Y(f(x), f(x_0)) < \varepsilon.$$

La funzione si dice continua su X se è continua in ogni punto di X.

Osservazione 1.28. Una funzione $f : X \to Y$ tra spazi metrici è continua se e solo se per ogni insieme aperto $U \subseteq Y$, la controimmagine $f^{-1}(U) \subseteq X$ è un insieme aperto.

Definizione 1.29 (Compattezza in uno spazio metrico)**.** *Sia (X, d) uno spazio metrico. Un sottoinsieme $K \subset X$ si dice compatto se ogni successione $(x_n) \subset K$ ammette una sottosuccessione (x_{n_k}) convergente ad un punto $x \in K$. In altre parole:*

$$K \text{ è compatto} \iff \forall (x_n) \subset K,\ \exists x \in K \text{ e } x_{n_k} \to x.$$

Definizione 1.30 (Insieme relativamente compatto)**.** *Sia (X, d) uno spazio metrico. Un insieme $A \subset X$ si dice relativamente compatto se la sua chiusura $\overline{A}$ è compatta, ovvero:*

$$\overline{A} \text{ compatto } \Rightarrow A \text{ relativamente compatto.}$$

Teorema 1.31 (Bolzano–Weierstrass)**.** *Sia (X, d) uno spazio metrico. Ogni successione $(x_n) \subset X$ contenuta in un insieme compatto $K \subset X$ ammette una sottosuccessione convergente ad un punto $x \in K$.*

$$(x_n) \subset K \text{ con } K \text{ compatto} \implies \exists\, x \in K,\ \exists\, (x_{n_k}) \text{ tale che } x_{n_k} \to x.$$

Gli spazi metrici forniscono un ambiente generale in cui è possibile definire concetti topologici come la convergenza, la continuità e la compattezza. Tuttavia, questa struttura non è sufficiente quando si vogliono studiare operatori tra spazi vettoriali, o formulare problemi variazionali e differenziali in maniera funzionale.
Per farlo, è necessario che lo spazio sia anche uno spazio vettoriale, e che la distanza tra due punti sia legata alla differenza tra vettori. In altre parole, vogliamo che la metrica sia indotta da una norma, cioè da una funzione che assegna a ciascun vettore una "lunghezza".

1.3 Spazi normati

Definizione 1.32 (Norma su uno spazio vettoriale)**.** *Sia X uno spazio vettoriale. Una funzione $\|\cdot\| : X \to \mathbb{R}$ si dice* norma *su X se, per ogni $x, y \in X$ e per ogni scalare α, valgono le seguenti proprietà:*

1. *(Positività)* $\|x\| \geq 0$, *e* $\|x\| = 0 \iff x = 0$;
2. *(Omogeneità)* $\|\alpha x\| = |\alpha| \cdot \|x\|$;
3. *(Disuguaglianza triangolare)* $\|x + y\| \leq \|x\| + \|y\|$.

Definizione 1.33 (Spazio normato)**.** *Uno* spazio normato *è una coppia $(X, \|\cdot\|)$, dove X è uno spazio vettoriale e $\|\cdot\|$ è una norma su X.*

La norma induce una metrica su X definita da:

$$d(x, y) := \|x - y\|.$$

Con questa distanza, ogni spazio normato è anche uno *spazio metrico*, e quindi dotato di una topologia naturale.

Esempio 1.34 (Spazio euclideo $\mathbb{R}^n$). Definiamo tre spazi normati differenti su $\mathbb{R}^n$:

- $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$

- $\|x\|_1 = \sum |x_i|$
- $\|x\|_\infty = \max |x_i|$

Esempio 1.35. Spazio delle funzioni continue $C[a,b]$ con norma uniforme:

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

è uno spazio vettoriale reale e normato, completo.

Esempio 1.36 (Spazi ℓ^p delle successioni). L'insieme delle successioni numeriche $x = (x_n)$ tali che:

$$\|x\|_p = \left(\sum_{n=1}^{\infty} |x_n|^p\right)^{1/p} < \infty$$

per $1 \le p < \infty$, è uno spazio normato completo.

Definizione 1.37 (Norme equivalenti)**.** *Siano $\|\cdot\|_1$ e $\|\cdot\|_2$ due norme definite su uno stesso spazio vettoriale X. Le due norme si dicono equivalenti se esistono costanti positive $c, C > 0$ tali che:*

$$c\|x\|_1 \le \|x\|_2 \le C\|x\|_1, \quad \forall x \in X.$$

In tal caso, si dice che $\|\cdot\|_1$ e $\|\cdot\|_2$ inducono la stessa topologia su X.

Teorema 1.38. *Se $\|\cdot\|$ e $\|\cdot\|'$ sono due norme qualsiasi su uno spazio vettoriale di dimensione finita X, allora sono equivalenti.*

Definizione 1.39. *Siano X e Y due spazi vettoriali normati, e sia $T : X \to Y$ una applicazione lineare. Si dice che T è* limitato *se esiste una costante reale positiva $M > 0$ tale che:*

$$\|T(x)\|_Y \le M\|x\|_X \quad \text{per ogni } x \in X.$$

Definizione 1.40. *Siano X e Y due spazi vettoriali normati. L'insieme di tutte le applicazioni lineari continue da X in Y si denota con $\mathcal{B}(X,Y)$. Gli elementi di $\mathcal{B}(X,Y)$ si chiamano applicazioni lineari limitate.*

Per le applicazioni lineari, la limitatezza è equivalente alla continuità come si evince dalla prossima proposizione:

Proposizione 1.41. *Siano X e Y due spazi vettoriali normati, e sia $T : X \to Y$ una applicazione lineare. Le seguenti affermazioni sono equivalenti:*

1. *T è uniformemente continua;*

2. *T è continua;*

3. *T è continua in 0;*

4. *esiste un numero reale positivo $M > 0$ tale che*

$$\|T(x)\|_Y \leq M \quad \text{per ogni } x \in X \text{ con } \|x\|_X \leq 1;$$

5. *esiste un numero reale positivo $M > 0$ tale che*

$$\|T(x)\|_Y \leq M \cdot \|x\|_X \quad \text{per ogni } x \in X.$$

Proposizione 1.42. *Siano X e Y due spazi normati. Se la funzione*

$$\|T\| := \sup\{\|T(x)\|_Y : \|x\|_X \leq 1\}$$

è definita per ogni $T \in \mathcal{B}(X, Y)$, allora $\|\cdot\|$ è una norma su $\mathcal{B}(X, Y)$.

Definizione 1.43 (Spazio di Banach)**.** *Uno spazio di Banach è uno spazio normato completo rispetto alla norma indotta dalla distanza.*

Esempio 1.44. L'esempio (1.35) è uno spazio di Banach.

1.4 Spazi di Hilbert

Come visto nella sezione precedente, gli spazi di Banach sono spazi vettoriali normati completi, che forniscono un contesto molto generale per l'analisi funzionale. Tuttavia, quando la norma è indotta da un prodotto scalare, si ottengono spazi con una struttura geometrica più ricca: gli spazi di Hilbert.
Gli spazi di Hilbert non sono altro che spazi di Banach in cui la norma deriva da un prodotto scalare. Questa struttura aggiuntiva consente di estendere molte proprietà dell'algebra lineare finito-dimensionale (autovalori, ortogonalità, proiezioni) a contesti infinito-dimensionali, rendendoli particolarmente adatti allo studio dei problemi agli autovalori, delle equazioni differenziali e degli operatori compatti autoaggiunti.

1.4.1 Proprietà fondamentali degli spazi di Hilbert

Sia H uno spazio vettoriale su $\mathbb{C}$.

Definizione 1.45. *Un prodotto interno è un'applicazione $\langle \cdot, \cdot \rangle : H \times H \to \mathbb{C}$ tale che:*

- $\forall x, y \in H, \langle x, y \rangle = \overline{\langle y, x \rangle}$

- $\forall x \in H, \langle x, x\rangle \geqslant 0$
- $\forall x, y \in H, \forall \alpha \in \mathbb{C}, \langle \alpha x, y\rangle = \alpha\langle x, y\rangle$
- $\forall x, y, z \in H, \langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$
- $\langle x, x\rangle = 0$ *se e solo se* $x = 0$.

Proposizione 1.46. *Sia H uno spazio vettoriale dotato di un prodotto interno allora*

- *Per ogni $z \in H$, la funzione $\langle \cdot, z\rangle : H \to \mathbb{C}$ è lineare, cioè per ogni $x, y \in H$ e per ogni $\alpha, \beta \in \mathbb{C}$ vale:*
$$\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle.$$
- *Per ogni $x \in H$, la funzione $\langle x, \cdot\rangle : H \to \mathbb{C}$ soddisfa: per ogni $y, z \in H$ e per ogni $\alpha, \beta \in \mathbb{C}$ vale:*
$$\langle x, \alpha y + \beta z\rangle = \overline{\alpha}\langle x, y\rangle + \overline{\beta}\langle x, z\rangle.$$

Osservazione 1.47. Se H è uno spazio vettoriale su $\mathbb{R}$ allora il prodotto interno è un prodotto scalare su H.

Teorema 1.48 (Disuguaglianza di Schwartz)**.** *Sia H uno spazio vettoriale su $\mathbb{C}$ dotato di prodotto interno. Allora for ogni $x, y \in H$*

$$|\langle x, y\rangle| \leq \sqrt{\langle x, x\rangle}\sqrt{\langle y, y\rangle}$$

Teorema 1.49. *Sia H spazio vettoriale dotato di prodotto interno $\langle \cdot, \cdot\rangle$, per ogni $x \in H$ denoto $\|x\| = \sqrt{\langle x, x\rangle}$. Allora $\|x\|$ è una norma.*

Corollario 1.50. *Sia H dotato di prodotto interno, allora è uno spazio metrico dotato della distanza $d(x, y) = \|x - y\|$*

Definizione 1.51 (Spazio di Hilbert)**.** *Sia H uno spazio vettoriale dotato di prodotto interno. Se lo spazio metrico (H,d) [1] è completo allora (H,d) è uno spazio di Hilbert.*

Esempio 1.52. $\mathbb{R}^n, \langle x, y\rangle = \sum_{i=1}^n x_i y_i$ è uno spazio di Hilbert reale .

Esempio 1.53. $\mathbb{C}^n, \langle x, y\rangle = \sum_{i=1}^n x_i \bar{y}_i$ è uno spazio di Hilbert complesso.

Esempio 1.54. $L^2(\Omega, \mathbb{C}) = \{f : \Omega \to \mathbb{C} \mid \int_\Omega |f(x)|^2\, dx < +\infty\}.\langle f, g\rangle = \int_\Omega f\bar{g}\, dx$ è uno spazio di Hilbert complesso.

Esempio 1.55. $L^2(\Omega, \mathbb{R}), \langle f, g\rangle = \int_\Omega fg\, dx$ è uno spazio di Hilbert reale.

[1] $d(x, y) = \|x - y\|$ con $\|\cdot\|$ indotto dal prodotto interno

Esempio 1.56 (Spazio di Sobolev $H^1(\Omega)$). Sia $\Omega \subset \mathbb{R}^n$ un aperto limitato. Si definisce lo spazio di Sobolev $H^1(\Omega)$ come

$$H^1(\Omega) := \left\{ u \in L^2(\Omega) \;\middle|\; \frac{\partial u}{\partial x_i} \in L^2(\Omega),\ \forall i = 1, \dots, n \right\},$$

dove le derivate sono intese in senso debole.

$$\langle u, v \rangle_{H^1(\Omega)} := \int_\Omega u(x)v(x)\,dx + \sum_{i=1}^{n} \int_\Omega \frac{\partial u}{\partial x_i}(x)\frac{\partial v}{\partial x_i}(x)\,dx.$$

Con questa struttura, $H^1(\Omega)$ è uno spazio di Hilbert.

Esempio 1.57. $l^2 = \left\{ \{a_n\}_{n\in\mathbb{N}} \subset \mathbb{C} \,|\sum_{n=1}^{\infty}| \, a_n\big|^2 < +\infty \right\}$ dotato di $\left\langle \{a_n\}_{n\in\mathbb{N}}, \{b_n\}_{n\in\mathbb{N}} \right\rangle = \sum_{n=0}^{\infty} a_n \bar{b}_n$, è uno spazio di Hilbert.

Definizione 1.58. *Sia H spazio vettoriale dotato di prodotto interno. Siano $x, y \in H$, si dicono che sono ortogonali se $\langle x, y \rangle = 0$.*

La relazione di ortogonalità è simmetrica. Per ogni $x \in H$ spazio vettoriale dotato di prodotto interno, $x^\perp = \{y \in H \mid \langle x, y \rangle = 0\}$. Per ogni $A \subset H, A^\perp = \{y \in H \mid \langle y, x \rangle = 0, \forall x \in A\}$. Si dimostra che è un sottospazio vettoriale di H.

Proposizione 1.59. *Sia H uno spazio vettoriale dotato di prodotto interno, $x \in H$, allora $x^\perp$ è un sottospazio vettoriale chiuso. Sia $M \subset H$,allora $M^\perp$ è un insieme chiuso.*

Definizione 1.60 (Insieme convesso in H)**.** *Sia H uno spazio vettoriale. Diciamo che $E \subset H$ è un insieme convesso se per ogni $x_1, x_2 \in E, y = tx_1 + (1-t)x_2 \in E$ per ogni $t \in [0, 1]$.*

Definizione 1.61 (Spazio separabile)**.** *Uno spazio di Hilbert H si dice separabile se esiste un insieme numerabile $\{e_n\}_{n\in\mathbb{N}} \subset H$ tale che il suo sottospazio vettoriale* $\mathrm{span}\{e_n\}$ *è denso in H, ovvero:*

$$\overline{\mathrm{span}\{e_n\}} = H.$$

In tal caso, la famiglia $\{e_n\}_{n\in\mathbb{N}}$ si dice densa in H.

Definizione 1.62 (Sistema ortonormale)**.** *Sia H uno spazio vettoriale dotato di prodotto interno. Sia $E = \{u_j\}_{j\in J}$.*
Diciamo che $E \subset H$ è un sistema ortonormale se $\langle u_i, u_j \rangle = \delta_{ij}$ per ogni $i, j \in J$. Inoltre si dice completo se $\langle x, u_j \rangle = 0$ per ogni $j \in J$ implica $x = 0$.

Teorema 1.63. *Ogni spazio di Hilbert H ammette un sistema ortonormale completo e numerabile, cioè esiste un sistema ortonormale $\{u_n\}_{n\in\mathbb{N}}$ tale che*

$$x = \sum_{n=1}^{\infty} \langle x, u_n \rangle u_n \text{ per ogni } x \in H.$$

Inoltre,

$$\|x\|^2 = \sum_{n=1}^{\infty} \langle x, u_n \rangle^2$$

$$\langle x, y \rangle = \sum_{n=1}^{\infty} \langle x, u_n \rangle \langle y, u_n \rangle$$

I numeri $\langle x, u_n \rangle$ *si chiamano coefficienti di Fourier di* x *rispetto alla base ortonormale* $\{u_n\}$ *e si indicano anche con* $\hat{x}(n)$.

Esempio 1.64. $\left\{ \frac{e^{ikx}}{\sqrt{2\pi}} \right\}_{k \in \mathbb{Z}}$ è un sistema ortonormale $L^2([-\pi, \pi], \mathbb{C})$.

Esempio 1.65. $\{\cos(kx), \sin(kx)\}_{k \in \mathbb{N} \cup \{0\}}$ è un sistema ortonormale in $L^2([-\pi, \pi], \mathbb{R})$.

1.4.2 Teorema di rappresentazione di Riesz

Iniziamo con un po' di risultati utili per il teorema.

Lemma 1.66 (Legge del parallelogramma)**.** *Per ogni* $x, y \in H, \|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2$.

Teorema 1.67. *Sia* $A \neq \emptyset$ *chiuso e convesso in uno spazio di Hilbert* H. *Quindi esiste uno e uno solo* $x_0 \in A$ *tale che*

$$\|x_0\| = \min_{x \in A} \|x\|_H \leq \|x\|_H, x \in A$$

Quindi, x_0 *realizza il minimo di* $\|\cdot\|_H$ *in* A *e* x_0 *è unico.*

Teorema 1.68. *Sia* A *un sottospazio vettoriale chiuso dello spazio di Hilbert* H. *Quindi esiste un'unica coppia di applicazioni lineari* $P : H \to A, Q : H \to A^{\perp}$ *tale che: per ogni* $x \in H, x = Px + Qx$ *e*

(i) $Px = x, Qx = 0$, *per ogni* $x \in A$.

(ii) Se $x \in A^{\perp}$, *allora* $Px = 0, Qx = x$.

(iii) $\|x - Px\| = \inf\{\|x - y\| \mid y \in A\}$ *per ogni* $x \in H$.

(iv) $\|x\|^2 = \|Px\|^2 + \|Qx\|^2$.

Corollario 1.69. *Se* $A \subset H$ *è un sottospazio chiuso di uno spazio di Hilbert* H *tale che* $A \neq H$, *allora esiste* $b \in H, b \neq 0$, *tale che* b *è ortogonale a* A.

Teorema 1.70 (Rappresentazione di Riesz)**.** *Sia H uno spazio di Hilbert. Per ogni funzionale lineare continuo (limitato) $f : H \to \mathbb{C}$, esiste uno e un solo $y \in H$ tale che per ogni $x \in H$*

$$f(x) = \langle x, y \rangle.$$

Inoltre vale:

$$\|f\|_{H^*} = \|y\|_H.^2$$

Dimostrazione. Esistenza: Nel caso $f(x) = 0$ per ogni $x \in H$, possiamo scegliere $y = 0$. Nel caso non banale: $f(x) \neq 0$, consideriamo: Ker$f = \{x \in H \mid f(x) = 0\}$. Ker$f$ è un sottospazio lineare chiuso, poiché f è continuo. In particolare $H = \ker f \oplus (\ker f)^\perp$. Quindi $\exists z \in (Ker f)^\perp$.
Cerchiamo $y \in (ker f)^\perp$ t.c. $f(y) = \langle y, y \rangle$. Cerchiamo $y = \eta z$, da cui $f(y) = f(\eta z) = \langle \eta z, \eta z \rangle = \eta \overline{\eta} \langle z, z \rangle$ da cui

$$\overline{\eta} = \frac{f(z)}{\|z\|^2}.$$

Definiamo, per ogni $x \in H$:

$$x' = x - \frac{f(x)}{\|y\|^2} y, \quad x'' = \frac{f(x)}{\|y\|^2} y.$$

Allora:

$$\begin{aligned} f(x') &= f\left(x - \frac{f(x)}{\|y\|^2} y\right) = f(x) - \frac{f(x)}{\|y\|^2} f(y) \\ &= f(x)(1 - \frac{f(y)}{\|y\|^2}) = f(x)(1 - \frac{\langle y, y \rangle}{\langle y, y \rangle}) = 0. \end{aligned}$$

Quindi $x' \in$ Ker$f$ e $x = x' + x''$ con $x' \perp x''$. Di conseguenza:

$$\begin{aligned} \langle x, y \rangle &= \langle x' + x'', y \rangle = \langle x', y \rangle + \langle x'', y \rangle = \langle x'', y \rangle \\ &= \left\langle \frac{f(x)}{\|y\|^2} y, y \right\rangle = \frac{f(x)}{\|y\|^2} \langle y, y \rangle = f(x). \end{aligned}$$

Unicità: se esiste un altro y'' tale che per ogni $x \in H$, $f(x) = \langle x, y'' \rangle$, allora:

$$\langle x, y - y'' \rangle = 0 \quad \text{per ogni } x \in H,$$

cioè $y - y'' = 0$, quindi $y = y''$. Infine, usando la disuguaglianza di Cauchy-Schwarz:

$$|f(x)| = |\langle x, y \rangle| \leq \|x\| \|y\| \Rightarrow \sup_{x \neq 0} \frac{|f(x)|}{\|x\|} \leq \|y\|.$$

[2] Eliminerò il pedice H per non appesantire la notazione

Dall'altra parte, per $x = y$ abbiamo:

$$f(y) = \langle y, y\rangle = \|y\|^2 \Rightarrow \frac{|f(y)|}{\|y\|} = \|y\|.$$

Quindi:

$$\|f\|_{H^*} = \sup_{x\neq 0} \frac{|f(x)|}{\|x\|} = \|y\|_H.$$

□

1.4.3 Teorema di Lax-Milgram

Definizione 1.71. *Sia H uno spazio vettoriale complesso, $a : H \times H \to \mathbb{C}$ è una forma sesquilineare se vale:*

- *linearità nel primo argomento:* $a(\alpha u_1 + \beta u_2, v) = \alpha a(u_1, v) + \beta a(u_2, v) \quad \forall u_1, u_2, v \in V, \quad \alpha, \beta \in \mathbb{C}$.
- *antilinearità nel secondo argomento:* $a(u, \alpha v_1 + \beta v_2) = \bar{\alpha} a(u, v_1) + \bar{\beta} a(u, v_2) \quad \forall u, v_1, v_2 \in V, \quad \alpha, \beta \in \mathbb{C}$.

Definizione 1.72. *Sia H uno spazio di Hilbert, sia $a : H \times H \to \mathbb{C}$ una forma sesquilineare, si dice continua (o limitata) se esiste $M > 0$ t.c. $|a(u, v)| \leq M\|u\|\|v\|$.*

Definizione 1.73. *Sia H uno spazio di Hilbert, sia $a : H \times H \to \mathbb{C}$ una forma sesquilineare, si dice che è coerciva se esiste $\alpha > 0$ t.c. $\forall u \in H \setminus \{0\} \quad |a(u, u)| \geq \alpha\|u\|^2$.*

Teorema 1.74 (Lax-Milgram)**.** *Sia H uno spazio di Hilbert, sia $a : H \times H \to \mathbb{C}$ una forma sesquilineare limitata e coerciva, $\forall F \in H^* \; \exists! u \in H$ tale che $a(v, u) = F(v) \, \forall v \in H$.*

Dimostrazione. Osserviamo che $\forall u \in H$, $a(\cdot, u)$ è un'applicazione lineare poichè a è sesquilineare.
Dall'ipotesi che a è limitata possiamo dedurre che $a(\cdot, u) \in H^*$.
Dal teorema di Riesz segue che $\exists! T_u \in H$ per cui $a(v, u) = \langle v, T_u\rangle$.
Definiamo $T : H \to H, \quad u \mapsto T_u$, in primis vogliamo dimostrare che è lineare.
Provo che $T(u_1 + u_2) = T(u_1) + T(u_2)$:
$\forall u_1, u_2 \in H \; a(v, u_1) = \langle v, T(u_1)\rangle$, $a(v, u_2) = \langle v, T(u_2)$. Inoltre, $a(v, u_1 + u_2) = \langle v, T(u_1 + u_2)\rangle \Rightarrow \langle v, T(u_1 + u_2)\rangle = \langle v, T(u_1)\rangle + \langle v, T(u_2)\rangle$, allora $T(u_1 + u_2) = T(u_1) + T(u_2)$.
Si prova allo stesso modo che $\forall \alpha \in \mathbb{C}$ e $\forall u \in H \; T(\alpha u) = \alpha T(u)$.
Dalla limitatezza di a, si ha che $|\langle v, T(u)\rangle| = |a(v, u)| \leq M\|v\|\|u\|$, sostituendo v con l'elemento $T(u)$ si ha che $\|T(u)\|^2 = a(T(u), u) \leq M\|T(u)\|\|u\|$ da cui $\|Tu\| \leq M\|u\|$, cioè T è limitata.

Dalla coercività di a, esiste $\alpha > 0$ tale che $\alpha\|u\|^2 \leq |a(u,u)|$, da questo si ha che $\alpha\|u\|^2 \leq |a(u,u)| = \langle u, T(u)\rangle \leq \|u\|\|T(u)\|$ da cui $\alpha\|u\| \leq \|Tu\|$.
Sostituendo $u_1 - u_2$, al posto di u, ricavo che $\forall u_1, u_2 \in H$ tale che $T(u_1) = T(u_2)$ allora $u_1 = u_2$ cioè che T è iniettiva.
Dimostro che $T(H)$ è chiuso, infatti se $T(u_j) \xrightarrow[j\to+\infty]{} w \in H$, allora $\{T(u_j)\}_{j\in\mathbb{N}}$ è di Cauchy, dalla definizione di successione di Cauchy e dalla linearità di T si ha che $\|T(u_j) - T(u_k)\| = \|T(u_j - u_k)\| < \epsilon$ per ogni $j, k > \bar{n}(\epsilon)$.
Dall'altro lato si ha che $\alpha\|u_j - u_k\| \leq \|T(u_j - u_k)\| < \epsilon$ per ogni $j, k > \bar{n}(\epsilon)$ e questo implica che $\|u_j - u_k\| < \frac{\epsilon}{\alpha}$ cioè che è di Cauchy, allora converge a $\bar{u} \in H$. Quindi dalla continuità di T si ha che $\lim_{j\to+\infty} T(u_j) = T(\bar{u}) \in T(H)$.
Quindi $T(H)$ è chiuso dal teorema di caratterizzazione sequenziale dei chiusi.
Ora dimostriamo che T è suriettiva: dal fatto che $T(H)$ è chiuso, dal teorema della proiezione ortogonale si ha che $H = T(H) \oplus T(H)^\perp$, se per assurdo $T(H) \subsetneq H$ esiste $z \in T(H)^\perp$ non nullo.
Sappiamo che si ha che $0 = |\langle T(z), z\rangle| = |a(z,z)| \geq \alpha\|z\|^2$ quindi $z = 0$, assurdo.
Concludendo, $\forall F \in H^*$ sappiamo che $\exists! w \in H$ t.c. $F(v) = \langle v, w\rangle$, dalla suriettività di T: $\exists! u$ tale che $w = T(u)$, quindi $a(v,u) = \langle v, T(u)\rangle = \langle v, w\rangle = F(v)$. □

1.5 Operatori compatti

Supponiamo di lavorare tra spazi di Hilbert anche se in alcune definizioni non è necessario.

1.5.1 Proprietà degli operatori compatti

Definizione 1.75 (Operatore compatto)**.** *Siano H_1 e H_2 spazi di Hilbert e sia $K : H_1 \to H_2$ un operatore lineare. K si dice compatto se per ogni insieme limitato B in H, l'immagine $T(B)$ è relativamente compatta in H_2 (cioè la sua chiusura è compatta in H_2).*

Vediamo alcune proprietà degli operatori compatti:

1. Gli operatori compatti sono operatori limitati cioè continui.
2. K è compatto $\Leftrightarrow K(B(0,1))$ è relativamente compatto.
3. Un operatore $K : H_1 \to H_2$ è compatto se e solo se per ogni successione $\{\phi_n\}$ limitata in H_1, la successione $\{K\phi_n\}$ ha una sottosuccessione convergente in H_2.
4. Combinazioni lineari di operatori compatti sono ancora compatti.

5. Siano $T_1 : H_1 \to H_2$ e $T_2 : H_2 \to H_3$ operatori lineari limitati. Se uno dei due operatori è compatto, allora la composizione $T_2 \circ T_1 : H_1 \to H_3$ è un operatore compatto.

6. Se $K_n : H_1 \to H_2$ è una successione di operatori compatti convergente in norma all'operatore $K : H_1 \to H_2$, allora K è compatto.

7. Sia $T : H_1 \to H_2$ un operatore lineare limitato con immagine $Im(T)$ di dimensione finita. Allora T è compatto.

8. L'operatore identità $I : H_1 \to H_2$ è compatto se e solo se H_1 ha dimensione finita.

Dimostriamo (3) :

- Se K è compatto la successione $\{K\phi_n\}$ è contenuta nell'immagine di un limitato, quindi in un relativamente compatto, quindi ammette una sottosuccesione convergente.

- Supponiamo invece che dall'immagine di ogni successione limitata si possa estrarre una sottosuccessione convergente. Per mostrare che K è compatto basta far vedere che $K(B(0,1))$ è relativamente compatto. Sia $\{\phi_n\} \subset K(B(0,1))$. Esiste una successione $\{\phi_n\} \subset B(0,1)$ tale che $\psi_n = K\phi_n$. Ma $\|\phi_n\| \leq 1$, quindi si può estrarre da $\{\psi_n\}$ una sottosuccessione convergente, cioè $K\left(B(0,1)\right)$ è relativamente compatto.

Una conseguenza importante di (8) è che se H_1 ha dimensione infinita e $K : H_1 \to H_2$ è un operatore compatto iniettivo, allora l'inverso di K non è limitato.
Infatti, se $K^{-1} : K(H_1) \to H_1$ fosse limitato, allora risulterebbe compatto l'operatore

$$I : K^{-1}K : H_1 \to H_1$$

e H_1 dovrebbe aver dimensione finita, contrariamente all'ipotesi.

1.5.2 Teoria spettrale degli operatori compatti negli spazi di Hilbert

Proposizione 1.76. *Siano H_1 e H_2 due spazi di Hilbert e sia $T : H_1 \to H_2$ un operatore lineare continuo. Esiste allora un unico operatore lineare $T^* : H_2 \to H_1$ tale che*

$$\langle u, Tv\rangle = \langle T^*u, v\rangle \text{ per ogni } u \in H_2, v \in H_1.$$

Inoltre

$$\|T^*\|_{\mathcal{B}(H_2,H_1)} = \|T\|_{\mathcal{B}(H_1,H_2)}$$

L'operatore T^* si chiama operatore aggiunto di T.

Vale anche la relazione $(S \circ T)^* = T^* \circ S^*$.

Definizione 1.77. *Un operatore lineare $T : H \to H$ si dice autoaggiunto se*

$$T = T^*$$

cioè se

$$\langle x, Ty\rangle = \langle Tx, y\rangle \text{ per ogni } x, y \in H.$$

Proposizione 1.78. *Se K: $H_1 \to H_2$ è un operatore compatto tra spazi di Hilbert, allora anche $K^* : H_2 \to H_1$ è compatto.*

Proposizione 1.79. *Sia $T : H_1 \to H_2$ un operatore lineare limitato tra due spazi di Hilbert. Valgono allora le seguenti identità:*

$$\overline{Im(T)} = \mathrm{Ker}(T^*)^\perp \quad e \quad \overline{Im\,(T^*)} = \mathrm{Ker}(T)^\perp$$

e, inoltre,

$$H_2 = \overline{Im(T)} \oplus \mathrm{Ker}(T^*) \quad e \quad H_1 = \overline{Im\,(T^*)} \oplus \mathrm{Ker}(T)$$

Un risultato fondamentale riguardante gli operatori compatti,è il seguente teorema:

Teorema 1.80 (Alternativa di Fredholm)**.** *Sia H uno spazio di Hilbert, sia $K : H \to H$ un operatore compatto e K^* il suo aggiunto.Sia $\lambda \in C^*$. Allora:*

1. $\mathrm{Ker}(\lambda I - K)$ *ha dimensione finita*
2. $Im(\lambda I - K)$ *è chiuso e* $\mathrm{Im}(\lambda I - K) = \mathrm{Ker}\,(\lambda I - K^*)^\perp$
3. $\mathrm{Ker}(\lambda I - K) = \{0\}$ *se e solo se* $Im(\lambda I - K) = H$
4. $\dim\big(\mathrm{Ker}(\lambda I - K)\big) = \dim\big(\mathrm{Ker}(\lambda I - K^*)\big)$.

Osservazione 1.81. La proprietà (3) fa somigliare gli operatori compatti a operatori tra spazi di dimensione finita.

Definizione 1.82. *Sia H uno spazio di Hilbert e sia $T : H \to H$ un operatore lineare. Si chiama risolvente di T l'insieme*

$$\rho(T) = \{\lambda \in \mathbb{C} : T - \lambda I \text{ ha inverso limitato su } H\}.$$

Si chiama invece spettro di T il complementare del risolvente

$$\sigma(T) = \mathbb{C}\backslash\rho(T).$$

Un elemento $\lambda \in \sigma(T)$ è un autovalore se $T - \lambda I$ non è iniettivo.
Se λ è un autovalore, gli elementi non nulli del nucleo $\mathrm{Ker}(T - \lambda I)$ *si dicono autovettori.*

Vale il seguente risultato:

Teorema 1.83. *Sia $T : H \to H$ un operatore lineare su uno spazio di hilbert H.*

- *Se $x_1, \ldots, x_n$ sono un insieme finito di autovettori, ognuno corrispondente ad un diverso autovalore, allora essi sono linearmente indipendenti. Se T è autoaggiunto tali autovettori sono a due a due ortogonali.*
- *Se T è autoaggiunto,*

$$\|T\| = \sup_{\|x\|=1} |\langle x, x\rangle| = \sup\{|\lambda| : \lambda \in \sigma(T)\}.$$

Teorema 1.84 (Teorema spettrale per operatori autoaggiunti)**.** *Sia H uno spazio di Hilbert e sia $K : H \to H$ un operatore compatto autoaggiunto.Allora:*

1. *$\sigma(K)\backslash\{0\}$ è composto da soli autovalori reali. K ha almeno un autovalore e ne ha al più una infinità numerabile con 0 come unico possibile punto di accumulazione.*
2. *Per ogni autovalore $\lambda \neq 0$ esiste un numero finito di autovettori linearmente indipendenti. Autovettori corrispondenti ad autovalori diversi sono ortogonali.*
3. *Ordiniamo gli autovalori in modo che sia $|\lambda_1| > |\lambda_2| > \cdots$. Se indichiamo con P_j la proiezione su* $\mathrm{Ker}(\lambda_j I - K)$*, si ha*

$$K = \sum_{j=1}^{\infty} \lambda_j P_j$$

4. *Esiste una successione $\{x_j\}_{j\in J}$ (con J finito o $J = \mathbb{N}$) tale che x_j è un autovettore per $K, \langle x_i, x_j\rangle = 0$ se $i \neq j$ e tale che per ogni $x \in H$ esiste $x_0 \in Ker(K)$ tale che*

$$x = x_0 + \sum_{j\in J} \langle x, x_j\rangle x_j$$

e

$$Kx = \sum_{j\in J} \lambda_j \langle x, x_j\rangle x_j$$

Se K è iniettivo, $\{x_j : j \in J\}$ è un sistema completo in X.

Vogliamo adesso introdurre la decomposizione a valori singolari, che è l'analogo del teorema spettrale per operatori compatti non autoaggiunti.

Proposizione 1.85. *Sia $K : H_1 \to H_2$ un operatore compatto tra spazi di Hilbert allora $K^* \circ K : H_1 \to H_1$ è compatto, autoaggiunto e ha tutti gli autovalori positivi.*

Pongo $\sigma_j = \sqrt{\lambda_j}$ dove λ_j sono gli autovalori di K^*K. e sono chiamati valori singolari per K.

Teorema 1.86. *Sia $K : H_1 \to H_2$ un operatore compatto tra spazi di Hilbert. Esistono un insieme di indici J (finito o $J = \mathbb{N}$), una successione di numeri reali positivi $\{\sigma_j\}_{j\in J}$ e due sistemi ortonormali $\{e_j\}_{j\in J}$ in H_1 e $\{f_j\}_{j\in J}$ in H_2, tali che:*

1. *La successione $\{\sigma_j\}_{j\in J}$ è monotona non crescente e, se $J = \mathbb{N}$,*

$$\lim_{j\to+\infty} \sigma_j = 0.$$

2. *$Ke_j = \sigma_j f_j$ e $K^* f_j = \sigma_j e_j$ per $j \in J$.*

3. *Per ogni $x \in H_1$ esiste $x_0 \in \mathrm{Ker}(K)$ tale che*

$$x = x_0 + \sum_{j\in J} \langle x, e_j\rangle e_j$$

e

$$Kx = \sum_{j\in J} \sigma_j \langle x, e_j\rangle f_j.$$

4. *Per ogni $y \in H_2$,*

$$K^* y = \sum_{j\in J} \sigma_j \langle y, f_j\rangle e_j.$$

Definizione 1.87 (Sistema singolare)**.** *La famiglia $\{\sigma_j, e_j, f_j\}_{j\in J}$ è detta sistema singolare per K.*

Capitolo 2

Quadro astratto per problemi agli autovalori compatti

In questo capitolo ci concentreremo sulla teoria dell'approsimazione spettrale relativa alla formulazione debole classica dei problemi agli autovalori, cioè sulla formulazione variazionale standard in spazi di Hilbert; non verrà invece trattata la formuulazione mista.
L'esposizione inizierà con alcuni richiami di teoria spettrale per operatori compatti, per poi introdurre i problemi agli autovalori in forma variazionale.
Successivamente verrà presentata la teoria generale di Babuška-Osborn, che costituisce il riferimento fondamentale per lo studio della convergenza degli autovalori e delle autofunzioni approssimati con il metodo di Galerkin.
Infine, mostreremo l'applicazione di tali strumenti al problema classico di Laplace.

2.1 Richiami di teoria spettrale per operatori compatti

Sia X uno spazio di Hilbert e sia $T : X \to X$ un operatore lineare compatto.
L'insieme risolvente $\rho(T) = \{z \in \mathbb{C} \mid (zI - T) \text{ è biettiva}\}$ e l'operatore risolvente è $(z - T)^{-1}$.
Lo spettro di T è $\sigma(T) = \mathbb{C} \setminus \rho(T)$ che è un insieme numerabile con limite non diverso da 0. Tutti i valori non zero in $\sigma(T)$ sono autovalori, zero può essere come no un autovalore.
Se λ è un autovalore diverso da 0 di T, allora la molteplicità ascendente α di $\lambda - T$ è il più piccolo intero tale che $\ker(\lambda - T)^{\alpha} = \ker(\lambda - T)^{\alpha+1}$.
La terminologia proviene dal fatto che esiste una definizione simile per la molteplicità discendente, che fa uso dell'immagine al posto del Kernel; per gli operatori compatti la molteplicità ascendente e discendente coincidono.
La dimensione di $\ker(\lambda - T)^{\alpha}$ è chiamata molteplicità algebrica di λ, e gli elementi di

$\ker(\lambda - T)^\alpha$ sono gli autovettori generalizzati di T associati a λ.
Un autovettore generalizzato è di ordine k se è in $\ker(\lambda - T)^k \setminus \ker(\lambda - T)^{k-1}$.
Gli autovettori generalizzati di ordine 1 sono chiamati autovettori di T associati a λ, e sono gli elementi di $\ker(\lambda - T)$. La dimensione di $\ker(\lambda - T)$ è chiamata la molteplicità geometrica di λ, che è sempre minore o uguale alla molteplicità algebrica.
Se T è autoaggiunto, la molteplicità ascendente di ogni autovalore è uguale a 1, questo implica che ogni autovettore generalizzato è un autovettore e che molteplicità algebrica e geometrica coincidono.
Data una curva liscia $\Gamma \subset \rho(T)$ che include $\lambda \in \sigma(T)$ e non altri elementi di $\sigma(T)$, la proiezione spettrale di Riesz associata a λ è definita da

$$E(\lambda) = \frac{1}{2\pi \mathrm{i}} \int_\Gamma (z - T)^{-1} \, \mathrm{d}z. \tag{2.1}$$

La definizione chiaramente non dipende dalla curva scelta, e può essere verificato che :

1. $E(\lambda) : X \to X$
2. $E(\lambda) \circ E(\lambda) = E(\lambda)$ (significa che è una proiezione)
3. $E(\lambda) \circ E(\mu) = 0$ se $\lambda \neq \mu$
4. $E(\lambda) \circ T = T \circ E(\lambda)$

Infine vale che $E(\lambda)X = \ker(\lambda - T)^\alpha$ (quindi lo spazio degli autovettori generalizzati è un sottospazio invariante per T). Questa ultima proprietà è molto importante per lo studio dell'approssimazione degli autovettori.
In generale se $\Gamma \subset \rho(T)$ include più autovalori $\lambda_1, \ldots, \lambda_n$, quindi abbiamo che

$$E(\lambda_1, \lambda_2, \ldots, \lambda_n) X = \bigoplus_{i=1}^{n} \ker(\lambda_i - T)^{\alpha_i}$$

dove α_i denota la molteplicità ascendente di $\lambda_i - T$, cosicchè la dimensione dell'immagine della proiezione spettrale è in generale la somma delle molteplicità algebriche degli autovalori che stanno in Γ.
Sia $T^* : X \to X$ che denota l'aggiunto di T. Allora $\lambda \in \sigma(T^*)$ se e solo se $\overline{\lambda} \in \sigma(T)$, dove $\overline{\lambda}$ denota il coniugato di λ.
Gli autovalori degli operatori autoaggiunti sono reali, infatti se λ è autovalore di T:
$\lambda\langle v, v\rangle = \langle Tv, v\rangle \overset{T=T^*}{=} \langle v, Tv\rangle = \overline{\lambda}\langle v, v\rangle$ da cui $\overline{\lambda} = \lambda$.
La molteplicità algebrica di $\lambda \in \sigma(T^*)$ è uguale alla molteplicità algebrica di $\overline{\lambda} \in \sigma(T)$ e la molteplicità ascendente di $\lambda - T^*$ è uguale a quella di $\overline{\lambda} - T$.

2.2 Problemi agli autovalori in formulazione variazionale

In questa sezione introduciamo alcuni risultati preliminari sui problemi agli autovalori posti in forma variazionale, concentrandoci sul caso simmetrico.
Siano V e H due spazi di Hilbert reali. Supponiamo $V \subset H$ e che l'inclusione sia continua e densa.
Sia $a : V \times V \to \mathbb{R}$ e $b : H \times H \to \mathbb{R}$ forme bilineari simmetriche e continue, consideriamo il problema seguente: trovare $\lambda \in \mathbb{R}$ e $u \in V$, con $u \neq 0$, tale che

$$a(u, v) = \lambda b(u, v) \quad \forall v \in V. \tag{2.2}$$

Supponiamo che a sia coerciva (o V-ellittica), cioè che esiste $\alpha > 0$ tale che

$$a(v, v) \geq \alpha \|v\|_V \quad \forall v \in V.$$

Per semplicità supponiamo che b sia un prodotto scalare su H, in molte applicazioni, H sarà $L^2(\Omega)$ e b il suo prodotto interno standard.
Un importante strumento per l'analisi di (2.2) è l'operatore $T : H \to H$: dato $f \in H$, le nostre ipotesi ci garantiscono la validità delle ipotesi del teorema di Lax-Milgram, quindi $\exists! Tf \in V$ tale che

$$a(Tf, v) = b(f, v) \quad \forall v \in V.$$

Dal fatto che noi siamo interessati al problema agli autovalori compatto, assumiamo che $T : H \to H$ è un operatore compatto che è conseguenza spesso è una conseguenza dell'inclusione compatta di V in H.
Da queste ipotesi possiamo facilmente ricavare che T è autoaggiunto rispetto a b in H. Infatti, prendiamo $f, g \in H$, scegliamo $v = Tg$ nella definizione di T, scambiando i ruoli di f e g e usando la simmetria di a, troviamo che $b(f, Tg) = b(g, Tf)$.
Richiamiamo il fatto che lo spettro $\sigma(T)$ di T è un insieme finito o numerabile con 0 come unico punto di accumulazione.
Tutti i valori positivi di $\sigma(T)$ sono autovalori con molteplicità finita e i loro reciproci sono esattamente gli autovalori di (2.2); inoltre le autofunzioni di (2.2) hanno lo stesso autospazio di quelli di T.
Sia $\lambda_k, k \in \mathbb{N}$ denota il k-esimo autovalore di (2.2) con la naturale numerazione

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k \leq \cdots$$

dove lo stesso autovalore si può ripetere più volte in accordo con la sua molteplicità.
Sia u_k il corrispettivo autovettore con la normalizzazione standard $b(u_k, u_k) = 1$ e

$E_k = \text{span}\{u_k\}$ denota l'autospazio associato.
Osserviamo anche per gli autovalori semplici la normalizzazione non individua u_k unicamente, ma solo a meno del segno.
É ben noto che le autofunzioni godono della proprietà di ortogonalità:

$$a(u_m, u_n) = b(u_m, u_n) = 0 \quad \text{se } m \neq n \tag{2.3}$$

altrimenti, per autovalori multipli, quando $\lambda_m = \lambda_n$ le autofunzioni u_m e u_n possono essere scelte tale che l'ortogonalità (2.3) si mantenga.
Il quoziente di Rayleigh è un importante strumento per lo studio degli autovalori: risulta che

$$\begin{aligned} \lambda_1 &= \min_{v \in V} \frac{a(v,v)}{b(v,v)}, & u_1 &= \arg\min_{v \in V} \frac{a(v,v)}{b(v,v)}, \\ \lambda_k &= \min_{v \in \left(\bigoplus_{i=1}^{k-1} E_i\right)^{\perp}} \frac{a(v,v)}{b(v,v)}, & u_k &= \arg \min_{v \in \left(\bigoplus_{i=1}^{k-1} E_i\right)^{\perp}} \frac{a(v,v)}{b(v,v)}. \end{aligned} \tag{2.4}$$

dove il simbolo "$\perp$" denota il complemento ortogonale in V rispetto al prodotto scalare indotto dalla forma bilineare b. Dall'ortogonalità (2.3), deriva che il complemento ortogonale può essere preso rispetto al prodotto scalare indotto da a.
La discretizzazione di Galerkin del problema (2.2) è basata sullo spazio finito-dimensionale $V_h \subset V$ e letta nel modo seguente: trovare λ_h e u_h, con $u_h \neq 0$ tale che

$$a(u_h, v) = \lambda_h b(u_h, v) \quad \forall v \in V_h. \tag{2.5}$$

Osservazione 2.1. Per ragioni storiche, adottiamo la notazione della h-version del metodo degli elementi finiti, intendendo con h un parametro che tende a zero. Tuttavia, salvo diversa indicazione esplicita, la teoria che descriviamo si applica a una generica approssimazione di Galerkin.

Dal fatto che $V_h \subset V$ è uno spazio di Hilbert, possiamo ripetere i commenti che abbiamo fatto per il problema (2.2), partendo dall'operatore discreto $T_h : H \to H$: dato $f \in H$, $T_h f \in V_h$ è unicamente definito da

$$a(T_h f, v) = b(f, v) \quad \forall v \in V_h$$

Dal fatto che V_h è finito-dimensionale, T_h è compatto; gli autovalori di T_h sono in corrispondenza 1-1 con quelli di (2.5) (gli autovalori sono gli inversi del nostro problema e gli autospazi sono gli stessi), ordiniamo gli autovalori discreti di (2.5) in modo seguente:

$$\lambda_{1,h} \leq \lambda_{2,h} \leq \cdots \leq \lambda_{k,h} \leq \cdots,$$

dove gli autovalori sono ripetuti in base alle loro molteplicità. Usiamo $u_{k,h}$ con la normalizzazione $b(u_{k,h}, u_{k,h}) = 1$ per denotare le autofunzioni discrete, e $E_{k,h} = \text{span}\{u_{h,k}\}$

per l'autospazio associato.
Le autofunzioni discrete soddisfano la stessa ortogonalità che quelli continui,

$$a\left(u_{m,h}, u_{n,h}\right) = b\left(u_{m,h}, u_{n,h}\right) = 0 \quad \text{se } m \neq n,$$

dove questa proprietà è un teorema se $\lambda_{m,h} \neq \lambda_{n,h}$ o una definizione quando $\lambda_{m,h} = \lambda_{n,h}$.
La proprietà del quoziente di Rayleigh può essere applicato anche agli autovalori discreti così:

$$\begin{aligned} \lambda_{1,h} &= \min_{v \in V_h} \frac{a(v,v)}{b(v,v)}, & u_{1,h} &= \arg\min_{v \in V_h} \frac{a(v,v)}{b(v,v)}, \\ \lambda_{k,h} &= \min_{v \in \left(\oplus_{i=1}^{k-1} E_{i,h}\right)^{\perp}} \frac{a(v,v)}{b(v,v)}, & u_{k,h} &= \arg \min_{v \in \left(\oplus_{i=1}^{k-1} E_{i,h}\right)^{\perp}} \frac{a(v,v)}{b(v,v)}. \end{aligned} \tag{2.6}$$

dove con il simbolo "$\perp$" si denota il complemento ortogonale in V_h.
Una conseguenza immediata dell'inclusione $V_h \subset V$ e del quoziente di Rayleigh è

$$\lambda_1 \leq \lambda_{1,h}$$

cioè il primo autovalore discreto fornisce sempre una stima dall'alto del primo autovalore continuo. Lo stesso discorso non si può fare direttamente perchè in generale non è vero che $\left(\oplus_{i=1}^{k-1} E_{i,h}\right)^{\perp}$ è un sottospazio di $\left(\oplus_{i=1}^{k-1} E_i\right)^{\perp}$.
Per questa ragione, richiamiamo la caratterizzazione min-max degli autovalori.

Proposizione 2.2. *Il k-esimo autovalore λ_k del problema (2.2) soddisfa*

$$\lambda_k = \min_{E \in V_k} \max_{v \in E} \frac{a(v,v)}{b(v,v)},$$

dove V_k denota l'insieme dei sottospazi di V di dimensione k.

Dimostrazione. Per dimostrare che $\lambda_k \geq \min_{E \in V_k} \max_{v \in E} \frac{a(v,v)}{b(v,v)}$ basta che dimostro che scelto $E = \oplus_{i=1}^{k} E_i$ valga $a(v,v)/b(v,v) \leq \lambda_k \ \forall v \in V$. Questo vale dalla normalizzazione e dall'ortogonalità delle autofunzioni.
La dimostrazione della disuguaglianza opposta fornisce anche l'informazione che il minimo si ottiene per $E = \oplus_{i=1}^{k} E_i$ e la scelta $v = u_k$.
É chiaro che se $E = \oplus_{i=1}^{k} E_i$ la scelta ottimale per v è u_k. Dall'altra parte se $E \neq \oplus_{i=1}^{k} E_i$ allora esiste $v \in E$ con v ortogonale a u_i per $i \leq k$ e quindi $a(v,v)/b(v,v) \geq \lambda_k$, che mostra che $E = \oplus_{i=1}^{k} E_i$ è la scelta ottimale per E. □

Si ha un risultato analogo per il problema discreto :

$$\lambda_{k,h} = \min_{E_h \in V_{k,h}} \max_{v \in E_h} \frac{a(v,v)}{b(v,v)}, \tag{2.7}$$

dove $V_{k,h}$ denota l'insieme dei sottospazi di V_h di dimensione k.
É allora una conseguenza immediata che un'approssimazione conforme $V_h \subset V$ implica che tutti gli autovalori vengano approssimati dall'alto,

$$\lambda_k \leq \lambda_{k,h} \quad \forall k,$$

poichè tutti i sottospazi $E_h \in V_{k,h}$ che compaiono nel principio min-max discreto appartengono anche a V_k, e quindi il minimo discreto viene calcolato su un insieme più piccolo rispetto a quello continuo.
La monotonia è una proprietà interessante ma non è abbastanza per mostrare la convergenza.
La definizione di convergenza per autovalori/autofunzioni è un concetto intuitivo, che richiede però un formalismo accurato. Prima di tutto, noi vorremmo che il k-esimo autovalore discreto converga al k-esimo autovalore continuo. Questo implica due fatti importanti:

- tutte le soluzioni sono ben approssimate e nessun autovalore spurio "inquina" lo spettro.
- la numerazione che abbiamo scelto per gli autovalori, inoltre, implica che essi vengano approssimati correttamente con la loro molteplicità.

La convergenza delle autofunzioni è un po' più delicata, poichè non possiamo semplicemente richiedere che $u_{k,h}$ converga a u_k in una norma oppurtuna. Questo tipo di convergenza non può essere garantito per almeno due buoni ragioni.
Anzitutto l'autospazio associato ad autovalori multipli può essere approssimato dagli autospazi di distinti autovalori distinti. Inoltre, anche nel caso degli autovalori semplici, la normalizzazione delle autofunzioni non è sufficiente a garantire la convergenza, poichè potrebbero avere il segno opposto. La definizione naturale di convergenza fa uso della nozione di gap tra spazi di Hilbert definito da :

$$\delta(E,F) = \sup_{\substack{u\in E \\ \|u\|_H=1}} \inf_{v\in F} \|u-v\|_H,$$

$$\hat{\delta}(E,F) = \max\big(\delta(E,F),\, \delta(F,E)\big).$$

Per ogni numero positivo k, pongo $m(k)$ la somma delle dimensioni dei primi k autospazi diversi.

Definizione 2.3. *Il problema agli autovalori discreto (2.5) converge a quello continuo (2.2), se, per ogni $\epsilon > 0$ e $k > 0$, allora esiste $h_0 > 0$ tale che, per ogni $h < h_0$ abbiamo*

$$\max_{1\leq i\leq m(k)} \big|\lambda_i - \lambda_{i,h}\big| \leq \varepsilon,$$

$$\hat{\delta}\Big(\oplus_{i=1}^{m(k)} E_i,\ \oplus_{i=1}^{m(k)} E_{i,h}\Big) \leq \varepsilon.$$

É importante notare che questa definizione include tutte le proprietà di cui abbiamo bisogno: convergenza di autovalori e autofunzioni con la corretta molteplicità, e assenza di soluzioni spurie.

Osservazione 2.4. La definizione di convergenza (2.3) non fornisce alcuna indicazione sulla velocità di convergenza.

Proposizione 2.5. *Il problema (2.5) converge a (2.2) nel senso della definizione 2.3 se e solo se vale la seguente convergenza in norma:*

$$\|T - T_h\|_{\mathcal{L}(H)} \to 0 \quad \text{quando } h \to 0. \tag{2.8}$$

Osservazione 2.6. La nostra ipotesi di compattezza può essere modificata assumendo che $T : V \to V$ sia compatto. In questo caso una convergenza in norma analoga a (2.8) in $\mathcal{L}(V)$ garantirebbe una convergenza degli autospazi analoga a quella della definizione (2.3), con le naturali modifiche.

Per dimostrare la convergenza in norma (2.8), è utile osservare che l'operatore discreto T_h può essere visto come $T_h = P_h T$ dove $P_h : V \to V_h$ è la proiezione ellittica associata alla forma bilineare a cioè è definita nel modo seguente: $\forall u \in V, P_h u \in V_h$ è l'unico elemento che soddisfa $a(P_h u, v_h) = a(u, v_h)$.
Questo fatto implica che $T - T_h$ può essere scritta come $(I - P_h)T$ dove I denota l'operatore identità.

Proposizione 2.7. *Sia $T : H \to V$ compatto e $P_h \to I$ fortemente da V a H, allora vale la convergenza in norma (2.8).*

Dimostrazione. Prima di tutto mostriamo che la successione $\left\{\|I - P_h\|_{\mathcal{L}(V,H)}\right\}$ è limitata. Definisco $c(h,u)$ che è tale che $\|(I - P_h)\, u\|_H = c(h,u)\|u\|_V$. Dalla convergenza (puntuale) forte deriva che $c(h,u) \to 0$. $\forall h > 0 \quad \|(I - P_h)u\|_H \leq \|u\|_H + \|P_h u\|_H \leq \|u\|_H + C\|u\|_H$ quindi $M(u) = \max_h c(h,u)$ è finito. Quindi dal teorema di Banach-Steinhaus esiste C tale che, per ogni h, $\|I - P_h\|_{\mathcal{L}(V,H)} \leq C$.
Consideriamo una successione $\{f_h\}$ tale che, per ogni h, $\|f_h\|_H = 1$ e $\|(T - T_h)\|_H = \|(T - T_h)\, f_h\|_H$.
Dal fatto che $\{f_h\}$ è limitato in H e T è compatto da H a V, esiste una sottosuccessione, che denotiamo di nuovo con $\{f_h\}$ tale che $T f_h \to w$ in V.
Ora dimostriamo che $\|(I - P_h)\, T f_h\|_H \to 0$ e concludiamo, infatti $\|(T - T_h)\|_H = \|(T - T_h)\, f_h\|_H = \|(I - P_h)\, T f_h\|_H$.
Dalla suriettività di T: $\exists v \in H$ t.c. $Tv = w$ e osserviamo che $T_h v = P_h T v = P_h w$. Per

costruzione $Tf_h \to w$ e per ipotesi di convergenza puntuale forte $P_h w \to w$ in H, da questo segue che per ogni $\epsilon > 0$, esiste h abbastanza piccolo tale che

$$\begin{aligned}\|(I - P_h)Tf_h\|_H &\leq \|(I - P_h)(Tf_h - w)\|_H + \|(I - P_h)w\|_H \\ &\leq C\|Tf_h - w\|_V + \|(I - P_h)w\|_H \\ &\leq \varepsilon.\end{aligned}$$

□

Osservazione 2.8. La stessa dimostrazione della proposizione può essere utilizzata per mostrare che, se T è compatto da V a V, allora una convergenza puntuale più forte di P_h all'operatore identità da V in V, è sufficiente a garantire la convergenza in norma

$$\|T - T_h\|_{\mathcal{L}(V)} \to 0 \quad \text{quando } h \to 0.$$

2.3 La teoria di Babuška-Osborn

In questa sezione tratteremo i risultati più importanti della teoria di Babuška-Osborn trattando il problema generico (non simmetrico) nel campo complesso $\mathbb{C}$.
Sia X uno spazio di Hilbert e sia $T : X \to X$ un operatore lineare compatto. Consideriamo una famiglia di operatori compatti $T_h : X \to X$ tale che

$$\|T - T_h\|_{\mathcal{L}(X)} \to 0 \quad \text{quando } h \to 0. \tag{2.9}$$

Nelle nostre applicazioni T_h sarà un operatore a rango finito.
Come conseguenza di (2.9), se $\lambda \in \sigma(T)$ è un autovalore non nullo di molteplicità algebrica m, allora esistono esattamente m autovalori discreti di T_h(contate con le relative molteplicità) che convergono a λ se h tende a zero.
Questo segue dal fatto ben noto che, data una curva chiusa $\Gamma \subset \rho(T)$ per h sufficientemente piccolo abbiamo che $\Gamma \subset \rho(T_h)$ e Γ include esattamente m autovalori di T_h contati con le relative molteplicità.
Più precisamente, per h sufficientemente piccolo ha senso considerare la proiezione spettrale discreta

$$E_h(\lambda) = \frac{1}{2\pi \mathrm{i}} \int_\Gamma (z - T_h)^{-1} \, \mathrm{d}z$$

e si dimostra che la dimensione di $E_h(\lambda)X$ è uguale a m. Va inoltre osservato che

$$\|E(\lambda) - E_h(\lambda)\|_{\mathcal{L}(X)} \to 0 \quad \text{se } h \to 0$$

che implica la convergenza degli autospazi generalizzati.

Teorema 2.9. *Assumiamo che la convergenza in norma (2.9) è verificata. Per ogni insieme compatto $K \subset \rho(T)$, allora esiste $h_0 > 0$ tale che, per ogni $h < h_0$ abbiamo che $K \subset \rho(T_h)$(assenza di autovalori spuri).*
Se λ è un autovalore non nullo di T con molteplicità algebrica pari a m, allora esistono m autovalori $\lambda_{1,h}, \lambda_{2,h}, \ldots, \lambda_{m,h}$ di T_h, ripetuti secondo la loro molteplicità algebrica, tali che ciascuno $\lambda_{i,h}$ converge a λ quando $h \to 0$.
Inoltre, il gap tra la somma diretta degli spazi propri generalizzati associati a $\lambda_{1,h}, \lambda_{2,h}, \ldots, \lambda_{m,h}$ e lo spazio proprio generalizzato associato a λ tende a zero quando $h \to 0$.

Ora riportiamo i risultati fondamentali della teoria di Babuška-Osborn che trattano l'ordine di convergenza degli autovalori e degli autovettori.
Una delle principali applicazioni della teoria di Babuśka-Osborn consiste nell'analisi della convergenza per problemi agli autovalori in forma variazionale. Si inizia con la generalizzazione del quadro della sezione precedente ai problemi agli autovalori variazionali non simmetrici.
Siano V_1 e V_2 due spazi di Hilbert complessi. Noi siamo interessati al seguente problema agli autovalori: trovare $\lambda \in \mathbb{C}$ e $u \in V_1$, con $u \neq 0$, tale che

$$a(u, v) = \lambda b(u, v) \quad \forall v \in V_2 \tag{2.10}$$

dove $a : V_1 \times V_2 \to \mathbb{C}$ e $b : V_1 \times V_2 \to \mathbb{C}$ due forme sesquilineari. Supponiamo che a sia continua cioè

$$|a(v_1, v_2)| \leq C \|v_1\|_{V_1} \|v_2\|_{V_2} \quad \forall v_1 \in V_1 \quad \forall v_2 \in V_2,$$

e b sia continua rispetto alla norma compatta: cioè esiste una norma $\|\cdot\|_{H_1}$ in V_1 tale che ogni successione limitata in V_1 ha una sottosuccessione di Cauchy rispetto a $\|\cdot\|_{H_1}$ e

$$|b(v_1, v_2)| \leq C \|v_1\|_{H_1} \|v_2\|_{V_2} \quad \forall v_1 \in V_1 \quad \forall v_2 \in V_2.$$

Per definire l'operatore soluzione, assumiamo le condizioni inf-sup:

$$\inf_{v_1 \in V_1} \sup_{v_2 \in V_2} \frac{|a(v_1, v_2)|}{\|v_1\|_{V_1} \|v_2\|_{V_2}} \geq \gamma > 0,$$
$$\sup_{v_1 \in V_1} |a(v_1, v_2)| > 0 \qquad \forall v_2 \in V_2 \setminus \{0\}.$$

Introduciamo $T : V_1 \to V_1$ e $T_* : V_2 \to V_2$ sono tali che valgono:

$$a(Tf, v) = b(f, v) \qquad \forall f \in V_1, \ \forall v \in V_2,$$
$$a(v, T_* g) = b(v, g) \qquad \forall g \in V_2, \ \forall v \in V_1.$$

Assumiamo che T e T_* sono operatori compatti; inoltre, l'aggiunto di T su V_1 è dato da $T^* = A^* \circ T_* \circ A^{*-1}$, dove $A : V_1 \to V_2$ è l'operatore lineare standard associato alla forma

sesquilineare a.
Una coppia (λ, u) è un'autocoppia del problema (2.10) se e solo se soddisfa $\lambda Tu = u$ e (μ, u) è un'autocoppia dell'operatore T con $\mu = \lambda^{-1}$. I concetti di molteplicità ascendente, molteplicità algebrica e autovettori generalizzati del problema (2.10) sono legati alle analoghe proprietà dell'operatore T.
Utilizzeremo anche il seguente problema agli autovalori aggiunto: trovare $\lambda \in \mathbb{C}$ e $u \in V_2$, con $u \neq 0$, tale che

$$a(v, u) = \lambda b(v, u) \quad \forall v \in V_1. \tag{2.11}$$

La discretizzazione del problema (2.10) consiste nel selezionare due spazi finito-dimensionali $V_{1,h}$ e $V_{2,h}$, e considero il problema seguente:trovare $\lambda_h \in \mathbb{C}$ e $v_{1,h} \in V_{1,h}$ con $v_{1,h} \neq 0$ tale che

$$a\left(v_{1,h}, v_2\right) = \lambda_h b\left(v_{1,h}, v_2\right) \quad \forall v_2 \in V_{2,h}. \tag{2.12}$$

Supponiamo che $\dim\left(V_{1,h}\right) = \dim\left(V_{2,h}\right)$ cosicchè (2.12) è un problema agli autovalori quadratico (cioè in forma quadrata).
Assumiamo che le condizioni inf-sup discrete siano soddisfatte,

$$\inf_{v_1 \in V_{1,h}} \sup_{v_2 \in V_{2,h}} \frac{|a(v_1, v_2)|}{\|v_1\|_{V_1} \|v_2\|_{V_2}} \geq \gamma > 0,$$
$$\sup_{v_1 \in V_{1,h}} |a(v_1, v_2)| > 0 \qquad \forall v_2 \in V_{2,h} \setminus \{0\}.$$

cosicchè gli operatori discreti T_h e $T_{*,h}$ possono essere definiti analogamente a T e T_*. É chiaro che la convergenza delle autosoluzioni di (2.12) a (2.10) può essere analizzato per mezzo della convergenza di T_h e $T_{*,h}$ a rispettivamente T e T_*.
Siamo ora pronti a presentare i quattro principali risultati della teoria. Per ciascun risultato, enunciamo un teorema riguardante l'approssimazione delle coppie autovalore-autovettore di T, seguito da un corollario che contiene le conseguenze per l'approsimazione delle autosoluzioni di (2.10).
Consideriamo l'autovalore λ di (2.10) ($\mu = \lambda^{-1}$ nel caso dell'operatore T) di molteplicità algebrica m e con molteplicità ascendente di $\mu - T$ uguale a α.
Indichiamo con X il dominio V_1.
Il primo teorema tratta l'approssimazione degli autovettori generalizzati.

Teorema 2.10. *Sia μ un autovalore diverso da 0 di T, sia $E = E(\mu)X$ il suo autospazio generalizzato e pongo $E_h = E_h(\mu)X = \oplus_{i=1}^{m} E(\mu_{i,h})\, X_h$. Allora*

$$\hat{\delta}\left(E, E_h\right) \leq C \left\| (T - T_h)_{|E} \right\|_{\mathcal{L}(X)}.$$

Il risultato afferma che la distanza fra gli autospazi generalizzati veri e discreti è controllata dall'errore d'operatore, limitato al sottospazio E. La costante C dipende solo dalla separazione spettrale di μ.

Corollario 2.11. *Sia λ un autovalore di (2.10), sia $E = E(\lambda^{-1})V_1$ il suo autospazio generalizzato e sia $E_h = E_h(\lambda^{-1})V_1$. Allora*

$$\hat{\delta}\left(E, E_h\right) \leq C \sup_{\substack{u \in E \\ \|u\|=1}} \inf_{v \in V_{1,h}} \|u - v\|_{V_1}.$$

Nel caso di autovalori multipli si osserva che conviene introdurre la media aritmetica degli autovalori approssimanti.

Teorema 2.12. *Sia μ un autovalore non nullo di T con molteplicità algebrica pari a m e sia $\widehat{\mu}_h$ la media aritmetica dei m autovalori discreti di T_h che convergono verso μ. Siano $\phi_1, \ldots, \phi_m$ una base di autovettori generalizzati in $E = E(\mu)X$ e $\phi_1^*, \ldots, \phi_m^*$ una base duale di autovettori generalizzati in $E^* = E^*(\overline{\mu})X$. Allora*

$$|\mu - \widehat{\mu}_h| \ \leq \ \frac{1}{m} \sum_{i=1}^{m} |((T - T_h)\phi_i, \phi_i^*)| + C \, \|(T - T_h)|_E\|_{\mathcal{L}(X)} \, \|(T^* - T_h^*)|_{E^*}\|_{\mathcal{L}(X)}.$$

Osserviamo che $E^* = E^*(\overline{\mu})X = \ker((T^* - \overline{\mu}I)^{\alpha})$ dove $\phi_1^*, \ldots, \phi_m^*$ soddisfa la condizione di biortogonalità: $(\Phi_i, \Phi_j^*) = \delta_{ij} \quad i, j = 1, \ldots, m$.
Questo teorema dice che l'errore sulla media degli autovalori discreti che approssimano μ è dominato dalla consistenza di T_h su E, più un termine che è il prodotto degli errori di T_h e del suo aggiunto sui relativi spazi propri.

Corollario 2.13. *Sia λ un autovalore di (2.10) e sia $\widehat{\lambda_h}$ la media aritmetica dei m autovalori discreti di (2.12) convergenti a λ. Allora*

$$|\lambda - \widehat{\lambda}_h| \leq C \sup_{\substack{u \in E \\ \|u\|=1}} \inf_{v \in V_{1,h}} \|u - v\|_{V_1} \sup_{\substack{u \in E^* \\ \|u\|=1}} \inf_{v \in V_{2,h}} \|u - v\|_{V_2},$$

dove E è lo spazio degli autovettori generalizzati associato a λ e E^* è lo spazio generalizzato aggiunto associato a λ.

Quindi l'errore sugli autovalori non dipende "direttamente" da λ, ma dalla capacità del metodo di approssimare:

- gli autovettori del problema primale (E in norma V_1)
- gli autovettori del problema aggiunto (E^* in norma V_2)

La stima dell'errore degli autovalori coinvolge la molteplicità ascendente α.

Teorema 2.14. *Siano $\phi_1, \dots, \phi_m$ una base dello spazio proprio generalizzato $E = E(\mu)X$ di T e $\phi_1^*, \dots, \phi_m^*$ una base duale. Allora, per $i = 1, \dots, m$, vale*

$$|\mu - \mu_{i,h}|^\alpha \;\le\; C \left\{ \sum_{j,k=1}^{m} |((T - T_h)\phi_j, \phi_k^*)| + \|(T - T_h)|_E\|_{\mathcal{L}(X)} \, \|(T^* - T_h^*)|_{E^*}\|_{\mathcal{L}(X)} \right\},$$

dove $\mu_{1,h}, \dots, \mu_{m,h}$ sono i m autovalori discreti (ripetuti secondo la loro molteplicità algebrica) che convergono a μ, ed E^* è lo spazio degli autovettori generalizzati di T^* associati a $\overline{\mu}$.

Osserviamo che se μ non è semplice, la convergenza può essere più lenta.

Corollario 2.15. *Con la notazione del teorema precedente, per $i = 1, \dots, m$ abbiamo*

$$|\lambda - \lambda_{i,h}|^\alpha \le C \sup_{\substack{u \in E \\ \|u\|=1}} \inf_{v \in V_{1,h}} \|u - v\|_{V_1} \sup_{\substack{u \in E^* \\ \|u\|=1}} \inf_{v \in V_{2,h}} \|u - v\|_{V_2} \tag{2.13}$$

dove E è lo spazio degli autovettori generalizzati associati a λ e E^* è lo spazio generalizzato aggiunto associato a λ.

Osservazione 2.16. Apparentemente le stime degli ultimi due corollari non sembrano conseguenze dei rispettivi teoremi. Nella prossima sezione mostreremo una prova di questi risultati, nel caso particolare del problema agli autovalori di Laplace.

L'ultimo risultato è più tecnico dei precedenti e completa il teorema 2.10 nella descrizione dell'approssimazione degli autovettori generalizzati. In particolare nel caso $k = \ell = 1$, il teorema si applica agli autovettori.

Teorema 2.17. *Sia $\{\mu_h\}$ una successione di autovalori discreti di T_h che converge a un autovalore non nullo μ di T. Si consideri una successione $\{u_h\}$ di vettori unitari in $\ker(\mu_h - T_h)^k$ per qualche $k \le \alpha$ (autovettori generalizzati discreti di ordine k). Allora, per ogni intero ℓ con $k \le \ell \le \alpha$, esiste un autovettore generalizzato $u(h)$ di T di ordine ℓ tale che*

$$\|u(h) - u_h\|_X^{\alpha/(\ell-k+1)} \;\le\; C \, \|(T - T_h)|_E\|_{\mathcal{L}(X)}.$$

Il teorema garantisce che ad ogni autovettore generalizzato discreto u_h associato a μ_h corrisponde un autovettore generalizzato dell'operatore continuo T, di ordine non minore, che approssima u_h.
La distanza tra i due è controllata dall'errore di approssimazione dell'operatore discreto rispetto a quello continuo. In tal modo, non solo gli autovalori discreti convergono a quelli esatti, ma viene preservata anche la struttura degli autospazi generalizzati, con una stima quantitativa della convergenza.

Corollario 2.18. *Sia $\{\lambda_h\}$ una successione di autovalori discreti di (2.12) che converge a un autovalore λ di (2.10). Si consideri una successione $\{u_h\}$ di autofunzioni unitarie in $\ker(\lambda_h^{-1} - T_h)^k$ per qualche $k \leq \alpha$ (autofunzioni generalizzate discrete di ordine k). Allora, per ogni intero ℓ con $k \leq \ell \leq \alpha$, esiste un autovettore generalizzato $u(h)$ di (2.10) di ordine ℓ tale che*

$$\|u(h) - u_h\|_{V_1}^{\alpha/(\ell-k+1)} \leq C \sup_{\substack{u \in E \\ \|u\|=1}} \inf_{v \in V_{1,h}} \|u - v\|_{V_1}.$$

Concludiamo la sezione con un'applicazione di questa teoria al caso dei problemi agli autovalori simmetrici formulati in forma variazionale.
Supponiamo $V_1 = V_2$ due identici spazi di Hilbert, lo denotiamo V, e siano a e b due forme bilineari simmetriche. Assumiamo che a sia coerciva:

$$a(v, v) \geq \gamma > 0 \quad \forall v \in V,$$

e che b può essere estesa a una forma bilineare su $H \times H$ con H spazio di Hilbert tale che l'inclusione $V \subset H$ è compatta. Inoltre, assumiamo che b è definita positiva su $V \times V$:

$$b(v, v) > 0 \quad \forall v \in V \backslash \{0\}.$$

Come abbiamo visto nelle sezioni precedenti, gli autovalori di (2.10) sono tutti positivi e possono essere ordinati in una successione tendente a infinito,

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k \leq \cdots$$

dove la ripetizione degli autovalori è coerente con la relativa molteplicità.
Sia $V_h \subset V$ lo spazio finito-dimensionale usato per l'approssimazione delle autocoppie e denoto $\lambda_{k,h}$ $(k = 1, \ldots, dim(V_h))$ gli autovalori discreti.
La condizione del min-max (proposizione 2.2 e la discussione successiva) e il corollario 2.15 danno il seguente risultato.

Teorema 2.19. *Per ogni k vale*

$$\lambda_k \leq \lambda_{k,h} \leq \lambda_k + C \sup_{\substack{u \in E \\ \|u\|=1}} \inf_{v \in V_h} \|u - v\|_V^2,$$

dove E denota l'autospazio associato a λ_k.

Il corollario (2.11), letto nel caso simmetrico , diventa il seguente.

Teorema 2.20. *Sia u_k un'autofunzione unitaria associata ad un autovalore λ_k di molteplicità m, tale che*

$$\lambda_k = \cdots = \lambda_{k+m-1}.$$

Siano $u_{k,h}, \dots, u_{k+m-1,h}$ le autofunzioni associate ai m autovalori discreti che convergono a λ_k. Allora esiste

$$w_{k,h} \in \text{span}\{u_{k,h}, \dots, u_{k+m-1,h}\}$$

tale che

$$\|u_k - w_{k,h}\|_V \ \leq \ C \sup_{\substack{u \in E \\ \|u\|=1}} \inf_{v \in V_h} \|u - v\|_V,$$

dove E denota l'autospazio associato a λ_k.

In realtà, si potrebbero ottenere stime più precise nel caso di autovalori multipli, in particolare nel caso un autovalore multiplo è associato a più autofunzioni in spazi di regolarità differente.
La stima (2.13) suggerirebbe che, in tal caso, l'autovalore viene approsimato con l'ordine di convergenza dettato dalla minore regolarità tra le autofunzioni; dall'altra parte è possibile che gli autovalori approssimati presentino velocità di convergenza differenti, a seconda delle diverse regolarità delle autofunzioni associate.

2.4 Il problema agli autovalori di Laplace

In questa sezione applichiamo la teoria di Babuška-Osborn all'analisi di convergenza dell'approssimazione con elementi finiti conformi del problema agli autovalori di Laplace. Richiamiamo il problema: dato $\Omega \subset \mathbb{R}^n$ e lo spazio di Sobolev reale $H_0^1(\Omega)$, cerchiamo gli autovalori $\lambda \in \mathbb{R}$ e le autofunzioni $u \in H_0^1(\Omega)$, con $u \neq 0$, tale che

$$(\nabla u, \nabla v) = \lambda(u, v) \quad \forall v \in H_0^1(\Omega).$$

La discretizzazione di Ritz-Galerkin si basa sullo spazio finito-dimensionale $V_h \subset H_0^1(\Omega)$ e consiste nel cercare gli autovalori $\lambda_h \in \mathbb{R}$ e le autofunzioni $u_h \in V_h$, con $u_h \neq 0$, tale che

$$(\nabla u_h, \nabla v) = \lambda(u_h, v) \quad \forall v \in V_h.$$

Denotiamo con $a(\cdot, \cdot)$ la forma bilineare $(\nabla \cdot, \nabla \cdot)$ che è il prodotto scalare stardard in $H_0^1(\Omega)$ e $(\cdot, \cdot)$ il prodotto scalare standard in $L^2(\Omega)$.
Siamo nel setting della sezione 2.2 con $V = H_0^1(\Omega)$ e $H = L^2(\Omega)$, verifichiamolo:

- a continua: $\forall u, v \in H_0^1(\Omega)\ \|a(u,v)\| \overset{Holder}{\leq} ||\nabla u||_{L^2(\Omega)} ||\nabla v||_{L^2(\Omega)} = ||u||_{H_0^1(\Omega)} ||v||_{H_0^1(\Omega)}$;
- a coerciva: direttamente dalla sua definizione;
- b continua: $\forall u, v \in L^2(\Omega)\ \|b(u,v)\| \overset{Holder}{\leq} ||u||_{L^2(\Omega)} ||v||_{L^2(\Omega)}$.

Usando la notazione della sezione precedente, il punto di partenza per l'analisi consiste in una definizione adeguata dell'operatore soluzione $T : X \to X$ e, in particolare, dello spazio funzionale X.
Siano V e H, rispettivamente $H_0^1(\Omega)$ e $L^2(\Omega)$. La definizione più naturale consiste nel prendere $X = V$ e definire $T : V \to V$, tramite

$$a(Tf, v) = (f, v) \quad \forall v \in V. \tag{2.14}$$

La definizione può essere estesa a $X = H$, poichè ha senso considerare la soluzione del problema forzato di Laplace con $f \in L^2(\Omega)$.
Quindi abbiamo due soluzioni possibili: $T_V : V \to V$ e $T_H : H \to H$ dove T_H è definita analogamente a T_V da $T_H f \in V \subset H$ e (2.14). Osserviamo che T_V e T_H sono autoggiunti come dimostrato dalle seguenti proposizioni.

Proposizione 2.21. *Se V è dotato della norma indotta dal prodotto scalare definito dalla forma bilineare a, allora T_V è autoaggiunto.*

Dimostrazione. Il risultato segue dalle uguaglianze

$$a\left(T_V x, y\right) = (x, y) = (y, x) = a\left(T_V y, x\right) = a\left(x, T_V y\right) \quad \forall x, y \in V.$$

□

Proposizione 2.22. *L'operatore T_H è autoaggiunto.*

Dimostrazione. Il risultato segue dalle uguaglianze

$$(T_H x, y) = (y, T_H x) = a\left(T_H y, T_H x\right) = a\left(T_H x, T_H y\right) = (x, T_H y) \quad \forall x, y \in H.$$

□

Osservazione 2.23. Le autosoluzioni degli operatori T_V e T_H sono le stesse quindi entrambe le funzioni possono essere usate per l'analisi.

Per l'analisi della convergenza assumiamo che V_h soddisfa le seguenti stime :

$$\inf_{v \in V_h} \|u - v\|_{L^2(\Omega)} \leq C h^{\min\{k+1, r\}} \|u\|_{H^r(\Omega)},$$

$$\inf_{v \in V_h} \|u - v\|_{H^1(\Omega)} \leq C h^{\min\{k, r-1\}} \|u\|_{H^r(\Omega)}.$$

Tali stime sono standard quando V_h contiene polinomi a tratti di grado al più k.

2.4.1 Analisi nel caso $T = T_V$

Mostriamo come usare i risultati della sezione precedente con la scelta $T = T_V$.
L'operatore discreto può essere definito nel modo seguente come $T_h : V \to V$ da $T_h f \in V_h \subset V$ e

$$a\left(T_h f, v\right) = (f, v) \quad \forall v \in V_h.$$

La stima d'errore standard per la soluzione dell'equazione di Laplace con termine sorgente implica che la convergenza in norma (2.8) è soddisfatta per tutti i domini ragionevoli. Se Ω è continuo-Lipschitz, per esempio, vale la prossima stima: esiste $\epsilon > 0$ tale che

$$\|Tf - T_h f\|_{H_0^1(\Omega)} \leq Ch^{\varepsilon} \|f\|_{H_0^1(\Omega)}.$$

Quindi le conseguenze del teorema (2.9) sono valide: tutti gli autovalori/autofunzioni continui sono ben approssimati e tutti gli autovalori/autofunzioni discreti approssimano alcuni autovalori/autofunzioni continui con la molteplicità corretta.
Ora vogliamo stimare la velocità di convergenza. Supponiamo che siamo interessati alla velocità di convergenza dell'approssimazione di λ e di avere che l'autospazio associato E è sottospazio vettoriale di $H^r(\Omega)$, che implica

$$\left\| (T - T_h)_{|E} \right\|_{\mathcal{L}(V)} = O\left(h^{\min\{k, r-1\}}\right). \tag{2.15}$$

Denotiamo con τ la quantità $\min\{k, r-1\}$. Il risultato di (2.20), per questa istanza particolare, può essere interpretato in questo modo:

Teorema 2.24. *Sia u un'autofunzione unitaria associata all'autovalore λ di molteplicità m, e siano $w_{1,h}, \ldots, w_{m,h}$ le autofunzioni associate ai m autovalori discreti che convergono a λ. Allora esiste*

$$u_h \in span\{w_{1,h}, \ldots, w_{m,h}\}$$

tale che

$$\|u - u_h\|_V \leq Ch^{\tau} \|u\|_{H^{1+\tau}(\Omega)}.$$

Proponiamo una dimostrazione del Corollario (2.15) in questo caso particolare.

Teorema 2.25. *Sia λ_h un autovalore convergente a λ. Allora vale il seguente doppio ordine di convergenza:*

$$\lambda \leq \lambda_h \leq \lambda + Ch^{2\tau}.$$

Dimostrazione. Dalla (2.15) e dal teorema (2.14) è chiaro che, poiché T è autoaggiunto, è sufficiente maggiorare il termine

$$\sum_{j,k=1}^{m} \left| \left((T - T_h)\Phi_j, \Phi_k \right)_V \right|,$$

dove $\{\Phi_1, \ldots, \Phi_m\}$ è una base dell'autospazio E.
Abbiamo infatti

$$\begin{aligned}\left|\left((T-T_h)u, v\right)_V\right| &= |a((T-T_h)u, v)| \\ &= \inf_{v_h \in V_h} |a((T-T_h)u, v - v_h)| \\ &\leq \|(T-T_h)u\|_V \inf_{v_h \in V_h} \|v - v_h\|_V \\ &\leq Ch^\tau \|u\|_V \, h^\tau \|v\|_{H^{1+\tau}(\Omega)} \\ &\leq Ch^{2\tau} \|u\|_V \|v\|_V,\end{aligned}$$

che vale per ogni $u, v \in E$, poiché $v = \lambda T v$ implica

$$\|v\|_{H^{1+\tau}(\Omega)} \leq C\|v\|_V.$$

□

2.4.2 Analisi per $T = T_H$

Definiamo l'operatore $T_h : H \to H$: $T_h f \in V_h \subset V \subset H$ tale che

$$a(T_h f, v) = (f, v) \quad \forall v \in V_h.$$

Come nel caso precedente, è immediato osservare che la convergenza in norma (2.8) è soddisfatta per ogni dominio ragionevole. Infatti se Ω è Lipschitz-continuo, abbiamo che esiste $\epsilon > 0$ tale che

$$\|Tf - T_h f\|_{L^2(\Omega)} \leq Ch^{1+\varepsilon} \|f\|_{L^2(\Omega)}.$$

Usando la definizione di τ del caso precedente, possiamo dedurre dal teorema (2.20) la stima di convergenza ottimale per le autofunzioni.

Teorema 2.26. *Sia u un'autofunzione unitaria associata all'autovalore λ di molteplicità m, e siano $w_{1,h}, \ldots, w_{m,h}$ le autofunzioni associate ai m autovalori discreti che convergono a λ. Allora esiste*

$$u_h \in \operatorname{span}\{w_{1,h}, \ldots, w_{m,h}\}$$

tale che

$$\|u - u_h\|_H \ \leq \ Ch^{1+\tau} \|u\|_{H^{1+\tau}(\Omega)}.$$

Mostriamo che la stima di convergenza ottimale doppia del (2.25) possiamo mostrarla anche in questo setting. Come in quel teorema dobbiamo stimare

$$\sum_{j,k=1}^{m} \left|\left((T - T_h)\, \phi_j, \phi_k\right)_H\right|$$

dove $\{\phi_1, \ldots, \phi_m\}$ è una base dell'autospazio E.

$$\begin{aligned}
|((T - T_h)u, v)| &= |(v, (T - T_h)u)| = |a(Tv, (T - T_h)u)| \\
&= |a((T - T_h)u, Tv)| = |a((T - T_h)u, Tv - T_h v)| \\
&\leq \|(T - T_h)u\|_V \, \|(T - T_h)v\|_V \\
&\leq Ch^{2\tau},
\end{aligned}$$

che è valida per ogni $u, v \in E$ con $\|u\|_H = \|v\|_H = 1$.

Capitolo 3

Studio del problema agli autovalori di Laplace in una dimensione

Sia $\Omega = (0, \pi)$. Consideriamo il seguente problema agli autovalori:
Trovare $u : \Omega \to \mathbb{R}$, con $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$, e un numero reale $\lambda \in \mathbb{R}$, tali che:

$$\begin{cases} -u''(x) = \lambda u(x), & \text{per ogni } x \in \Omega, \\ u(0) = u(\pi) = 0. \end{cases} \tag{3.1}$$

Questa è la cosiddetta formulazione forte del problema agli autovalori per l'operatore di Laplace in una dimensione, con condizioni di Dirichlet omogenee.
Questa formulazione richiede che la funzione u sia sufficientemente regolare (almeno due volte derivabile) affinché l'equazione differenziale sia soddisfatta punto per punto.
Questa ipotesi di regolarità può risultare troppo restrittiva infatti in molti casi applicativi, la soluzione cercata non è sufficientemente regolare, oppure i dati del problema non consentono una regolarità così elevata.
Per superare queste difficoltà, si introduce una formulazione debole (o variazionale) che consente di ridurre i requisiti di regolarità, lavorando in spazi funzionali più generali, e allo stesso tempo mantenendo il significato matematico e fisico del problema.
In un secondo momento, si può riformulare il problema in forma mista, introducendo una variabile ausiliaria, utile sia dal punto di vista teorico che numerico. Inoltre questa formulazione permette di riscrivere il problema in modo da evitare di lavorare con la derivata seconda della soluzione, che può essere difficile da gestire.
Tornando all'equazione (3.1) nella prossima proposizione dimostriamo che gli autovalori sono i quadrati dei numeri interi $\lambda_k = 1, 4, 9, 16, ...$ e i corrispondenti autospazi sono generati dalle autofunzioni $u_k(x) = \sin(kx)$ per $k = 1, 2, 3, 4, \ldots$

Proposizione 3.1. *L'autocoppia soluzione della (3.1) è della forma* $(k^2, \sin(kx))_{k \in \mathbb{N}^+}$.

Dimostrazione. Dall'analisi sappiamo che $e^{\alpha x}$ è soluzione di $u''(x) + \lambda u(x) = 0$ se e solo se α è soluzione di $\alpha^2 + \lambda = 0$, detta equazione caratteristica [1].
La soluzione dell'EQC è della forma $\alpha_{1,2} = \frac{\pm\sqrt{-4\lambda}}{2}$. Dividiamo i 3 casi:

1. se $\lambda < 0$ dalla condizione al bordo troviamo la soluzione banale;
2. se $\lambda = 0$, stessa situazione precedente;
3. se $\lambda > 0$ le soluzioni dell'EQC sono $\alpha_{1,2} = \pm i\sqrt{\lambda}$ e lo spazio vettoriale delle soluzioni dell'equazione differenziale è della forma $c_1 \cos(\sqrt{\lambda}x) + c_2 \sin(\sqrt{\lambda}x)$.
   Applicando la condizione al bordo $u(0) = 0$, troviamo $c_1 = 0$.
   Applicando la seconda condizione al bordo, troviamo $c_2 \sin(\sqrt{\lambda\pi}) = 0$, per la soluzione non banale dobbiamo avere $\sqrt{\lambda} = k \in \mathbb{N}^+$ da cui $\lambda = k^2$. Da cui la tesi.

□

3.1 Formulazione debole classica

Una classica approssimazione del problema (3.1) mediante il metodo degli elementi finiti si basa sulla sua formulazione debole classica.
Sia $V = H_0^1(\Omega)$, moltiplichiamo per $v \in V$, e integriamo per parti, segue che ora il problema è trovare $\lambda \in \mathbb{R}$ e $u \in V$ non nulla tale che

$$\int_0^\pi u'(x)v'(x)\mathrm{d}x = \lambda \int_0^\pi u(x)v(x)\mathrm{d}x \quad \forall v \in V, \tag{3.2}$$

L'approssimazione di Galerkin di (3.2) è basato su $V_h \subset V$, sottospazio vettoriale di dimensione finita, e consiste nel trovare gli autovalori discreti $\lambda_h \in \mathbb{R}$ e le autofunzioni $u_h \in V_h$ non nulle tale che

$$\int_0^\pi u_h'(x)v'(x)\mathrm{d}x = \lambda_h \int_0^\pi u_h(x)v(x)\mathrm{d}x \quad \forall v \in V_h. \tag{3.3}$$

Osservazione 3.2. Siamo nel setting della sezione 2.4 in $1D$.

3.1.1 Metodo degli elementi finiti lineari

Il metodo degli elementi finiti è un caso particolare del metodo di Galerkin, in cui il sottospazio V_h è costruito a partire da una partizione del dominio.
Definiamo la griglia

$$\mathcal{T}_h = \{K_i = [x_i, x_{i+1}]\}_{i=0}^{N+1},$$

[1] In modo sintetico: "EQC"-

che costituisce una partizione del dominio $[0,\pi]$, gli intervalli K_i vengono detti elementi. Consideriamo il caso in cui gli i nodi sono equispaziati (figura 3.1):

Figura 3.1: *Discretizzazione uniforme dell'intervallo* $[0,\pi]$ *in* $N+2$ *nodi equispaziati.*

dove $h = \frac{\pi}{n+1}$, viene detta ampiezza della griglia.
Sia $V_h = S_0^1(\mathcal{T}_h) = \left\{ v \in C^0([0,\pi]) \,\middle|\, v|_{[x_i,x_{i+1}]} \in \mathbb{P}_1 \text{ per ogni } i = 0,\dots,N \text{ e } v(0) = v(\pi) = 0 \right\}$ dove $\mathbb{P}_1$ denota lo spazio dei polinomi di grado al più 1.
Come base di $S_0^1(\mathcal{T}_h)$ scegliamo le funzioni $\varphi_i \in S_0^1(\mathcal{T}_h)$ tali che

$$\varphi_i(x_j) = \delta_{ij}, \qquad i,j = 1,\dots,N.$$

La sua espressione è data da:

$$\varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{h} & \text{per } x_{i-1} \le x \le x_i \\ \frac{x_{i+1} - x}{h} & \text{per } x_i \le x \le x_{i+1} \\ 0 & \text{altrimenti.} \end{cases} \tag{3.4}$$

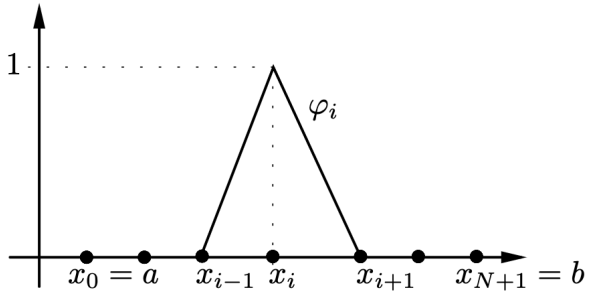


Figura 3.2: *La funzione base di* $S_0^1(\tau_h)$ *relativa al nodo* x_i *dove* $a = 0$,$b = \pi$.

Come si osserva dall'espressione (3.4), le funzioni di base φ_i e φ_{i+1}, definite su ciascun intervallo $[x_i, x_{i+1}]$, hanno la stessa forma su tutti gli elementi, poiché i nodi della griglia sono equispaziati. Nella pratica si possono ottenere le due funzioni di base φ_i e φ_{i+1} trasformando due funzioni di base $\widehat{\varphi}_0$ e $\widehat{\varphi}_1$ costuite una volta per tutte su un intervallo di riferimento, tipicamente l'intervallo [0,1].
A tal fine basta sfruttare il fatto che il generico intervallo $[x_i$,$x_i + h]$ della decomposizione

di $(0, \pi)$ può essere ottenuto a partire dall'intervallo [0,1], tramite la trasformazione lineare $\phi : [0, 1] \to [x_i, x_i + h]$ definita come

$$x = \phi(\xi) = x_i + \xi h$$

Se definiamo le due funzioni di base $\widehat{\varphi}_0$ e $\widehat{\varphi}_1$ su $[0, 1]$ come

$$\widehat{\varphi}_0(\xi) = 1 - \xi, \quad \widehat{\varphi}_1(\xi) = \xi, \tag{3.5}$$

le funzioni di base φ_i e φ_{i+1} su $[x_i, x_i + h]$ saranno semplicemente date da

$$\varphi_i(x) = \widehat{\varphi}_0(\phi^{-1}(x)), \quad \varphi_{i+1}(x) = \widehat{\varphi}_1(\phi^{-1}(x))$$

essendo $\phi^{-1}(x) = \frac{(x - x_i)}{h}$.

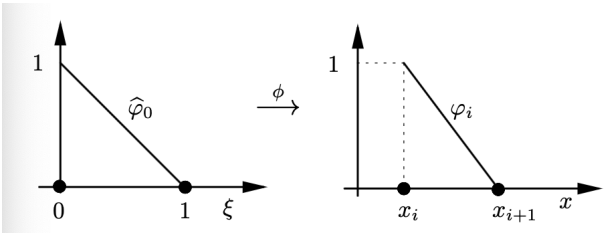


Figura 3.3: *La funzione di base φ_i in $[x_i, x_i + h]$ e la corrispondente funzione di base $\widehat{\varphi}_0$ sull'elemento di riferimento.*

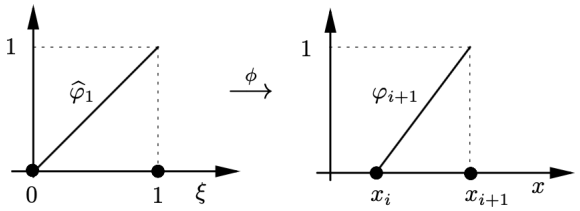


Figura 3.4: *La funzione di base φ_{i+1} in $[x_i, x_i + h]$ e la corrispondente funzione di base $\widehat{\varphi}_1$ sull'elemento di riferimento.*

Torniamo alla (3.3), poichè $u_h \in S_0^1(\mathcal{T}_h)$, allora si può scrivere come combinazione lineare degli elementi della base: $u_h = \sum_{j=1}^{N} u_j \, \varphi_j$.
Si dimostra facilmente che l'uguaglianza (3.3) vale per ogni $v \in V_h$ se e solo se vale per ogni elemento della base, quindi pongo $v = \varphi_i$.
Sostituendo (3.3) diventa :

$$\sum_{j=1}^{N} u_j \int_0^{\pi} \varphi_j'(x) \, \varphi_i'(x) \, dx = \lambda_h \sum_{j=1}^{N} u_j \int_0^{\pi} \varphi_j(x) \, \varphi_i(x) \, dx$$

Ponendo $a_{ij} = \int_0^\pi \varphi_j'(x)\varphi_i'(x)\mathrm{d}x$ e $m_{ij} = \int_0^\pi \varphi_j(x)\varphi_i(x)\mathrm{d}x$ dove $A = \{a_{ij}\}_{i,j=1}^N$ è detta matrice di rigidezza e $M = \{m_{ij}\}_{i,j=1}^N$ è detta matrice di massa.
Ponendo $x = (u_1, u_2, \ldots, u_N)^T$ trovo il seguente problema algebrico:

$$Ax = \lambda Mx. \tag{3.6}$$

Con la partizione da noi scelta :

$$a_{ij} = \tfrac{1}{h} \cdot \begin{cases} 2 & \text{per } i = j, \\ -1 & \text{per } |i-j| = 1, \\ 0 & \text{altrimenti} \end{cases} \qquad m_{ij} = h \cdot \begin{cases} 4/6 & \text{per } i = j, \\ 1/6 & \text{per } |i-j| = 1, \\ 0 & \text{altrimenti} \end{cases}$$

Sia $I_h : C^0([0,\pi]) \to S_0^1(\mathcal{T}_h)$ l'operatore di interpolazione nodale definito da

$$(I_h u)(x_i) = u(x_i), \qquad i = 1, \ldots, N.$$

Nel nostro caso, poichè $u_k(x) = \sin(kx)$ è l'autofunzione esatta del problema continuo, allora la corrispondente autofunzione discreta è data dal suo interpolante nodale:

$$u_{k,h}(x) = I_h u_k(x) = \sum_{i=1}^N u_k(x_i)\,\varphi_i(x) = \sum_{i=1}^N \sin(kx_i)\,\varphi_i(x).$$

Ora calcoliamo il corrispondente autovalore discreto: volendo calcolare il k-esimo autovalore, sappiamo che $(x)_i = sin(kih)$.
Dall'equazione matriciale (3.6) sappiamo che $(Ax)_i = \frac{1}{h}\left(2x_i - x_{i-1} - x_{i+1}\right)$ e $(Mx)_i = h\left(4x_i + x_{i-1} + x_{i+1}\right)$ per $i = 2, \ldots, N-1$.
Calcolo le combinazioni trigonometriche: $x_{i+1} + x_{i-1} = 2\sin(ikh)\cos(kh)$, quindi $2x_i - x_{i-1} - x_{i+1} = 2(1-\cos(kh))\sin(ikh)$.
Quindi $(Ax)_i = 2(1-\cos(kh))\sin(ikh)/h$ e $(Mx)_i = h(2+\cos(kh))\sin(ikh)/3$.
Da cui uguagliando per ogni i, si trova che

$$\lambda_{k,h} = \left(6/h^2\right)\frac{1-\cos kh}{2+\cos kh} \tag{3.7}$$

È immediato dedurre le stime ottimali (per $h \to 0$)

$$\|u_k - u_{k,h}\|_V = O(h) \quad |\lambda_k - \lambda_{k,h}| = O\left(h^2\right) \tag{3.8}$$

con $u_k(x) = \sin(kx)$ e $\lambda_k = k^2$.

Osservazione 3.3. Poichè nel nostro caso gli autospazi $E \subset H^r(\Omega)$ per r arbitrariamente grande, allora τ definito nel capitolo 2 è uguale al grado p dei polinomi a tratti che definiscono le funzioni continue di V_h.
Abbiamo visto che valgono le seguenti stime:

$$\|u_k - u_{k,h}\|_V \overset{2.24}{=} O\left(h^p\right), \ \|u_k - u_{k,h}\|_{L^2(\Omega)} \overset{2.26}{=} O\left(h^{p+1}\right), \ |\lambda_k - \lambda_{k,h}| \overset{2.25}{=} O\left(h^{2p}\right). \tag{3.9}$$

Osservazione 3.4. Anche se non è esplicito la stima (3.8) dipende da k.
In particolare, la stima degli autovalori può essere più precisa osservando che

$$\lambda_{k,h} = k^2 + \left(k^4/12\right) h^2 + O\left(k^6 h^4\right), \quad \text{per } h \to 0.$$

Questa proprietà ha un chiaro significato fisico: poiché le autofunzioni presentano un numero crescente di oscillazioni all'aumentare della frequenza, è necessaria una mesh sempre più fine per mantenere l'errore di approssimazione entro la stessa accuratezza.

Osservazione 3.5. Un'importante conseguenza di (3.7) è che gli autovalori sono approssimati dall'alto (vedi 2.19).
Questo comportamento, che è legato alla cosiddetta proprietà del minimo-massimo, può essere enunciato nel modo seguente:

$$\lambda_k \leq \lambda_{k,h} \leq \lambda_k + C(k)h^2.$$

Esperimenti numerici

k	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$
λ_1	1.010194634143	1.002849135094	1.000755477752	1.000194681908	1.000049425112	1.000012452439
λ_2	4.164959042832	4.045738903270	4.012098545171	4.003115637252	4.000790848682	4.000199241998
λ_3	9.848419050035	9.232802215904	9.061340168458	9.015779033651	9.004004066868	9.001008687728
λ_4	18.720276613098	16.741006720586	16.194264392504	16.049896539765	16.012656576948	16.003188062445
λ_5	31.643900506006	26.824161285023	25.475510741321	25.121902149993	25.030905328668	25.007783704252

Tabella 3.1: Autovalori approssimati $\lambda_{k,h}$ calcolati con FEM lineare

Questa tabella rappresenta i primi 5 autovalori discreti, che come dalla teoria convergono a k^2 al crescere di n che è il numero di nodi interni della partizione $\mathcal{T}_h$.
Poichè l'errore sugli autovalori è di ordine $O(h^2)$, ogni dimezzamento del passo h(ovvero ogni raddoppio di n) comporta una riduzione dell'errore di circa un fattore di 4, il che per gli autovalori λ_1 e λ_2 equivale a guadagnare approssimativamente una cifra decimale significativa.
Da (3.8) l'errore degli autovalori va come Ch^2, infatti il prossimo grafico ce lo conferma e come evidenziato da (3.4) C dipende da k e più è grande k più la costante C è grande come si può vedere da (3.5) quindi la convergenza è più lenta per autovalori associati a k più alti.
Come mostrato nella tabella 3.2, possiamo osservare che l'ordine di convergenza degli autovalori approssimati $\lambda_{h,k}$ è prossimo al valore teorico 2 per tutti i valori di k considerati.

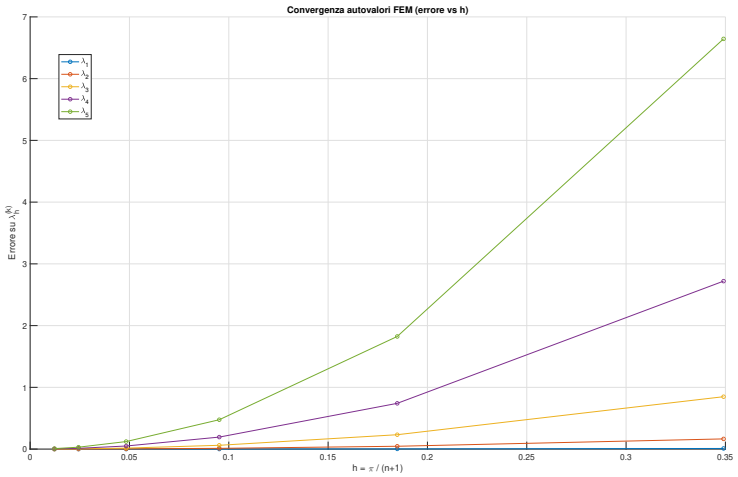


Figura 3.5: *Convergenza degli autovalori: errore vs h*

k	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$
λ_1	2.004510	2.001254	2.000330	2.000085	2.000021
λ_4	2.016935	2.004938	2.001316	2.000339	2.000086
λ_9	2.033346	2.010803	2.002941	2.000760	2.000193
λ_{16}	2.044815	2.018395	2.005176	2.001349	2.000343
λ_{25}	2.032393	2.026983	2.007983	2.002101	2.000535

Tabella 3.2: Ordine di convergenza degli autovalori

Dall'osservazione 3.8 la norma energia dell'errore va come Ch con C costante che dipende da k e lo conferma la figura 3.6.
Infatti come si può osservare dalla tabella 3.3 l'ordine è 1.

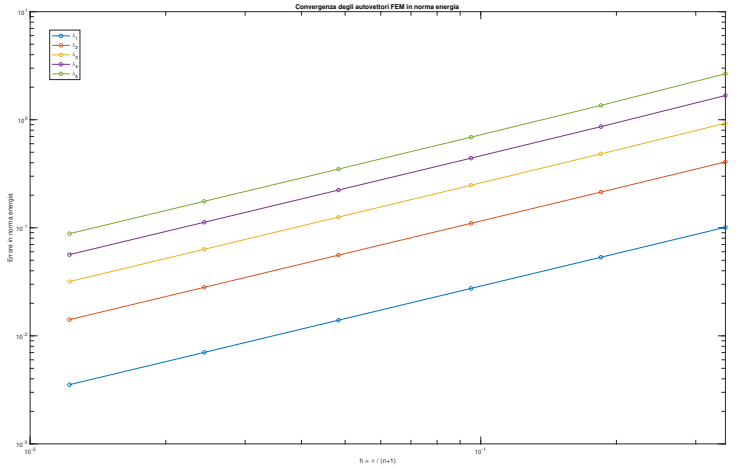


Figura 3.6: *Convergenza degli autovettori: norma energia dell'errore vs h*

k	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$
λ_1	1.002867	1.000788	1.000207	1.000053	1.000013
λ_2	1.011372	1.003145	1.000826	1.000212	1.000054
λ_3	1.025049	1.007050	1.001858	1.000477	1.000121
λ_4	1.042412	1.012459	1.003300	1.000847	1.000215
λ_5	1.059526	1.019277	1.005147	1.001323	1.000335

Tabella 3.3: Ordine di convergenza autovettori in norma energia

Nel caso della norma L^2, gli autovettori approssimati con elementi lineari convergono all'autofunzione esatta con ordine $O(h^2)$. Questo è coerente con la teoria che vedremo. Tale comportamento è confermato anche numericamente, come mostrato nella tabella 3.4.

k	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$
λ_1	2.029570	2.022142	2.017202	2.013751	2.011245
λ_2	2.118777	2.088932	2.069063	2.055181	2.045103
λ_3	2.264712	2.199799	2.155612	2.124462	2.101760
λ_4	2.440806	2.346129	2.273744	2.220418	2.180780
λ_5	2.543597	2.491947	2.409632	2.337204	2.279509

Tabella 3.4: Ordine di convergenza autovettori in norma L^2

Come è possibile osservare dal seguente grafico la funzione FEM oscilla in modo spuriamente amplificato rispetto alla soluzione esatta, infatti per autovalori elevati (cioè k vicino a n), ossia per le alte frequenze, la funzione approssimata ha componenti spurie che non fanno parte dell'autospazio vero.
Questo è causato dal fatto che gli spazi FEM lineari non riescono a rappresentare in modo accurato le alte frequenze, come mostrato nell'osservazione 3.4.
Infatti come possiamo notare dalla tabella 3.5 l'errore per i 5 autovalori discreti più grandi (per $n = 128$) è dell'ordine di 10^3.

k	$\lambda_{k,h}$	λ_k	Errore
128	20224.032917	16384	3.84e+03
127	20197.079555	16129	4.07e+03
126	20152.281242	15876	4.28e+03
125	20089.822582	15625	4.46e+03
124	20009.959394	15376	4.63e+03

Tabella 3.5: *Errore autovalori*

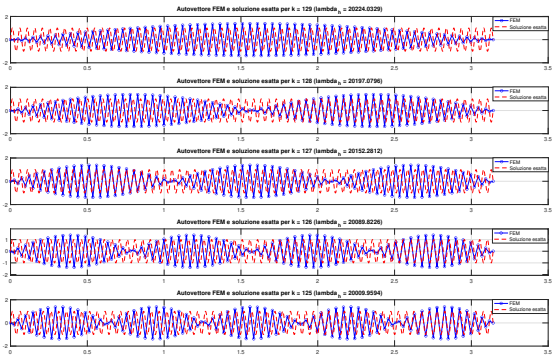


Figura 3.7: *Autofunzioni FEM vs esatte per* $k = 124, \ldots, 128$

Il prossimo grafico ci fa vedere il buon comportamento della FEM lineare per i primi 5 autovalori per $n = 10$.
Come dalla teoria per gli autovalori λ_3, λ_4, λ_5 si comporta peggio.

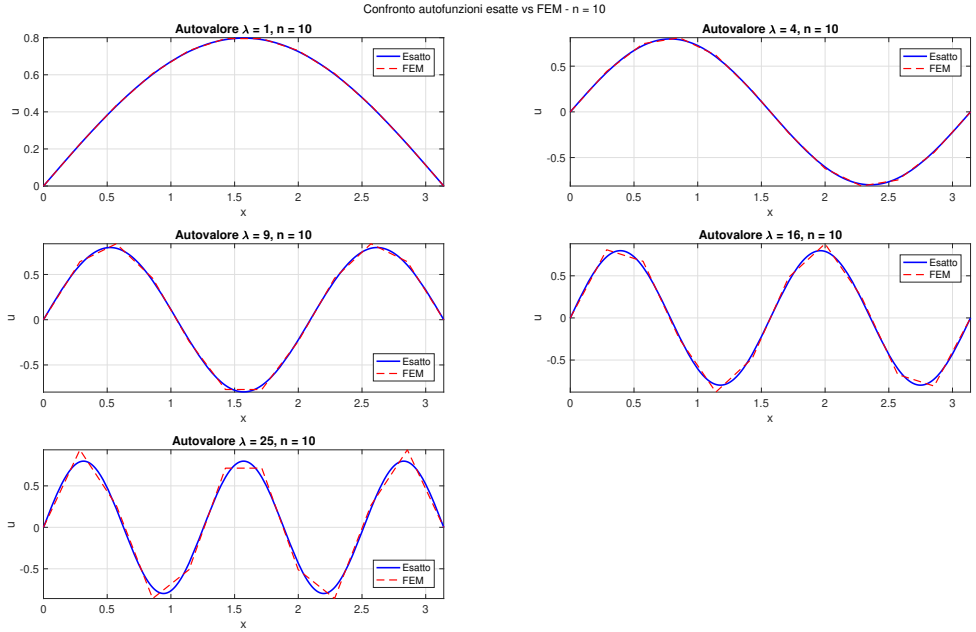


Figura 3.8: *Autofunzioni esatte vs FEM*

3.1.2 Metodo degli elementi finiti quadratici

Sia $V_h = S_0^2(\mathcal{T}_h) = \left\{ v \in C^0([0,\pi]) \,\middle|\, v|_{[x_i,x_{i+1}]} \in \mathbb{P}_2 \text{ per ogni } i = 0, \ldots, N \text{ e } v(0) = v(\pi) = 0 \right\}$ dove $\mathbb{P}_2$ denota lo spazio dei polinomi di grado al più 2.

Le funzioni di $S_0^2(\mathcal{T}_h)$ sono polinomi compositi, di grado 2 su ciascun intervallo di $\mathcal{T}_h$, il terzo sarà il punto medio dello stesso (ad indice dispari).

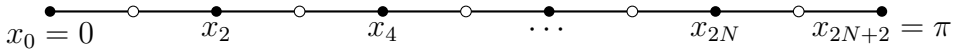


Figura 3.9: *Partizione del dominio* $[0, \pi]$*, con nodi* x_i *e punti medi interni.*

Come nel caso degli elementi finiti lineari, trattiamo il caso in cui i nodi sono equispaziati con passo h.
Come base di $S_0^2(\mathcal{T}_h)$, scegliamo le funzioni $\varphi_i \in S_0^2(\mathcal{T}_h)$ tale che $\varphi_i(x_j) = \delta_{ij}, i, j = 1, \ldots, 2N+1$.
Sono quindi funzioni quadratiche a tratti che assumono valore 1 nel nodo a cui sono associate e sono nulle nei restanti nodi:

$$(i \text{ pari}) \quad \varphi_i(x) = \begin{cases} \frac{(x-x_{i-1})(x-x_{i-2})}{2h^2} & \text{se } x_{i-2} \le x \le x_i \\ \frac{(x_{i+1}-x)(x_{i+2}-x)}{2h^2} & \text{se } x_i \le x \le x_{i+2} \\ 0 & \text{altrimenti.} \end{cases}$$

$$(i \text{ dispari}) \quad \varphi_i(x) = \begin{cases} \frac{(x_{i+1}-x)(x-x_{i-1})}{h^2} & \text{se } x_{i-1} \le x \le x_{i+1} \\ 0 & \text{altrimenti.} \end{cases}$$

Come nel caso degli elementi finiti lineari, per descrivere la base è sufficiente fornire l'espressione delle funzioni di base sull'intervallo di riferimento $[0, 1]$ e poi trasformare queste ultime tramite la (3.5). Abbiamo:

$$\widehat{\varphi}_0(\xi) = (1-\xi)(1-2\xi), \quad \widehat{\varphi}_1(\xi) = 4(1-\xi)\xi, \quad \widehat{\varphi}_2(\xi) = \xi(2\xi-1)$$

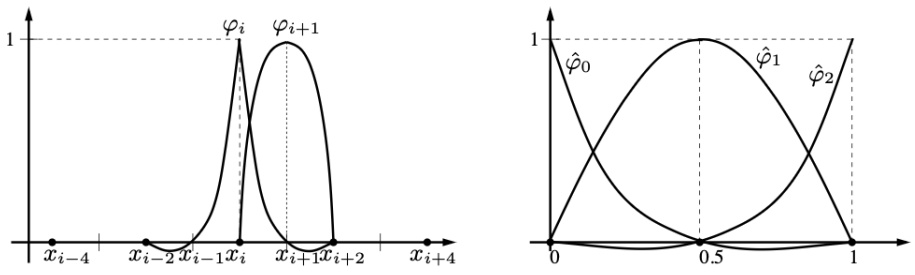


Figura 3.10: *Funzioni di base di* $S_0^2(\mathcal{T}_h)$*(a sinistra) e loro corrispondenti sull'intervallo di riferimento(a destra)*

Torniamo alla (3.3), poichè $u_h \in S_0^2(\mathcal{T}_h)$, allora si può scrivere come combinazione lineare degli elementi della base: $u_h = \sum_{j=1}^{2N+1} u_j \, \varphi_j$, pongo $v = \varphi_i$.

Sostituendo a (3.3) diventa :

$$\sum_{j=1}^{2N+1} u_j \int_0^{\pi} \varphi_j'(x)\,\varphi_i'(x)\,dx = \lambda_h \sum_{j=1}^{2N+1} u_j \int_0^{\pi} \varphi_j(x)\,\varphi_i(x)\,dx$$

Ponendo $a_{ij} = \int_0^{\pi} \varphi_j'(x)\varphi_i'(x)\mathrm{d}x$ e $m_{ij} = \int_0^{\pi} \varphi_j(x)$.
Per costruire le matrici globali del sistema lineare agli elementi finiti, si ricorre all'uso delle cosiddette matrici locali, definite su ciascun elemento della discretizzazione. Sia $K = [x_i, x_i + h]$ un generico elemento dell'intervallo $[0, \pi]$. In corrispondenza di ciascun elemento si considerano tre funzioni di base locali ottenute prima. Le matrici locali di massa e rigidezza si ottengono trasformando gli integrali sul riferimento $\hat{K} = [0, 1]$ tramite la mappa affine $x(\xi) = x_i + h\xi$.
Si ha:

$$M_{ij}^{(K)} = \int_K \varphi_i(x)\varphi_j(x)\,dx = h\int_0^1 \hat{\varphi}_i(\xi)\hat{\varphi}_j(\xi)\,d\xi,\ i,j = 1,2,3$$

$$A_{ij}^{(K)} = \int_K \varphi_i'(x)\varphi_j'(x)\,dx = \frac{1}{h}\int_0^1 \hat{\varphi}_i'(\xi)\hat{\varphi}_j'(\xi)\,d\xi,\ i,j = 1,2,3$$

dove $\varphi_i(x) = \hat{\varphi}_i\left(\frac{x-x_i}{h}\right)$ e si è utilizzata la regola del cambiamento di variabile per l'integrale. L'assemblaggio globale delle matrici A e M si effettua sommando i contributi locali nei corrispondenti indici globali:

$$A_{r,s} = \sum_K A_{i,j}^{(K)}, \qquad M_{r,s} = \sum_K M_{i,j}^{(K)},$$

dove $r = \mathrm{glob}(i, K)$, $s = \mathrm{glob}(j, K)$ sono gli indici globali associati ai gradi di libertà locali i, j dell'elemento K.

Esperimenti numerici

k	$n = 8$	$n = 16$	$n = 32$	$n = 64$	$n = 128$	$n = 256$
λ_1	1.000032766086	1.000002060211	1.000000128958	1.000000008063	1.000000000505	1.000000000036
λ_2	4.002048562170	4.000131064342	4.000008240843	4.000000515832	4.000000032253	4.000000002020
λ_3	9.022486886740	9.001478254003	9.000093633109	9.000005871943	9.000000367310	9.000000022966
λ_4	16.120357238038	16.008194248679	16.000524257370	16.000032963372	16.000002063327	16.000000129011
λ_5	25.432690817153	25.030733808184	25.001990984224	25.000125603200	25.000007868734	25.000000492091

Tabella 3.6: Autovalori approssimati $\lambda_{k,h}$ calcolati con FEM quadratica

La tabella mostra i primi cinque autovalori discreti $\lambda_{k,h}$, ottenuti mediante il metodo degli elementi finiti utilizzando funzioni quadratiche a tratti.

Come atteso dalla teoria spettrale, i valori approssimati convergono rapidamente ai corrispondenti autovalori esatti $\lambda_k = k^2$, al crescere del numero di elementi.
Ci si aspetta un ordine di convergenza pari a 4, come confermato da (3.7).
Ogni raddoppio del numero di nodi interni (ossia dimezzamento di h) porta a una riduzione dell'errore di circa 16 volte. Quindi poichè $\log(16) \approx 1.2$, quindi per ogni raffinamento si guadagnano in media più di una cifra significativa. Guardando la tabella λ_2 per $n = 64$ ha già 7 cifre significative corrette e bisezionando 2 volte h si guadagnano altre 3 cifre significative.

Come evidenziato dalle curve superiori (verde e viola) sono più in alto perché gli autovalori associati agli indici più alti convergono più lentamente, come previsto dalla teoria.
Come mostrato nella tabella 3.7, possiamo osservare che l'ordine di convergenza degli autovalori approssimati $\lambda_{k,h}$ è prossimo al valore teorico 4 per tutti i valori di k considerati.

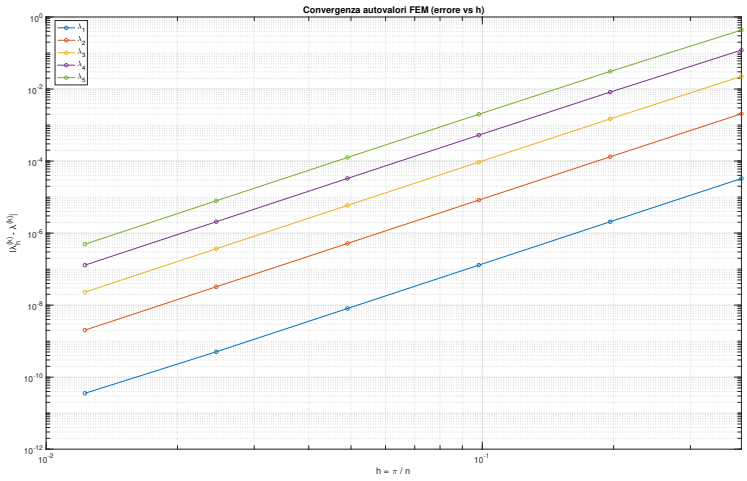


Figura 3.11: *Errore autovalori vs h*

k	$n = 8 \to 16$	$n = 16 \to 32$	$n = 32 \to 64$	$n = 64 \to 128$	$n = 128 \to 256$
λ_1	3.991339	3.997819	3.999412	3.997099	3.828216
λ_2	3.966265	3.991339	3.997819	3.999414	3.997106
λ_3	3.927118	3.980732	3.995109	3.998769	3.999443
λ_4	3.876567	3.966265	3.991339	3.997819	3.999412
λ_5	3.815438	3.948273	3.986537	3.996598	3.999136

Tabella 3.7: Ordini di convergenza degli autovalori

Dall'osservazione 3.9 l'errore in norma energia è dell'ordine di Ch^2 dove C dipende da k, infatti il seguente grafico lo conferma.
La tabella 3.8 ci conferma che l'ordine di convergenza degli autovettori in norma energia è 2.

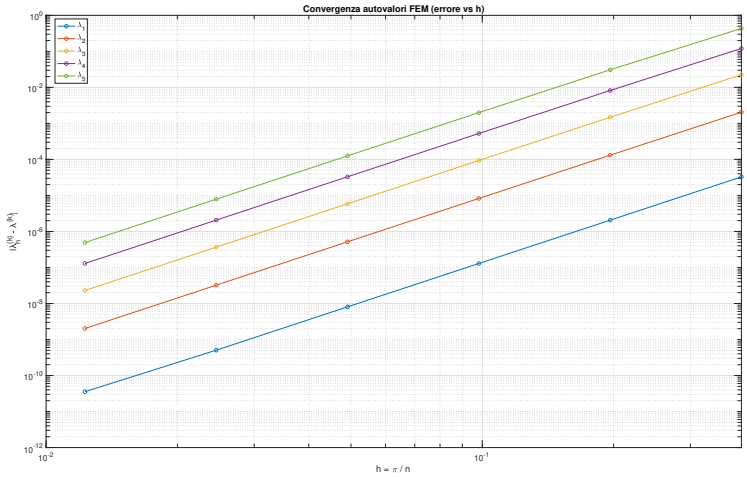


Figura 3.12: *Errore autovettori in norma energia vs h*

k	$n = 8 \to 16$	$n = 16 \to 32$	$n = 32 \to 64$	$n = 64 \to 128$	$n = 128 \to 256$
λ_1	1.997936	1.999480	1.999870	1.999968	1.999992
λ_2	1.992011	1.997936	1.999480	1.999870	1.999967
λ_3	1.983062	1.995418	1.998834	1.999707	1.999927
λ_4	1.972669	1.992011	1.997936	1.999480	1.999870
λ_5	1.963929	1.987838	1.996795	1.999189	1.999797

Tabella 3.8: Ordine di convergenza degli autovettori

Come abbiamo visto nel caso lineare l'errore in norma L^2 guadagna un ordine come conferma la tabella 3.9.

k	$n = 8 \to 16$	$n = 16 \to 32$	$n = 32 \to 64$	$n = 64 \to 128$	$n = 128 \to 256$
λ_1	2.990701	2.997636	2.999407	2.999851	2.999963
λ_2	2.965375	2.990701	2.997636	2.999407	2.999851
λ_3	2.932391	2.979672	2.994718	2.998667	2.999666
λ_4	2.908836	2.965375	2.990701	2.997636	2.999407
λ_5	2.929622	2.949037	2.985655	2.996318	2.999073

Tabella 3.9: Ordine di convergenza degli autovettori in norma L^2

La figura 3.13 mostra un confronto grafico tra le autofunzioni esatte e quelle approssimate mediante il metodo degli elementi finiti quadratici, nel caso in cui $n = 10$.
Osserviamo :

- Per $k = 1$, $k = 2$, le curve sono quasi indistinguibili: l'approssimazione numerica coincide visivamente con quella esatta.

- Per $k = 3, 4, 5$, l'errore aumenta leggermente,come atteso, a causa della maggiore oscillazione delle autofunzioni, ma l'errore resta comunque molto contenuto: i massimi e i minimi coincidono quasi perfettamente.

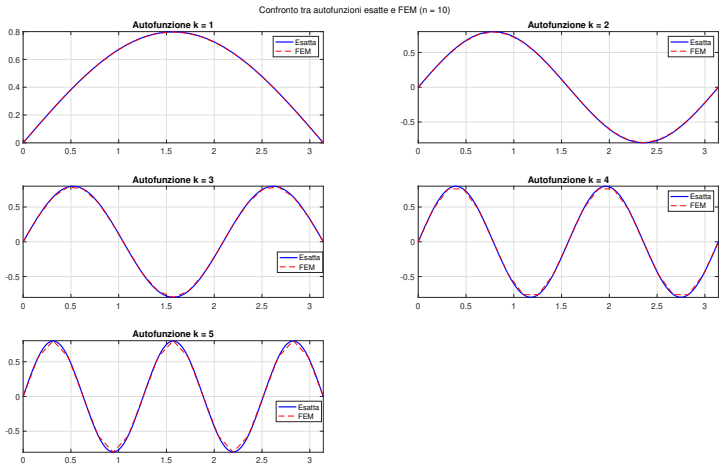


Figura 3.13: *Autovettori esatti vs autovettori FEM per n=10*

Contrariamente a quanto ci si potrebbe aspettare, nel caso della FEM quadratica , per autovalori elevati (cioè k vicino a n), le autofunzioni approssimate mostrano comportamenti spuri anche più marcati rispetto al caso lineare. Come si osserva dalla figura, le soluzioni discrete presentano oscillazioni artificiali più ampie, rispetto alle corrispondenti soluzioni esatte.
Questo fenomeno è riconducibile al fatto che, per k elevati, anche lo spazio $S_0^2(\mathcal{T}_h)$ diventa insufficientemente ricco per rappresentare in modo fedele frequenze così alte. Anzi, poiché i gradi di libertà aumentano, lo spazio discreto può "ospitare" componenti spurie a frequenza più elevata , e quindi tende a introdurre più errori oscillatori locali.
Anche la tabella 3.10 conferma che l'errore sugli autovalori più alti è significativo (dell'ordine di 10^5), e non migliora sostanzialmente rispetto al caso lineare, anzi peggiora.

k	$\lambda_{k,h}$	λ_k	Errore
128	141054.588577	16384	1.25e+05
127	140906.720561	16129	1.25e+05
126	140660.825258	15876	1.25e+05
125	140317.725100	15625	1.25e+05
124	139878.562255	15376	1.25e+05

Tabella 3.10: *Errore sugli autovalori*

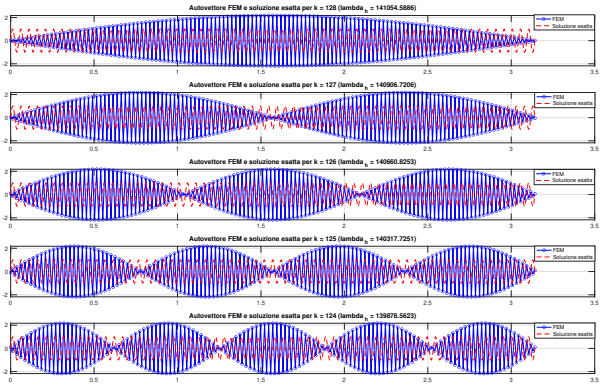


Figura 3.14: *Autovettori FEM quadratici vs esatti per* $k = 124, \dots, 128$

Prima di introdurre la formulazione debole mista, osserviamo che la sua trattazione teorica si basa su risultati più generali, che non verranno discussi in questa sede.
In questa sezione ci limiteremo a presentare la struttura variazionale del problema e le relative approssimazioni numeriche, mentre per una trattazione completa della teoria si rimanda al capitolo dedicato nella review di riferimento [5]

3.2 Formulazione debole mista

Riscriviamo l'equazione di Laplace (3.1) come sistema del primo ordine: dato $\Omega =]0, \pi[$, trovare gli autovalori λ e le autofunzioni u con $u \neq 0$, tale che, per qualche s,

$$\begin{cases} s(x) - u'(x) = 0 & \text{in } \Omega, \\ s'(x) = -\lambda u(x) & \text{in } \Omega, \\ u(0) = u(\pi) = 0. \end{cases} \tag{3.10}$$

Questa è la formulazione forte mista del problema agli autovalori di Laplace in 1D.

Osservazione 3.6. Nel problema (3.10) sono coinvolte due funzioni: s e u.
Nella formulazione del problema, abbiamo esplicitamente indicato che le autofunzioni di nostro interesse sono quelle rappresentate da u.
Questa osservazione è utile perchè nel caso continuo sono nel rapporto di uno a uno, invece nel discreto questo non potrebbe più valere.
In particolare, definiamo la molteplicità geometrica di λ come il numero di autofunzioni u relative a λ linearmente indipendenti; in generale, potrebbe accadere che ci siano più s associati allo stesso u, e non vogliamo tener conto della molteplicità di s quando valutiamo la molteplicità di λ.

Dati $\Sigma = H^1(\Omega)$ and $U = L^2(\Omega)$, una forma variazionale del problema misto (3.10) è trovare $\lambda \in \mathbb{R}$ e $u \in U$ con $u \neq 0$, tale che, per qualche $s \in \Sigma$,

$$\int_0^\pi s(x)t(x)\mathrm{d}x + \int_0^\pi u(x)t'(x)\mathrm{d}x = 0 \quad \forall t \in \Sigma$$
$$\int_0^\pi s'(x)v(x) = -\lambda \int_0^\pi u(x)v(x)\mathrm{d}x \quad \forall v \in U$$

La sua discretizzazione di Galerkin è basata sugli spazi discreti $\Sigma_h \subset \Sigma$ e $U_h \subset U$ ed è la seguente : trovare $\lambda_h \in \mathbb{R}$ e $u_h \in U_h$ e $u_h \neq 0$, tale che, per qualche $s_h \in \Sigma_h$,

$$\int_0^\pi s_h(x)t(x)\,dx + \int_0^\pi u_h(x)t'(x)\,dx = 0 \quad \forall t \in \Sigma_h, \tag{3.11a}$$
$$\int_0^\pi s_h'(x)v(x)\,dx = -\lambda_h \int_0^\pi u_h(x)v(x)\,dx \quad \forall v \in U_h. \tag{3.11b}$$

Se $\Sigma_h = Span\{\varphi_1, \ldots, \varphi_{N_s}\}$ e $U_h = Span\{\psi_1, \ldots, \psi_{N_u}\}$, sia $s_h(x) = \sum_{k=1}^{N_s} x_k\,\varphi_k(x)$ e $u_h(x) = \sum_{j=1}^{N_u} y_j\,\psi_j(x)$.
Ci concentriamo sulla prima equazione di (3.11) e sostituisco s_h e u_h nell'equazione e uso come funzione test $t(x) = \varphi_l$ per $l = 1, \ldots, N_s$. L'equazione diventa

$$\sum_{k=1}^{N_s} x_k \int_0^\pi \varphi_k(x)\,\varphi_l(x)\,dx + \sum_{j=1}^{N_u} y_j \int_0^\pi \psi_j(x)\,\varphi_l'(x)\,dx = 0.$$

Ponendo $a_{lk} = \int_0^\pi \varphi_k(x)\varphi_l(x)\mathrm{d}x$ e $b_{jl} = \int_0^\pi \varphi_l'(x)\psi_j(x)\mathrm{d}x$, la prima equazione è equivalente a $Ax + B^\top y = 0$ dove $x = [x_1,\, x_2,\, \ldots,\, x_{N_s}]^\top$ e $y = [y_1,\, y_2,\, \ldots,\, y_{N_u}]^\top$.
Passando alla seconda equazione di (3.11) e sostituendo s_h e u_h nell'equazione e uso come funzione test $t(x) = \psi_i$ per $l = 1, \ldots, N_u$. L'equazione diventa

$$\sum_{k=1}^{N_s} x_k \int_0^\pi \varphi_k'(x)\,\psi_i(x)\,dx = -\lambda \sum_{j=1}^{N_u} y_j \int_0^\pi \psi_j(x)\,\psi_i(x)\,dx.$$

Ponendo $m_{ij} = \int_0^\pi \psi_j(x)\psi_i(x)\mathrm{d}x$, la seconda equazione è equivalente a $Bx = -\lambda My$. Mettendo entrambe le equazioni scritte in forma matriciale insieme l'equazione (3.11) è equivalente a

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \mathrm{x} \\ \mathrm{y} \end{pmatrix} = -\lambda \begin{pmatrix} 0 & 0 \\ 0 & M \end{pmatrix} \begin{pmatrix} \mathrm{x} \\ \mathrm{y} \end{pmatrix}.$$

3.2.1 Schema $P1 - P0$

Definiamo la griglia

$$\mathcal{T}_h = \{K_i = [x_i, x_{i+1}]\}_{i=0}^{N+1},$$

che costituisce una partizione del dominio $[0, \pi]$.
Come nel caso della formulazione debole classica, consideriamo il caso dei nodi equispaziati (figura 3.15).

Figura 3.15: *Griglia uniforme sull'intervallo* $[0, \pi]$ *costituita da* N *intervalli equispaziati*

Introduciamo lo schema di ordine minore per la risoluzione del problema.
Pongo $\Sigma_h = S^1(\mathcal{T}_h)$ e $U_h = S^0(\mathcal{T}_h)$, se N è il numero di sottointervalli, allora $N_s = N+1$ e $N_u = N$.
Come abbiamo osservato nella sezione precedente abbiamo che le soluzioni esatte sono $\lambda_k = k^2$ e $u_k(x) = \sin(kx)(k = 1, 2, \ldots)$, osserviamo che abbiamo $s_k(x) = k\cos(kx)$. La soluzione approssimata di s è la funzione continua lineari a tratti che nei nodi è $s_{k,h}(ih) = k\cos(kih)$, e gli autovalori e gli autovettori discreti sono dati da

$$\lambda_{k,h} = \left(6/h^2\right)\frac{1-\cos kh}{2+\cos kh}, \quad u_{h,k}|_{]ih,(i+1)h[} = \frac{s_{k,h}(ih) - s_{k,h}((i+1)h)}{h\lambda_{k,h}},$$

con $k = 1, \ldots, N$.
È piuttosto sorprendente che gli autovalori discreti siano esattamente gli stessi della sezione precedente. C'è in realtà una leggera differenza nel numero di gradi di libertà: qui N è il numero di intervalli, mentre nella sezione precedente N rappresentava il numero di nodi interni, cioè qui si calcola un valore in più con lo schema misto sulla stessa griglia. D'altra parte, le autofunzioni sono diverse, come è naturale che sia, poiché qui sono costanti a tratti, mentre lì erano funzioni continue lineari a tratti.
Si può dimostrare che se consideriamo la soluzione esatta $u_k(x) = \sin(kx)$, allora si ha

$$\int_{ih}^{(i+1)h} (u_k(x) - u_{k,h}(x))\,dx = \frac{\lambda_h - \lambda}{\lambda_h}\int_{ih}^{(i+1)h} u_k(x)\,dx.$$

In particolare, risulta che $u_{h,k}$ non è la proiezione L^2 di u_k nello spazio delle funzioni costanti a tratti.
Come base di Σ_h prendo $\{\varphi_1, \ldots, \varphi_{N+1}\}$ dove

$$\varphi_1(x) = \begin{cases} \frac{x_2 - x}{h} & \text{se } x \in K_2 \\ 0 & \text{altrove} \end{cases} \quad \varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{h} & \text{se } x \in K_i \\ \frac{x_{i+1} - x}{h} & \text{se } x \in K_{i+1} \\ 0 & \text{altrimenti} \end{cases} \quad \varphi_{N+1}(x) = \begin{cases} \frac{x - x_N}{h} & \text{se } x \in K_{N+1} \\ 0 & \text{altrimenti} \end{cases}$$

Come base di U_h prendo $\{\psi_1, \dots, \psi_N\}$ dove per $i = 1, \dots, N$ vale

$$\psi_i(x) = \begin{cases} 1 & \text{se } x \in K_{i+1} \\ 0 & \text{altrimenti} \end{cases}$$

Passando alla rappresentazione matriciale:

- $A \in \mathbb{R}^{(N+1)\times(N+1)}$ si assembla dalla matrice locale che è della forma

$$A_{\text{loc}} = \frac{h}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

- $$B = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{bmatrix} \in \mathbb{R}^{N\times(N+1)}$$

- $$M = \begin{bmatrix} h & 0 & 0 & \cdots & 0 \\ 0 & h & 0 & \cdots & 0 \\ 0 & 0 & h & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & h \end{bmatrix} \in \mathbb{R}^{N\times N}$$

Errori teorici noti:

1. Errore sulla variabile di flusso $s_{h,k}$:

$$\|s_k - s_{k,h}\|_{L^2(\Omega)} \;\leq\; Ch\,\|s_k\|_{H^1(\Omega)} \tag{3.12}$$

   cioè di ordine 1 in L^2.
   In 1D e su griglie uniformi si osserva spesso superconvergenza $O(h^2)$ che è un fenomeno numerico dovuto alla simmetria e alla regolarità delle soluzioni trigonometriche.

2. Errore sulla variabile primaria $u_{h,k}$:

$$\|u_k - u_{k,h}\|_{L^2(\Omega)} \;\leq\; Ch\,\|u_k\|_{H^1(\Omega)} \tag{3.13}$$

   cioè di ordine 1 in L^2.

3. Errore sugli autovalori $\lambda_{k,h}$:

$$|\lambda_k - \lambda_{k,h}| = O(h^2). \tag{3.14}$$

La variabile $u_h \in S^0(\mathcal{T}_h)$ è una funzione costante a tratti: approssima u in media su ogni elemento, ma non è una buona approssimazione puntuale e non vive in $H_0^1(\Omega)$.
Per avere una migliore approssimazione delle autofunzioni definiamo

$$\tilde{u}_h(x) := \int_0^x s_h(\xi)\, d\xi, \qquad x \in [0, \pi].$$

e la chiamiamo funzione ricostruita (o approssimazione continua) della soluzione primaria u.
Per il teorema fondamentale del calcolo integrale $\tilde{u}_h(x)' = s_h(x)$ ed è continua. Poichè $s_h(x)$ sta in $L^2(\Omega)$ allora $\tilde{u}_h(x) \in H^1(\Omega)$ e per costruzione $\tilde{u}_h(0) = 0$.
Ponendo $t = 1$ nella prima equazione di (3.11) allora $t' = 0$, quindi $\int_0^\pi s_h(x)\, dx = 0$. Ne segue $\tilde{u}_h(x) \in H_0^1(\Omega)$. Questa funzione porta a una stima migliore della variabile primaria come conferma la prossima osservazione.

Osservazione 3.7. $\|u - \tilde{u}_h\|_{L^2(\Omega)} \leq C\, h^2\, \|u\|_{H^2(\Omega)}$ cioè di ordine 2 in L^2.

3.2.2 Esperimenti numerici

k	$N = 8$	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
λ_1	1.012916	1.003217	1.000803	1.000201	1.000050	1.000013
λ_2	4.209547	4.051664	4.012867	4.003214	4.000803	4.000201
λ_3	10.080291	9.263131	9.065245	9.016276	9.004067	9.001017
λ_4	19.453667	16.838190	16.206657	16.051470	16.012855	16.003213
λ_5	33.262830	27.064923	25.505923	25.125749	25.031390	25.007845

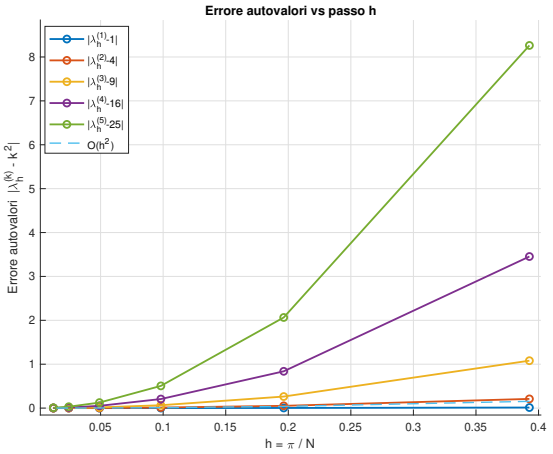
Tabella 3.11: Autovalori approssimati con il metodo $P1 - P0$

Nella tabella 3.11 riportiamo gli autovalori approssimati $\lambda_{k,h}$ ottenuti con il metodo misto $P1 - P0$ per diversi valori di N (numero di sottointervalli). Come atteso, le aprossimazioni si avvicinano ai corrispondenti valori esatti $1, 4, 9, 16, 25$ al crescere di N. Per k più grandi, con griglie grossolane gli errori iniziali sono elevati (es. per $k = 5, 33.26$ con $N = 8$), perchè l'autofunzione è più oscillante e la mesh non la risolve a sufficienza. Raffinando la griglia l'errore cala rapidamente e la legge quadratica si rende evidente.
In questa tabella ogni $\lambda_{k,h}$ sovrastima il valore esatto e descresce verso λ_k al crescere

di N. Questo comportamento (approssimazioni "dall'alto") è in generale frequente nei problemi simmetrici; è bene comunque notare che, in generale, per formulazioni miste non è universalmente garantito come nella formulazione classica.
Tutti gli autovalori discreti sono reali e positivi e si allineano ai corrispondenti k^2 senza produrre valori "spuri", a conferma della stabilità dello schema $P_1 - P_0$ in questo setting $1D$.

k	$N = 8 \to 16$	$N = 16 \to 32$	$N = 32 \to 64$	$N = 64 \to 128$	$N = 128 \to 256$
λ_1	2.005433	2.001383	2.000347	2.000087	2.000022
λ_2	2.020041	2.005433	2.001383	2.000347	2.000087
λ_3	2.037569	2.011843	2.003088	2.000780	2.000195
λ_4	2.042780	2.020041	2.005433	2.001383	2.000347
λ_5	2.000548	2.029098	2.008371	2.002153	2.000542

(a) *Ordine di convergenza degli autovalori* $(P1 - P0)$



(b) *Errore degli autovalori al variare di* h $(P1 - P0)$.

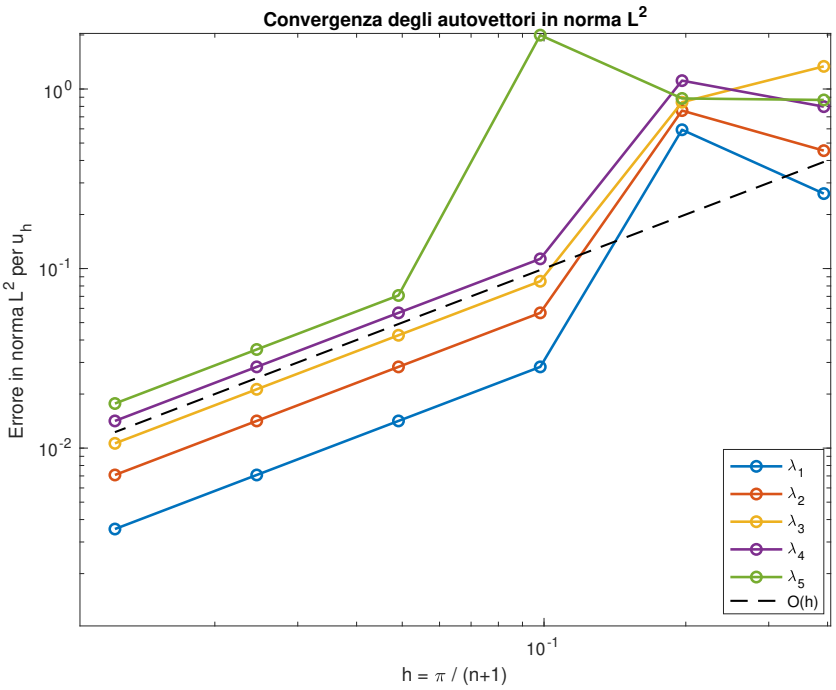
Figura 3.16: *Confronto tra (a) l'ordine di convergenza e (b) l'andamento dell'errore degli autovalori per il metodo* $(P1 - P0)$.

La tabella 3.16a conferma numericamente quanto visto nell'equazione (3.14).
Dalla stima di convergenza $|\lambda - \lambda_h| = O(h^2)$ segue che, al dimezzare del passo di discretizzazione h, l'errore sugli autovalori si riduce di circa un fattore 4.
Infatti come si può notare dal primo autovalore (vedi tabella 3.11) ad ogni dimezzamento del passo l'autovalore discreto guadagna una cifra significativa. La figura 3.16b mostra l'andamento degli errori tra autovalori esatti e discreti all'infittirsi della mesh, per ogni autovalore.
Tutte le curve descrescono quando h si riduce (mesh più fine) segno che gli autovalori

discreti $\lambda_{h,k}$ convergono ai valori esatti $\lambda_k = k^2$. Per h sufficientemente piccolo, le curve risultano "parallele" al riferimento $O(h^2)$.
Gli errori iniziali crescono con l'indice k: le autofunzioni più oscillanti richiedono una griglia più fine per entrare nel regime $O(h^2)$. Per esempio la curva di λ_5 parte con errori molto maggiori e si comporta quadraticamente solo per h più piccoli.
Gli errori sono positivi e decrescenti; non si osservano anomalie tipiche di instabilità (autovalori negativi o fuori scala), indice della buona stabilità del metodo in 1D.

k	$N = 8 \to 16$	$N = 16 \to 32$	$N = 32 \to 64$	$N = 64 \to 128$	$N = 128 \to 256$
λ_1	-1.181059	4.387115	0.999935	0.999984	0.999996
λ_2	-0.741328	3.742206	0.999739	0.999935	0.999984
λ_3	0.662833	3.312893	0.999413	0.999853	0.999963
λ_4	-0.481307	3.297299	0.999857	0.999739	0.999935
λ_5	-0.027835	-1.171089	4.815981	0.999593	0.999898

(a) *Ordine di convergenza degli autovettori (P1–P0).*



(b) *Errore degli autovettori al variare di h (P1–P0).*

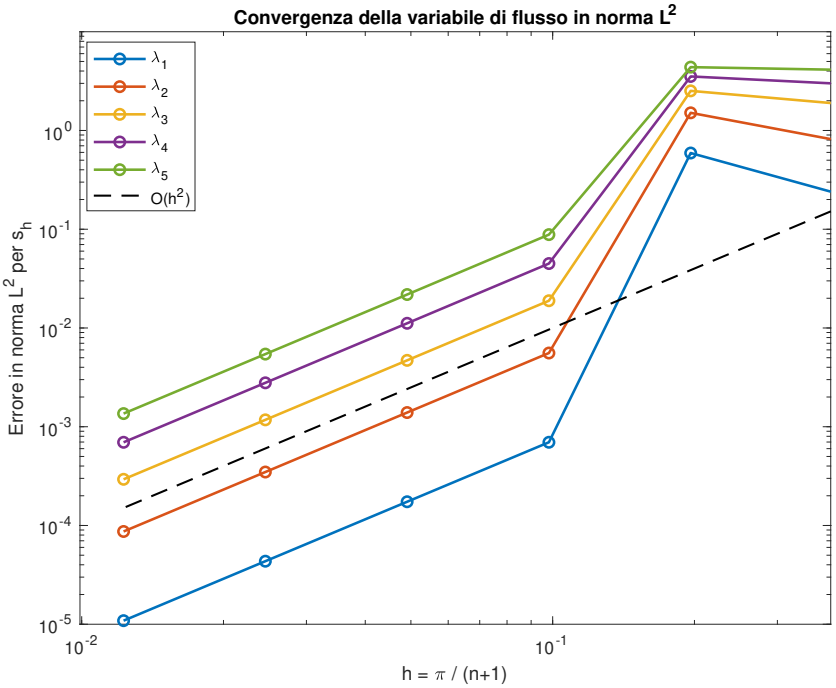
Figura 3.17: *Confronto tra (a) l'ordine di convergenza e (b) l'andamento dell'errore per il metodo P1–P0.*

La figura 3.17 mette a confronto due prospettive sul comportamento del metodo misto $P1 - P0$ applicato al calcolo degli autovettori del problema agli autovalori considerato. Nella parte (a) viene riportata la stima numerica dell'ordine di convergenza ottenuta al

diminuire del passo, mentre nella parte (b) si mostra l'andamento effettivo dell'errore in norma L^2 al variare di h.
Si osserva innanzitutto che, per griglie sufficientemente fitte, l'ordine di convergenza tende ad assestarsi intorno al valore 1 per tutti gli autovalori considerati, quindi abbiamo confermato numericamente quanto espresso in (3.13). Questo comportamento è chiaramente visibile anche nel grafico (b), in cui le curve relative ai diversi autovettori risultano sostanzialmente parallele e mostrano una pendenza unitaria su scala logaritmica, a conferma della disequazione (3.13).
Va però notato che sulle griglie più grossolane compaiono alcune irregolarità: in particolare, i valori dell'ordine riportati nella tabella presentano ocillazioni e, in alcuni casi, risultano del tutto fuori scala.
Questi fenomeni non sono indicativi di un'instabilità del metodo, ma piuttosto riflettono il fatto che su mesh molto larghe ci si trova ancora in una fase pre-asintotica, in cui il comportamento teorico atteso non è ancora dominante.

k	$N = 8 \to 16$	$N = 16 \to 32$	$N = 32 \to 64$	$N = 64 \to 128$	$N = 128 \to 256$
λ_1	-1.315567	9.730412	2.000623	2.000156	2.000039
λ_2	-0.895245	8.080138	2.002491	2.000623	2.000156
λ_3	-0.414076	7.059686	2.005605	2.001401	2.000350
λ_4	-0.234108	6.293549	2.009962	2.002491	2.000623
λ_5	-0.086062	5.633894	2.015563	2.003892	2.000973

(a) *Ordine di convergenza della variabile di flusso ($P1$–$P0$).*



(b) *Errore della variabile di flusso al variare di h ($P1$–$P0$).*

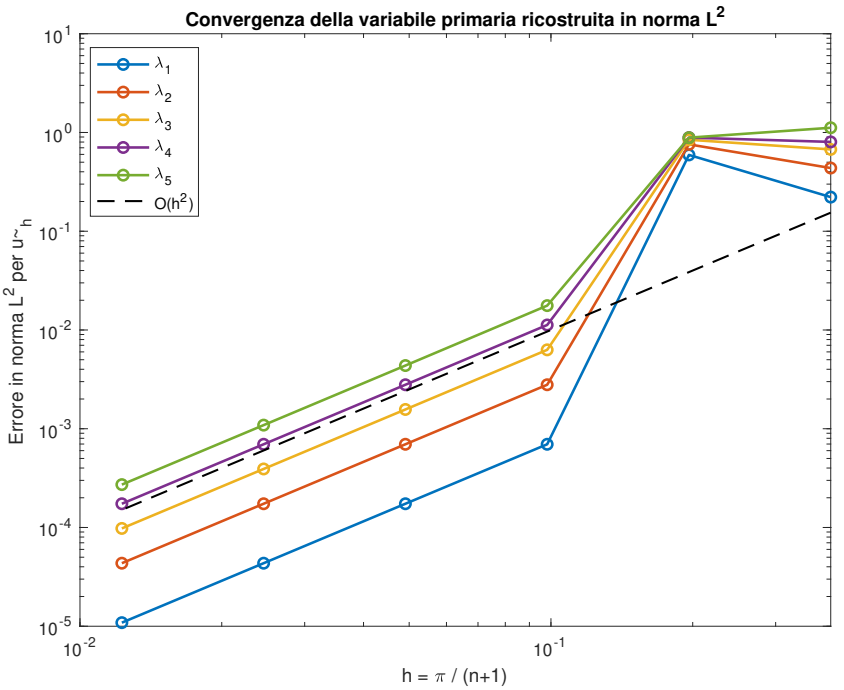
Figura 3.18: *Confronto tra (a) l'ordine di convergenza e (b) l'andamento dell'errore per il metodo $P1$–$P0$.*

Nella figura (3.18) è riportato il confronto tra l'ordine di convergenza (3.18a) e l'andamento dell'errore (3.18b) della variabile di flusso per il metodo misto $P1 - P0$.
La tabella (3.18a) mostra l'ordine sperimentale di convergenza calcolato al raddoppio del numero di suddivisioni N.
Si nota che, per valori molto piccoli di N, i risultati siano irregolari e in alcuni casi non significativi: si tratta del cosidetto regime pre-asintotico, in cui la discretizzazione è troppo grossolana per catturare correttamento l'andamento delle autofunzioni. A partire da $N = 32$, invece l'ordine di convergenza si stabilizza attorno al valore 2 per tutti i primi autovalori considerati.
Questo comportamento conferma che la variabile di flusso, approssimata tramite funzioni

lineari continue, converge in norma L^2 con ordine quadratico, come atteso dalla teoria. Differenze tra autovalori diversi si riflettono soltanto nelle costanti moltiplicative dell'errore, che risultano più elevate per autofunzioni più oscillanti.
Il grafico (3.18b) rappresenta l'errore in norma L^2 al variare della dimensione della mesh h. In scala $log - log$ le curve associate ai diversi autovalori risultano quasi parallele tra loro e alla retta di riferimento di pendenza 2, a conferma del comportamento quadratico già evidenziato nella tabella.
Le curve appaiono verticalmente sfalsate: questo significa che, a parità di passo h, l'errore risulta maggiore per gli autovalori di indice più alto, in ragione della maggiore complessità delle autofunzioni corrispondenti.

k	$N = 8 \to 16$	$N = 16 \to 32$	$N = 32 \to 64$	$N = 64 \to 128$	$N = 128 \to 256$
λ_1	-1.418915	9.733577	2.000652	2.000163	2.000041
λ_2	-0.795741	8.086012	2.002607	2.000652	2.000163
λ_3	-0.325376	7.067337	2.005866	2.001467	2.000367
λ_4	-0.144519	6.303008	2.010427	2.002607	2.000652
λ_5	0.333067	5.649497	2.016289	2.004074	2.001018

(a) *Ordine di convergenza della variabile primaria ricostruita* $(P1 - P0)$.



(b) *Errore della variabile primaria ricostruita al variare di* h $(P1 - P0)$.

Figura 3.19: *Confronto tra (a) l'ordine di convergenza e (b) l'andamento dell'errore della variabile primaria ricostruita per il metodo* $(P1 - P0)$.

Nella figura 3.19 viene analizzato l'andamento della variabile primaria ricostruita per il metodo misto $P1 - P0$, confrontando l'ordine di convergenza (3.19a) e l'errore in funzione del passo di discretizzazione (3.19b).

La tabella (3.19a) riporta l'ordine sperimentale di convergenza al raddoppio del numero di sottointervalli N. Come già osservato per la variabile di flusso, anche in questo caso i valori relativi a griglie molto grossolane (ad esempio $N = 8$ e $N = 16$) risultano non affidabili, con oscillazioni marcate e incoerenti con il comportamento atteso.

A partire da $N = 32$, invece, l'ordine si stabilizza rapidamente attorno al valore 2 per tutti i primi autovalori, confermando che la variabile primaria, ricostruita a partire dal

sistema misto, converge in norma L^2 con ordine quadratico. Questo è coerente con la teoria discussa precedentemente.
Il grafico (3.19b) mostra l'errore della variabile primaria ricostruita in norma L^2 al variare della dimensione della mesh h. In scala $log - log$, le curve associate ai diversi autovalori presentano una pendenza pressochè parallela alla retta di riferimento $O(h^2)$ evidenziando ancora una volta la convergenza di ordine 2.
Le curve risultano sfalsate verticalmente, indicando errori maggiori per autovalori associati a funzioni più oscillanti, ma mantenendo invariato l'ordine di convergenza.
Nel complesso, la figura (3.19) mette in evidenza come la ricostruzione della variabile primaria del metodo $P1 - P0$ permetta di ottenere un stima estremamente più accurata, con una convergenza quadratica stabile e uniforme per tutti i primi autovalori. Ciò conferma l'efficacia della procedura di ricostruzione e l'aderenza del comportamento numerico alle previsioni teoriche.

3.3 Schema $P1 - P1$

Questo schema non è stabile per l'approsimazione del problema di Laplace unidimensionale. In particolare, è stato dimostrato che produce risultati accettabili per soluzioni regolari, sebbene non sia convergente nel caso di dati singolari. Sebbene le autofunzioni del problema che consideriamo siano regolari (anzi, sono analitiche), il $P1 - P1$ non dà buoni risultati, come mostreremo in questa sezione.
Data la solita partizione uniforme di $[0, \pi]$ di passo h, pongo $\Sigma_h = S^1(\mathcal{T}_h)$ e $U_h = S^1(\mathcal{T}_h)$, se N è il numero di sottointervalli, allora $N_s = N_u = N + 1$.
Come base di $S^1(\mathcal{T}_h)$ prendo quella vista anche nella sezione precedente. Con questo schema mettendo entrambe le equazioni di (3.11) in forma matriciale otteniamo:

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \mathrm{x} \\ \mathrm{y} \end{pmatrix} = -\lambda \begin{pmatrix} 0 & 0 \\ 0 & A \end{pmatrix} \begin{pmatrix} \mathrm{x} \\ \mathrm{y} \end{pmatrix}.$$

dove $A \in \mathbb{R}^{(N+1)\times(N+1)}$ si assembla dalla matrice locale che è della forma

$$A_{\text{loc}} = \frac{h}{6} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

e $B \in \mathbb{R}^{(N+1)\times(N+1)}$ si assembla dalla matrice locale che è della forma

$$B_{\text{loc}} = \frac{1}{2} \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}.$$

Esperimenti numerici

	$N = 8$	$N = 16$	$N = 32$	$N = 64$	$N = 128$
0	0.0000	-0.0000	-0.0000	-0.0000	-0.0000
1	1.0001	1.0000 (4.1)	1.0000 (4.0)	1.0000 (4.0)	1.0000 (4.0)
4	3.9660	3.9981 (4.2)	3.9999 (4.0)	4.0000 (4.0)	4.0000 (4.0)
9	7.4257	8.5541 (1.8)	8.8854 (2.0)	8.9711 (2.0)	8.9928 (2.0)
9	8.7603	8.9873 (4.2)	8.9992 (4.1)	9.0000 (4.0)	9.0000 (4.0)
16	14.8408	15.9501 (4.5)	15.9971 (4.1)	15.9998 (4.0)	16.0000 (4.0)
25	16.7900	24.5524 (4.2)	24.9780 (4.3)	24.9987 (4.1)	24.9999 (4.0)
36	38.7154	29.7390 (-1.2)	34.2165 (1.8)	35.5415 (2.0)	35.8846 (2.0)
36	39.0906	35.0393 (1.7)	35.9492 (4.2)	35.9970 (4.1)	35.9998 (4.0)
49		46.7793	48.8925 (4.4)	48.9937 (4.1)	48.9996 (4.0)

Tabella 3.12: Autovalori calcolati (ordine di convergenza)

Esatti	Calcolati
	-0.0000000000
1	1.0000000000
4	3.9999999999
9	8.9998815658
9	8.9999999992
16	15.9999999971
25	24.9999999784
36	35.9981051039
36	35.9999999495
49	48.9999998977

Tabella 3.13: Autovalori calcolati con lo schema $P1 - P1$ con $N = 1000$

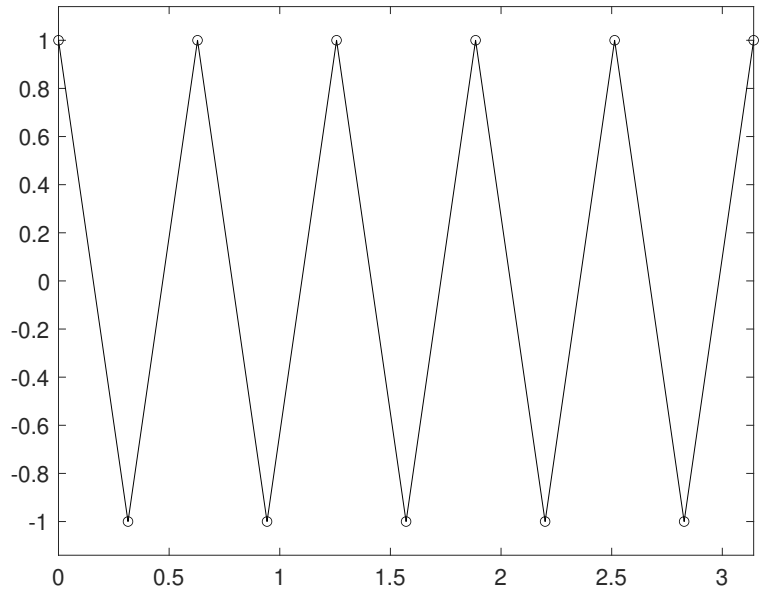


Figura 3.20: *Autofunzione u_h associata a $\lambda_h = 0$*

I risultati numerici degli esperimenti al crescere di N sono elencati nella tabella 3.12. Anzitutto è chiaro che i valori corretti son ben approssimati e per questi l'ordine di convergenza è 4.
Dall'altra parte, ci sono alcune soluzioni spurie che ora descriviamo in maggior dettaglio.
La frequenza discreta nulla è legata al fatto che lo schema non soddisfa la condizione di inf-sup.
Le autofunzioni corrispondenti sono $s_h(x) = 0$ e $u_h(x)$ come rappresentata in figura 3.20 nel caso $N = 10$.
La funzione u_h è ortogonale in $L^2(\Omega)$ a tutte le derivate delle funzioni in Σ_h, e l'esistenza di u_h in questo caso mostra, in particolare, che questo schema non soddisfa la classica condizione di inf-sup.
Osserviamo che $\lambda_h = 0$ è un vero autovalore del nostro problema discreto anche se la funzione corrispondente s_h si annulla, poichè l'autofunzione che ci interessa è u_h.
Oltre alla frequenza nulla, ci sono altre soluzioni spurie: la prima varia tra 7.4257 e 8.9928 nei calcoli riportati nella tabella (3.12), ed aumenta al crescere di N.
Purtroppo questa frequenza spuria rimane limitata e sembra convergere a 9 (il che implica una molteplicità discreta errata per l'autovettore esatto $\lambda = 9$) come mostrato nella tabella (3.13) dove presentiamo i risultati del calcolo per $N = 1000$.
La stessa situazione si verifica per l'altro valore spurio delle tabelle 3.12 e 3.13, che sembra convergere a un valore vicino a 36.

La situazione è in realtà più complessa e interessante: succede questo: $\lambda_{h,4k} \to \lambda_{3k}$ per $k = 0, 1, 2, ...$
Questo fatto è evidente dal seguente grafico.

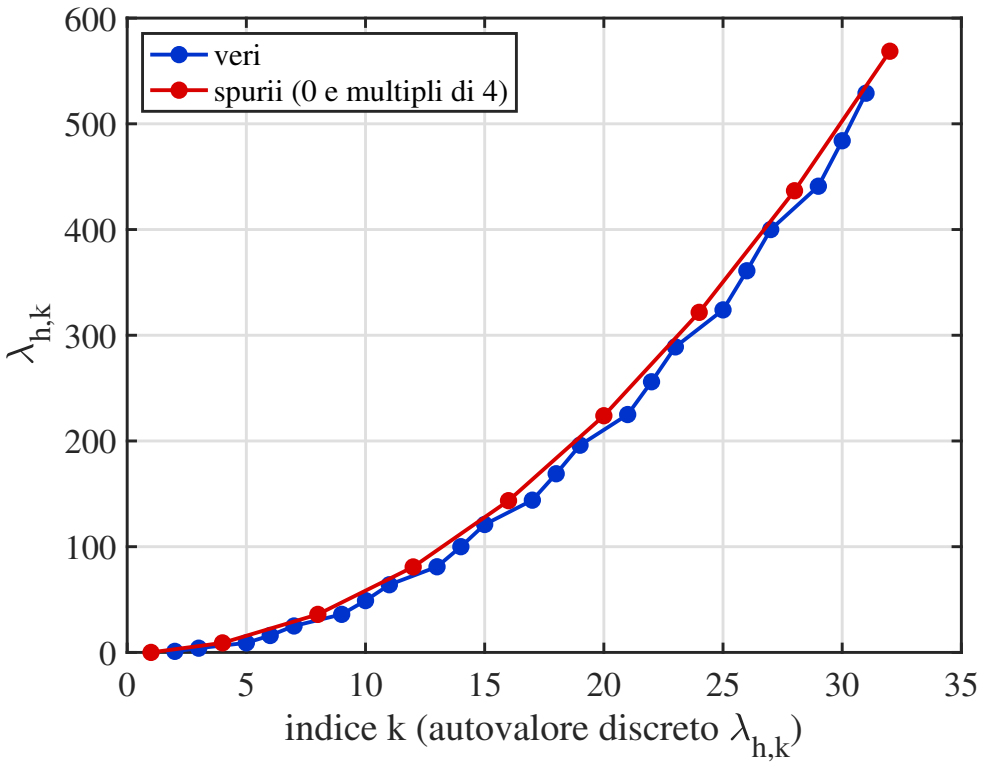


Figura 3.21: *Autovalori discreti del metodo P1-P1*

Quindi lo schema $P1 - P1$ crea una molteplicità sbagliata, ad esempio $\lambda = 9$ risulta di molteplicità 2 invece che 1, $\lambda = 36$ risulta di molteplicità 2 invece che 1, ecc.
Nella prossima figura vengono mostrate le autofunzioni corrispondenti a $\lambda_{h,4}$ e $\lambda_{h,8}$.

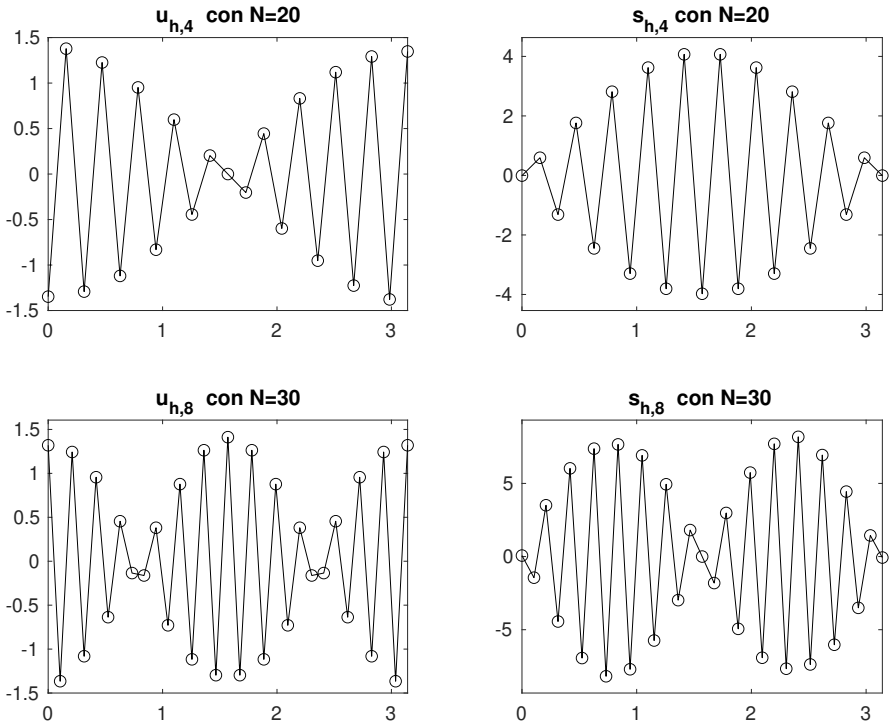


Figura 3.22: *$P1 - P1$ autofunzioni spurie corrispondenti a $\lambda_{h,4}$ (sopra,$N = 20$) e $\lambda_{h,8}$ (sotto,$N = 30$); u_h (sinistra) e s_h (destra)*

3.3.1 Schema $P2 - P0$

Ora discuteremo lo schema $P2 - P0$, che è noto essere instabile per il corrispondente problema di sorgente.

Data la solita partizione uniforme di $[0, \pi]$ di passo h, la coppia di spazi è volutamente sbilanciata: $s_h \in S^2(\mathcal{T}_h)$ e $u_h \in S^0(\mathcal{T}_h)$.

Le funzioni di $S^2(\mathcal{T}_h)$ sono polinomi compositi, di grado 2 su ciascun intervallo di $\mathcal{T}_h$, il terzo sarà il punto medio dello stesso.

La base lagrangiana globale di $S^2(\mathcal{T}_h)$ è $\{\Phi_1, \Phi_2, \ldots, \Phi_{2N+1}\}$ dove ciascuno Φ_j soddisfa $\Phi_j(x_i) = \delta_{ij}$. Invece come base di $S^0(\mathcal{T}_h)$ prendiamo $\{\psi_1, \psi_2, \ldots, \psi_N\}$ definiti come nella settosezione precedente.

Quindi la rappresentazione matriciale diventa:

$$\begin{bmatrix} A & B^\top \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = -\lambda \begin{bmatrix} 0 & 0 \\ 0 & M \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

$$A_{\ell k} = \int_0^\pi \Phi_k(x)\,\Phi_\ell(x)\,dx, \qquad B_{j\ell} = \int_0^\pi \psi_j(x)\,\Phi'_\ell(x)\,dx = \Phi_\ell(x_{j+1}) - \Phi_\ell(x_j), \qquad M = h\,I.$$

Esperimenti numerici

k	$N = 8$	$N = 16$	$N = 32$	$N = 64$	$N = 128$
λ_1	5.706113	5.923836	5.980783	5.995185	5.998795
λ_2	19.880026	22.824452	23.695343	23.923131	23.980738
λ_3	36.706515	48.379811	52.480872	53.612347	53.902582
λ_4	51.876446	79.520106	91.297808	94.781372	95.692526
λ_5	63.614025	113.181915	138.816486	147.045140	149.250630
λ_6	71.666643	146.826058	193.519245	209.923488	214.449386
λ_7	76.305120	178.640420	253.804391	282.851496	291.134446
λ_8	77.814669	207.505784	318.080423	365.191231	379.125489
λ_9		232.846145	384.842541	456.244471	478.217232
λ_{10}		254.456099	452.727660	555.265945	588.180560

Tabella 3.14: Autovalori calcolati con schema $P2 - P0$

I risultati dei calcoli numerici su una sequenza di mesh via via più raffinate sono riportati nella tabella (3.14).

In questo caso non ci sono soluzioni spurie, ma gli autovalori calcolati risultano errati di un fattore 6. Più precisamente, essi convergono correttamente verso sei volte le soluzioni esatte.

L'errore è quindi dovuto a un difetto strutturale del metodo, non alla presenza di autovalori non fisici.

k	$N = 8 \to 16$	$N = 16 \to 32$	$N = 32 \to 64$	$N = 64 \to 128$
λ_1	1.9	2.0	2.0	2.0
λ_2	1.8	1.9	2.0	2.0
λ_3	1.6	1.9	2.0	2.0
λ_4	1.4	1.8	1.9	2.0
λ_5	1.2	1.7	1.9	2.0
λ_6	1.1	1.6	1.9	2.0
λ_7	0.9	1.5	1.9	2.0
λ_8	0.8	1.4	1.8	1.9
λ_9		1.3	1.8	1.9
λ_{10}		1.2	1.7	1.9

Tabella 3.15: Ordine di convergenza di $\lambda_{h,k}$ rispetto a $6\lambda_k$

Il tasso di convergenza (rispetto a $6k^2$) riportato in tabella, mostra che gli autovalori più bassi tendono rapidamente a una convergenza di ordine 2 rispetto a $6k^2$. Per autovalori più alti il comportamento è inizialmente meno regolare, ma si stabilizza anch'esso intorno a 2 per valori di N sufficientemente grandi.

Le autofunzioni corrispondenti ai primi due autovalori sono riportate in figura (3.23): esse mostrano un comportamento analogo a quello osservato in letteratura per il problema alle derivate parziali con termine sorgente.
In particolare si nota che le autofunzioni u_h costituiscono una corretta approssimazione di u (vedi figura 3.24), mentre le funzioni s_h presentano componenti spurie chiaramente associate a una "bolla" in ciascun elemento.
Questo comportamento è legato al fatto che l'ellitticità del nucleo discreto non risulta soddisfatta a causa della presenza delle funzioni bolla nello spazio $S^2(\mathcal{T}_h)$).

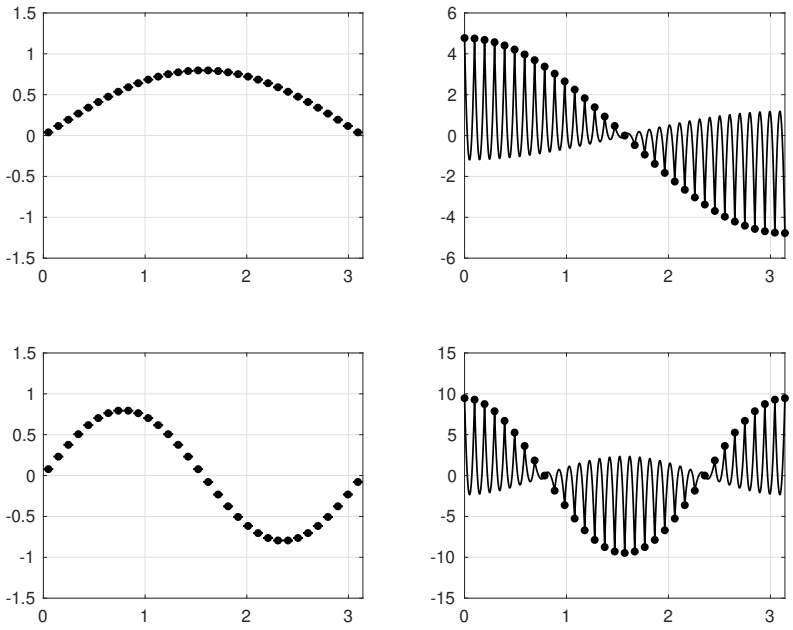


Figura 3.23: *Autofunzioni corrispondenti al primo (sopra) e al secondo (sotto) autovalore discreto; u_h(a sinistra) e s_h(a destra)*

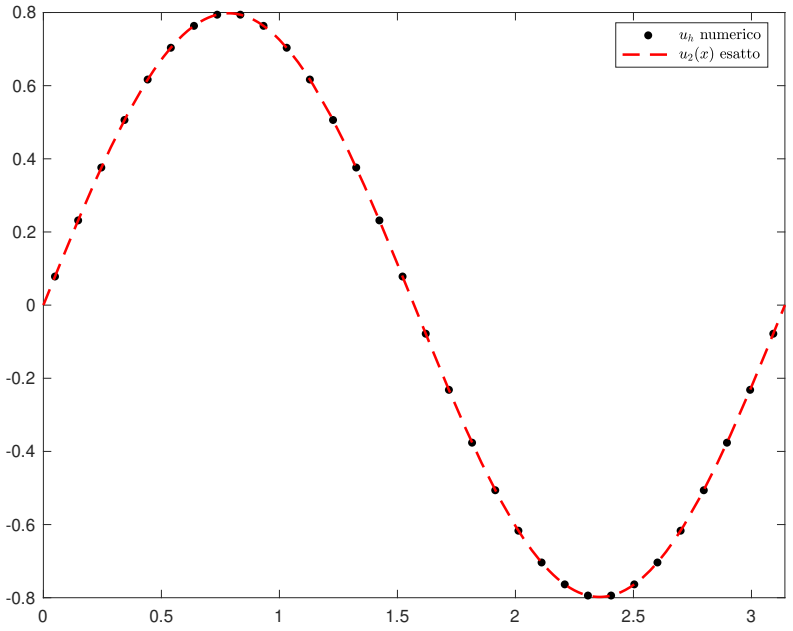


Figura 3.24: *Confronto tra la seconda autofunzione discreta ed esatta*

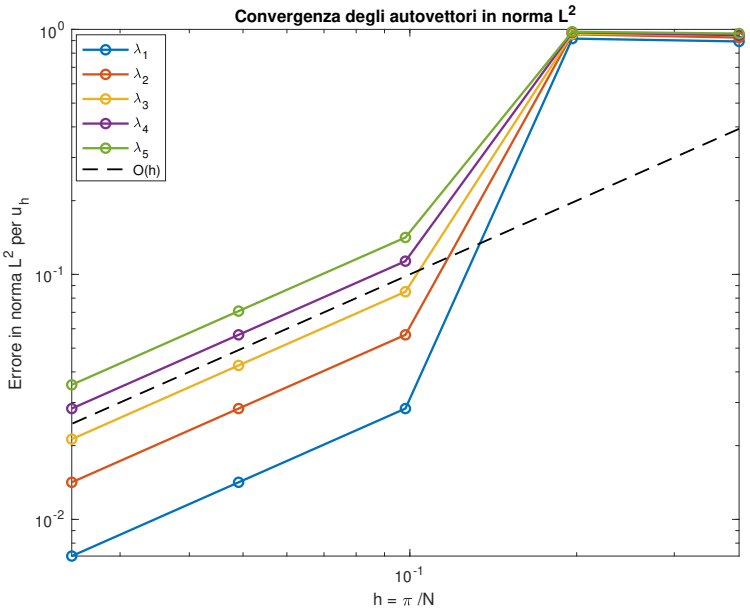
Dalla tabella (3.25a) si osserva che, al crescere del numero di elementi, l'ordine tende a valori prossimi a 1, confermando la convergenza lineare prevista teoricamente per il metodo $P2 - P0$.

Il grafico 3.25b conferma visivamente questo comportamento: le curve degli errori per i diversi autovettori mostrano una pendenza allineata con la retta di riferimento $O(h)$, indicando che la convergenza in norma L^2 avviene con ordine 1.

Inoltre, si nota che l'errore è più contenuto per i primi autovettori e tende a crescere per quelli di indice maggiore, come tipicamente accade nei problemi approssimati con elementi finiti.

k	$N = 8 \to 16$	$N = 16 \to 32$	$N = 32 \to 64$	$N = 64 \to 128$
λ_1	-0.038030	5.015835	0.999935	0.999984
λ_2	-0.044048	4.074228	0.999739	0.999935
λ_3	-0.012994	3.496430	0.999413	0.999853
λ_4	-0.039313	3.103094	0.998957	0.999739
λ_5	-0.024513	2.789479	0.998370	0.999593

(a) *Ordine di convergenza degli autovettori ($P2$–$P0$)*



(b) *Errore degli autovettori discreti in norma L^2 in funzione di h*

Figura 3.25: *Confronto tra (a) ordine di convergenza e (b) andamento dell'errore per il metodo $P2$–$P0$.*

Capitolo 4

Il problema agli autovalori generalizzato

Nei capitoli precedenti abbiamo introdotto gli stumenti funzionali necessari allo studio dei problemi agli autovalori per poi inquadrare la teoria astratta e discutere il caso del problema agli autovalori di Laplace in una dimensione.
In questo capitolo affrontiamo invece il passaggio alla discretizzazione numerica, che nel metodo degli elementi finiti porta in modo naturale alla formulazione del problema agli autovalori generalizzato:

$$Ax = \lambda Mx,$$

dove A viene detta matrice di rigidezza e M matrice di massa.
L'obbiettivo del capitolo è quello di presentare le proprietà principali di questo problema, il legame con la formulazione standard, e i metodi numerici che ne permettono l'approssimazione.
In particolare, partiremo da un richiamo al metodo delle potenze e alle sue varianti, per poi estendere l'analisi al caso generalizzato e discutere i relativi risultati di convergenza. Infine, concluderemo con alcune implementazioni numeriche, applicate ad esempi derivati dal metodo degli elementi finiti.

4.1 Problema agli autovalori classico

Definizione 4.1. *Sia $A \in \mathbb{C}^n$, una coppia (λ, x) con $\lambda \in \mathbb{C}, x \in \mathbb{C}^n$ con x non nullo, è detta autocoppia se*

$$Ax = \lambda x \tag{4.1}$$

Ogni scalare $\lambda \in \mathbb{C}$ che verifica questa equazione è detto autovalore, mentre il vettore x è detto autovettore di A relativo a λ.
Il problema agli autovalori classico consiste nella'approssimare questa autocoppia.

4.1.1 Localizzazione degli autovalori

Il problema degli autovalori si configura, dal punto di vista algebrico, come un problema di natura non lineare, la cui risoluzione diretta spesso risulta impraticabile in presenza di matrici di grandi dimensioni. Per questo motivo, nella pratica numerica si ricorre a metodi iterativi.
In tale contesto, avere a disposizione una stima preliminare della posizione degli autovalori nel piano complesso rappresenta uno strumento prezioso: conoscere l'area in cui essi si collocano permette di valutare e spesso anche di prevedere la velocità e l'efficacia della convergenza dei metodi iterativi usati.
Inoltre permette anche di dedurre importanti proprietà qualitative della matrice: ad esempio, stabilire se gli autovalori sono tutti reali e positivi, e dunque se la matrice è definita positiva.
Per questo motivo, la teoria mette a disposizione semplici strumenti di localizzazione spettrale come ad esempio la proposizione di Hirsch e il primo teorema di Gershgorin, illustrati qui sotto.

Proposizione 4.2 (di Hirsch)**.** *Sia $A \in \mathbb{C}^{n\times n}$ e $\|\cdot\|$ norma matriciale indotta. Allora $spec(A) \subset \{z \in \mathbb{C}\ t.c.\ |z| \leq \|A\|\}$.*

Definizione 4.3 (Cerchi di Gershgorin)**.** *Sia $A \in \mathbb{C}^{n\times n}$. I cerchi del piano complesso*

$$G_i^{(r)} := \left\{ z \in \mathbb{C}\ t.c\ |z - A_{ii}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} |A_{ij}| \right\}, \quad i = 1, \ldots, n,$$

di centro A_{ii} e raggio $\displaystyle\sum_{\substack{j=1 \\ j\neq i}}^{n} |A_{ij}|$ sono detti cerchi di Gershgorin per righe.

Teorema 4.4 (Primo teorema di Gershgorin)**.** *Sia $A \in \mathbb{C}^{n\times n}$. Allora*

$$\mathrm{spec}(A) \subset \bigcup_{i=1}^{n} G_i^{(r)},$$

cioè gli autovalori di A sono contenuti nell'unione dei cerchi di Gershgorin per righe.

Questo teorema vale anche per A^T, e quindi per le colonne di A:

$$\text{spec}\left(A^T\right) \subset \bigcup_{i=1}^{n} \mathcal{G}_i^{(c)},$$

dove l'apice "c" indica che sono cerchi di Gerschgorin per colonne.
Si ha dunque che

$$\text{spec}(A) \subset \left(\bigcup_{i=1}^{n} \mathcal{G}_i^{(r)}\right) \cap \left(\bigcup_{i=1}^{n} \mathcal{G}_i^{(c)}\right).$$

4.1.2 Metodo delle potenze

Il metodo delle potenze è un metodo iterativo particolarmente adatto per il calcolo dell'autovalore dominante della matrice $A \in \mathbb{R}^{n\times n}$, ovvero quello di modulo massimo λ_1, ed il relativo autovettore associato.
Sia $A \in \mathbb{R}^{n\times n}$, l'iterazione del metodo delle potenze consiste in:

Algorithm 1 Metodo delle potenze

1: Scegli un vettore iniziale $x^{(0)}$ con $\|x^{(0)}\| = 1$
2: **for** $i = 0, 1, 2, \ldots$ fino a convergenza **do**
3: $\quad y \leftarrow Ax^{(i)}$
4: $\quad x^{(i+1)} \leftarrow \dfrac{y}{\|y\|}$
5: $\quad \lambda^{(i+1)} \leftarrow \left(x^{(i+1)}\right)^H Ax^{(i+1)}$
6: **End**

dove $\lambda^{(i+1)}$ è calcolato applicando il quoziente di Rayleigh in cui non è presente il denominatore, in quanto il vettore $x^{(i+1)}$ ha norma unitaria.
Indicata con (λ_1, x_1) l'autocoppia dominante di A, il metodo delle potenze costruisce due successioni $\left\{x^{(k)}\right\}_{k\in\mathbb{N}}$ e $\left\{\lambda^{(k)}\right\}_{k\in\mathbb{N}}$ tali che

$$x^{(k)} \to x_1, \quad \lambda^{(k)} \to \lambda_1, \quad k \to +\infty.$$

Per far convergere il metodo, λ_1 deve essere un autovalore semplice in modulo.
Nel caso A sia diagonalizzabile, sia $A = X\Lambda X^{-1}$ con $X = [x_1, \ldots, x_n]$ matrice avente cone colonne gli autovettori di A e $|\lambda_i|$ in ordine decrescente. Ponendo $\xi = X^{-1}x^{(0)}$ e supponendo $(\xi)_1 \neq 0$ vale il seguente teorema:

Teorema 4.5. *Siano $|\lambda_i|, i = 1, \ldots, n$ ordinati in modo decrescente, e $|\lambda_1| > |\lambda_2|$, con $|\lambda_1|$ semplice e sia x_1 l'autovettore corrispondente.Allora esiste una costante $C > 0$, tale che*

$$\left\|x^{(i)} - x_1\right\|_2 \leq C \left|\frac{\lambda_2}{\lambda_1}\right|^i, \quad i \geq 1$$

dove $x^{(i)}$ *è normalizzato in modo opportuno.*

Invece per gli autovalori vale che :

$$\left|\lambda^{(k)} - \lambda_1\right| \leq C \left|\frac{\lambda_2}{\lambda_1}\right|^{2k}$$

cioè la convergenza è quadratica in $|\frac{\lambda_2}{\lambda_1}|$.
Si noti che solo l'autovalore converge in modo quadratico,mentre l'autovettore approssimato converge in modo lineare.

4.1.3 Metodo delle potenze shiftate

Per approssimare autovalori diversi da quello dominante esistono metodi derivanti dal metodo delle potenze.
Infatti, dato $\sigma \in \mathbb{C}$ con $\sigma \notin spec(A)$, è possibile approssimare l'autovalore di A più vicino a σ in modulo. Più precisamente, sia σ tale che esiste $\lambda_m \in spec(A)$ che soddisfa

$$|\lambda_m - \sigma| < |\lambda_i - \sigma|, \quad \forall i = 1, \ldots, n, i \neq m.$$

In particolare, questa disuguaglianza stretta implica che l'autovalore λ_m più vicino al parametro σ abbia molteplicità uno in modulo. Allora è possibile usare il metodo delle potenze dove al prodotto $Ax^{(i)}$ viene sostituita l'operazione $(A - \sigma I)^{-1}x^{(i)}$.
L'iterazione del metodo delle potenze inverse shiftate consiste in:

Algorithm 2 Metodo delle potenze shiftate

1: Scegli un vettore iniziale $x^{(0)}$ con $\|x^{(0)}\| = 1$
2: **for** $i = 0, 1, 2, \ldots$ fino a convergenza **do**
3: $\quad y \leftarrow (A - \sigma I)^{-1}x^{(i)}$
4: $\quad x^{(i+1)} \leftarrow \dfrac{y}{\|y\|}$
5: $\quad \lambda^{(i+1)} \leftarrow (x^{(i+1)})^H A x^{(i+1)}$
6: **End**

L'algoritmo sfrutta il fatto che gli autovettori di A e $(A-\sigma I)^{-1}$ rimangono invariati. Più precisamente, (λ, x) è una autocoppia di A se e solo se $(1/(\lambda - \sigma), x)$ è una autocoppia di $(A - \sigma I)^{-1}$.
La velocità di convergenza degli autovalori e degli autovettori dipende da quanto σ è

vicino all'autovalore cercato, rispetto agli altri autovalori. Se gli autovalori λ_j sono ordinati in modo che $|\lambda_1 - \sigma| < |\lambda_2 - \sigma| \leq \ldots \leq |\lambda_n - \sigma|$ allora la velocità di convergenza degli autovettori è del tipo $O\left(\left(|\lambda_1 - \sigma| / |\lambda_2 - \sigma|\right)^k\right)$ invece quella degli autovalori è del tipo $O\left(\left(|\lambda_1 - \sigma| / |\lambda_2 - \sigma|\right)^{2k}\right)$.

4.2 Il problema agli autovalori generalizzato

Definizione 4.6. *Siano $A, M \in \mathbb{R}^{n\times n}$ simmetriche, una coppia (λ, x) con $\lambda \in \mathbb{C}$, $x \in \mathbb{C}^n$ con x non nulla è detta autocoppia se*

$$Ax = \lambda Mx \tag{4.2}$$

La coppia (A, M) è detta pencil.
Ogni autovalore $\lambda \in \mathbb{C}$ che verifica questa equazione è detto autovalore, mentre il vettore x è detto autovettore della pencil (A, M) relativo a λ.
Il problema agli autovalori generalizzato consiste nell'approssimazione di questa autocoppia.
Gli autovalori $\lambda_1, \ldots, \lambda_n$ sono le radici del polinomio $P(\lambda) = det(\lambda M - A)$.

Osservazione 4.7. Il problema (4.1) è un caso particolare di (4.2) con $M = I$.

Nel problema agli autovalori generalizzato di cerca di ricondurci ad una coppia di matrici equivalenti, il cui studio è più semplice dello studio di (A, M).

4.2.1 Equivalenza tra coppie di matrici

Date $A, M \in \mathbb{R}^{n\times n}$ simmetriche, per la ricerca delle autocoppie (λ, x) che soddisfano la (4.2) non si lavora quasi mai direttamente sulle matrici di partenza.
Più frequentemente si ricorre a trasformazioni che portano a coppie di matrici equivalenti, sulle quali i calcoli risultano più semplici e stabili. Presentiamo quindi la nozione di equivalenza tra coppie di matrici e vediamo in quali casi due pencil possono considerarsi equivalenti.

Definizione 4.8. *Due pencil (A_1, M_1) e (A_2, M_2) si dicono equivalenti se esistono due matrici invertibili F, G tali che*

$$A_2 = FA_1G \qquad M_2 = FM_1G \tag{4.3}$$

Osservazione 4.9. Siano (A_1, M_1) e (A_2, M_2) due pencil equivalenti, allora i loro autovalori sono uguali.

Dimostrazione. Sia $P_2(\lambda) = det(\tau M_2 - A_2)$ il polinomio caratteristico di (A_2, M_2).

$$P_2(\lambda) = \det(\tau M_2 - A_2) = \det(\lambda F M_1 G - F A_1 G) = \det(F)\det(\lambda M_1 - A_1)\det(G).$$

Sfruttando il fatto che F e G sono invertibili, l'equazione precedente implica che $P_2(\lambda) = 0$ se e solo se $\det(\lambda M_1 - A_1) = 0$. Cioè λ è autovalore di (A_1, M_1) se e solo se è autovalore di (A_2, M_2). □

Osservazione 4.10. Con le stesse assunzioni precedenti, x è un autovettore di (A_2, M_2) relativo a λ sse $y = Gx$ è un autovettore di (A_1, M_1) relativo a λ.

Dimostrazione. Poichè x è autovalore di (A_2, M_2) allora vale che $(\lambda M_2 - A_2)x = 0$. Quest'ultima vale se e solo se $(F(\lambda M_1 - A_1)G)x = 0$. Ponendo $y = Gx$ si ha

$$(F(\lambda M_1 - A_1))y = 0.$$

Poichè F è invertibile questo implica che

$$(\lambda M_1 - A_1)y = 0$$

cioè che y è autovettore di (A_1, M_1). □

Definizione 4.11. *Due pencil si dicono congruenti se vale la (4.3) con $F = G^*$.*

Cerchiamo ora di capire, data una pencil (A, M), quale sia la scelta migliore di una pencil equivalente.
Per alcune pencil (A, M) esiste una matrice invertibile G tale che

$$G^* A G = \operatorname{diag}(\phi_1, \ldots, \phi_n) = \Phi$$
$$G^* M G = \operatorname{diag}(\psi_1, \ldots, \psi_n) = \Psi$$

C'è la possibilità di rendere uniche tali matrici. Infatti è possibile normalizzare la coppia di matrici, prendendo $\Psi = I$ e scegliendo Φ di conseguenza.
Nel problema (4.1) avere una matrice $A \in \mathbb{R}^{n\times n}$ simmetrica rende la ricerca degli autovalori molto più semplice, mentre nel problema agli autovalori generalizzato ci troviamo davanti a situazioni più complesse anche se le matrici A e M sono simmetriche.
Infatti se ad esempio M è invertibile allora risolvere il problema agli autovalori generalizzato

$$Ax = \lambda M x$$

è equivalente a risolvere il problema standard

$$M^{-1}A = \lambda x$$

quindi i problemi riscontrati nel calcolo degli autovalori di $M^{-1}A$ influenzano la risoluzione di questo problema.
Noi ci limiteremo a trattare il caso in cui almeno M è definita positiva superando le difficoltà di cui abbiamo appena parlato.

Teorema 4.12. *Se M è definita positiva allora esiste una matrice invertibile F tale che (A, M) è equivalente a (F^TAF, I).*

Dimostrazione. Poichè M è s.d.p. si può scrivere come $M = M^{1/2}M^{1/2}$.
Moltiplicando il problema generalizzato a sinistra per $M^{-1/2}$ si trova che

$$M^{-1/2}Ax = \lambda M^{1/2}x,$$

ponendo $y = M^{1/2}x$ troviamo che

$$M^{-1/2}AM^{-1/2}y = \lambda y$$

Pongo $F = M^{-1/2}$ da cui la tesi. □

Teorema 4.13. *Sia A simmetrica e M s.d.p., allora (A, M) ha n radici reali $\lambda_1, \ldots, \lambda_n$ nell'intervallo $[-\|M^{-1}A\|, \|M^{-1}A\|]$ e n autovettori indipendenti $z_1, \ldots, z_n$. Inoltre si ha che*

$$z_i^T M z_j = 0$$

ovvero z_i e z_j sono M-ortogonali se $\lambda_i \neq \lambda_j$. Se $\lambda_i = \lambda_j$,z_i e z_j possono essere comunque essere scelti M-ortogonali.

4.2.2 Riduzione esplicita in forma standard

Per risolvere il problema agli autovalori generalizzati si cerca di non lavorare direttamente su una pencil data (A, M), ma di ricondursi a una pencil equivalente.
In particolare è possibile cercare di ricondursi alla forma standard cioè

$$Ax = \lambda Mx \tag{4.4}$$

$$\Downarrow$$

$$\hat{A}x = \lambda x \tag{4.5}$$

Questo può essere fatto in diversi modi:

(i) Si moltiplica per (4.4) per M^{-1} ottenendo un'espressione equivalente:

$$M^{-1}Ax = \lambda x$$

che è esattamente la (4.5) con $\hat{A} = M^{-1}A$.

(ii) Poichè M è s.d.p. per il teorema spettrale si può scrivere come: $M = G\Delta G^T$, da cui:
$A - \lambda M = A - \lambda(G\Delta G^T) = G(G^T AG - \lambda\Delta^2)G^T = G\Delta\left(\Delta^{-1}G^T AG\Delta^{-1} - \lambda I\right)\Delta G^T$
Pongo $\hat{A} = \Delta^{-1}G^T AG\Delta^{-1}$, si ha che :

$$\left(G\Delta(\hat{A} - \lambda I)\Delta G^T\right)x = 0$$

e ponendo $y = \Delta G^T x$, dall'invertibilità di $G\Delta$ si riduce a $(\hat{A} - \lambda I)y = 0$

(iii) Poichè M è s.d.p. è possibile calcolarne la decomposizione di Cholesky: $M = LL^T$.
Sostituendo a (4.4) si ha che:
$A - \lambda LL^T = L\left(L^{-1}AL^{-T} - \lambda I\right)L^T$
Poniamo $\hat{A} = L^{-1}AL^{-T}$ e ripetiamo il ragionamento di prima.

Osservazione 4.14. Per i metodi ii),iii) nel processo di riduzione di M è necessario tenere memoria delle matrici di passaggio L, G solo se si è interessati al calcolo degli autovettori.

4.2.3 Riduzione implicita in forma standard

Le tre riduzioni esaminate nella sezione precedente possono essere fatte anche in forma implicita.
Tale tecnica viene usata in particolare quando le matrici A, M sono sparse e di grandi dimensioni. Nel caso i) abbiamo ridotto il problema al problema degli autovalori classico di matrice $M^{-1}A$. Supponendo che questa matrice non abbia stuttura buona, sfrutto la sparsità di M e A in questo modo:

1. si calcola $b = Ax$

2. si risolve il sistema lineare $Mv = b$ con tecniche note.

La riduzione del caso ii) non viene quasi mai usata in maniera esplicita perchè la matrice di trasformazione G non conserva la sparsità di M. Invece il caso iii) fatto in maniera implicita, si divide in 3 passaggi:

1. Risoluzione del sistema $L^T v_1 = x$

2. calcolo del vettore $b = Av_1$

3. Risoluzione del sistema $Lv_2 = b$.

4.3 Descrizione del metodo delle potenze generalizzato e risultati di convergenza

È possibile usare altri metodi per risolvere il problema agli autovalori generalizzato anche quando non è possibile o è troppo difficile fattorizzare M.
In questa sezione mostreremo due metodi iterativi usati per approssimare gli autovalori di una pencil (A, M) senza necessariamente ricorrere alla fattorizzazione.
Consideriamo il metodo delle potenze e il metodo delle potenze inverse che si basano rispettivamente sulle equazioni:

$$Mv_{k+1} = (A - \sigma_k M)\, v_k \tau_k \tag{4.6}$$

$$(A - \sigma_k M)\, u_{k+1} = M u_k \tau_k \tag{4.7}$$

dove:

- (A, M) è la pencil relativa al problema iniziale,
- τ_k è una costante di normalizzazione.

Se A e M possono essere fattorizzate, sotto particolari ipotesi, i due metodi hanno lo stesso costo e in tale circostanza si preferisce usare il primo rispetto al secondo perchè è più naturale.
Molti risultati relativi al metodo delle potenze standard possono essere estesi alla pencil (A, M) adottando una norma appropriata al posto della norma standard.
Nel caso in cui M sia simmetrica definita positiva, è possibile definire la norma-M^{-1} di un vettore, definita come segue:

$$\|x\|_{M^{-1}} = \sqrt{x^* M^{-1} x}.$$

Teorema 4.15. *Scelti arbitrariamente $u \neq 0$ e σ esiste un autovalore λ di (A, M) tale che*

$$|\lambda - \sigma| \leq \|(A - \sigma M)u\|_{M^{-1}} / \|Mu\|_{M^{-1}}$$

Poichè sappiamo che nel caso standard una scelta ottimale di σ nel metodo delle potenze inverse è rappresentata dal quoziente di Rayleigh, analizziamo le proprietà di quest'ultimo nel caso generalizzato.

Teorema 4.16. *Sia M s.d.p. e gli autovalori della pencil sono ordinati in maniera decrecente, allora il quoziente di Rayleigh gode delle seguenti proprietà:*

1. $\rho(\alpha u) = \rho(u)$, $\alpha \neq 0$,

2. $\rho(u) \in [\lambda_n, \lambda_1]$ *quando u varia nella sfera unitaria,*

3. $\rho(u) = \frac{2(Au-\rho(u)Mu)^*}{u^*Mu}$ *dunque* $\rho(u)$ *è stazionario nei punti che sono autovettori di (A,M)*

4. 4. $\|(A-\sigma M)u\|^2_{M^{-1}} \geq \|Au\|^2_{M^{-1}} - |\rho|^2\|Mu\|^2_{M^{-1}}$ *e vale l'uguaglianza se e solo se* $\sigma = \rho(u)$.

Andiamo ora a definire la quantità $r(u)$ detto residuo tramite la segue:

$$r(u) = Au - \rho(u)Mu$$

e osserviamo che questa è una decomposizione ortogonale di Au rispetto al prodotto scalare indotto da M^{-1} in quanto $\langle r(u), Mu\rangle_{M^{-1}} = 0$.
Sostituendo il quoziente di Rayleigh nella (4.7) ottengo la seguente iterazione:

$$\left(A - \rho^{(k)}M\right) x^{(k+1)} = Mx^{(k)}\tau^{(k)}$$

dove $\tau^{(k)}$ è scelta in modo che $\left\|Mx^{(k)}\right\|_{M^{-1}} = 1$ per ogni k e $\rho^{(k)} = \rho(x^k)$. Passiamo ora a presentare due risultati fondamentali che descrivono il comportamento di convergenza del metodo.

Teorema 4.17. $\left\|\left(A - \rho^{(k+1)}M\right) x^{(k+1)}\right\|_{M^{-1}} \leq \left\|\left(A - \rho^{(k)}M\right) x^{(k)}\right\|_{M^{-1}}$ *per ogni k*

Teorema 4.18. *Per ogni iterato iniziale* $x^{(0)} \neq 0$ *si ha che* $\lim_{k\to\infty} \rho^{(k)} = \rho$ *inoltre, esiste un'autocoppia* (λ, z) *tale che* $(\rho^{(k)}, x^{(k)})$ *converge all'autocoppia* (λ, z) *con velocità di convergenza cubica (cioè* $e^{(k+1)} \leq C\left(e^{(k)}\right)^3$*).*

4.4 Applicazioni numeriche

In questa sezione presentiamo l'implementazione numerica del metodo delle potenze generalizzato shiftato applicato al problema agli autovalori di Laplace in una dimensione. Si considera dunque il nostro solito problema:

$$\begin{cases} -u''(x) = \lambda u(x), & x \in (0,\pi) \\ u(0) = u(\pi) = 0 \end{cases} \tag{4.8}$$

i cui autovalori esatti sono $\lambda_k = k^2$ e le corrispondenti autofunzioni sono $u_k(x) = sin(kx)$, con $k = 1, 2, \ldots$.
Questo problema può essere affrontato sia mediante la formulazione classica che mediante la formulazione mista.

In entrambi i casi, dopo la discretizzazione agli elementi finiti, si perviene ad un problema agli autovalori generalizzato della forma:

$$Au = \lambda Mu \tag{4.9}$$

dove le matrici A e M dipendono dalla formulazione scelta: nel caso classico risultano matrici simmetriche definite positive(rigidezza e massa), mentre nel caso misto si ottiene un sistema a blocchi di tipo saddle-point, che può tuttavia essere ricondotto ad un problema agli autovalori generalizzato s.p.d.(cioè con la matrice a destra s.p.d.).

4.4.1 Formulazione variazionale standard lineare

Sappiamo bene che (4.8) lo possiamo ricondurre ad un problema agli autovalori generalizzato del tipo (4.9) dove

$$A = 1/h \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \\ & & & -1 & 2 \end{bmatrix}$$

e

$$M = h \begin{bmatrix} \frac{4}{6} & \frac{1}{6} & & & \\ \frac{1}{6} & \frac{4}{6} & \frac{1}{6} & & \\ & \frac{1}{6} & \frac{4}{6} & \frac{1}{6} & \\ & & \ddots & \ddots & \ddots \\ & & & \frac{1}{6} & \frac{4}{6} \end{bmatrix}$$

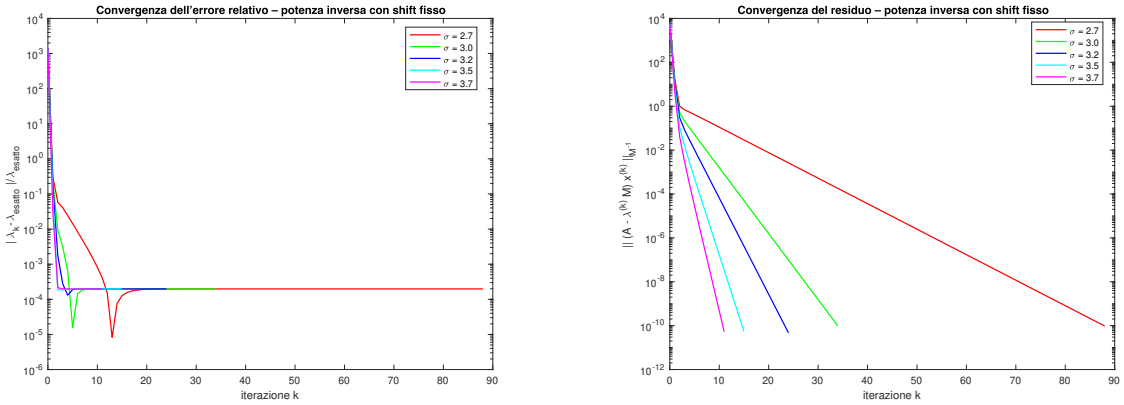
Grazie al primo teorema di Gershgorin possiamo facilmente dimostrare che gli autovalori della matrice A si trovano in $[1/h, 3/h]$ e quelli della matrice M si trovano in $[1/2h, 5/6h]$ quindi sono entrambe definite positive. Ora mostreremo la funzione matlab per la risoluzione del nostro problema agli autovalori generalizzato in questa formulazione, usando il metodo delle potenze generalizzato shiftate la cui equazione fondamentale è (4.7).

```
function [lambda, u, res_list, lam_list,err_rel_list,k] =
    inverse_power_rayleigh_spd1(A, M, x0, s, tol, maxit,lambda_esatto)
% Inverse power con shift fisso per Ax = lambda Mx (A,M simmetriche; M
      0).
% Restituisce:
%   lambda   = stima finale dell'autovalore
%   u        = autovettore normalizzato in norma M (u' M u = 1)
%   res_list = || (A - lambda^(k) M) x^(k) ||_{M^{-1}} ad ogni
    iterazione
%   lam_list = quoziente di Rayleigh ad ogni iterazione
%   err_rel_list=errore relativo ad ogni iterazione
%   k        = numero di iterazioni eseguite
```

```
10
11      % Cholesky M = L * L' (L triangolare inf.)
12      L=Cholesky(M);
13      Lt = L';
14
15      % Normalizza in norma M: x0' M x0 = 1
16      x = x0 / sqrt(x0' * M * x0);
17
18      res_list = zeros(maxit,1);
19      lam_list = zeros(maxit,1);
20      err_rel_list=zeros(maxit,1);
21      E=A-s*M;
22      a=diag(E);
23      b=[0;diag(E,-1)];
24      c=diag(E,1);
25      [L_s,U_s]=LU_tridiag(a,b,c);
26
27      for k = 1:maxit
28          % Quoziente di Rayleigh corrente
29          rho = (x' * A * x);
30          err_rel=abs(rho-lambda_esatto)/abs(lambda_esatto);
31          err_rel_list(k)=err_rel;
32          % Criterio di arresto sull'errore relativo
33          if err_rel_list(k) < tol
34              break;
35          end
36          lam_list(k) = rho;
37
38          % Residuo r^(k) = (A - rho M) x^(k)
39          r = (A - rho * M) * x;
40
41          % Norma M^{-1}: ||r||_{M^{-1}}
42          res_list(k) = norm(RisolviTriangInf(L,r));
43          % Criterio di arresto sul residuo
44          if res_list(k) < tol
45              break;
46          end
47          % Risolvi: (A - sigma M) y = M x  (rhs = L * (L' * x))
48          rhs = L * (Lt * x);
49          v=RisolviTriangInf(L_s,rhs);
50          y=RisolviTriangSup(U_s,v);
51          % Normalizza in norma M
52          x = y / sqrt(y' * M * y);
53      end
54      res_list = res_list(1:k);
55      lam_list = lam_list(1:k);
56      err_rel_list= err_rel_list(1:k);
57
58      % Output finali
59      u = x;
60      lambda = (x' * A * x);
61  end
```

Nella figura (4.1) è riportato il caso in cui il numero di nodi interni è pari a 128. In particolare, le scelte di shift effettuate hanno l'obiettivo di individuare l'autocoppia corrispondente al secondo autovalore, $\lambda_2 = 4$.

Nel grafico a sinistra (4.1a) è mostrata la convergenza dell'errore relativo sugli autovalori $|\lambda^{(k)} - \lambda_{esatto}|/\lambda_{esatto}$ al variare delle iterazioni. Si osserva che per tutti i valori di shift



(a) *Convergenza dell'errore relativo* **(b)** *Convergenza del residuo*

Figura 4.1: *Confronto tra convergenza dell'errore e del residuo (formulazione classica).*

σ considerati l'errore decresce rapidamente fino a stabilizzarsi fino a valori dell'ordine di 10^{-4}, compatibili con la precisione numerica della discretizzazione.
Nel grafico di destra (4.1b) è riportata invece la convergenza del residuo $\left\|\left(A-\rho^{(k)}M\right)x^{(k)}\right\|_{M^{-1}}$. Questo perchè indicando $r^{(k)} = (A-\rho^{(k)}M)x^{(k)}$, sappiamo che $\left\|r^{(k)}\right\|_{M^{-1}} \sim C\cdot\rho^k$, in scala log: $\log\left\|r^{(k)}\right\|_{M^{-1}} \approx \log C + k\log\rho$. Questa è l'equazione di una retta in k e la pendenza della retta è $log(\rho)$ che è negativa poichè $\rho = |\frac{\mu_2}{\mu_1}|$ dove $\mu_j = \frac{1}{\lambda_j-\sigma}$ dove λ_1 e λ_2 sono rispettivamente il primo e il seconodo autovalore più vicino a μ. Da questo possiamo dedurre (e il grafico ce lo conferma) che i valori di σ più vicini all'autovalore cercato portano a un decadimento più rapido del residuo.

4.4.2 Formulazione variazionale mista P1-P0

Sappiamo bene che (4.8) riscrivendo in :

$$\begin{cases} s(x) - u'(x) = 0 & \text{in } \Omega, \\ s'(x) = -\lambda u(x) & \text{in } \Omega, \\ u(0) = u(\pi) = 0. \end{cases} \tag{4.10}$$

lo possiamo ricondurre ad un problema agli autovalori generalizzato con matrici a blocchi del tipo:

$$\begin{bmatrix} A & B^\top \\ B & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = -\lambda \begin{bmatrix} 0 & 0 \\ 0 & M \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

ponendo N il numero di sottointervalli, allora

$$A = \begin{bmatrix} 2 & 1 & & & & \\ 1 & 4 & 1 & & & \\ & 1 & 4 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 4 & 1 \\ & & & & 1 & 2 \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)} \qquad B = \begin{bmatrix} 1 & -1 & 0 & & & \\ 0 & 1 & -1 & 0 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 0 & 1 & -1 \end{bmatrix} \in \mathbb{R}^{N\times(N+1)}$$

$$M = \begin{bmatrix} h & & & \\ & h & & \\ & & \ddots & \\ & & & h \end{bmatrix} \in \mathbb{R}^{N\times N}$$

Sviluppo il sistema a blocchi ed è equivalente a

$$\begin{cases} Ax + B^\top y = 0, \\ Bx = -\lambda M y. \end{cases}$$

Dalla seconda ricavo: $y = -\frac{1}{\lambda} M^{-1} B x$.
Sostituendo nella prima equazione trovo che

$$B^T M^{-1} B x = \lambda A x$$

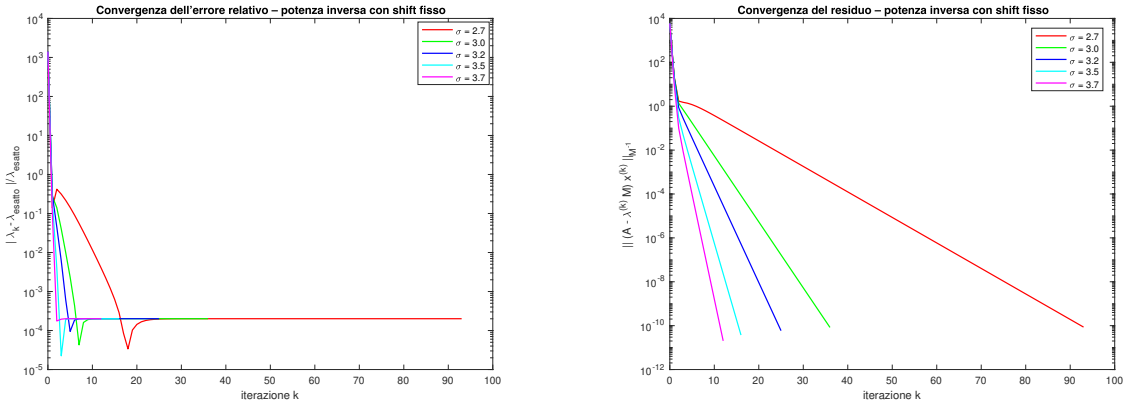
e ponendo $G = B^T M^{-1} B$ ho l'equazione

$$Gx = \lambda A x$$

. Sempre dal primo teorema di Gershgorin possiamo osservare che A è s.d.p. quindi ritrovo il problema generalizzato s.p.d. come promesso.
Ora mostreremo la funzione matlab per la risoluzione di questo problema agli autovalori generalizzato s.p.d. usando il metodo delle potenze generalizzato shiftate con shift fisso come nella sottosezione precedente.

```
function [lambda, u, res_list, lam_list,err_rel_list,k] =
    inverse_power_rayleigh_spd2(A,B, M, x0, s, tol, maxit,lambda_esatto)
    % Cholesky A= L * L' (L triangolare inf.)
    L=Cholesky(A);
    Lt = L';

    % Normalizza in norma M: x0' M x0 = 1
    x = x0 / sqrt(x0' * A * x0);

    res_list = zeros(maxit,1);
```

```
    lam_list = zeros(maxit,1);
    err_rel_list=zeros(maxit,1);
    h=M(1,1);
    G=B'*(1/h*B);
    E=G-s*A;
    a=diag(E);
    b=[0;diag(E,-1)];
    c=diag(E,1);
    [L_s,U_s]=LU_tridiag(a,b,c);


    for k = 1:maxit
        % Quoziente di Rayleigh corrente
        rho = (x' * G * x);
        err_rel=abs(rho-lambda_esatto)/abs(lambda_esatto);
        err_rel_list(k)=err_rel;
        if err_rel_list(k) < tol
            break;
        end
        lam_list(k) = rho;

        % Residuo r^(k) = (G - rho A) x^(k)
        r = (G - rho * A) * x;

        % Norma A^{-1}: ||r||_{A^{-1}}
        res_list(k) = norm(RisolviTriangInf(L,r));
        % Criterio di arresto sul residuo
        if res_list(k) < tol
            break;
        end

        % Risolvi: (G - sigma A) y = A x  (rhs = L * (L' * x))
        rhs = L * (Lt * x);
        v=RisolviTriangInf(L_s,rhs);
        y=RisolviTriangSup(U_s,v);
        % Rinormalizza in norma M
        x = y / sqrt(y' * A * y);
    end
    res_list = res_list(1:k);
    lam_list = lam_list(1:k);
    err_rel_list= err_rel_list(1:k);
    y=-1/rho*(1/h*(B*x));

    % Output finali
    u = y;
    lambda = (x' * G * x);
end
```



(a) *Convergenza dell'errore relativo* **(b)** *Convergenza del residuo*

Figura 4.2: *Confronto tra convergenza dell'errore e del residuo (formulazione classica).*

Nella figura (4.2) è riportato il caso in cui il numero di sottointervalli è pari a 128. In particolare, le scelte di shift effettuate hanno l'obbiettivo di individuare l'autocoppia corrispondente al secondo autovalore, $\lambda_2 = 4$.

Un'analisi del tutto analoga può essere svolta per la formulazione mista $P1-P0$: anche in questo caso l'errore relativo converge fino a valori compatibili con la precisione numerica della discretizzazione e il residuo descresce linearmente in scala logaritmica, con velocità di convergenza data dalla scelta dello shift.

Bibliografia

[1] E. Francini, S. Vessella, *Cenni di analisi funzionale e teoria degli operatori compatti*, Dispense del corso, Università di Firenze, 2010.

[2] J. Nielsen, *An Introduction to Compact Operators*, Honours Seminar, Lakehead University, Thunder Bay (Ontario), 2017.

[3] W. Rudin, *Real and Complex Analysis*, McGraw–Hill, New York, 1987.

[4] M. Manetti, *Algebra lineare per matematici*, Dipartimento di Matematica Guido Castelnuovo, Sapienza Università di Roma, 2024.

[5] D. Boffi, *Finite element approximation of eigenvalue problems*, *Acta Numerica*, vol. 19, pp. 1–120, 2010.

[6] A. Quarteroni, *Numerical Models for Differential Problems*, Springer, Milano, 2009.

[7] V. Simoncini, D. Palitta, *Dispense del corso di Calcolo Numerico*, Modulo di Algebra Lineare Numerica, Università di Bologna, 2022.

[8] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1998.

[9] L. W. Johnson, R. D. Riess, *Numerical Analysis*, Addison–Wesley Publishing Company, Reading (MA), 1982.