Dispense del corso di Metodi Numerici per l'Ingegneria
Appunti di teoria sul Metodo agli Elementi Finiti (FEM)

Mario Putti
Dipartimento di Metodi e Modelli Matematici
per le Scienze Applicate

9 giugno 2014

Indice

1 Introduzione

2 Equazioni differenziali alle derivate parziali

Ci occuperemo in queste note della soluzione numerica di equazioni differenziali alle derivate parziali derivanti da leggi di conservazione. Queste equazioni sono anche chiamate "equazioni in forma di divergenza" dal fatto che l'operatore di divergenza traduce in termini matematici il concetto di conservazione nello spazio. Per esempio, l'equazione di convezione e diffusione, che rappresenta il bilancio di massa di una componente disciolta in acqua che si muove in un campo di moto $\vec{v}$ si scrive come:

$$\frac{\partial u}{\partial t} = \text{div}\, D\nabla u - \text{div}(\vec{v}u) + f \qquad \text{in } \Omega \in \mathbb{R}^3$$

dove u rappresenta la concentrazione del soluto, t è il tempo, div $= \partial/\partial x + \partial/\partial y + \partial/\partial z$ è l'operatore di divergenza (x, y, e z sono le 3 direzioni coordinate spaziali, D è il tensore di dispersione-diffusione, $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z)^T$ è l'operatore di gradiente spaziale (un vettore). Per la risoluzione di tale equazione bisogna specificare le condizioni iniziali e al contorno. Per fare ciò, assumiamo che il contorno $\Gamma = \partial\Omega$ del dominio Ω sia dato dall'unione di tre pezzi Γ_D, Γ_N e Γ_C, per cui abbiamo:

$$\begin{aligned}
u(x,0) &= u_o(x) && x \in \Omega, && t = 0 && \text{(condizioni iniziali)}\\
u(x,t) &= g_o(x) && x \in \Gamma_D, && t > 0 && \text{(condizioni al contorno di Dirichlet)}\\
D\nabla u(x,t)\cdot\vec{n} &= q_N(x) && x \in \Gamma_N, && t > 0 && \text{(condizioni al contorno di Neumann)}\\
(\vec{v}u + D\nabla u(x,t))\cdot\vec{n} &= q_c(x) && x \in \Gamma_C, && t > 0 && \text{(condizioni al contorno di Cauchy)}
\end{aligned}$$

dove $\vec{n}$ è il vettore unitario normale al contorn Γ di Ω e diretto verso l'esterno. Formalmente questa è un'equazione "parabolica".

E' infatti possibile classificare le equazioni differenziali alle derivate parziali (PDE-partial differential equation in analogia con la classificazione delle coniche in geometria di $\mathbb{R}^n$. Per fare ciò scriviamo la generica PDE come:

$$F(x, y, z, u, u_x, u_y, u_z, u_{xx}, u_{xy}, u_{yy}, u_{xz}, u_{zz}, u_{yz}) = 0 \tag{1}$$

dove abbiamo indicato con u_x e u_{xx} le derivate parziali prime e seconde di $u(x, y, z)$ lungo la direzione x, con ovvia estensione alle altre direzioni. Se la F è una funzione lineare della u e delle sue derivate, l'equazione si dice lineare, e si può quindi scrivere in $\mathbb{R}^2$ come:

$$a(x,y) + b(x,y)u + c(x,y)u_x + d(x,y)u_y + e(x,y)u_{xx} + f(x,y)u_{yy} + = 0$$

Tale equazione è a coefficienti variabili, al contrario del caso in cui tutti i coefficienti sono costanti. L'ordine di una PDE è uguale all'ordine della derivata di grado massimo che compare nell'equazione. Ad esempio:

Figura 1: Curva γ con sistema di riferimento locale solidale con la curva

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \qquad 2^o \text{ grado (eq. di Laplace)}$$

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0 \qquad 1^o \text{ grado (eq. di trasporto o convezione)}$$

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \qquad 2^o \text{ grado (eq. di diffusione)}$$

Prendiamo quindi un'equazione a coefficienti costanti considerando per semplicità un dominio bidimensionale e le derivate di ordine massimo (quelle di ordine inferiore non entrano nella caratterizzazione):

$$au_{xx} + bu_{xy} + cu_{yy} + e = 0 \tag{2}$$

Cerchiamo una curva $\gamma(x, y)$ sufficientemente regolare per cui la (2) diventa una equazione alle derivate ordinarie (ODE-ordinary differential equation), e quindi più semplice da risolvere. In realtà' vedremo che non in tutti i casi questo procedimento aiuta la soluzione della PDE.

Scriviamo la curva γ in coordinate parametriche definendo una coordinata σ solidale con la curva γ (Fig. 1), che è quindi definita dalle equazioni parametriche:

$$\begin{aligned} x &= x(\sigma) \\ y &= y(\sigma) \end{aligned}$$

Per la regola di derivazione di funzione composta possiamo scrivere le derivate delle quantità sopra definite sul sistema di riferimento locale:

$$\begin{aligned} \frac{du_x}{d\sigma} &= \frac{\partial u_x}{\partial x}\frac{dx}{d\sigma} + \frac{\partial u_x}{\partial y}\frac{dy}{d\sigma} = u_{xx}\frac{dx}{d\sigma} + u_{xy}\frac{dy}{d\sigma} \\ \frac{du_y}{d\sigma} &= \frac{\partial u_y}{\partial x}\frac{dx}{d\sigma} + \frac{\partial u_y}{\partial y}\frac{dy}{d\sigma} = u_{xy}\frac{dx}{d\sigma} + u_{yy}\frac{dy}{d\sigma} \end{aligned}$$

Ricavando u_{xx} dal sistema precedente e sostituendolo in (2) si ottiene:

$$u_{xy}\left[a\left(\frac{dy}{dx}\right)^2 - b\frac{dy}{dx} + c\right] - \left(a\frac{du_x}{dx}\frac{dy}{dx} + c\frac{du_y}{dx} + e\frac{dy}{dx}\right) = 0$$

Questa equazione è soddisfatta sulla curva γ (è una riscrittura della PDE su $\gamma(\sigma)$). Quindi, scegliendo $\gamma(\sigma)$ in maniera tale da azzerare il primo addendo tra parentesi quadra si ottiene una equazione alle derivate ordinarie nelle incognite u_x e u_y. Si pone cioè:

$$a\left(\frac{dy}{dx}\right)^2 - b\frac{dy}{dx} + c = 0$$

Quindi, l'equazione della curva $\gamma(\sigma)$ si ottiene risolvendo l'ODE che si ricava dalla precedente, e cioè:

$$\frac{dy}{dx} = \frac{b \pm \sqrt{b^2 - 4ac}}{2a}$$

Si vede immediatamente che si possono avere diverse famiglie di curve, dette curve caratteristiche, in funzione del segno del discriminante $b^2 - 4ac$. In analogia con le coniche in $\mathbb{R}^n$ si pone:

- $b^2 - 4ac < 0$: due soluzioni complesse coniugate: equazione di tipo ellittico;
- $b^2 - 4ac = 0$: una soluzione reale: equazione di tipo parabolico;
- $b^2 - 4ac > 0$: due soluzioni reali distinte: equazione di tipo iperbolico.

Vediamo subito alcuni esempi di classificazione di PDE:

- Equazione di Laplace:

  $$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

  $a = c = 1$ $b = 0 \Rightarrow b^2 - 4ac < 0$ è un'equazione ellittica.

- equazione delle onde:

  $$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0$$

  $a = 1$ $b = 0$ $c = -1 \Rightarrow b^2 - 4ac > 0$ è un'equazione iperbolica.

- $$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \qquad\qquad 2^o \text{ grado (eq. di diffusione)}$$

  $a = 1$ $b = c = 0 \Rightarrow b^2 - 4ac = 0$ è un'equazione parabolica.

3 Equazioni ellittiche

3.1 Il caso monodimensionale

Partiamo con un semplice esempio di un problema di Cauchy monodimensionale, formato da un'equazione differenziale del secondo ordine e due condizioni al bordo del dominio:

Problema 3.1 (differenziale)**.**
Trovare $u(x)$ che soddisfa al problema di Cauchy:

$$\begin{aligned} -u''(x) &= f(x), \\ u(0) = u(1) &= 0, \end{aligned} \tag{D}$$

dove abbiamo indicato la derivata prima con $u' = du/dx$ e quella seconda con $u'' = d^2u/dx^2$, e la funzione $f(x)$ è sufficientemente continua perchè l'equazione abbia senso. Si vede immediatamente che il problema è ben posto (ha soluzione unica): basta integrare due volte e imporre le condizioni al bordo, come mostrato nei seguenti passi:

$$\begin{aligned} -\int u''(x)\,dx &= \int f(x)\,dx; \\ -u'(x) &= c_1 + \int f(t)\,dt; \\ -\int u'(x)\,dx &= \int c_1\,dx + \int \left(\int f(t)\,dt \right) dx; \\ u(x) &= c_2 + c_1 x - \int_0^x F(t)\,dt, \end{aligned}$$

dove abbiamo definito il funzionale (funzione di funzione) lineare:

$$F(t) = \int_0^t f(s)\,ds. \tag{3}$$

Usando le condizioni al bordo si calcolano in maniera univoca le costanti c_1 e c_2, ottenendo la soluzione:

$$u(x) = x \left(\int_0^1 F(t)\,dt \right) - \int_0^x F(t)\,dt,$$

che è evidentemente univocamente determinata, mostrando che il problema di Cauchy (D) è ben posto.

Integrando per parti la (3), si ottiene:

$$\int_0^x F(t)\,dt = [tF(t)]_0^x - \int_0^x tF'(t)\,dt = \int_0^x (x-t) f(t)\,dt,$$

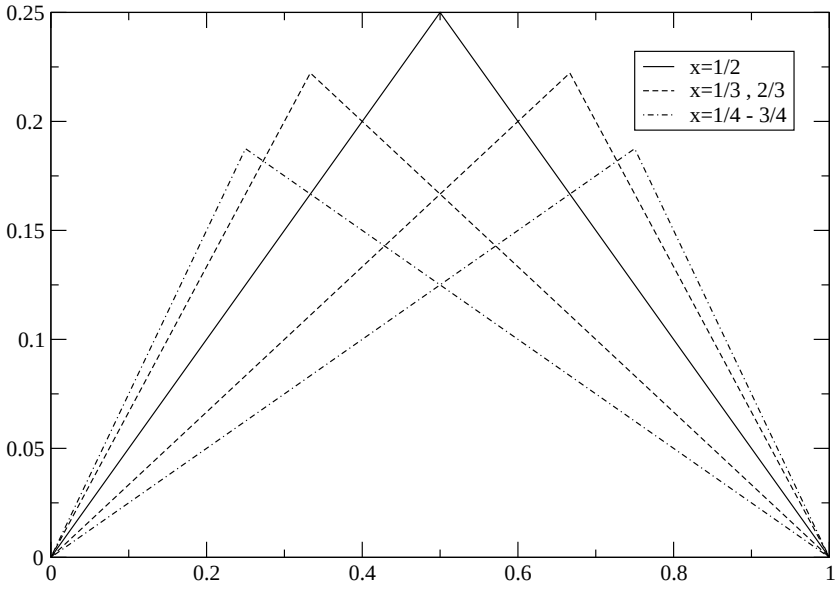

Figura 2: La funzione di Green in corrispondenza a diversi valori di x

.

da cui la soluzione del problema (D) si può scrivere come:

$$u(x) = x \int_0^1 (1-t) f(t)\, dt - \int_0^x (x-t) f(t); dt.$$

Definendo la funzione di Green $G(x,t)$, data da:

$$G(x,t) = \begin{cases} t(1-x), & \text{se } 0 \le t \le x, \\ x(1-t), & \text{se } x \le t \le 1. \end{cases},$$

la soluzione può essere scritta in forma più compatta come:

$$u(x) = \int_0^1 G(x,t) f(t)\, dt.$$

La funzione di Green ha le seguenti proprietà:

- è lineare a t fissato e viceversa;
- è simmetrica, e cioè $G(x,t) = G(t,x)$;
- è continua;
- è non negativa, nulla solo agli estremi dell'intervallo $[0,1]$;

- $\int_0^1 G(x,t)\,dt = \frac{1}{2}x(1-x)$.

La funzione di Green in questione è rappresentata graficamente in Figura 2 per diversi valori di x.

Dal punto di vista fisico, tale problema può rappresentare ad esempio una corda elastica vincolata agli estremi e soggetta ad un carico distribuito, nel qual caso $u(x)$ rappresenta lo spostamento verticale, $\sigma(x)$ rappresenta la tensione sulla corda, e il coefficiente E è il modulo di Young. Il problema (D) si può quindi scrivere come:

$$\begin{aligned} \sigma(x) &= Eu'(x) && \text{Legge di Hook;} \\ -\sigma'(x) &= g(x) && \text{Equilibrio elastico;} \\ u(0) &= u(1) = 0 && \text{Condizioni al bordo.} \end{aligned} \tag{4}$$

Oppure si può pensare a $u(x)$ come la temperatura di una barra soggetta ad una sorgente di calore $g(x)$. In questo caso, indicando con k la conducibilità termica del materiale e con $q(x)$ il flusso di calore, il modello diventa:

$$\begin{aligned} q(x) &= -ku'(x) && \text{Legge di Fourier;} \\ q(x) &= g(x) && \text{Conservazione dell'energia;} \\ u(0) &= u(1) = 0 && \text{Condizioni al bordo.} \end{aligned} \tag{5}$$

Nello stesso modo si può pensare all'equazione della diffusione di una sostanze disciolta in un liquido, nel qual caso si utilizza la legge di Fick, oppure al flusso dell'acqua in un mezzo poroso, nel qual caso si usa la legge di Darcy, eccetera. Più in generale, si può dire che queste equazioni rappresentano il modello di un moto a potenziale.

3.2 Formulazione variazionale

In questo paragrafo, discuteremo brevemente l'approccio variazionale alla soluzione del problema (D), che forma la base per il metodo agli elementi finiti. Per fare questo, introduciamo lo spazio (di funzioni $v(x)$) V, lineare e normato, definito dalla seguente[1]:

$$\begin{aligned} V([0,1]) = \{\, v(x) : &\text{ dove } v(x) \text{ è una funzione continua e limitata nell'intervallo } [0,1], \\ &v'(x) \text{ è una funzione continua a tratti e limitata nell'intervallo } [0,1], \\ &\text{e } v(0) = v(1) = 0 \,\}\,. \end{aligned}$$

In questo spazio si può definire un prodotto interno (o prodotto scalare) tra i suoi elementi:

$$(v,w) = \int_0^1 v(x)w(x)\,dx,$$

[1]Si confrontino le definizioni seguenti con le analoghe definizioni dell'algebra lineare e degli spazi vettoriali.

che definisce il funzionale (o forma) quadratico $F : V \to \mathbb{R}$:

$$F(v) = \frac{1}{2}(v', v') - (f, v) + c.$$

Si possono definire ora il problema di minimizzazione (M) e quello variazionale (V), che vedremo sono in qualche senso da specificare meglio equivalenti al problema differenziale di partenza (D).

Problema 3.2 (di minimizzazione)**.**
Trovare $u \in V$ tale che:

$$F(u) \leq F(v) \qquad \forall v \in V. \tag{M}$$

Problema 3.3 (variazionale)**.**
Trovare $u \in V$ tale che:

$$(u', v') = (f, v) \qquad \forall v \in V. \tag{V}$$

Osservazione 3.4. Con riferimento al problema elastico mostrato in (4), si noti che la quantità $F(v)$ rappresenta l'energia potenziale totale associata allo spostamento ammissibile $v(x) \in V$; il termine $\frac{1}{2}(v', v')$ è l'energia elastica del sistema e il termine (f, v) il potenziale delle forze esterne. Da questa osservazione si deduce immediatamente che il problema (M) è la formulazione del noto "principio della minimizzazione dell'energia potenziale", mentre il problema (V) è la formulazione del "principio dei lavori virtuali".

Equivalenza tra le varie formulazioni (D), (M), (V).

(D)⇒(V) *Dimostrazione* - Dimostriamo che la soluzione di (D) è anche soluzione di (V). Per fare questo basta moltiplicare l'equazione differenziale per una funzione arbitraria $v \in V$ e integrare sul dominio, ottenendo immediatamente:

$$-\int_0^1 u''(x)v(x)\,dx = \int_0^1 f(x)v(x)\,dx,$$

che può essere riscritta usando la notazione di prodotto scalare:

$$-(u'', v) = (f, v).$$

Integrando per parti il primo membro si ottiene:

$$-(u'', v) = -u'(1)v(1) + u'(0)v(0) + (u', v') = (u', v'),$$

che può essere scritto quindi, notando che $v(0) = v(1) = 0$:

$$(u', v') = (f, v) \qquad \forall v \in V \tag{6}$$

che dimostra la tesi. □

(V)⇔(M) *Dimostrazione* - Adesso vogliamo dimostrare che (V) e (M) hanno la stessa soluzione. Supponiamo quindi che $u(x)$ sia soluzione di (V) e sia $v(x) \in V$; calcoliamo la differenza $w(x) = v(x) - u(x) \in V$. Abbiamo facilmente che:

$$\begin{aligned} F(v) = F(u+w) &= \frac{1}{2}(u'+w', u'+w') - (f, u+w) + c \\ &= \frac{1}{2}(u', u') - (f, u) + c + (u', w')(f, w) + \frac{1}{2}(w', w') \geq F(u), \end{aligned}$$

poichè da (6) $(u', w') - (f, w) = 0$ e $(w', w') \geq 0$. Quindi, siccome w è una funzione arbitraria, questo dimostra che u è punto di minimo del funzionale $F(u)$ e quindi è soluzione del problema (M). Anche il contrario è vero. Infatti, se u fosse soluzione di (M), allora per ogni $v \in V$ e $\epsilon \in \mathbb{R}$, si ha:

$$F(u) \leq F(u + \epsilon v),$$

poichè $u + \epsilon v \in V$. Definiamo la funzione differenziabile

$$g(\epsilon) := F(u+\epsilon v) = \frac{1}{2}(u', u') + \epsilon(u', v') + \frac{\epsilon^2}{2}(v', v') - (f, u) - \epsilon(f, v) + c,$$

che ha un minimo in $\epsilon = 0$, da cui $g'(0) = 0$. Facendo i conti si ottiene:

$$g'(0) = (u', v') - (f, v),$$

che dimostra che u è soluzione di (V).

E' anche facile vedere che la soluzione di (V) è unica. Infatti, se $u_1 \in V$ e $u_2 \in V$ sono due soluzioni di (V), allora:

$$\begin{aligned} (u_1', v') &= (f, v) \qquad \forall v \in V; \\ (u_2', v') &= (f, v) \qquad \forall v \in V. \end{aligned}$$

Sottraendo membro a membro e prendendo $v = u_1' - u_2'$, si ottiene immediatamente che

$$\int_0^1 (u_1' - u_2')^2 \, dx = 0,$$

da cui, per la linearità dell'operatore di derivata, risulta che $(u_1 - u_2)(x) = cost$, e siccome $u(0) = u(1) = 0$, tale costante è nulla, da cui la tesi. □

(V)⇒(D). Per dimostrare la tesi ci serve il seguente (caso particolare del) lemma fondamentale del calcolo delle variazioni:

Lemma 3.5. *Sia* $g \in C^0([0,1])$ *e*

$$\int_0^1 g(x) \cdot \phi(x)\, dx = 0 \qquad \forall \phi(x) \in V([0,1]),$$

allora $g(x) = 0$ *per ogni* $x \in [0,1]$.

Per esempio, lo spazio $V([0,1])$ potrebbe essere C^2. *Dimostrazione* - Sia $g(x) \in C^0([0,1])$. Una funzione in $V([0,1])$ è continua e ha le derivate continue (a tratti). Prendiamo allora $r(x)$ una funzione che si annulli in $x = 0$ e $x = 1$ e che sia positiva in $(0,1)$ (ad es. $r = x(1-x)$). Prendiamo quindi $\phi(x) = r(x)g(x)$ una funzione che si annulla agli estremi. Ovviamente $\phi(x) \in V([0,1])$. Si ha quindi:

$$0 = \int_0^1 g(x)\phi(x)\, dx = \int_0^1 r(x)g^2(x)\, dx$$

La funzione integranda è non negativa, quindi deve essere nulla. Siccome $r(x) \neq 0$ per $x \in (a,b)$, dovrà necessariamente essere $g(x) = 0$ per ogni $x \in [a,b]$. Si noti che tale dimostrazione si può estendere con uno sforzo minimo anche a funzioni di $\mathbb{R}^d$. □

Dimostrazione - Ora, per verificare che (V)⇒ (D), assumiamo che $u \in V$ sia soluzione del problema (V). Allora:

$$\int_0^1 u'v'\, dx - \int_0^1 fv\, dx = 0 \qquad \forall v \in V.$$

Assumendo ora che u'' esista e sia continua, possiamo integrare per parti:

$$\int_0^1 u'v'\, dx - \int_0^1 fv\, dx = [u'v]_0^1 - \int_0^1 u''v\, dx - \int_0^1 fv\, dx = 0,$$

da cui, usando le condizioni al contorno omogenee, otteniamo:

$$-\int_0^1 (u'' + f)\, v\, dx = 0 \qquad \forall v \in V.$$

Ora, essendo V uno spazio di funzioni che sono C_0^∞ a tratti (escludendo quindi punti isolati dove si possono avere discontinuità nelle derivate, ma tali discontinuità non contribuiscono agli integrali che dobbiamo calcolare avendo supporto di misura nulla), si può applicare (tratto per tratto) il Lemma 3.5, per cui dovrà essere per forza:

$$-u'' + f = 0.$$

□

Abbiamo quindi dimostrato l'equivalenza del problema variazionale con il problema differenziale. Si noti però che questo è vero sotto l'ipotesi di derivata seconda di u continua. Si può concludere questo paragrafo osservando che tale ipotesi non è richiesta nel problema variazionale, visto che esso richiede solo l'uso di derivate prime. Infatti, utilizzando l'integrazione per parti abbiamo diminuito l'ordine massimo delle derivate presenti nel nostro problema, richiedendo di fatto una grado di continuità inferiore alla nostra soluzione. Riassumendo, le soluzioni del problema differenziale sono sempre anche soluzioni del problema variazionale. Viceversa, le soluzioni del problema variazionale sono soluzioni del problema differenziale se imponiamo una sufficiente continuità nelle derivate seconde.

3.3 Equazioni di Eulero-Lagrange

Si può estendere ad un contesto più generale gli argomenti sopra esposti, per arrivare all'equazione di Eulero-Lagrange del calcolo delle variazioni. Di seguito mostriamo la derivazione di tali equazioni nel caso monodimensionale.

Si vuole trovare una funzione $u(x)$ che soddisfi alle condizioni $u(0) = u(1) = 0$ e minimizzi il funzionale:

$$F(u) = \int_0^1 L(x, u(x), u'(x))\, dx.$$

Assumiamo che L sia sufficientemente continuo in modo tale che le derivate parziali fatte rispetto a x, u e u' esistano. Se $u(x)$ è punto di minimo, allora ogni sua perturbazione deve aumentare il valore di $F(u)$, cioè:

$$F(u) \leq F(u + \epsilon v) \qquad \forall \epsilon \in \mathbb{R} \text{ e } \forall v(x).$$

Sia quindi $w(x) = u(x) + \epsilon v(x)$. Si noti che $v(x)$ dovrà essere presa in modo tale da soddisfare le condizioni $v(0) = v(1) = 0$. Allora:

$$F(w) = F(\epsilon) = \int_0^1 L(\epsilon, x, w(x), w'(x))\, dx.$$

La variazione prima di $F(w)$ è data da:

$$\frac{dF}{d\epsilon} = \frac{d}{d\epsilon} \int_0^1 L(\epsilon, x, w(x), w'(x))\, dx = \int_0^1 \frac{d}{d\epsilon} L(\epsilon, x, w(x), w'(x))\, dx.$$

Usando la regola di derivazione della funzione composta si ottiene:

$$\begin{aligned} \frac{dL(\epsilon)}{d\epsilon} =& \frac{\partial L}{\partial x}\frac{dx}{d\epsilon} + \frac{\partial L}{\partial w}\frac{\partial w}{\partial \epsilon} + \frac{\partial L}{\partial w'}\frac{\partial w'}{\partial \epsilon} \\ =& \frac{\partial L}{\partial w} v(x) + \frac{\partial L}{\partial w'} v'(x). \end{aligned}$$

Quindi:

$$\frac{dF}{d\epsilon} = \int_0^1 \left(\frac{\partial L}{\partial w} v(x) + \frac{\partial L}{\partial w'} v'(x) \right) \, dx.$$

Per $\epsilon = 0$ si ha che $w = u$ e quindi $F(w)|_{\epsilon=0}$ deve essere minima. Quindi:

$$\frac{dF}{d\epsilon}|_{\epsilon=0} = \int_0^1 \left(\frac{\partial L}{\partial w} v(x) + \frac{\partial L}{\partial w'} v'(x) \right) \, dx = 0.$$

Ora, usando il teorema di integrazione per parti otteniamo:

$$\int_0^1 \frac{\partial L}{\partial w} v(x) \, dx + v(x) \frac{\partial L}{\partial w'}|_0^1 - \int_0^1 \frac{d}{dx} \frac{\partial L}{\partial w'} v(x) \, dx = 0.$$

Raccogliendo $v(x)$, notando che il termine agli estremi si annulla perchè si $v(0) = v(1) = 0$, l'applicazione del lemma fondamentale (Lemma 3.5) porta alla cosiddetta equazione di Eulero-Lagrange:

$$\frac{\partial L}{\partial w} - \frac{d}{dx} \left[\frac{\partial L}{\partial w'} \right] = 0.$$

Tale equazione determina la condizione necessaria (non sufficiente) per l'esistenza di un minimo del funzionale $F(u) = \int_0^1 L(x, u, u') \, dx$. se $L(x, u, u')$ è una funzione convessa di u e u', allora l'equazione di Eulero-Lagrange è anche condizione sufficiente.

Esempio 3.6. Come esempio, consideriamo il cosiddetto integrale di Dirichlet:

$$D(x, u, u') = \int_0^1 \frac{1}{2} (u')^2 \, dx.$$

Cerchiamo il minimo di $D(u)$ nella classe di funzioni continue con derivata continua ($e \in C^1([0, 1])$). L'equazione di Eulero-Lagrange la ricaviamo calcolando le derivate di $L(x, u, u') = (u')^2/2$, e cioè:

$$\frac{\partial L}{\partial u} = 0; \qquad \frac{d}{dx} \left[\frac{\partial L}{\partial u'} \right] = \frac{d}{dx} \frac{1}{2} (2u') = u''(x),$$

da cui si ricava subito:

$$u''(x) = 0,$$

e cioè l'equazione di Laplace in una dimensione. Come conseguenza, la soluzione dell'equazione di Laplace è proprio il punto di minimo (il funzionale è convesso) del funzionale $D(x, u, u')$.

Esempio 3.7. Modifichiamo ora il funzionale di Dirichlet scrivendolo come:

$$D(x, u, u') = \int_0^1 \left[\frac{1}{2} (u')^2 - fu \right] \, dx.$$

Derivando opportunamente il funzionale le derivate di $L(x, u, u') = (u')^2/2 + fu$, otteniamo:

$$\frac{\partial L}{\partial u} = f(x); \qquad \frac{d}{dx} \left[\frac{\partial L}{\partial u'} \right] = \frac{d}{dx} (2u') = u''(x).$$

L'equazione di Eulero-Lagrange coincide quindi con l'equazione di Poisson monodimensionale:

$$-u''(x) = f(x).$$

3.4 Formulazione agli elementi finiti

La costruzione di un metodo numerico che risolva il problema (V) si può ricondurre essenzialmente al problema di trovare un opportuno sottospazio $V_h \subset V$ di dimensione finita. Per esempio, possiamo scegliere di lavorare con funzioni lineari a tratti che interpolino la soluzione vera $u(x)$ (incognita) in maniera opportuna. Per fare questo consideriamo una "griglia computazionale o mesh" e cioè una partizione dell'intervallo $I = [0, 1]$ in sotto intervalli aventi estremi di coordinata x_i, $i = 0, 1, \ldots, n + 1$, quindi tali che il generico $I_i = [x_i, x_{i-1}]$ sotto intervallo abbia lunghezza $h_i = x_i - x_{i-1}$, e sia $h = \max_i h_i$ la dimensione caratteristica della mesh (Fig. 3). Costruiamo lo spazio V_h come lo spazio delle funzioni v lineari a tratti, quindi continue con derivata continua a tratti e che appartengono a V, tali che $v(0) = v(1) = 0$. Ricordando l'interpolazione di Lagrange [3], possiamo costruire queste funzioni utilizzando delle funzioni di base per lo spazio V_h che scegliamo per comodità con supporto[2] in ciascun I_i, e lineari a tratti. Queste sono quindi univocamente definita dalla seguente condizione:

$$\phi_j(x_i) = \begin{cases} 1, & \text{se } i = j, \\ 0, & \text{se } i \neq j. \end{cases} .$$

La funzione $v \in V_h$ si può quindi costruire tramite una combinazione lineare delle funzioni di base sui valori nodali:

$$v(x) = \sum_{j=1}^{n} v_j \phi_j(x), \tag{7}$$

dove il coefficiente $v_j = v(x_j)$ è appunto il valore di v in ciascun nodo della mesh. Si osservi che utilizzando una griglia con $n+2$ nodi (estremi inclusi) possiamo definire n funzioni di base, per cui lo spazio V_h risulta avere dimensione n, oltre che essere ovviamente uno spazio lineare. Si noti anche che $V_h = \text{span}\,(\phi_1, \ldots, \phi_n)$.

Possiamo ora scrivere la formulazione del seguente metodo agli *elementi finiti* (FEM, Finite Element Method) nei seguenti modi equivalenti:

[2]Il supporto di una funzione è quel sottoinsieme del dominio dove la funzione è diversa da zero.

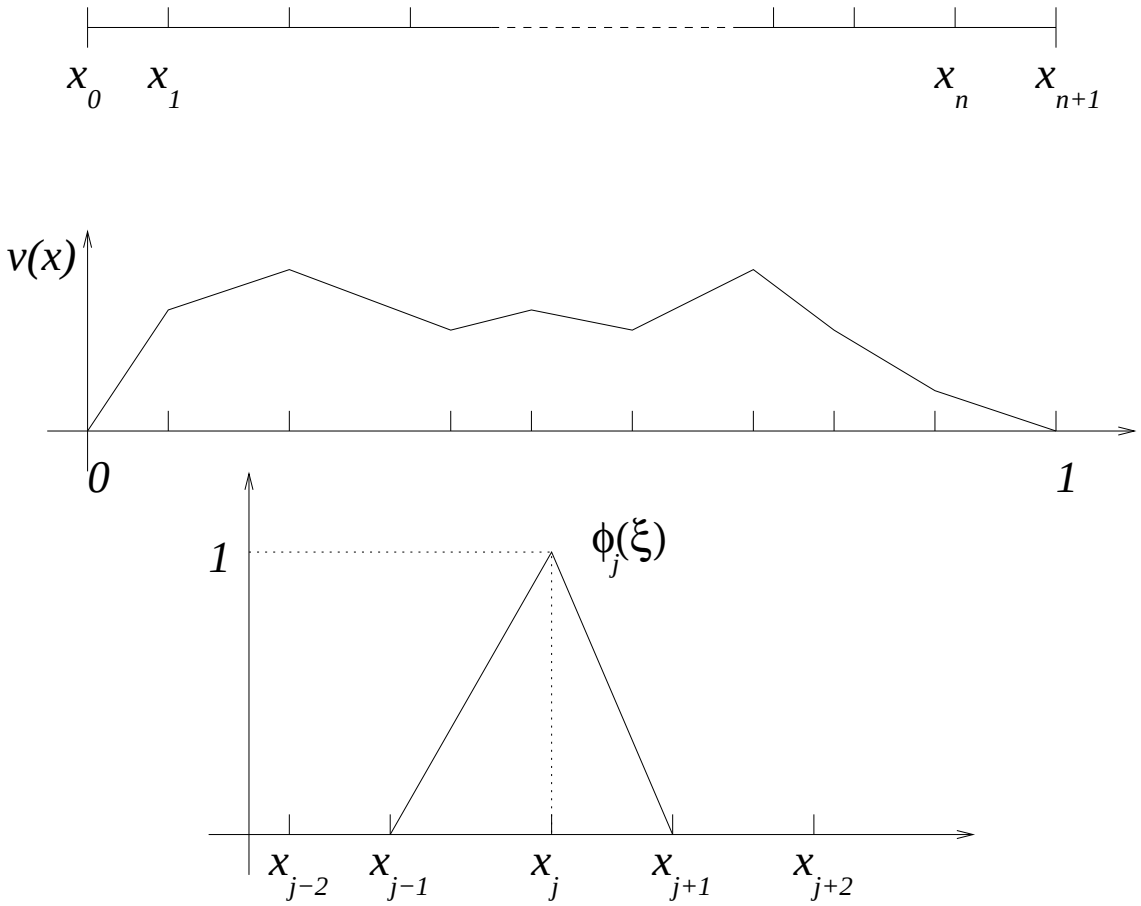

Figura 3: Griglia o mesh computazionale nell'intervallo $I = [0, 1]$ (in alto); esempio di una funzione $v(x) \in V_h$ (centro); esempio di una funzione di base $\phi_j(x)$ (in basso).

Problema 3.8 (metodo di Ritz)**.**
Trovare $u_h \in V_h$ tale che:

$$F(u_h) \leq F(v) \qquad \forall v \in V_h. \tag{Mh}$$

Problema 3.9 (metodo di Galerkin)**.**
Trovare $u_h \in V_h$ tale che:

$$(u_h', v') = (f, v) \qquad \forall v \in V_h. \tag{Vh}$$

Usando la combinazione lineare di funzioni di base per esprimere la generica funzione $v \in V_h$, e cioè l'eq. (7, si vede subito che se $u_h \in V_h$ soddisfa l'equazione (Vh), in particolare soddisfa anche:

$$(u_h', \phi_i') = (f, \phi_i) \qquad i = 1, \ldots, n, \tag{8}$$

e siccome anche u_h può essere scritta come combinazione lineare delle funzioni di base, e cioè:

$$u_h(x) = \sum_{j=1}^{n} u_j \phi_j(x) \qquad u_j = u_h(x_j), \qquad u_h'(x) = \sum_{j=1}^{n} u_j \phi_j'(x) \tag{9}$$

si ottiene immediatamente:

$$\sum_{j=1}^{n} \left(\phi_i', \phi_j'\right) u_j = (f, \phi_i) \qquad i = 1, \ldots, n, \tag{10}$$

che è un sistema lineare $n \times n$. In forma matriciale esso può essere scritto come:

$$Au = b$$

dove la matrice $A_{[n\times n]} = \{a_{ij}\} = \{\left(\phi_i', \phi_j'\right)\}$ è detta matrice di *rigidezza*, il vettore delle incognite è $u_{[n\times 1]} = \{u_i\}$ e il vettore termine noto è $b_{[n\times 1]} = \{b_i\} = \{(f, \phi_i)\}$.

Gli elementi a_{ij} e b_i sono facilmente calcolabili. Infatti, si osservi che $a_{ij} = 0$ per $\mid i - j \mid > 1$, essendo in tale caso i supporti di ϕ_i e di ϕ_j hanno intersezione nulla, per cui $\phi_i(x)\phi_j(x) = 0$ e anche $\phi_i'(x)\phi_j'(x) = 0$. Quindi per $i = 1, \ldots, n$:

$$a_{ii} = (\phi_i', \phi_i') = \int_{x_{i-1}}^{x_i} \frac{1}{h_i^2}\, dx + \int_{x_i}^{x_{i+1}} \frac{1}{h_{i+1}^2}\, dx = \frac{1}{h_i} + \frac{1}{h_{i+1}},$$

e per $i = 2, \ldots, n$:

$$a_{i,i-1} = a_{i-1,i} = \left(\phi_i', \phi_{i-1}'\right) = \left(\phi_{i-1}', \phi_i'\right) = -\int_{x_{i-1}}^{x_i} \frac{1}{h_i^2}\, dx = -\frac{1}{h_i}.$$

La matrice A è quindi tridiagonale e simmetrica. Si può dimostrare anche che A è anche positiva definita. Infatti, per ogni $v(x) = \sum_{j=1}^{n} c_j \phi_j(x)$ si ha immediatamente che:

$$\sum_{i,j=1}^{n} c_i \left(\phi_i', \phi_j'\right) c_j = \left(\sum_{i=1}^{n} c_i \phi_i', \sum_{j=1}^{n} c_j \phi_j'\right) = (v', v') \geq 0.$$

Nella precedente, si verifica l'uguaglianza solo nel caso in cui $v'(x) \equiv 0$, e cioè $v(x) = \text{cost}$, ma tale costante risulta nulla perchè $v(0) = v(1) = 0$. Quindi, raggruppando in un vettore $c = c_i$ le costanti della generica combinazione lineare (7), possiamo scrivere:

$$\left(\sum_{i=1}^{n} c_i \phi_i', \sum_{j=1}^{n} c_j \phi_j'\right) = \langle c, Ac \rangle > 0 \qquad \forall c \in \mathbb{R}^n,\ c \neq 0,$$

che dimostra la positiva definizione di A, e quindi anche che il sistema ha soluzione unica. Si noti che la matrice A è sparsa, cioè ha solo pochi elementi non nulli. Questa è una caratteristica importante, che dipende in maniera fondamentale dal fatto che le funzioni di base hanno supporto compatto e locale. Nel nostro caso monodimensionale infatti ciascuna di esse è diverse da zero solo in due sotto intervalli contigui. Questa caratteristica dovrà essere mantenuta in tutti gli schemi agli elementi finiti, anche in dimensione spaziale maggiore di uno.

Nel caso speciale di griglia uniforme, e cioè con $h_i = h = 1/(n+1)$, e una funzione forzante costante $f(x) = \text{cost}$ il sistema ha la forma speciale:

$$\frac{1}{h^2} \begin{bmatrix} 2 & -1 & 0 & \dots & \dots & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & \dots & \dots & 0 \\ 0 & -1 & 2 & -1 & 0 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & 0 & -1 & 2 & -1 \\ 0 & \dots & \dots & \dots & \dots & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ . \\ . \\ . \\ u_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_n \end{bmatrix},$$

con $b_i = f$.

3.4.1 Studio dell'errore e della convergenza del metodo FEM

Consistenza, convergenza e stabilità La convergenza di del metodo FEM [3] parte dai concetti di consistenza e stabilità di uno schema numerico. Si dice che uno schema è consistente

[3] Lo studio della convergenza di uno schema di discretizzazione spaziale è fondamentale non solo da un punto di vista teorico, ma anche per procedere a confronti tra i diversi schemi e quindi poter scegliere lo schema ottimale per il problema che si deve risolvere. Inoltre, da un confronto tra la convergenza teorica e quella sperimentale, non solo su problemi sufficientemente semplici per avere una soluzione analitica, necessaria per poter calcolare l'errore, si hanno utilissime indicazione sulla correttezza dell'implementazione particolare che si sta utilizzando.

se l'errore commesso dallo schema avendo sostituito la soluzione esatta tende a zero al tendere a zero del passo di discretizzazione. Uno schema è stabile se piccole variazioni dei dati portano a piccole variazioni nei risultati numerici.

Nel nostro caso, indicando con $L(u, f) = 0$ l'equazione alle derivate parziali, dove L è l'operatore differenziale, u la soluzione "vera", f i dati del problema, e con $L_h(u_h, f_h) = 0$ lo schema numerico, con L_h l'operatore discreto, u_h la soluzione numerica, e f_h l'approssimazione numerica dei dati del problema, si dice che lo schema numerico "converge" se

$$\|u - u_h\| \to 0 \qquad h \to 0,$$

dove $\|\cdot\|$ è una norma funzionale opportuna.

Un metodo numerico di discretizzazione si dice "consistente" se

$$L_h(u, g) \to 0 \qquad h \to 0,$$

e si dice "fortemente consistente" se:

$$L_h(u, g) = 0 \qquad \forall h.$$

Spesso non è agevole provare la convergenza di uno schema direttamente, ma si utilizza il risultato fondamentale (teorema di equivalenza) per cui uno schema consistente è convergente se e solo se è anche stabile [8].

D'altro canto, lo studio diretto della convergenza è utile non solo perchè si riesce a verificare che uno schema funziona, ma anche perchè si riesce a quantificare la velocità di convergenza, (in realtà l'ordine con cui l'errore tende a zero al tendere a zero di h) e quindi dall'analisi parallela del costo computazionale del metodo si riesce a prevedere i tempi di calcolo necessari per risolvere un dato problema con un prefissato errore.

Stima dell'errore del metodo FEM Sia $u \in V$ la soluzione al problema (D) e $u_h \in V_h$ la soluzione al problema (Vh). Siccome la (V) vale per ogni $v \in V$ e $V_h \subset V$, la stessa equazione vale anche per le funzioni $v \in V_h$. Sostituendo la stessa $v \in V_h$ nella (V) e nella (Vh), e sottraendo membro a membro si ottiene subito:

$$\begin{aligned} (u', v') &= (f, v) && \forall v \in V_h \\ (u_h', v') &= (f, v) && \forall v \in V_h \\ ((u' - u_h'), v') &= 0 && \forall v \in V_h, \end{aligned} \tag{11}$$

che dimostra che lo schema è fortemente consistente.

Ora dobbiamo usare la nozione di norma di funzione. Pensando alla norma euclidea di vettori, una semplice sua estensione fornisce la seguente norma per funzioni:

$$\|w\| = (w, w)^{\frac{1}{2}} = \left(\int_0^1 w^2 \, dx \right)^{\frac{1}{2}}.$$

Si può verificare facilmente che il simbolo $(v, w) = \int_0^1 vw\ dx$ definisce un prodotto scalare tra due funzioni v e w (soddisfa alle proprietà fondamentali), e dà luogo alla norma funzionale sopra definita. Si ricorda in particolare la proprietà definita come disuguaglianza di Cauchy:

$$| (v, w) | \leq \|v\| \, \|w\|$$

A questo punto è facile dimostrare il seguente risultato:

$$\|(u - u_h)'\| \leq \|(u - v)'\| \qquad \forall v \in V_h \tag{12}$$

Dimostrazione - Assumiamo $\|(u - u_h)'\| \neq 0$. Nel caso la norma fosse nulla, il risultato seguirebbe direttamente.

Sia quindi $v \in V_h$ una funzione arbitraria e chiamiamo $w = u_h - v$, una funzione anch'essa appartenente a V_h e arbitraria. Sfruttando il fatto che dalla (11) il termine $((u - u_h)', w') = 0$ e quindi può essere sommato arbitrariamente, si ottiene:

$$\begin{aligned} \|(u - u_h)'\|^2 &= ((u - u_h)', (u - u_h)') + ((u - u_h)', w') \\ &= ((u - u_h)', (u - u_h + w)') = ((u - u_h)', (u - v)') \\ &\leq \|(u - u_h)'\| \, \|(u - v)'\| \, . \end{aligned}$$

Il risultato segue dividendo per $\|(u - u_h)'\|$, che è stato assunto non nullo. □

Si può anche dimostrare che $\|v\| \leq \|v'\|$ per ogni $v \in V_h$, ovvero:

$$\int_0^1 v^2\ dx \leq \int_0^1 (v')^2\ dx \qquad \forall v \in V_h.$$

Si noti che questo è vero perchè si è richiesto nella definizione di V_h che le funzioni assumano valori nulli agli estremi dell'intervallo. Infatti:

$$v(x) = v(0) + \int_0^x v'(t)\ dt = \int_0^x v'(t)\ dt,$$

da cui, usando ancora la disuguaglianza di Cauchy:

$$| v(x) | \leq \int_0^1 | v' |\ dx \leq \left(\int_0^1 1^2\ dx \right)^{\frac{1}{2}} \left(\int_0^1 | v' |^2\ dx \right)^{\frac{1}{2}} \leq \left(\int_0^1 | v' |^2\ dx \right)^{\frac{1}{2}},$$

da cui, integrando tra 0 e 1, si ottiene infine:

$$\int_0^1 | v(x) |^2\ dx \leq \int_0^1 \left(\int_0^1 | v'(x) |^2\ dx \right)\ dy = \int_0^1 | v'(x) |^2\ dx.$$

L'applicazione di tale risultato alla funzione $v = u - u_h$ dimostra che

$$\|u - u_h\| \leq \|(u - u_h)'\| \leq \|(u - v)'\| \qquad \forall v \in V_h \tag{13}$$

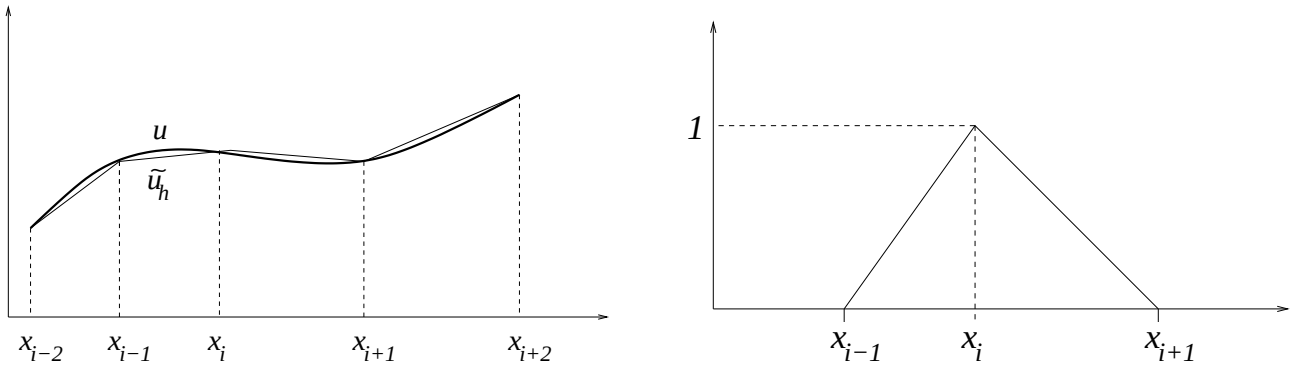

Figura 4: Interpolante $\tilde{u}_h$ (a sinistra), funzione di base $\phi_i(x)$ (a destra).

Osservazione 3.10. Il risultato precedente ci dice che u_h' è la migliore approssimazione di u' in V_h, essendo l'errore minore rispetto a quello commesso da qualsiasi altra funzione $v \in V_h$.

Il risultato precedente ci dice che se riusciamo a dare una maggiorazione della norma della differenza tra u e una qualsiasi funzione $v \in V_h$, questa maggiora anche l'errore del metodo FEM. Quindi si può ricavare una stima quantitativa dell'errore $\|(u - u_h)'\|$ passando attraverso la stima dell'errore commesso andando a prendere al posto di v in (12) una funzione opportuna, Per fare questo, scegliamo di lavorare con la funzione $\tilde{u}_h \in V_h$, una "interpolante" lineare a tratti che interpola la u sui punti della griglia. Si dice che la funzione $\tilde{u}_h$ è una interpolante di $u(x)$, ovvero $\tilde{u_h}$ interpola $u(x)$ nei nodi $x_i,\ i = 0, \ldots, n+1$, se valgono le seguenti relazioni:

$$\tilde{u}_h(x_i) = u(x_i) \qquad\qquad i = 0, \ldots, n+1.$$

Il concetto è mostrato graficamente in Figura 4, a sinistra.

E' facile definire tale interpolante utilizzando i polinomi di Lagrange [3, 8]. Di seguito, per completezza, ricaviamo tutti i risultati senza ricorrere a tali polinomi. Si vuole dunque interpolare una funzione $v(x)$ generica tramite un polinomio lineare a tratti. Tale polinomio si può scrivere come:

$$P_1(x) = \sum_{i=1}^{n} a_i \phi_i(x).$$

La generica funzione di base nel nodo i-esimo è data da:

$$\phi_i(x) = \begin{cases} \frac{x - x_i}{x_i - x_{i-1}}, & \text{se } x_{i-1} \le x \le x_i, \\ \frac{x_{i+1} - x}{x_{i+1} - x_i}, & \text{se } x_i \le x \le x_{i+1}. \end{cases}$$

Tale funzione, mostrata in Figura 4 (a destra), è effettivamente una funzione di base dell'interpolazione. Le seguenti proprietà, che mostrano che $P_1(x)$ è proprio il polinomio cercato, sono

facilmente verificabili:

$$\phi_i(x) = \begin{cases} 1, & \text{se } x = x_i, \\ 0, & \text{se } x = x_j, \quad i \neq j. \end{cases}$$
$$P_1(x_i) = a_i = v(x_i)$$
$$P_1'(x_i) = v'(x_i)$$

Definiamo ora l'errore di interpolazione $e(x) = v(x) - P_1(x)$. Si noti che essendo $P_1(x)$ lineare a tratti, $P_1''(x) = 0$ in tutto l'intervallo $[0, 1]$. E' chiaro che $e(x_i) = 0$ in tutti i punti di appoggio (i nodi della griglia) $x_i, \quad i = 0, \ldots, n+1$. Quindi, per il teorema di Rolle, esistono n punti $\eta_i,\ i = 1, \ldots, n$ con $\eta_i \in [x_i, x_{i+1}]$ dove $e'(\eta_i) = 0$. Possiamo quindi scrivere per $x_i \leq x \leq x_{i+1}$:

$$e'(x) = \int_{\eta_i}^{x} e''(t)\, dt = \int_{\eta_i}^{x} v''(t)\, dt,$$

da cui:

$$\begin{aligned} \mid e'(x) \mid &\leq \int_{x_i}^{x_{i+1}} \mid v''(t) \mid dt = \int_{x_i}^{x_{i+1}} 1 \cdot \mid v''(t) \mid dt \leq \text{(per la disuguaglianza di Cauchy)} \\ &\leq \left(\int_{x_i}^{x_{i+1}} 1^2\, dt \right)^{\frac{1}{2}} \left(\int_{x_i}^{x_{i+1}} \mid v''(t) \mid^2 dt \right)^{\frac{1}{2}} = h^{\frac{1}{2}} \left(\int_{x_i}^{x_{i+1}} \mid v''(t) \mid^2 dt \right)^{\frac{1}{2}}, \end{aligned} \tag{14}$$

da cui si ricava subito:

$$\mid e'(x) \mid^2 \leq h \left(\int_{x_i}^{x_{i+1}} \mid v''(t) \mid^2 dt \right).$$

Integrando la precedente tra x_i e x_{i+1}, si ha:

$$\int_{x_i}^{x_{i+1}} \mid e'(x) \mid^2 dx \leq h^2 \int_{x_i}^{x_{i+1}} \mid v''(t) \mid^2 dt.$$

Per valutare $e(x)$, si noti che $e(x) = \int_{x_i}^{x} e'(t)\, dt$. Quindi, usando la (14) e integrando, si ottiene:

$$\mid e(x) \mid \leq h^{\frac{3}{2}} \left(\int_{x_i}^{x_{i+1}} \mid v''(t) \mid^2 dt \right)^{\frac{1}{2}},$$

da cui si ricava:

$$\int_{x_i}^{x_{i+1}} \mid e(x) \mid^2 dx \leq h^4 \int_{x_i}^{x_{i+1}} \mid v''(t) \mid^2 dt.$$

Sommando ora su tutti i sotto intervalli che formano la griglia computazionale si ottengono le seguenti stime dell'errore di interpolazione:

$$\begin{aligned}\left(\int_0^1 \mid e(x) \mid^2\right)^{\frac{1}{2}} &\leq h^2 \left(\int_0^1 \mid v''(x) \mid^2 dx\right)^{\frac{1}{2}} \\ \left(\int_0^1 \mid e'(x) \mid^2\right)^{\frac{1}{2}} &\leq h \left(\int_0^1 \mid v''(x) \mid^2 dx\right)^{\frac{1}{2}}\end{aligned}$$

ovvero, in termini di norme:

$$\begin{aligned}\|v - P_1(x)\| &\leq h^2 \|v''(x)\| \\ \|v' - P_1'(x)\| &\leq h \|v''(x)\|\end{aligned}$$

Utilizzando ora (12) e (13) si ottengono le seguenti stime dell'errore:

$$\|u - u_h\| \leq h \|u''\| \tag{15}$$

$$\|(u - u_h)'\| \leq h \|u''\| \tag{16}$$

che dimostrano che, se la derivata seconda della soluzione vera ha norma limitata, l'errore dello schema FEM converge a zero al tendere a zero del passo di griglia h. Si noti che con qualche sforzo in più si può dimostrare che l'errore sulla soluzione $u(x)$ converge a zero quadraticamente, anzichè linearmente come l'errore sulla derivata, e cioè:

$$\|u - u_h\| \leq h^2 \|u''\| . \tag{17}$$

Di nuovo se la soluzione vera non ha derivata seconda limitata l'ordine quadratico è perso.

Osservazione 3.11. Per poter apprezzare quest'ultimo fatto, bisognerebbe ricorrere alla definizione di integrale di Lebesgue. Tale definizione esula dagli ambiti di queste note e si rimanda il lettore interessato a testi più specializzati [7, 6]. Per il momento basti pensare a funzioni continue, con derivate sufficientemente lisce non necessariamente limitate il cui quadrato abbia integrale finito (funzioni di quadrato sommabile), in modo tale che i prodotti scalari e le norme integrali usate siano ben definite.

Osservazione 3.12. Dalla stima dell'errore si può ricavare una stima dell'indice di condizionamento spettrale della matrice di rigidezza A (simmetrica e definita positiva) del metodo FEM. Infatti si ha:

$$\kappa(A) = \frac{\lambda_1}{\lambda_N} = Ch^{-2}$$

dove si è indicato rispettivamente con λ_1 e con λ_N gli autovalore massimo e minimo di A, e la costante C non dipende da h. Se si utilizzasse il metodo del gradiente coniugato per risolvere il sistema lineare, sarebbe possibile stimare l'indice di condizionamento e quindi il numero di iterazioni necessarie al metodo del gradiente coniugato per ottenere una soluzione con una prefissata tolleranza. Analogamente, è possibile tramite questo risultato stabilire il variare del numero di iterazioni impiegate dal gradiente coniugato per arrivare alla convergenza al variare della dimensione della mesh.

Alcuni esempi semplici Si consideri il problema di Cauchy:

$$\begin{aligned} -u''(x) &= q \qquad\qquad x \in [0,1], \\ u(0) &= u(1) = 0. \end{aligned}$$

Si consideri il funzionale:

$$F(u) = \int_0^1 \left[\frac{1}{2}(u')^2 - qu\right]\, dx,$$

con una soluzione approssimata data da:

$$u_n(x) = \sum_{j=1}^{n} a_j \phi_j(x).$$

La minimizzazione del funzionale (metodo di Ritz) richiede che la soluzione sia un punto di stazionarietà per $F(u)$, talchè si ottiene un sistema lineare (uguale a quello che si otterrebbe con l'approccio di Galerkin), la cui i-esima equazione è data da:

$$\frac{\partial F}{\partial a_i} = \int_0^1 \left[\left(\sum_{j=1}^{n} a_j \phi_j'(x)\right)\phi_i'(x) - q\phi_i(x)\right]\, dx = 0.$$

Dobbiamo ora scegliere le funzioni di base $\phi_i(x) \in V_h$.

Esempio 3.13. Scegliamo come funzioni di base la base canonica dello spazio dei polinomi di grado n:

$$\phi_i(x) = x^i \qquad\qquad i = 0, 1, \ldots, n-1.$$

La nostra soluzione numerica può quindi essere scritta:

$$u_n(x) = x(x-1)\sum_{i=1}^{n} a_i x^{i-1}$$

dove i primi due monomi sono stati inseriti per poter soddisfare le condizioni al contorno. Si noti che si hanno le seguenti funzioni:

$$\begin{aligned} \phi_1(x) &= x(x-1) \\ \phi_1'(x) &= 2x - 1 \\ \ldots &\quad \ldots \\ \phi_i(x) &= x(x-1)x^{i-1} = x^{i-1} - x^i \\ \phi_i'(x) &= (i+1)x^i - ix^{i-1} \\ \ldots &\quad \ldots \end{aligned}$$

Per $n = 1$ si ha che $i = 1$, da cui:

$$u_n(x) = x(x-1)a_1$$
$$u_n'(x) = 1(x-1)a_1$$
$$\begin{aligned}\frac{\partial F}{\partial a_1} &= \int_0^1 \left[a_1(2x-1)^2 - qx(x-1)\right]\, dx \\ &= \int_0^1 \left[a_1(4x^2+1-4x) - qx^2 + qx\right]\, dx = 0,\end{aligned}$$

da cui si ricava immediatamente $a_1 = -q/2$, che sostituito nella soluzione numerica mi fornisce:

$$u_n(x) = -x(x-1)\frac{q}{2}.$$

Derivando due volte si vede che immediatamente la $u_n(x)$ soddisfa l'equazione differenziale di partenza, e quindi è la sua soluzione esatta e certamente $a_2 = a_3 = \ldots = a_n = 0$.

Esempio 3.14. Sia

$$u_n(x) = \sum_{i=1}^{n} a_i \sin(i\pi x)$$

da cui le funzioni di base sono individuate da:

$$\phi_i(x) = \sin(i\pi x) \qquad \phi_i'(x) = i\pi\cos(i\pi x).$$

Il sistema lineare (metodo di Ritz) diventa:

$$\frac{\partial F}{\partial a_i} = \int_0^1 \left[\left(\sum_{j=1}^{n} a_j\phi_j'(x)\right)\phi_i'(x) - q\phi_i(x)\right]\, dx = 0,$$

da cui, risolvendo per a_1 nel caso $n = 1$, si ottiene:

$$\frac{\partial F}{\partial a_1} = \int_0^1 \left[a_1\pi^2\cos^2(\pi x) - -q\sin(\pi x)\right]\, dx = 0,$$

che fornisce la seguente espressione:

$$a_1 = \frac{\int_0^1 q\sin(\pi x)\, dx}{\int_0^1 \pi^2\cos(\pi x)\, dx} = \frac{4}{\pi^3}q.$$

La soluzione numerica è quindi data da:

$$u_n(x) = \frac{4}{\pi^3}q\sin(\pi x)$$

Un confronto tra la soluzione numerica e quella analitica è data nella seguente tabella:

x	u/q	u_n/q
0.00	0.00	0.00
0.25	0.09375	0.09122
0.50	0.125	0.12901
0.75	0.09375	0.09122
1.00	0.00	0.00

Esempio 3.15. Sia:

$$u_n(x) = \sum_{i=1}^{n} a_i \sin(2\pi i x)$$

In questo caso risulta $a_1 = a_2 = \ldots = a_n = 0$. In realtà lo spazio V_h generato dalle funzioni di base $\phi_i(x) = \sin(2\pi i x)$ non contiene la soluzione analitica del problema, e quindi lo schema calcola la soluzione identicamente nulla.

3.5 Estensione al caso multidimensionale

3.5.1 Operatori differenziali

Sia $\Omega \subset \mathbb{R}^d$, e u una funzione $u : \Omega \to \mathbb{R}$.

Il gradiente. Il gradiente di u è un vettore d-dimensionale delle derivate prime di u:

$$\nabla u = (\frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_d}).$$

Divergenza. Dato un campo vettoriale $q(x) \in \mathbb{R}^d$, si definisce la divergenza del vettore q come il prodotto scalare tra i vettori q e l'operatore di ∇, quindi:

$$\operatorname{div} q = \langle \nabla, q \rangle = \nabla \cdot q = \frac{\partial q_1}{\partial x_1} + \ldots + \frac{\partial q_n}{\partial x_n}.$$

Laplaciano. Si definisce il laplaciano di u la funzione:

$$\Delta u = \operatorname{div} \nabla u = \langle \nabla, \nabla u \rangle = \nabla \cdot \nabla u.$$

curl (rotore) Il rotore di q il prodotto vettoriale tra i vettori gradiente e q. Come tale, è definito solo per $d = 3$, e risulta essere:

$$\operatorname{rot} q = \nabla \times q = (\frac{\partial q_3}{\partial x_2} - \frac{\partial q_3}{\partial x_3}, \frac{\partial q_2}{\partial x_3} - \frac{\partial q_3}{\partial x_1}, \frac{\partial q_2}{\partial x_1} - \frac{\partial q_1}{\partial x_2}).$$

Derivate di ordine superiore. Introduciamo la notazione "multi-indice": sia $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d) \in \mathbb{N}^d$ un multi-indice di ordine $k \mid \alpha \mid = k = \sum_{i=1}^{d} \alpha_i$. Sia:

$$\partial^\alpha u = \frac{\partial^{\mid \alpha \mid} u}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}.$$

Dato un intero $k \geq 0$, denotiamo con ∂^k l'insieme di tutte le derivate di u di ordine k: $\partial^k u = \{\partial^\alpha, \mid \alpha \mid = k\}$.

Derivatata debole. Come abbiamo visto nel caso mono-dimensionale, avremo spesso bisogno di condizioni di continuità "rilassate" per le funzioni e le loro derivate. Si parla allora di "derivata debole" o "nel senso delle distribuzioni" o "generalizzata" sfruttando il teorema di integrazione per parti.

Definizione 3.16. Date due funzioni $u, v : \Omega \to \mathbb{R}$ e un multi-indice α. Allora $v = \partial^\alpha u$ è derivata debole di u se per ogni funzione continua $\phi \in C^\infty(\Omega)$ a supporto compatto (quindi nulla in $\partial\Omega$) si ha:

$$\int_\Omega v\phi \, dx = (-1)^{|\alpha|} \int_\Omega u \partial^\alpha \phi \, dx.$$

E' chiaro che la derivata debole concide con la derivata classica, quando quest'ultima esiste.

Teorema di Gauss o della divergenza. Lo strumento principale che utilizzeremo in questo paragrafo è la formula (o lemma) di Green, ovvero il procedimento di integrazione per parti multidimensionale. La formula di Green deriva dal teorema della divergenza (o di Gauss): dato un dominio Ω compatto e con bordo $\Gamma = \partial\Omega$ sufficientemente liscio e un campo vettoriale $\vec{F}(x) \in \Omega$, si ha che:

$$\int_\Omega \operatorname{div} \vec{F} \, dx = \int_\Gamma \vec{F} \cdot \vec{n} \, ds, \tag{18}$$

dove $\vec{n}$ è il vettore normale unitario esterno a Γ, dx denota la misura di volume su Ω (in $\mathbb{R}^d$) e ds la misura di superficie su Γ (in $\mathbb{R}^{d-1}$), e $\vec{F} \cdot \vec{n}$ indica il prodotto scalare tra due vettori di $\mathbb{R}^d$. Applichiamo il teorema di Gauss ad un campo vettoriale opportuno, $\vec{F} = v\vec{q}$, dato dal prodotto di un campo vettoriale $\vec{q}(x)$ per una funzione $v(x)$. Utilizzando la regola di derivazione del prodotto dopo aver sviluppato componente per componente il prodotto scalare, si ottiene:

$$\int_\Omega \nabla v \cdot \vec{q} \, dx = \int_\Gamma v \, \vec{q} \cdot \vec{n} \, ds - \int_\Omega v \, \operatorname{div} \vec{q} \, dx.$$

Nel caso particolare in cui $\vec{q} = \nabla w$, si ottiene la prima identità o lemma di Green:

$$\int_\Omega \nabla v \cdot \nabla w \, dx = \int_\Gamma v \, \nabla w \cdot \vec{n} \, ds - \int_\Omega v \, \Delta w \, dx, \tag{19}$$

che intuitivamente può essere pensata come una formula di integrazione per parti in domini multidimensionali, notando che v è una primitiva di ∇v, e $\Delta w = \operatorname{div} \nabla w$ è la derivata di ∇w.

3.5.2 Formulazioni deboli e FEM

Consideriamo ora l'equazione di Poisson nel caso d-dimensionale, con $d = 2$ o 3:

Problema 3.17 (differenziale)**.**
Trovare $u(x)$ che soddisfa al problema al contorno:

$$\begin{aligned} -\Delta u &= f(x), && x \in \Omega \subset \mathbb{R}^d \\ u(x) &= 0 && x \in \Gamma, \end{aligned} \tag{20}$$

dove $\Omega \subset \mathbb{R}^d$ è un dominio limitato di $\mathbb{R}^d = \{x = [x_1, x_2, \ldots, x_d], x_i \in \mathbb{R}\}$ avente contorno $\Gamma = \partial\Omega$, assunto sufficientemente liscio, e Δ è l'operatore Laplaciano definito da:

$$\Delta = \operatorname{div} \nabla = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}.$$

La formulazione variazionale per il problema (20) si scrive nel modo seguente:

Problema 3.18 (variazionale)**.**
Trovare $u \in V$ tale che:

$$a(u, v) = (f, v) \qquad \forall v \in V, \tag{21}$$

dove:

$$\begin{aligned} a(u, v) &= \int_\Omega \nabla u \cdot \nabla v \, dx \\ (f, v) &= \int_\Omega f \, v \, dx \\ V &= \left\{ v(x) : v \text{ è continua in } \Omega, \frac{\partial v}{\partial x_i} \text{ sono continue in } \Omega \; \forall i, \text{ e } v(x) = 0 \text{ per } x \in \Gamma \right\}. \end{aligned}$$

Per vedere come tale formulazione variazionale segue dal problema differenziale di partenza, moltiplichiamo la (20) per una funzione test arbitraria $v(x) \in V$ e integriamo su Ω. Usando la formula di Green si ottiene:

$$(f, v) = -\int_\Omega (\Delta u) v \, dx = -\int_\Gamma v \nabla u \cdot \vec{n}; ds + \int_\Omega \nabla u \cdot \nabla v \, dx = a(u, v),$$

dove l'integrale al bordo è nullo perchè $v(x) = 0$ per $x \in \Gamma$. In modo del tutto analogo al caso mono-dimensionale, si vede che:

- la soluzione del problema variazionale è soluzione del problema differenziale se si assume che $u(x)$ sia sufficientemente regolare;

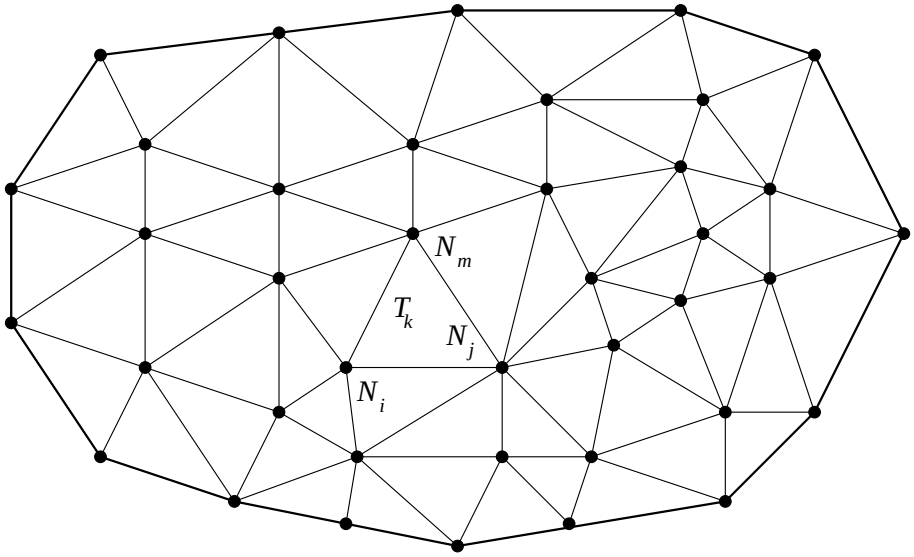

Figura 5: Un esempio di una triangolazione ammissibile di un dominio Ω. Il contorno $\Gamma = \partial\Omega$ è evidenziato con la linea più spessa.

- il problema variazionale è equivalente al seguente problema di minimizzazione:

  Problema 3.19 (di minimizzazione)**.**
  Trovare $u \in V$ tale che:

  $$F(u) \leq F(v) \qquad \forall v \in V, \tag{22}$$

  dove:

  $$F(v) = \frac{1}{2} a(u, v) - (f, v).$$

Bisogna ora definire opportunamente le funzioni di base, e per fare ciò bisogna prima costruire la griglia computazionale, cioè un opportuno partizionamento del dominio Ω. Nel caso bi-dimensionale (i.e., $d = 2$, $\Omega \subset \mathbb{R}^2$) possiamo procedere definendo una triangolazione che partiziona Ω in un insieme $\mathcal{T}_h$ di triangoli T_k con le seguenti proprietà:

- $\mathcal{T}_h$ è formata da n nodi (i vertici dei triangoli, indicati con il simbolo N_i, $i = 1, \ldots, n$, che è sostanzialmente il vettore delle coordinate dell'i-esimo nodo) e m triangoli (indicati con T_k, $k = 1, \ldots, m$);
- $\Omega = \bigcup_{T_k \in \mathcal{T}_h} T_k = T_1 \cup T_2 \ldots \cup T_m$;
- $T_i \cap T_j = e_{ij}$, $i \neq j$, dove e_{ij} indica il lato in comune ai triangoli T_i e T_j;

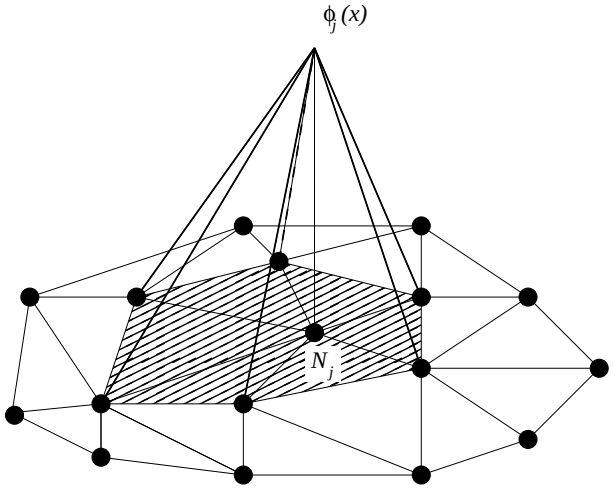

Figura 6: Funzione di base lineare $\phi_j(x) \in V_h$.

- nessun vertice di nessun triangolo giace all'interno di un lato;
- i triangoli di "bordo o di contorno" hanno almeno un vertice nel contorno $\Gamma = \partial\Omega$.

Un esempio di triangolazione ammissibile è riportato in Figura 5. Si noti come perchè le stime teoriche di convergenza siano efficaci, bisogna che la geometria del contorno non vari al variare della mesh. Per questo motivo si è disegnato un dominio con un contorno formato da una spezzata (lineare a tratti).

Si introduce ora il parametro di mesh h definito da:

$$h = \max_{T_i \in \mathcal{T}_h} \text{diam } (T_i), \tag{23}$$

dove il diametro del triangolo T_i indicato con diam (T_i) è il lato di lunghezza massima di T_i. Lo spazio di funzioni finito-dimensionale V_h è quindi definito da:

$$V_h = \{v(x) : v \text{ è continua in } \Omega, v|_{T_i} \text{ è lineare su ciascun } T_i \in \mathcal{T}_h, v(x) = 0 \text{ per } x \in \Gamma\}.$$

dove $v|_{T_i}$ è la restrizione della funzione test $v(x)$ al triangolo T_i[4]. Si noti che $V_h \subset V$. Per usare l'interpolazione lagrangiana consideriamo come punti di appoggio i nodi N_i della triangolazione escludendo quelli di contorno dove $v(x) = 0$. Le funzioni di base $\phi_i(x)$, $i = 1, \ldots, n$ sono quindi definite dalle seguenti condizioni:

$$\phi_i(x_j) = \begin{cases} 1, & \text{se } i = j, \\ 0, & \text{se } i \neq j. \end{cases} \qquad i, j = 1, \ldots, n$$

[4]La restrizione a T_i di $v(x)$ è una funzione definita in T_i e coincidente con $v(x)$ in T_i

Esse sono funzioni piramidali, come mostrato in Figura 6, che hanno come supporto tutti gli elementi che hanno il nodo j in comune. La generica funzione $v \in V_h$ può essere rappresentata nel seguente modo:

$$v(x) = \sum_{j=1}^{n} \eta_j \phi_j(x), \qquad \eta_j = v(x_j),$$

e infine si può scrivere il problema agli elementi finiti alla Galerkin come:

Problema 3.20 (metodo di Galerkin)**.**
Trovare $u_h \in V_h$ tale che:

$$a(u_h, v) = (f, v) \qquad \forall v \in V_h. \tag{24}$$

Sostituendo ora l'espansione in termini delle funzioni di base (esattamente come fatto nella (9) nel caso 1D) si trova il seguente sistema lineare:

$$\sum_{j=1}^{n} a(\phi_i, \phi_j)\, u_j = (f, \phi_i) \qquad i = 1, \ldots, n, \tag{25}$$

che fornisce il sistema lineare degli elementi finiti, che in forma matriciale può essere scritto di nuovo come:

$$Au = b$$

dove ora la matrice di rigidezza, il vettore delle incognite e il vettore termini noti sono dati da:

$$A_{[n\times n]} = \{a_{ij}\} \qquad a_{ij} = a(\phi_i, \phi_j) = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \, d\Omega \tag{26}$$

$$u_{[n\times 1]} = \{u_i\}, \qquad b_{[n\times 1]} = \{b_i\} \qquad b_i = (f, \phi_i) = \int_\Omega f\phi_i \, d\Omega. \tag{27}$$

Notiamo che l'espressione del prodotto scalare ora coinvolge un integrale multidimensionale definito sul dominio Ω. Procedendo in modo del tutto analogo al caso 1D, si dimostra che la matrice A è simmetrica, sparsa e definita positiva.

Nel caso di dominio quadrato discretizzato con triangoli rettangoli aventi i cateti di lunghezza h, come mostrato in Figura 7, la matrice diventa penta-diagonale, e il sistema assume la seguente forma:

$$\begin{bmatrix} 4 & -1 & 0 & 0 & 0 & -1 & \ldots & \ldots & \ldots & 0 \\ -1 & 4 & -1 & 0 & 0 & 0 & -1 & \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & -1 & \ldots & -1 & 4 & -1 & \ldots & -1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & \ldots & -1 & \ldots & 0 & -1 & 4 & -1 \\ 0 & \ldots & \ldots & \ldots & -1 & 0 & 0 & 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ u_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_n \end{bmatrix},$$

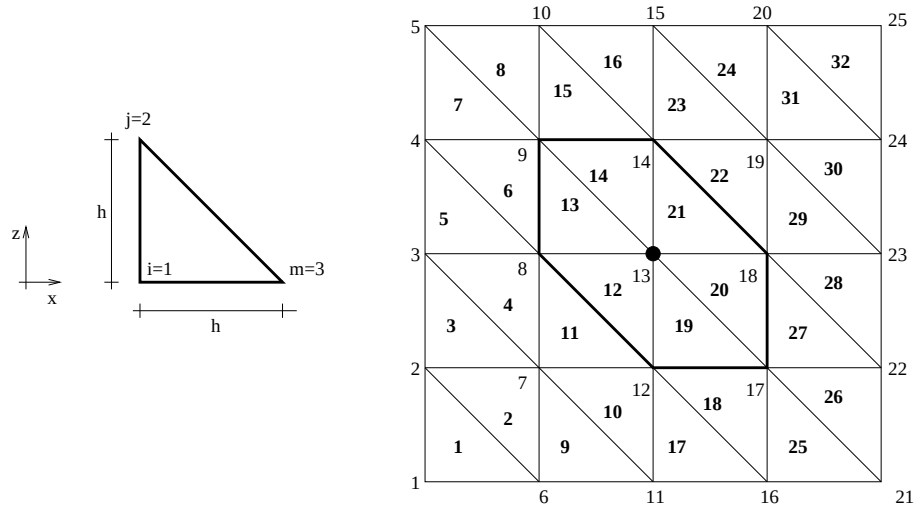

Figura 7: Triangolazione regolare su dominio quadrato.

dove, nel caso $f = cost$, si ha:

$$b_i = fh^2$$

Si riconosce che la matrice di rigidezza coincide con la matrice di rigidezza del metodo alle differenze finite (FD) del secondo ordine (il cosiddetto "stencil a 5 punti") [3], mentre con facili calcoli si vede che il termine noto è diverso da quello calcolato da FD.

Osservazione 3.21. Si noti che la costruzione della matrice di rigidezza può procedere secondo un procedimento che calcola la matrice di rigidezza locale di ciascun triangolo e poi assembla i diversi contributi nella matrice di rigidezza globale. Tale procedura, chiamata assemblaggio, comune a tutti gli schemi agli elementi finiti, permette di calcolare le quantità necessarie elemento per elemento, e ciò rende il calcolo della matrice A particolarmente efficiente e flessibile soprattutto in presenza di domini irregolari ed eterogeneità spaziali nei coefficienti della equazione alle derivate parziali da discretizzare. Un altro vantaggio considerevole di questa procedura è quello di rendere facilmente gestibile raffinamenti localizzati della griglia. Talchè è possibile immaginare di costruire una griglia computazionale in cui il passo h_j della mesh diventa via via più piccolo in corrispondenza a zone del dominio in cui si prevede di avere una derivata seconda della soluzione analitica particolarmente grande (ad esempio vicino a sorgenti puntuali). E ciò per rendere l'errore del metodo il più possibile uniforme spazialmente, utilizzando formule analoghe alla (17) estese al caso multidimensionale. Le griglie triangolari sono particolarmente adatte a questo scopo.

3.5.3 Convergenza del metodo FEM nel caso multidimensionale

Gli argomenti riportati nella sezione 3.4.1 possono essere estesi al caso multidimensionale con complicazioni tecniche che esulano dallo scopo di queste note. In particolare, è possibile di-

mostrare, analogamente al caso 1D, che $u_h \in V_h$ è la miglior approssimazione della soluzione esatta nel senso che:

$$\|\nabla u - \nabla u_h\| \leq \|\nabla u - \nabla v\| \qquad \forall v \in V_h,$$

dove la norma è definita qui con:

$$\|\nabla v\| = a\,(v,v)^{\frac{1}{2}} = \left(\int_\Omega v^2 + |\nabla v|^2 \, d\Omega\right)^{\frac{1}{2}},$$

che dimostra la proprietà di ottimalità della soluzione di Galerkin u_h rispetto alla norma usata.

Procedendo come prima, possiamo usare le interpolazioni lagrangiane a tratti e usare le stime dell'errore di interpolazione per ottenere:

$$\|\nabla u - \nabla \tilde{u}_h\| \leq Ch.$$

Infine, con dettagli tecnici non trascurabili, è possibile dimostrare che:

$$\|u - u_h\| = \left(\int_\Omega (u - u_h)^2 \, d\Omega\right)^{\frac{1}{2}} \leq Ch^2,$$

che è l'equivalente della (17) del caso 1D. Qui di deve assumere però la "regolarità" della triangolazione al tendere a zero del parametro di mesh h dato in (23). Più precisamente, si deve richiedere che ogni triangolo non tende a degenerare con il raffinamento progressivo, cioè nessun angolo di nessun triangolo tende a zero per $h \to 0$.

Ottimalità della soluzione al problema variazionale. Per apprezzare meglio l'ottimalità della soluzione u_h dichiarata in precedenza, prendiamo un caso particolare. Applichiamo il metodo di Galerkin alla seguente equazione:

$$\begin{aligned} -\Delta u + u &= f \qquad x \in \Omega \\ u &= 0 \qquad x \in \Gamma = \partial\Omega \end{aligned}$$

Il problema di variazionale diventa:

Problema 3.22. Problema variazionale Trovare $u \in H_0^1(\Omega)$ tale che:

$$a\,(u,v)_V = \langle f, v\rangle \qquad \forall v \in H_0^1(\Omega) \tag{28}$$

dove

$$a\,(u,v)_V = \int_\Omega \left[\nabla u \cdot \nabla v + uv\right] \, dx$$

Il corrispondente problema agli elementi finiti diventa dunque:

Problema 3.23. Problema FEM Trovare $u_h \in V_h(\Omega) \subset H_0^1(\Omega)$ tale che:

$$a\left(u_h, v\right)_V = \langle f, v \rangle \qquad \forall v \in V_h(\Omega). \tag{29}$$

Sottraendo (29) da (28), si ottiene l'equazione che mostra la consistenza forte dello schema, e cioè:

$$a\left(u - u_h, v\right)_V = 0 \qquad \forall v \in V_h(\Omega)$$

che sancisce l'ortogonalità della funzione errore a tutte le funzioni di $V_h(\Omega)$ rispetto al prodotto scalare $a\left(\cdot,\cdot\right)_V$. Questo equivale a dire che u_h è la proiezione ortogonale di u su $V_h(\Omega)$ rispetto al prodotto scalare $a\left(\cdot,\cdot\right)$. In altre parole, notando che il prodotto scalare è proprio quello di $H^1(\Omega)$, u_h è caratterizzato da un norma H^1 minima in confronto con qualsiasi altra funzione di $V_h(\Omega)$, e cioè:

$$\|u - u_h\|_{H^1(\Omega)} \leq \|u - v\|_{H^1(\Omega)} \qquad \forall v \in V_h(\Omega)$$

D'ora in avanti si ometteranno i pedici nei simboli di norma o prodotto scalare quando lo spazio di definizione è dettato dal contesto.

3.6 Problema di Neumann: condizioni al contorno naturali e essenziali

Si consideri ora il seguente problema al contorno:

$$\begin{aligned} -\Delta u + u &= f && \text{in } \Omega, \\ \nabla u \cdot \vec{n} &= g && \text{in } \Gamma = \partial\Omega. \end{aligned} \tag{30}$$

Moltiplicando la prima equazione per una funzione test $v \in V$, e integrando sul dominio, si ottiene:

$$-\int_\Omega (\Delta u v - u v)\, d\Omega = \int_\Omega f v\, d\Omega. \tag{31}$$

Applicando il lemma di Green al primo termine nell'integrale del primo membro, si ottiene:

$$\int_\Omega uv\, d\Omega - \int_\Gamma \nabla u \cdot \vec{n} v\, d\Gamma + \int_\Omega \nabla u \cdot \nabla v\, d\Omega = \int_\Omega f v\, d\Omega,$$

dove ancora $\Gamma = \partial\Omega$ è la frontiera di Ω. Notiamo che il primo termine dell'equazione precedente contiene esattamente il termine di Neumann sul bordo: $\nabla u \cdot \vec{n}$. Nel caso di condizioni al contorno di Dirichlet omogenee, avevamo richiesto che la funzione v fosse nulla al bordo con la conseguenza che l'integrale su Γ era nullo. In questo caso, invece, dobbiamo richiedere che le funzioni test siano diverse da zero al bordo. Infatti, possiamo dare la seguente formulazione variazionale:

Problema 3.24 (variazionale)**.**
Trovare $u \in V$ tale che:

$$a(u,v) = (f,v) + \langle g, v\rangle \qquad \forall v \in V, \tag{32}$$

dove:

$$\begin{aligned} a(u,v) &= \int_\Omega (\nabla u \cdot \nabla v + uv)\, dx \\ (f,v) &= \int_\Omega f\, v\, dx \\ \langle g, v\rangle &= \int_\Gamma g\, v\, dx \\ V &= \left\{ v(x) : v \text{ è continua in } \Omega, \frac{\partial v}{\partial x_i} \text{ sono continue in } \Omega\ \forall i \right\}, \end{aligned}$$

che si può dimostrare essere equivalente al seguente problema di minimizzazione:

Problema 3.25 (minimizzazione)**.**
Trovare $u \in V$ tale che:

$$F(u) \le F(v) \qquad \forall v \in V \tag{33}$$

dove:

$$F(v) = \frac{1}{2} a(v,v) - (f,v) - \langle g, v\rangle .$$

Si noti che, assumendo la u sufficientemente regolare, e applicando all'indietro il lemma di Green alla (32), si ottiene

$$\int_\Omega (-\Delta u + u - f)\, v\, d\Omega + \int_\Gamma (\nabla u \cdot \vec{n} - g)\, v\, d\Gamma = 0 \qquad \forall v \in V. \tag{34}$$

Si noti chele funzioni $v \in V$ sono non nulle al contorno, per cui possiamo imporre le seguenti due condizioni:

$$\int_\Omega (-\Delta u + u - f)\, v\, d\Omega = 0 \qquad \forall v \in V,$$

e

$$\int_\Gamma (\nabla u \cdot \vec{n} - g)\, v\, d\Gamma = 0 \qquad \forall v \in V.$$

Variando v nello spazio V (dove v non si annulla in Γ), si ottiene applicando il lemma 3.5:

$$-\Delta u + u - f = 0 \qquad \text{in } \Omega,$$

e

$$\nabla u \cdot \vec{n} - g = 0 \qquad \text{in } \Gamma;$$

che ci dice che le condizioni al contorno del problema originale sono soddisfatte.

Osservazione 3.26. Le condizioni al contorno di Neumann non appaiono esplicitamente nella formulazione variazionale, che si differenzia dal caso di condizioni al contorno di Dirichlet solo dal fatto che le funzioni test non sono più nulle al bordo. Possiamo quindi dire che le condizioni di Dirichlet vanno imposte esplicitamente (richiedendo appunto l'annullarsi delle funzioni test al bordo di Dirichlet), mentre le condizioni di Neumann sono "naturali" nella formulazione variazionale. Si noti che se fosse $g = 0$ (flusso nullo al contorno), il termine $\langle g, v\rangle$ sparirebbe: se non si impone esplicitamente alcuna condizione al bordo nella formulazione variazionale, si assume implicitamente condizioni al bordo di Neumann nulle. E' facile ora ricavare una formulazione variazionale per un problema misto in cui si impongano condizioni al contorno di Dirichlet in una porzione della frontiera, e condizioni di Neumann nella porzione complementare.

Siamo ora in grado di scrivere la seguente formulazione agli elementi finiti:

Problema 3.27 (Galerkin)**.**
Trovare $u_h \in Vh$ tale che:

$$a\left(u_h, v\right) = \left(f, v\right) + \langle g, v\rangle \qquad \forall v \in V_h, \tag{35}$$

dove:

$$\begin{aligned} a\left(u,v\right) &= \int_\Omega \left(\nabla u_h \cdot \nabla v + u_h v\right) dx \\ \left(f, v\right) &= \int_\Omega f\, v\, dx \\ \langle g, v\rangle &= \int_\Gamma g\, v\, dx \\ V_h &= \left\{v(x) : v \text{ è continua in } \Omega, v|_{T_k} \text{ è lineare } \ \forall T_k \in \mathcal{T}_h\right\}. \end{aligned}$$

Osservazione 3.28. La formulazione pratica procede quindi come nel caso del problema di Dirichlet omogeneo. Si noti ancora che le condizioni al contorno di Dirichlet, essendo imposte esplicitamente, sono soddisfatte in maniera "forte", mentre quelle di Neumann, essendo imposte in modo variazionale, sono soddisfatte in maniera "debole". Questo si osserva nella pratica quando si vanno a sperimentare le convergenze teoriche, dove si vede che l'errore puntuale nei nodi di Dirichlet è praticamente nullo (inferiore o uguale alla tolleranza usata per la soluzione del sistema lineare), mentre l'errore sulla soluzione nei nodi di Neumann tende a zero come h^2 (si veda la (17)).

Osservazione 3.29. Nella pratica l'imposizione delle condizioni al contorno di Dirichlet non omogenee avviene direttamente nella matrice del sistema lineare. Si procede nel modo seguente. Supponiamo che il nodo i-esimo sia contenuto nel contorno di Dirichlet. Si procede allora alla costruzione della i-esima riga del sistema lineare senza tenere in considerazione il fatto che essa corrisponde ad un nodo di Dirichlet. Una volta costruita tutta la matrice di rigidezza, si impone direttamente sul nodo i-esimo che u_i sia uguale esattamente al valore imposto, chiamiamolo $\bar{u}_i$. Questo si può ottenere in due modi.

I modo. Si azzerano tutti gli elementi extra diagonali della i-esima riga della matrice, e si impone pari a uno l'elemento diagonale e pari al valore imposto il corrispondente elemento del termine noto. Facendo così, si sostituisce alla i-esima equazione l'equazione:

$$u_i = \bar{u}_i$$

però contemporaneamente la matrice di rigidezza non è più simmetrica, e va opportunamente cambiata per mantenere la simmetria.

II modo, detto di "penalty". Si sostituisce al termine diagonale della i-esima riga un valore molto elevato λ e si impone il corrispondente elemento del vettore termini noti pari a $\lambda\bar{u}_i$. In questo modo si ottiene la seguente equazione i-esima:

$$\sum_{k=1}^{i-1} a_{ik}u_k + \lambda u_i + \sum_{k=i+1}^{n} a_{ik}u_k = \lambda\bar{u}_i.$$

A primo membro, però, tutti i termini extra diagonali (rappresentati dalle due sommatorie) sono effettivamente trascurabili rispetto al termine diagonale s λ è sufficientemente grande. Quindi l'equazione precedente corrisponde in pratica a:

$$\lambda u_i = \lambda\bar{u}_i,$$

che evidentemente impone in maniera corretta la condizione di Dirichlet. Un valore utilizzabile per λ è ad esempio $\lambda = 10^{40}$.

3.7 Tipologia di Elementi finiti

Finora abbiamo visto esclusivamente elementi finiti che ammettono funzioni di base lineari (e.g. triangoli in $\mathbb{R}^2$). E' intuitivo pensare che si possano definire su elementi di questa forma funzioni di base polinomiali a tratti con grado più elevato. Per esempio, è immediato definire in 1-D funzioni di base quadratiche: ogni elemento finito sarà formato da 3 nodi, necessari per valutare le tre costanti che formano una parabola utilizzando la proprietà di interpolazione delle funzioni di base (Vedasi Fig. 8. Evidentemente, un elemeno finito quadratico su un triangolo è

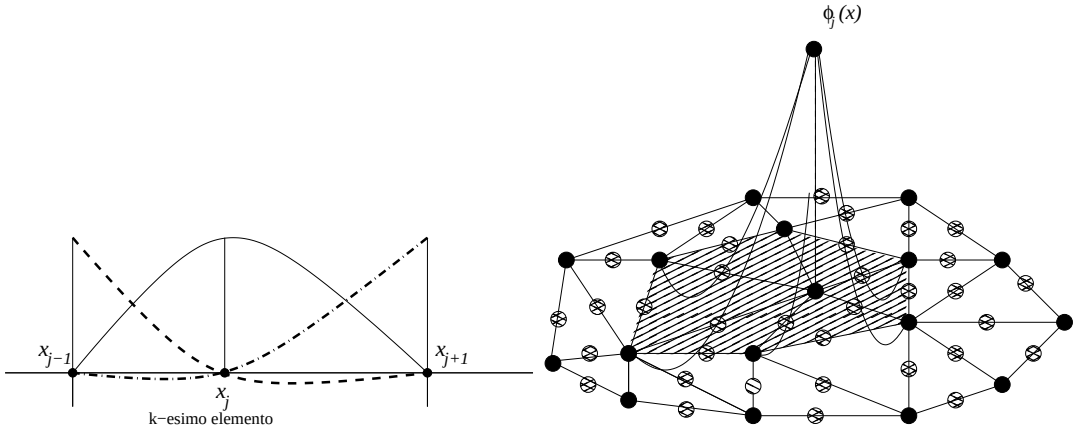

Figura 8: Funzioni di base per elementi finiti quadratici in 1D (a sinistra) e in 2D (a destra)

definito da 6 nodi dove imporre le condizioni di interpolazione per poter definire i 6 coefficienti che definiscono una parabola in $\mathbb{R}^2$.

Questa definitione basata su triangolazioni del dominio garantisce naturalmente (cioè senza trasformazioni particolari dell'elemento) la continuità della rappresentazione della soluzione, cioè di u_h ai bordi degli elementi, e quindi in tutto Ω. Ciò non è però più vero nel caso si usino forme gemoetriche degli elementi diverse dai simplessi (sottointervalli, triangoli, tetraedri). Per risolvere questo problema si introduce quindi una trasformazione di ciascun elemento ai fini solo della determinazione delle funzioni di interpolazione della soluzione e da usarsi essenzialmente per il calcolo degli integrali (26).

3.7.1 Elementi isoparametrici

Prendiamo in considerazione l'esempio semplificato di elementi finiti di forma quadrata, come l'esempio di Figura 9, dove sono rappresentati due elementi adiacenti. In questo caso, la continuità di u_h è assicurata se prendiamo delle funzioni di base cosidette bilineari, e cioè lineari separatamente in x ed in y. La loro espressione può genericamente essere scritta come:

$$\phi_i(x, y) = (a_i + b_i x)(c_i + d_i y)$$

Infatti, è facile vedere che al bordo degli elementi, per esempio per $x = 1$, la rappresentazione della funzione di base dipende solo da y ed è lineare, per cui i due nodi del lato in comune sono sufficienti a determinarne i coefficienti, e analogamente per gli altri bordi. I coefficienti da determinare per ciascun elemento sono 4, e 4 sono i nodi su cui imporre la condizione di interpolazione, quindi il loro calcolo è immediato. L'estensione diretta di questo procedimanto per funzioni di base formate da polinomi di ordine superiore al primo, ma sempre separatamente lungo x e lungo y, è immediata.

Nel caso di elementi quadrangolari ma di forma non quadrata o non rettangolare, quindi con lati non allineati alle direzioni coordinate, la continuità delle funzioni di base bilineari non

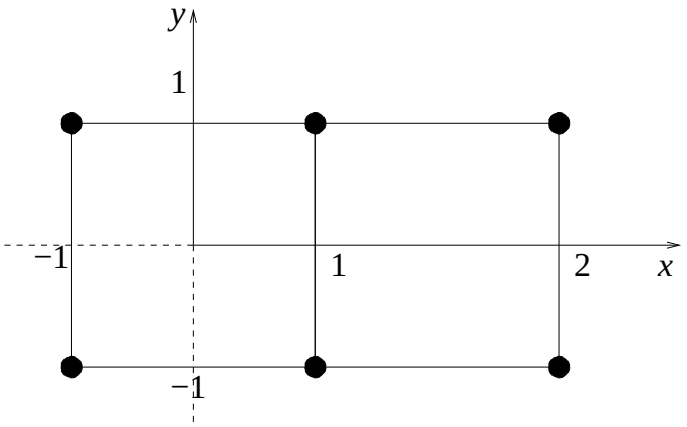

Figura 9: Elemento quadrato con funzioni di forma bilineari

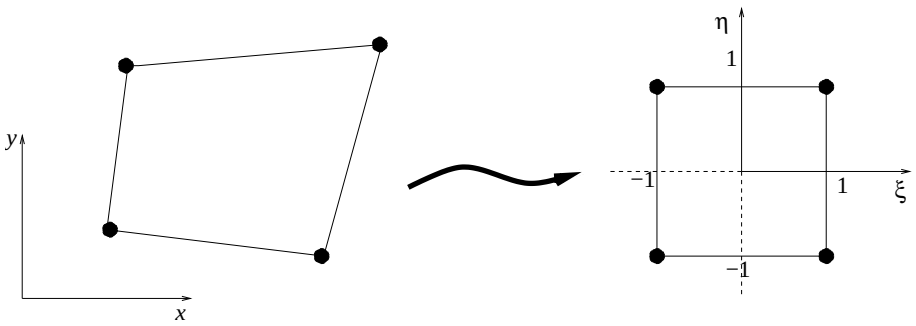

Figura 10: Elemento quadrilatereo generico e il suo trasformato

è più verificata. Un modo per ricavare funzioni di base continue al bordo è quello di trasformare ogni elemento quadrilatero in un elemento quadrato di riferimento tramite un cambio locale di coordinate (trasformazione conforme - Fig. 10). Sull'elemento di riferimento è possibile definire funzioni di base bilinieare col procedimento sopradescritto e quindi il calcolo degli integrali che formano la (26).

Vediamo un esempio nel caso di funzioni di base bilineari, con riferimento alla figura 10. E' facilmente verificabile che la trasformazione $(x, y) \to (\eta, \xi)$ è:

$$\begin{aligned} x &= \tfrac{1}{4}\left[(1-\xi)(1-\eta)x_i + (1+\xi)(1-\eta)x_j + (1+\xi)(1+\eta)x_m + (1-\xi)(1+\eta)x_k\right] \\ y &= \tfrac{1}{4}\left[(1-\xi)(1-\eta)y_i + (1+\xi)(1-\eta)y_j + (1+\xi)(1+\eta)y_m + (1-\xi)(1+\eta)y_k\right]. \end{aligned}$$

Ricordiamo che si vogliono calcolare gli integrali di tipo (26). Per fare questo dobbiamo calcolare lo Jacobiano della trasformazione:

$$J = \begin{bmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} \\ \frac{\partial x}{\partial \eta} & \frac{\partial y}{\partial \eta} \end{bmatrix},$$

e l'integrale della generica funzione $f(x, y)$ è calcolabile come:

$$\int_{\Omega^e} f(x,y)\mathrm{det}J \; dxdy = \int_{-1}^{1}\int_{-1}^{1} f(\eta,\xi) \; d\eta d\xi,$$

che può essere calcolate numericamente ad esempio con le formule di Gauss a 4 punti, ricordando comunque che per una funzione scalare $f(x, y)$ si hanno le seguenti relazioni tra i gradienti (e quindi le derivate):

$$\nabla_{(\xi,\eta)} f(x,y) = J\nabla f(x,y).$$

3.8 Equazione di diffusione e trasporto

Siamo ora in grado di affrontare un problema differenziale più complicato. Si consideri l'equazione (ellittica) di diffusione e trasporto seguente:

$$\begin{aligned} -\operatorname{div}(D\nabla u) + \operatorname{div}\left(\vec{\beta}u\right) &= f && \text{in } \Omega \\ u &= 0 && \text{in } \Gamma_D \\ D\nabla u \cdot \vec{n} &= g && \text{in } \Gamma_N, \end{aligned} \tag{36}$$

dove D è il coefficiente di diffusione (scalare e strettamente positivo) e $\vec{\beta}(x)$ è un campo vettoriale. Dal punto di vista applicativo, questa equazione rappresenta per esempio il trasporto di una sostanza disciolta in un fluido che si muove con il campo di moto $\vec{\beta}(x)$.

Procediamo dunque allo sviluppo di una formulazione variazionale per questo problema. Moltiplicando per una funzione test e integrando sul dominio, si ottiene:

$$-\int_\Omega \operatorname{div} D\nabla u\, v\, d\Omega + \int_\Omega \operatorname{div}(\vec{\beta}u)v\, d\Omega = \int_\Omega f\, v\, d\Omega.$$

Applicando ora il lemma di Green solo al primo termine del primo membro, si ottiene:

$$-\int_{\Gamma_N} g\, v\, d\Gamma + \int_\Omega D\nabla u \cdot \nabla v\, d\Omega + \int_\Omega \operatorname{div}(\vec{\beta}u)\, v\, d\Omega = \int_\Omega f\, v\, d\Omega,$$

da cui si può ricavare direttamente il seguente metodo agli elementi finiti:

Problema 3.30 (Galerkin)**.**
Trovare $u_h \in V_h$ tale che:

$$a(u_h, v) = (f, v) + \langle g, v\rangle \qquad \forall v \in V_h, \tag{37}$$

dove:

$$\begin{aligned} a(u_h, v) &= \int_\Omega \left(D\nabla u_h \cdot \nabla v + \operatorname{div}(\vec{\beta}u_h)v \right) dx \\ (f, v) &= \int_\Omega f\, v\, dx \\ \langle g, v\rangle &= \int_\Gamma g\, v\, d\Gamma \\ V_h &= \{v(x) : v \text{ è continua in } \Omega, v(x) = 0 \text{ in } \Gamma_D, v|_{T_k} \text{ è lineare } \forall T_k \in \mathcal{T}_h;\}. \end{aligned}$$

Osservazione 3.31. Si vede immediatamente che $a(u, v) \neq a(v, u)$, per cui questo metodo di Galerkin non ha un'equivalente di Ritz. In altre parole, non esiste in questo caso un metodo di minimizzazione, ma solo un metodo di ortogonalizzazione, cioè di Galerkin. Conseguenza di questo fatto è che il sistema lineare derivante dalla discretizzazione agli elementi finiti sarà sparso ma non simmetrico.

Procedendo nel modo consueto, si arriva dunque al sistema lineare seguente:

$$(A + B)u = c,$$

dove A è la matrice simmetrica di rigidezza, e B rappresenta la matrice non simmetrica del trasporto, date da:

$$\begin{aligned} A = \{a_{ij}\} \qquad & a_{ij} = \int_\Omega D\nabla\phi_j \cdot \nabla\phi_i\, d\Omega \\ B = \{b_{ij}\} \qquad & b_{ij} = \int_\Omega \operatorname{div}\left(\vec{\beta}\phi_j\right)\phi_i\, d\Omega \\ c = \{c_i\} \qquad & c_i = \int_\Omega f\,\phi_i\, d\Omega + \int_{\Gamma_n} g\,\phi_i\, d\Gamma_N. \end{aligned}$$

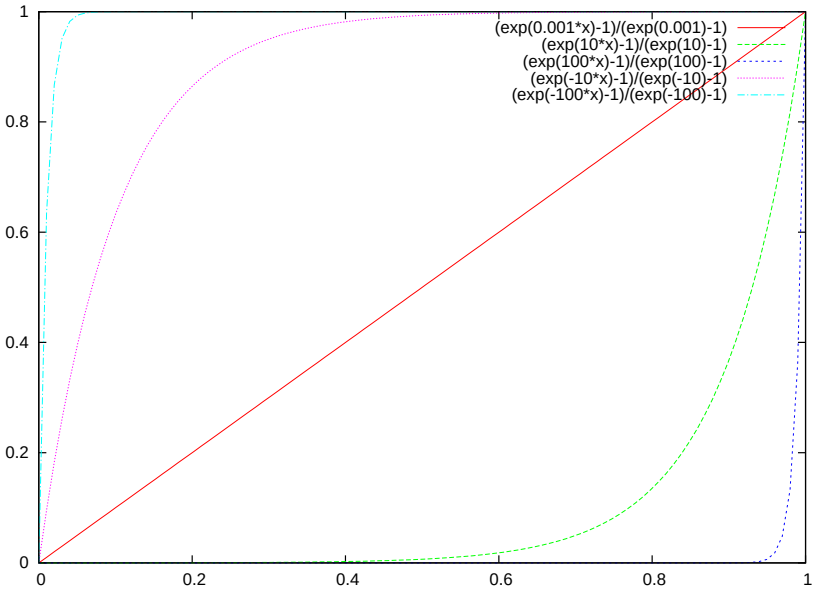

Figura 11: Andamento della soluzione del problema di trasporto (convezione-diffusione) per diversi valori del rapporto b/D (convezione/diffusione).

3.8.1 Caso monodimensionale

Consideriamo il seguente problema monodimensionale:

$$\begin{aligned} -Du'' + bu' &= 0, \qquad 0 < x < 1, \\ u(0) &= 0; \qquad u(1) = 1. \end{aligned} \tag{38}$$

Ricaviamo la soluzione analitica di tale problema. L'equazione caratteristica è data da:

$$-D\lambda^2 + b\lambda = 0,$$

che ha radici pari a $\lambda_1 = 0$ e $\lambda_2 = b/D$, per cui la soluzione generale è data da:

$$u(x) = c_1 e^{\lambda_1 x} + c_2 e^{\lambda_2 x} = c_1 + c_2 e^{bx/D}.$$

Imponendo le condizioni al contorno si ottiene immediatamente:

$$u(x) = \frac{e^{\frac{b}{D}x} - 1}{e^{\frac{b}{D}} - 1},$$

che è mostrata nel grafico di Figura 11 per diversi valori di b/D. Si vede che per valori bassi del rapporto b/D la soluzione è praticamente lineare, mentre per valori grandi la soluzione presenta un andamento esponenziale marcato, caratterizzato da zone del dominio dove si hanno gradienti spaziali elevati.

Procedendo alla formulazione agli elementi finiti con funzioni test lineari, e ricordando che in un caso monodimensionale con griglia regolare si ha equivalenza tra il metodo FEM e il metodo FD (si veda il paragrafo 3.4, si ricava la seguente equazione alle differenze per il nodo i-esimo:

$$\frac{D}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \frac{b}{2h}(u_{i+1} - u_{i-1}) = 0. \tag{39}$$

Si nota che tale equazione corrisponde ad una discretizzazione "centrata" alle differenze finite sia per la derivata seconda che per la derivata prima. Si veda a tal proposito l'Appendice A.

Introducendo ora il numero di Péclet di griglia, un numero adimensionale che indica il rapporto tra i flussi convettivi e quelli diffusivi, dato dal rapporto:

$$\mathbb{P}\mathrm{e} = \frac{|b|h}{D},$$

si arriva alla seguente equazione alle differenze (per $b > 0$):

$$(\mathbb{P}\mathrm{e} - 2)u_{i+1} + 4u_i - (\mathbb{P}\mathrm{e} + 2)u_{i-1} = 0 \qquad i = 1, \ldots, n-1. \tag{40}$$

Si può procedere alla soluzione analitica di tale equazione alle differenze imponendo una soluzione del tipo $u_i = \lambda^i$. Sostituendo si ottiene:

$$(\mathbb{P}\mathrm{e} - 2)\lambda^2 + 4\lambda - (\mathbb{P}\mathrm{e} + 2) = 0,$$

da cui si ricava:

$$\lambda_{1,2} = \frac{-2 \pm \sqrt{2 + \mathbb{P}\mathrm{e}^2 - 2}}{\mathbb{P}\mathrm{e} - 2} = \begin{cases} (2 + \mathbb{P}\mathrm{e})(2 - \mathbb{P}\mathrm{e}), \\ 1. \end{cases}$$

La soluzione generale della (40) è la combinazione lineare:

$$u_i = c_1\lambda_1^i + c_2\lambda_2^i$$

con le costanti che vanno ricavate dall'imposizione delle condizioni al contorno. Questo porta alla fine alla seguente soluzione dell'equazione alle differenze:

$$u_i = \frac{1 - \left(\frac{2+\mathbb{P}\mathrm{e}}{2-\mathbb{P}\mathrm{e}}\right)^i}{1 - \left(\frac{2+\mathbb{P}\mathrm{e}}{2-\mathbb{P}\mathrm{e}}\right)^n} \qquad i = 0, 1, \ldots, n,$$

che fornisce la soluzione del problema discretizzato agli elementi finiti (o alle differenze finite) per ogni nodo della griglia computazionale.

Da questa equazione si vede immediatamente che lo schema risulta oscillante nel caso $\mathbb{P}\mathrm{e} > 2$, perchè in tal caso il numeratore diventa negativo, e ovviamente assume valori oscillanti a seconda

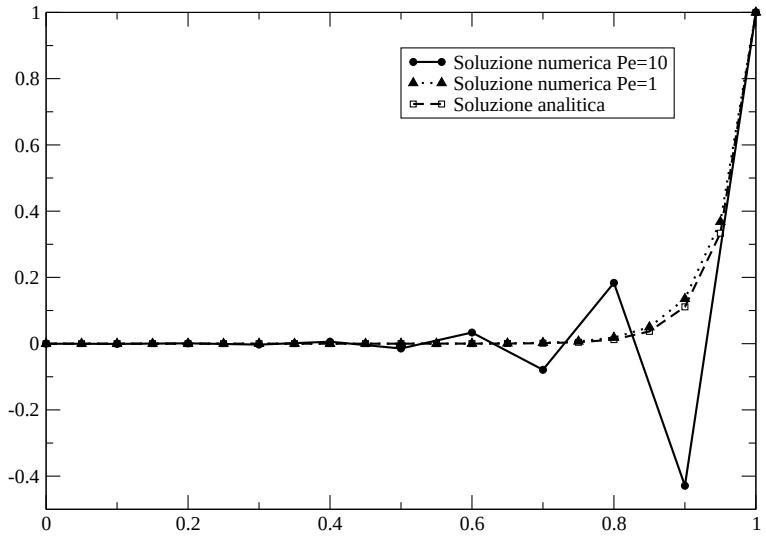

Figura 12: Comportamento dello schema alle differenze finite per la soluzione dell'equazione di convezione e diffusione a confronto con la soluzione analitica nel caso di $\mathbb{P}\mathrm{e} = 0.5$ e $\mathbb{P}\mathrm{e} = 2$.

che i sia pari o dispari. Tale comportamento è illustrato in Figura 12, dove si vede che per $\mathbb{P}\mathrm{e} > 2$ si verificano oscillazioni, mentre per valori inferiori lo schema risulta stabile.

Per cercare di correggere la situazione proviamo ad usare una discretizzazione decentrata per la derivata prima. La nuova discretizzazione diventa quindi:

$$\frac{D}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \frac{b}{h}(u_{i+1} - u_i) = 0,$$

dove il termine convettivo è ora discretizzato con una differenza del prim'ordine in avanti al posto della differenza centrale del secondo ordine usata in precedenza.

Procedendo nello stesso modo di prima si arriva all'equazione alle differenze:

$$(\mathbb{P}\mathrm{e} - 1)u_{i+1} - (\mathbb{P}\mathrm{e} - 2)u_i - u_{i-1} = 0 \qquad i = 1, \ldots, n-1.$$

La soluzione dell'equazione caratteristica corrispondente è:

$$\lambda_{1,2} = \frac{\mathbb{P}\mathrm{e} - 2 \pm \sqrt{(\mathbb{P}\mathrm{e} - 2)^2 + 4(\mathbb{P}\mathrm{e} - 1)}}{2(\mathbb{P}\mathrm{e} - 1)} = \begin{cases} (1)(1 - \mathbb{P}\mathrm{e}), \\ 1, \end{cases}$$

che porta alla soluzione:

$$u_i = \frac{1 - \left(\frac{1}{1-\mathbb{P}\mathrm{e}}\right)^i}{1 - \left(\frac{1}{1-\mathbb{P}\mathrm{e}}\right)^n} \qquad i = 0, 1, \ldots, n,$$

che risulta instabile per $\mathbb{P}\mathrm{e} < 1$. Se invece si usa la differenza decentrata all'indietro (upwind), si ottiene:

$$\frac{D}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \frac{b}{h}(u_i - u_{i-1}) = 0,$$

Rifacendo i conti si vede immediatamente che la soluzione numerica non mostra oscillazioni per nessun valore di $\mathbb{P}\mathrm{e}$ ed è incondizionatamente stabile. Con facili passaggi algebrici, l'equazione alle differenze precedente può essere scritta come:

$$\frac{D}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \frac{b}{2h}(u_{i+1} - u_{i-1}) + \frac{bh}{2}(\frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2}) = 0,$$

da cui si evince che la formulazione "upwind" equivale alla formulazione centrata a cui si è aggiunta una "diffusione numerica" pari a $h/2$. In questo caso, il nuovo numero di Péclet diventa:

$$\mathbb{P}\mathrm{e} = \frac{bh}{D + bh/2},$$

che è sempre inferiore o uguale a 2 per qualsiasi valore di D e $b(> 0)$. Per stabilizzare lo schema, si ricorre dunque all'aggiunta di un termine spurio (in pratica sto risolvendo una equazione diversa da quella di partenza), che però tende a zero al tendere a zero del parametro di griglia, per cui la consistenza dello schema è assicurata. Questo modo di procedere è un modo usato molto spesso per stabilizzare schemi numerici. In questo caso, quello che si paga è la diminuita accuratezza dello schema a griglia fissata, in questo caso evidenziata da una soluzione molto più regolare di quella analitica, con un fronte più diffuso. Il termine aggiuntivo, in effetti, equivale ad aver risolto un'equazione con il coefficiente di diffusione pari a $D = D + h/2$.

Nel caso degli elementi finiti si può pensare d procedere allo stesso modo, di aggiungere cioè al un termine diffusivo opportunamente creato. Per fare questo, aggiungiamo alla nostra forma bilineare un termine sempre bilineare proporzionale a:

$$\int_\Omega \left(\vec{\beta} \cdot \nabla u\right) \left(\vec{\beta} \cdot \nabla v\right) \, d\Omega,$$

che in pratica corrisponde ad un termine di diffusione numerica applicata solo lungo le linee di corrente, cioè solo lungo la direzione del campo di moto $\vec{\beta}(x)$. Il metodo agli elementi finiti, chiamato anche SD-FEM (Streamline Diffusion Finite Elements), diventa dunque:

Problema 3.32 (Streamline Diffusion)**.**
Trovare $u_h \in V_h$ tale che:

$$a_h(u_h, v) = (f, v) \qquad \forall v \in V_h, \tag{41}$$

dove:

$$\begin{aligned} a_h(u_h, v) &= \int_\Omega \left[D\nabla u_h \cdot \nabla v + \mathrm{div}(\vec{\beta} u_h) v + \tau \mathbb{P}\mathrm{e}_h \left(\vec{\beta} \cdot \nabla u_h \right) \left(\vec{\beta} \cdot \nabla v \right) \right] \, dx \\ (f, v) &= \int_\Omega f \, v \, dx \\ V_h &= \{ v(x) : v \text{ è continua in } \Omega, v(x) = 0 \text{ in } \Gamma_D, v|_{T_k} \text{ è lineare } \forall T_k \in \mathcal{T}_h; \}, \end{aligned}$$

dove $\mathbb{P}\mathrm{e}_h$ è il numero di Péclet di griglia definito elemento per elemento da:

$$\mathbb{P}\mathrm{e}_h = \frac{|\vec{\beta}_k| h_k}{D^{(k)}}$$

con $D^{(k)}$ il coefficiente di diffusione $\vec{\beta}_k$ il vettore velocità ancora assunti costanti nell'elemento T_k, ma potenzialmente variabili da elemento a elemento.

Si riconosce immediatamente che $a_h(\cdot, \cdot) \to a(\cdot, \cdot)$ per $h \to 0$, è il termine aggiuntivo di diffusione numerica introdotto lungo la direzione $\vec{\beta}$ della velocità, e τ è un coefficiente che deve essere tarato caso per caso in maniera da ottenere uno schema in cui le oscillazioni numeriche siano minimizzate o annullate, senza contemporaneamente introdurre una quantità eccessiva di diffusione numerica. Le figure 13 e 14 riportano esempi esemplificativi dei tipici andamenti della soluzione nei casi stabili e instabili.

Si noti infine che esistono schemi molto più avanzati dello schema Streamline Upwind, basati essenzialmente nell'introduzione di diffusione numerica in minima quantità e solo quando e dove necessario. Tali algoritmi non vengono trattati in queste note. Il lettore interessato può riferirsi ai testi [5, 7, 6] e alle referenze citate in essi.

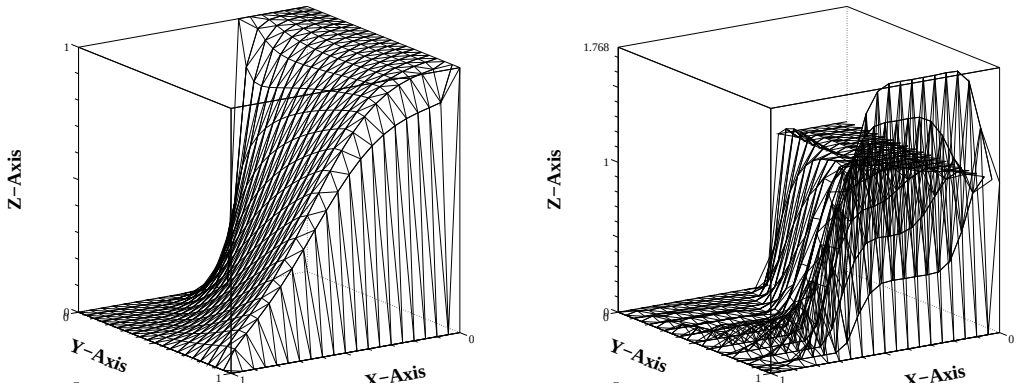

Figura 13: Problema di diffusione e convezione risolto con Galerkin lineare (P1) senza stabilizzazione. Nel grafico a sinistra si mostra il caso con $D = 0.1$ mentre in quello a destra vi è il caso con $D = 0.01$. In ambedue i casi il vettore velocità è $\vec{\beta} = (1, 3)^T$, talchè il numero di Péclet di griglia è costante e pari a $\mathbb{P}\mathrm{e}_h = 1$ e $\mathbb{P}\mathrm{e}_h = 10$ nei due casi.

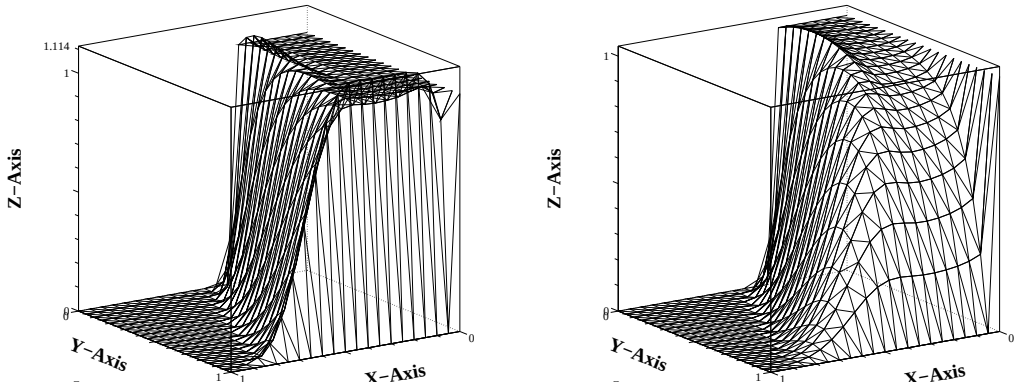

Figura 14: Problema di diffusione e convezione risolto con Galerkin lineare (P1) con stabilizzazione SD nel caso $D = 0.01$ e $\vec{\beta} = (1, 3)^T$ ($\mathbb{P}\mathrm{e}_h = 10$). A sinistra si riporta il caso con $\tau = 0.01$ e a destra il caso con $\tau = 1.0$. Si noti la diminuzione di oscillazioni e l'aumento della diffusione numerica.

3.9 Teoria matematica degli elementi finiti

3.9.1 Richiami di analisi funzionale

Si consideri uno spazio misurabile (Ω, Σ, μ) con misura nonnegativa μ, che indicheremo semplicemente Ω. In generale, Ω è un sottoinsieme di $\mathbb{R}^d$, con $d = 1, 2, 3$, aperto e limitato, con chiusura $\Gamma = \partial\Omega$ sufficientemente liscia (e.g., Lipschitz). Il fatto che lo spazio sia misurabile, ci permette di poter dire "quanto grande è una funzione" e quindi di poter fare "confronti" e "stime" tra funzioni. Per fare questo si intruduce il concetto di "norma" di una funzione la cui definizione può essere un'estensione diretta della definizione di norma per spazi vettoriali a dimensione finite. In analogia, date due funzioni u e v in uno spazio $V \subset \mathbb{R}$, $(u, v : V \to \mathbb{R})$ definiamo il prodotto scalare in uno spazio V:

Definizione 3.33 (prodotto scalare (in campo reale))**.** Un prodotto scalare tra due funzioni u e v dfinite in un dominio $V \subset \mathbb{R}$ è una forma bilineare $(\cdot, \cdot) : V \times V \to \mathbb{R}$ che soddisfa alle proprietà:

1. simmetria: $(u, v) = (v, u)$;
2. linearità (nel primo argomento): $(\alpha u, v) = \alpha\,(u, v)$, $(u + v, w) = (u, w) + (v, w)$;
3. definizione positiva: $(u, u) =\geq 0$, $(u, u) = 0 \Leftrightarrow u = 0$.

Definizione 3.34 (norma di funzione)**.** Data una funzione u definita in un dominio $V \subset \mathbb{R}$ è una forma lineare $\|\cdot\| : V \to \mathbb{R}$ che soddisfa alle proprietà:

1. $\|u\| > 0$, $\|u\| = 0$ se e solo se $u = 0$;
2. dato $\alpha \in \mathbb{R}$, $\|\alpha u\| = |\ \alpha\ |\, \|u\|$;
3. disuguaglianza triangolare: $\|u + v\| \leq \|u\| + \|v\|$;

Si parla di seminorma, indicata con $|\cdot|$, quando la prima delle proprietà precedenti è sostituita con la richiesta che sia $\|u\| \geq 0$.

Oltre alla disuguaglianza triangolare, useremo moltissimo la disuguaglianza di Cauchy-Schwartz:

$$|\ (u, v)\ | \leq \|u\|\, \|v\| \tag{42}$$

Dimostrazione. La disugaglianza di Cauchy-Schwarz si può dimostrare come segue. La dimostrazione è banale se $v = 0$. Assumiamo qiundi $v \neq 0$. Sia $\lambda \in \mathbb{R}$, con $\lambda = \langle v, u\rangle \,/\, \langle v, v\rangle$. La funzione $z = u - \lambda v$ è la proiezione ortogonale di u lungo v, per cui $\langle v, u - \lambda v\rangle = \langle u, v\rangle - \lambda\,\langle u, u\rangle = 0$. La positività e la simmetria del prodotto scalare implicano che:

$$0 \leq \langle u - \lambda v, u - \lambda v\rangle = \langle u, u - \lambda v\rangle - \lambda\,\langle v, u - \lambda v\rangle = \langle u, u - \lambda v\rangle = \langle u, u\rangle - \lambda\,\langle u, v\rangle\,,$$

da cui sommando $\lambda \langle u, v \rangle$ ad ambo i membri:

$$\lambda \langle u, v \rangle \leq \langle u, u \rangle$$

e moltiplicando per $\langle u, u \rangle$ si ottiene:

$$\langle u, v \rangle^2 \leq \langle u, u \rangle \langle v, v \rangle .$$

□

Avremo a che fare con le seguenti:

funzioni continue; lo spazio delle funzioni continue $C^0(\Omega)$ è:

$$C^0(\Omega) = \{u : \Omega \to \mathbb{R} : u \text{ è continua e limitata}\},$$

con la norma:

$$\|u\|_\infty = \sup_{x \in \Omega} | \, u(x) \, |; \tag{43}$$

funzioni limitate; lo spazio delle funzioni limitate potrà essere caratterizzato da:

$$L^\infty(\Omega) = \{u : \Omega \to \mathbb{R} : u \text{ è misurabile e } \mu\text{-q.o. limitata}\},$$

possiamo ancora definire la norma precedente (43);

funzioni integrabili; dato $0 < p < \infty$, lo spazio delle funzioni integrabili è:

$$L^p(\Omega) = \left\{ u : \Omega \to \mathbb{R} : u \text{ è misurabile e } \int_\Omega | \, u \, |^p \, d\mu < +\infty \right\};$$

dato $1 \leq p < \infty$, possiamo definire la norma:

$$\|u\|_{L^p(\Omega)} = \|u\|^p = \left(\int_\Omega | \, u(x) \, |^p \, d\mu \right)^{\frac{1}{p}}; \tag{44}$$

nel caso $0 < p < 1$, la norma non è definibile, ma si usa la metrica della distanza:

$$d_p(u, v) = \int_\Omega | \, u(x) - v(x) \, |^p \, d\mu;$$

funzioni derivabili; $C^k(\Omega)$ è lo spazio delle funzioni derivabili dato da:

$$C^k(\Omega) = \left\{ u : \Omega \to \mathbb{R} : \forall \alpha, \mid \alpha \mid \leq k : \partial^\alpha u \text{ è continua in } \overline{\Omega} \right\};$$

una norma ammissibile è definita da:

$$\|u\| = \sum_{0 \leq |\alpha| \leq k} \|\partial^\alpha u\|_\infty$$

Spazi di Sobolev. Si indica con $W^{k,p}(\Omega)$ lo spazio di Sobolev:

$$W^{k,p}(\Omega) = \{u : \Omega \to \mathbb{R} : u \in L^p(\Omega); \partial^\alpha u \in L^p(\Omega) \forall \alpha : \mid \alpha \mid \leq k\}$$

con la norma definita da:

$$\|u\|_{k,p} = \left(\sum_{0 \leq |\alpha| \leq k} \int_\Omega \mid \partial^\alpha u \mid^p \right)^{\frac{1}{p}},$$

se $p < \infty$, mentre per $p = \infty$ la norma è:

$$\|u\|_{k,\infty} = \max_{0 \leq |\alpha| \leq k} \|\partial^\alpha u\|_\infty .$$

Gli spazi $L^2(\Omega)$ e $H^1(\Omega)$. In questa sezione si riporta una breve e incompleta descrizione degli spazi funzionali di interesse per lo studio del metodo agli elementi finiti. Lo spazio $L^2(\Omega)$ è lo spazio delle funzioni "di quadrato sommabile" rispetto alla misura di Lebuesgue, tali cioè che, dato $f : \Omega \to \mathbb{R}$ l'integrale seguente è finito:

$$\int_\Omega \mid f(x) \mid^2 dx < \infty.$$

Se associamo a tale spazio il prodotto scalare:

$$\langle u, v \rangle_{L^2(\Omega)} = \int_\Omega u(x) v(x) \, dx,$$

e la norma indotta:

$$\|u\|_{L^2(\Omega)} = \left(\int_\Omega u(x) v(x) \, dx \right)^{1/2},$$

allora $L^2(\Omega)$ è uno spazio di Hilbert.

Dato un sottospazio V di $L^2(\Omega)$, avremo a che fare con funzionali (forme) lineari $\mathcal{F} : \mathcal{V} \to \mathbb{R}$, e forme bilineari $B : V \times V \to \mathbb{R}$. Dato uno sottospazio lineare $V(\Omega) \subset \Omega$, un operatore $a(\cdot, \cdot)$ definisce una forma bilineare in $V \times V$ se:

$$\begin{aligned} a : V \times V &\to \mathbb{R}, \\ a(u, v) &= a(v, u), \\ a(\alpha u + \beta v, w) &= \alpha a(u, w) + \beta a(v, w), \\ a(u, \alpha v + \beta w) &= \alpha a(u, v) + \beta a(u, w). \end{aligned}$$

La forma bilineare $a(\cdot, \cdot)$ definisce un prodotto scalare su $V(\Omega)$ se simmetrica e:

$$a(v, v) > 0 \qquad \forall v \in V, \quad v \neq 0.$$

La norma associata con tale prodotto scalare diventa:

$$\|v\|_V = (a(v, v))^{\frac{1}{2}}.$$

Il prodotto scalare soddisfa la disuguaglianza di Cauchy-Schwartz:

$$|a(u, v)| \leq \|u\|_V \|v\|_V.$$

Uno spazio lineare V dotato di prodotto scalare e norma corrispondente è detto spazio di Hilbert se è completo, cioè se ogni successione di Cauchy[5] converge rispetto a $\|\cdot\|_V$.

Per esempio, lo spazio delle funzioni di quadrato sommabile nell'intervallo $\Omega = [a, b]$:

$$L^2(I) = \left\{ v(x) : I \to \mathbb{R} \text{ tali che } \int_a^b v^2 \, dx < \infty \right\}$$

è uno spazio di Hilbert e può essere dotato del prodotto scalare:

$$(u, v) = \int_a^b u(x)v(x) \, dx$$

con norma corrispondente data da:

$$\|u\|_{L^2(I)} = \|u\|_2 = \left(\int_a^b [u(x)]^2 \, dx \right)^{\frac{1}{2}}.$$

Osservazione 3.35. In riferimento all'Osservazione 3.11, si noti che lo spazio L^2 è proprio l'insieme di tutte le funzioni a quadrato sommabile menzionate in tale osservazione.

[5]Una successione di funzioni $v_i \in V$ è detta di Cauchy se esiste $\epsilon > 0$ tale che $\|v_i = v_j\|_V < \epsilon$ per i e j sufficientemente grandi. Si dice che v_i converge a v se $\|v - v_i\|_V \to 0$ per $i \to \infty$.

Esempio 3.36. La funzione $v(x) = x^{-\alpha}$, $x \in I = [0,1]$, appartiene a $L^2(I)$ solo per valori di $\alpha < 1/2$.

Introduciamo ora lo spazio naturale che contiene la soluzione dei nostri problemi differenziali ellittici: lo spazio di Hilbert $H^1(I) = \{v : v \text{ e } v' \text{ appartengono a } L^2(I)\}$. Il prodotto scalare è dato da:

$$(u,v)_{H^1(I)} = \int_a^b \left[u(x)v(x) + u'(x)v'(x)\right]\, dx$$

e norma data da:

$$\|u\|_{H^1(I)} = \left(\int_a^b \left[u(x)^2 + u'(x)^2\right]\, dx\right)^{\frac{1}{2}}$$

Si noti che lo spazio V definito nel paragrafo 3.2 è uno spazio di Hilbert opportuno, ed è di solito indicato con:

$$V(I) = H_0^1(I) = \left\{v(x) : \mathbb{R} \to \mathbb{R} \text{ tali che } v(x) \in L^2(I), v'(x) \in L^2(I) \text{ e } v(0) = v(1) = 0\right\}$$

Si noti che il pedice 0 nel simbolo dello spazio di Hilbert viene usato per denotare il fatto che le funzioni sono nulle al bordo del dominio. L'apice 1 viene usato per denotare l'ordine massimo delle derivate che sono di quadrato sommabile (in questo caso le derivate prime).

Tutte queste nozioni sono facilmente estendibili al caso multidimensionale. Gli spazi sopra definiti per funzioni definite su un dominio limitato $\Omega \in \mathbb{R}^d$ con contorno $\Gamma = \partial\Omega$ sufficientemente liscio, sono dati da:

$$L^2(\Omega) = \left\{v(x) : \Omega \to \mathbb{R} \text{ tali che } \int_\Omega v(x)^2 < \infty\right\}$$
$$H^1(\Omega) = \left\{v(x) : \Omega \to \mathbb{R} \text{ tali che } v(x) \in L^2(\Omega) \text{ e } \frac{\partial v(x)}{\partial x_i} \in L^2(\Omega) \text{ per } i = 1, \ldots, d\right\}$$
$$H^k(\Omega) = \left\{v(x) : \Omega \to \mathbb{R} \text{ tali che } v(x) \in L^2(\Omega) \text{ e } \partial^\alpha v \in L^2(\Omega) \text{ per ogni } \mid \alpha \mid \leq k\right\}$$

con i seguenti prodotti scalari (e conseguenti norme):

$$(u,v)_{L^2(\Omega)} = \int_\Omega uv\, dx \qquad \|u\|_{L^2(\Omega)} = \left(\int_\Omega u^2\, dx\right)^{\frac{1}{2}}$$
$$(u,v)_{H^1(\Omega)} = \int_\Omega \left[uv + \nabla u \cdot \nabla v\right]\, dx \qquad \|u\|_{H^1(\Omega)} = \left(\int_\Omega \left[u^2 + |\nabla u|^2\right]\, dx\right)^{\frac{1}{2}}$$

Useremeo spesso la seguente seminorma:

$$\mid v \mid_{H^k(\Omega)} = \left(\int_\Omega \mid \nabla u \mid^2 dx\right)^{\frac{1}{2}} = \left(\int_\Omega \mid \partial u \mid^2 dx\right)^{\frac{1}{2}} = \|\partial v\|_{L^2(\Omega)}\,.$$

Chiaramente se necessario si possono definire gli spazi di Hilbert di ordine maggiore di 1:

Lo spazio $H_0^1(\Omega)$ e la disuguaglianza di Poincaré. Definiamo con i lpedice "0" lo spazio di funzioni che si annullano al bordo di Ω, che soddisfano cioè a condizioni di Dirichlet omogene al bordo:

$$H_0^1(\Omega) = \{v(x) \in H^1(\Omega) \text{ tali che } v(x) = 0 \in \Gamma\}.$$

con lo stesso prodotto scalare e la stessa norma definita per $H^1(\Omega)$. In generale indicheremo per ogni funzione $u \in H^1(\Omega)$ la seminorma:

$$|u|_1 = \left(\int_\Omega \nabla u \cdot \nabla u \, dx \right)^{\frac{1}{2}};$$

tale quantità risulta una seminorma perchè vale zero per ogni funzione costante $u = cost$. Nel caso che $u \in H_0^1(\Omega)$, le condizioni al bordo omogenee trasformano tale quantità in una norma equivalente a quella standard $\|\cdot\|_{H^1(\Omega)}$, come si vede da un'applicazione ovvia della diseguaglianza di Poincarè:

Lemma 3.37 (Poincarè)**.** *Sia $\Omega \subset \mathbb{R}^d$ un insieme aperto limitato. Allora esiste una costante C dipendente solo da Ω e tale che*

$$\|u\|_{L^2(\Omega)} \leq C \|\nabla u\|_{L^2(\Omega)}$$

per ogni $u \in H_0^1(\Omega)$.

Corollario 3.38. *La norma del gradiente:*

$$\|\nabla u\| = \left(\int_\Omega \nabla u \cdot \nabla u \, dx \right)^{\frac{1}{2}}$$

è una norma equivalente, (ai fini della topologia indotta e quindi delle convergenze) alla norma usuale $\|u\|_{H^1(\Omega)}$.

Infatti:

$$\|\nabla u\|_2^2 \leq \|u\|_{H^1}^2 = \|u\|_2^2 + \|\nabla u\|_2^2 \leq (1 + C^2) \|\nabla u\|_2^2$$

e quindi date u e v in $H_0^1(\Omega)$ possiamo definire un prodotto scalare equivalente all'usuale come:

$$(u, v)_{H_0^1(\Omega)} = \int_\Omega \nabla u \cdot \nabla v \, dx.$$

In generale la traccia di una funzione su $\partial\Omega$ (il suo valore nel contorno $\partial\Omega$) non è sempre definita (si pensi ad esempio a $\sin(1/x)$ per $x = 0$). D'altro canto, la traccia di una funzione di $H^1(\Omega)$ esiste sempre per cui ha senso parlare di H_0^1.

3.9.2 Teorema di Lax-Milgram

Definizione 3.39. Sia V uno spazio di Hilbert e $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ una forma bilineare. Si dice che la forma bilineare è:

- *continua* se esiste una costante $\gamma > 0$ tale che:

$$| a(u,v) | \leq \gamma \|u\|_V \|v\|_V \qquad \forall u, v \in V; \tag{45}$$

- *V-ellittica o coerciva* se esiste una costante $\alpha > 0$ tale che

$$a(v,v) \geq \alpha \|v\|_V^2 \qquad \forall v \in V. \tag{46}$$

Analogamente, una forma lineare $f(\cdot) : V \to \mathbb{R}$ si dice *continua* se esiste una costante $\Lambda > 0$ tale che:

$$| f(v) | \geq \Lambda \|v\|_V \qquad \forall v \in V.$$

Osservazione 3.40. La coercività dell'operatore continuo è generalmente ereditata dall'operatore discreto FEM. Tale proprietà è di fondamentale importanza per ottenere le stime di convergenza, ma è in generale una condizione assai limitante, e si trovano sperimentalmente casi in cui la coercività dello schema non è garantita ma lo schema mostra convergenza ottimale.

Nel caso $V = \mathbb{R}^n$, la coercività di $a(,)$ implica che la matrice (operatore lineare associato ad $a(,)$) $A = \{a_{ij}\}$, $a_{ij} = a(\phi_i, \phi_j)$, è definita positiva.

Per le forme lineari in spazi di Hilbert esiste il seguente teorema:

Teorema 3.1 (di rappresentazione di Riesz)**.** *Per ogni forma lineare continua $\phi_u(\cdot)$ in uno spazio di Hilbert V esiste un unico $u \in V$ tale che $\phi_u(v) = a(u,v)$ per ogni $v \in V$. Inoltre, $\|u\| = \|\phi_x\|$.*

Indichiamo con V^* lo spazio (duale di V) formato da tutte le forme lineari da V a $\mathbb{R}$. Il teorema di Riesz dice che ogni elemento di V^* può essere scritto univocamente nella forma $\phi_u(v) = a(u,v)$. In altre parole, la mappa $\Phi : V \to V^*$ definita da $\Phi(u) = \phi_u(v)$ è un'isomorfismo. Una conseguenza del teorema di Riesz è il teorema di Lax Milgram per forme bilineari continue e coercive:

Teorema 3.2 (Lax-Milgram)**.** *Sia $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ una forma bilineare continua e coerciva. Per ogni forma lineare $f(v) : V \to \mathbb{R}$, esiste una unica funzione $u \in V$ tale che:*

$$a(u,v) = f(v) \qquad \forall v \in V.$$

Dimostrazione. (Cenni) Dal teorema di Riesz possiamo definire una mappa lineare continua $A : V \to E$ definita da:

$$a(u,v) = (A(u), v) \qquad \text{con} \qquad \|A(u)\| \leq C \|u\|_V .$$

Possiamo quindi associare ad una forma lineare $f(\cdot) \in V^*$ una funzione di V tale che $f(v) = (f, v)$. Quindi bisogna dimostrare che la soluzione del problema $A(u) = f$ in V è unica. Dal teorema di punto fisso di Banach è facile dimostrare l'esistenza e l'unicità della soluzione dimostrando che la mappa $T : E \to E$:

$$T(u) = u - \epsilon A(u) + \epsilon f$$

è una contrazione per ϵ sufficientemente piccolo. □

3.10 Formulazione astratta del metodo FEM per equazioni ellittiche

3.10.1 Formulazione debole

Sia V uno spazio di Hilbert con prodotto scalare $(\cdot,\cdot)_V$ e norma indotta $\|\cdot\|_V$. Si indichi con $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ una forma bilineare continua e V-ellittica e con $f(\cdot) : V \to \mathbb{R}$ una forma lineare in V continua. L'esensione diretta di quanto fatto prima ci porta a formulare i seguenti problemi di mininimizzazione e variazionale:

Problema 3.41 (M)**.** Trovare $u \in V$ tale che

$$F(u) = \min_{v \in V} F(v),$$

dove

$$F(v) = \frac{1}{2} a(v,v) - f(v) \qquad \forall v \in V. \tag{47}$$

Problema 3.42 (V)**.** Trovare $u \in V$ tale che

$$a(u,v) = f(v) \qquad \forall v \in V. \tag{48}$$

Dimostriamo il seguente:

Teorema 3.3. *Se la forma bilinieare è simmetrica, e cioè $a(u,v) = a(v,u)$, i problemi 3.41 e 3.42 sono rquivalenti nel senso che $u \in V$ è soluzione del problema 3.41 se e solo se ̀soluzione del problema 3.42. Inoltre esiste una unica soluzione che soddisfa alla stima di stabilità:*

$$\|u\|_V \leq \frac{\Lambda}{\alpha}$$

Dimostrazione. L'esistenza della soluzione è una semplice applicazione del teorema di Lax-Milgram.

Per l'equivalenza dei due problemi, si procede esattamente come nel caso 1-D riportato in precedenza. A tal fine, proviamo prima che se $u \in V$ è soluzione del problema 3.41 allora è soluzione del problema 3.42. Sia quindi $v \in V$ e $\epsilon \in \mathbb{R}$ arbitrari. La condizione che $u \in V$ è un punto di minimo per F si può scrivere come:

$$F(u) \leq F(u + \epsilon v) \qquad \forall \epsilon \in \mathbb{R}.$$

La funzione $g(\epsilon) = F(u + \epsilon v)$ avrà quindi un minimo per $\epsilon = 0$ quindi sarà $g'(0) = 0$. Esplicitando la F e sfruttando la simmetria di $a(,)$ si ottiene:

$$\begin{aligned} g(\epsilon) &= \frac{1}{2}a(u + \epsilon v, u + \epsilon v) - f(u + \epsilon v) \\ &= \frac{1}{2}a(u,u) - f(u) + \epsilon a(u,v) - \epsilon f(v) + \frac{\epsilon^2}{2}a(v,v); \end{aligned}$$

il risultato richiesto è provato osservando che:

$$g'(0) = a(u,v) - f(v).$$

Assumiamo ora che $u \in V$ sia soluzione di 3.42. Dobbiamo dimostrare che per tale u $F(u) \leq F(u+v)$ per qualsiasi $v \in V$. Osserviamo subito che:

$$F(u+v) = \frac{1}{2}a(u,u) - f(u) + a(u,v) - f(v) + \frac{1}{2}a(v,v),$$

da cui il risultato segue per la coercività di $a(,)$.

L'ultimo punto del teorema, e cioè la stabilità della soluzione, si deduce prenendo $v = u$ in (48) e la coercività di $a(,)$ e la continuità di $f()$ producono immediatamente:

$$\alpha \|u\|_V \leq a(u,u) = f(u) \leq \Lambda \|u\|_V.$$

L'unicità della soluzione segue direttamente da quest'ultima diseguaglianza. Infatti, se u_1 e u_2 sono due soluzioni di 3.42, allora:

$$a(u_1 - u_2, v) = 0 \qquad \forall v \in V.$$

La stima di stabilità con $f() = 0$ e $\Lambda = 0$ implica che $\|u_1 - u_2\| = 0$, da cui $u_1 = u_2$. □

Il risultato importante del teorema d Lax-Milgram è quindi che la continuità e la coercività della forma bilineare implicano l'esistenza e l'unicità della soluzione. Esistono però delle equazioni alle derivate parziali le cui forme bilineari associate non sono coercive ma soddisfano ad una relazione più debole ma che garantisce lo stesso l'esistenza e l'unicità della soluzione. Tale condizione, chiamata di inf-sup, si può definire come:

Definizione 3.43. Si dice che una forma bilineare $a(,)$ soddisfa la condizione *inf-sup* in V se esiste $\alpha > 0$ tale che:

$$\sup_{v \in V} \frac{a(u,v)}{\|v\|_V} \geq \alpha \|u\|_V \qquad \forall u \in V; \tag{49}$$

e contemporaneamente:

$$\sup_{u \in V} \frac{a(u,v)}{\|u\|_V} \geq \alpha \|v\|_V \qquad \forall v \in V; \tag{50}$$

Ovviamente, se $a(,)$ è simmetrica, le due condizioni sopra sono equivalenti. Inoltre, la (49) (e allo stesso modo la (50)) si può riscrivere come:

$$\inf_{u \in V} \sup_{v \in V} \frac{a(,)}{\|u\|_V \|v\|_V} > 0 \tag{51}$$

da cui il nome inf-sup. Tale condizione è anche spesso chiamata condizione di Babuska-Brezzi-Ladyzenskaya, dal nome delle persone che l'hanno discussa per primi [2]. E' chiaro quindi che questa è una condizione molto importante che è certamente soddisfatta se la forma bilineare è coerciva. Infatti:

Lemma 3.44. *Se $a(\cdot,\cdot)$ è coerciva, allora soddisfa alla condizione 51.*

Dimostrazione. La coercività di $a(,)$ implica:

$$a(u,u) \geq \alpha \|u\|_V^2, \qquad \forall u \in V.$$

Possiamo quindi scrivere:

$$\sup_{v \in V} \frac{a(u,v)}{\|v\|_V} \geq \frac{a(u,u)}{\|u\|_V} \geq \alpha \|u\|_V.$$

□

Osservazione 3.45. La condizione inf-sup pù essere scritta in termini dell'operatore lineare A associato a $a(,)$ e del suo aggiunto A^*. Infatti, sia:

$$A : V' \to V \qquad\qquad A^* : V \to V',$$

con V' lo spazio duale di V (rispetto alla forma lineare $f()$) e gli operatori lineari A e A^* sono definiti da:

$$(Au, v)_{V' \times V} = a(u,v) \qquad (u, AV)_{V \times V'} = a(u,v),$$

allora la condizione (51) è equivalente alle:

$$\|Au\|_{V'} \geq \alpha \|u\|_V \qquad \forall u \in V; \tag{52}$$

$$\|A^* v\|_{V'} \geq \alpha \|v\|_V \qquad \forall v \in V. \tag{53}$$

Nel caso visto prima in cui $V = \mathbb{R}^n$, è chiaro che la condizione di coercività implica che la matrice A è definita positiva, mentre la condizione inf-sup indica che la matrice A è invertibile.

Possiamo ora dimostrare il seguente teorema di stabilità.

Teorema 3.4. *La forma bilineare continua $a(,)$ sodddisfa la condizione inf-sup se e solo se l'operatore A è biunivoco.*

Si noti che il fatto che A è biunivoco implica che il problema (3.42) ha soluzione unica per ogni $f()$, e cioè A ha inversa continua $\|u\|_V \leq C \|f()\|_{V'}$.

Dimostrazione. Se $a(,)$ soddisfa la condizione inf-sup, le condizioni (52) e (53) mostrano rispettivamente che gli operatori A e A^* sono iniettivi. Per dimostrare la biunivocità basta quindi dimostrare che l'immagine di A $R(A)$ è chiusa. Sia $Au_n \to w$, allora:

$$\|A(u_n - u)m\|_{V'} \geq \alpha \|u_n - u_m\|_V ,$$

che mostra che $\{u_n\}$ è una successione di Cauchy e quindi converge a un elemento $u \in V$. Per la continuità di A si ha subito che $w = Au \in R(A)$.

D'altro canto, se A è biunvoca, allora anche A^* lo è e quindi ha inversa continua. □

3.10.2 Formulazione FEM

La formulazione FEM si ottiene dalla formulazione variazionale vista prima utilizzando al posto di V unospazio finito-dimensionale $V_h \subset V$. L'insieme delle funzioni $\{\phi_1, \ldots, \phi_n\}$ sia una base per V_h cosicché ogni $v \in V_h$ può essere espressa come combinazione lineare delle ϕ_i:

$$v = \sum_{j=1}^{n} \xi_i \phi_i(x). \tag{54}$$

Abbiamo quindi:

Problema 3.46 (FEM, metodo di Ritz)**.** Trovare $u_h \in V_h$ rale che:

$$F(u_h) \leq F(v) \qquad \forall v \in V_h. \tag{55}$$

o equivalentemente:

Problema 3.47 (FEM, metodo di Galerkin)**.** Trovare $u_h \in V_h$ rale che:

$$a(u_h, v) = (f, v) \qquad \forall v \in V_h. \tag{56}$$

Usando la rappresentazione di u_h con le funzioni di base ϕ_i:

$$u_h = \sum_{j=1} n u_j \phi_j, \qquad u_j \in \mathbb{R},$$

otteniamo il seguente sistema lineare:

$$\sum_{j=1}^{n} u_j a\left(\phi_j, \phi_i\right) = f\left(\phi_i\right), \qquad i = 1, \ldots, n,$$

ovvero in forma matriciale:

$$Au = b,$$

dove:

$$u = \{u_i\}, \qquad A = \{a_{ij}\},\ a_{ij} = a\left(\phi_j, \phi_i\right), \qquad b = b_i,\ b_i = f\left(\phi_i\right).$$

La matrice A è detta *matrice di rigidezza*. Abbiamo il seguente:

Teorema 3.5. *La matrice di rigidezzza A è simmetrica e definita positiva.*

Dimostrazione. La simmetria deriva dalla simmetria della forma bilineare.

Usando la (54) si ottiene:

$$a\left(v, v\right) = a\left(\sum_{i=1}^{n} \xi_i \phi_i, \sum_{i=1}^{n} \xi_i \phi_i\right) = \sum_{i,j=1}^{n} \xi_i a\left(\phi_i, \phi_j\right) \xi_j = \xi \cdot A\xi,$$

dove $\xi = \{\xi_i\}$ è un vettore di $\mathbb{R}^n$ e il punto denota il prodotto scalare tra vettori. Dalla condizione di coercività (46) segue che:

$$\xi \cdot A\xi = a\left(v, v\right) \geq \alpha \left\|v\right\|_V^2 > 0$$

se $v \neq 0$, e cioè se $\xi \neq 0$. □

Teorema 3.6. *I due problemi* (3.46) *e* (3.47) *sono equivalenti e ammettono soluzione unica $u_k \in V_h$. Inoltre si ha la seguente stima di stabilità:*

$$\left\|u_h\right\|_V \leq \frac{\Lambda}{\alpha}.$$

Dimostrazione. Esistenza e unicità derivano direttamente dal teorema 3.5. Scegliendo $v = u_h$ in (56) e usando li condizioni di coercività di $a\left(,\right)$ e di conftinuità di $f\left(\right)$ si ottiene:

$$\alpha \left\|u_h\right\|_V^2 \leq a\left(u_h, u_h\right) \leq \Lambda \left\|u_h\right\|_V.$$

La stabilità si ottiene quindi dividendo per $\left\|u_h\right\|_V \neq 0$. □

Il risultato successivo, noto come Lemma di Céa, dice che l'errore relativo a u_h è ottimale in V_h a meno di costanti.

Teorema 3.7. *Si indichi con $u \in V$ la soluzione del problema 3.42 e con $u_h \in V_h$, con $V_h \subset V$, la soluzione di 3.47. Allora:*

$$\|u - u_h\|_v \leq \frac{\gamma}{\alpha} \|u - v\|_V \qquad \forall v \in V_h.$$

Dimostrazione. Poiché $V_h \subset V$, sottraendo membro a membro le (48) e (56), si ottiene immediatamente che lo schema FEM di Galerkin, e per l'equivalenza anche quello di Ritz, sono fortemente consistenti:

$$a(u - u_h, v) = 0 \qquad \forall v \in V_h. \tag{57}$$

Prendiamo quindi $w = u_h - v$ e notiamo che $w \in V_h$. Allora $v = u_h - w$, e per la consistenza e la coercività si ha:

$$\begin{aligned} \alpha \|u - u_h\|_V^2 \leq a(u - u_h, u - u_h) &= a(u - u_h, u - u_h) + a(u - u_h, v) \\ &= a(u - u_h, u - v) \leq \gamma \|u - u_h\|_V \|u - v\|_V . \end{aligned}$$

Il risultato segue dividendo per $\|u - u_h\|_V$. □

Questo risultato ci dice che una stima quantitativa dell'errore si può ottenere usando una opportuna $v \in V_h$. Nel nostro caso utilizzeremo per v una opportuna funzione interpolante di u che appartenga a V_h. Per esempio, possiamo prendere $v = \Pi_h u$ dve Π_h è un interpolatore lineare a tratti.

La norma energia. Se la forma bilineare $a(,)$ è simmetrica, si può definre una nuova norma in V:

$$\|v\|_a^2 = a(v, v), \qquad v \in V.$$

Tale norma è equivalente alla norma classica in V:

$$\sqrt{\alpha} \|v\|_V \leq \|v\|_a \leq \sqrt{\gamma} \|v\|_V ,$$

con prodotto scalare corrispondente dato da:

$$(u, v)_a = a(u, v) .$$

Questa norma si chiama *norma energia*, i risultati precedenti ci dicono che u_h è la proiezione di u su V_h rispetto al prodotto scalare $(\cdot, \cdot)_a$, e che u_h è la migliore approssimazione di u in norma energia.

Operatori non coercivi. Si può notare che nella dimostrazione del teorema 3.7 non abbiamo usato la simmetria della forma $a(,)$. Il lemma di Céa si può estendere anche a operatori non coercivi e si può particolarizzare per operatori simmetrici. Per operatori non cercivi però bisogna invocare la condizione inf-sup su ogni spazio V_h, per cui deve esistere un coefficiente $\beta > 0$ tale per cui:

$$\sup_{v \in V_h} \frac{a(u,v)}{\|v\|_V} \geq \alpha \|u\|_V \qquad \forall u \in V_h.$$

La seconda condizione inf-sup deriva dalla precedente perchè V_h è finito-dimensionale. Se la costante α è indipendente da h, allora abbiamo:

Teorema 3.8 (Lemma di Céa)**.** *Sia $u \in V$ soluzione del problema 3.42 e $u_h \in V_h$, con $V_h \subset V$, soluzione di 3.47. Allora abbiam le seguenti stime:*

1. *Se $a(,)$ non è coerciva, allora:*

$$\|u - u_h\|_v \leq \left(1 + \frac{\|a(,)\|}{\alpha}\right) \inf_{v \in V_h} \|u - v\|_V ,$$

dove

$$\|a(,)\| = \sup_{v \in V_h, v \neq 0} \frac{a(v,v)}{\|v\|_V^2};$$

2. *se $a(,)$ è continua e coerciva allora:*

$$\|u - u_h\|_v \leq \frac{\gamma}{\alpha} \inf_{v \in V_h} \|u - v\|_V ;$$

3. *se $a(,)$ è anche simmetrica, allora*

$$\|u - u_h\|_v \leq \sqrt{\frac{\gamma}{\alpha}} \inf_{v \in V_h} \|u - v\|_V .$$

Dimostrazione. Riportiamo solo la dimostrazione del primo punto, essendo quella degli altri punti immediata. Sia $v \in V_h$. Usando la condizione inf-sup e la consistenza, si ha, analogamente al caso coercivo:

$$\alpha \|v - u_h\|_V \leq \sup_{w \in V_h} \frac{a(v - u_h, w)}{\|w\|_V} = \sup_{w \in V_h} \frac{a(v - u, w)}{\|w\|_V} \leq M \|v - u\|_v ,$$

dove

$$M = \|a(,)\| = \sup_{v \in V_h, v \neq 0} \frac{a(v,v)}{\|v\|_V^2};$$

la dimostrazione segue dalla disuguaglianza triangolare. □

Corollario 3.48. *Sia $\{V_h\}$ una sequenza di sottospazi finito-dimensionali di V, indicizzati da un parametro h. Assumendo che per $h \to 0$ sia:*

$$\inf_{v_h \in V_h} \|v - v_h\|_V \to 0,$$

allora, se $\alpha = \inf_h \alpha_h > 0$, u_h converge a u in V.

Le dimostrazioni del Lemma di Céa e del suo corollario sono immediate. L'ipotesi del corollario serve perchè non ne abbiamo ancora dimostrato la validità. Tale ipotesti dovrà essere presa in considerazione per la determinazione delle stime di convergenza quantitative.

Esempio 3.49. Sia $V = H_0^1(\Omega)$, $\Omega \subset \mathbb{R}^2$ e consideriamo:

$$a(v, w) = \int_\Omega \nabla v \cdot \nabla w \, dx; \qquad f(v) = \int_\Omega f v \, dx,$$

conf $f \in L^s(\Omega)$. La forma bilineare è chiaramente simmetrica e continua. La V-ellitticità (o coercività) deriva dall'applicazione del lemma di Poincaré 3.37. Dai risultati precedenti risulta quindi:

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch,$$

se u è sufficientemente liscia.

Esempio 3.50. Si consideri il problema di convezione-diffusione in $\mathbb{R}^2$:

$$\begin{aligned} -\mu \Delta u + \operatorname{div}(\beta u) + u &= f \qquad in \Omega \\ u &= 0 \qquad in \partial\Omega \end{aligned}$$

dove $\beta = (\beta_1, \beta_2)$ è un vettore di $\mathbb{R}^2$. Moltiplicando per $v \in V = H_0^1(\Omega)$, integrando su Ω e applicando il lemma di Green al primo termine, si ottiene:

$$a(u, v) = f(v) \qquad \forall v \in V,$$

dove:

$$a(v, w) = \int_\Omega \left(\nabla v \cdot \nabla w + \operatorname{div}(\beta v) w\right) \, dx, \qquad f(v) = \int_\Omega f v \, dx.$$

Si assuma $\mu = 1$ e che $|\beta|/\mu$ sia piccolo. Il problema è coercivo. Infatti, applicando il lemme di Green al secondo termine otteniamo:

$$\begin{aligned} \int_\Omega \operatorname{div}(\beta v) v \, dx = \\ \int_\Gamma (\beta \cdot n) v^2 \, ds - \int_\Omega \operatorname{div}(\beta v) v \, dx \\ = - \int_\Omega \operatorname{div}(\beta v) v \, dx, \end{aligned}$$

da cui risulta:

$$\int_\Omega \operatorname{div}(\beta v) v \, dx = 0.$$

per cui:

$$a(v,v) = \int_\Omega \left(|\nabla v|^2 + v^2 \right) \, dx = \|v\|^2_{H^1(\Omega)}.$$

Si può quindi formulare un problema FEM: trovara $u_h \in V_h$ tale che:

$$a(u_h, v) = f(v) \qquad \forall v \in V_h.$$

La matrice di rigidezza risultante non è più simmetrica ma la coercività garantisce l'esistenza e l'unicità della soluzione. Abbiamo quindi la seguente stima dell'errore ($\alpha = 1$):

$$\|u - u_h\|_{H^1(\Omega)} \leq \gamma \|u - v\|_{H^1(\Omega)} \qquad \forall v \in V_h.$$

Esempio 3.51. Sia u la funzione temperatura in un corpo conduttore di forma $\Omega \in \mathbb{R}^3$. Il flusso di calore in ogni punto è dato dalla legge di Fourier:

$$q_i(x) = -k_i(x) \frac{\partial u}{\partial x_i} \qquad x \in \Omega; i = 1, 2, 3, ;$$

la conservazione dell'energia è data da:

$$\operatorname{div} q = \sum_{i=1}^{3} \frac{\partial}{\partial x_i} \left(k_i(x) \frac{\partial u}{\partial x_i} \right) = f \qquad x \in \Omega;$$

questo è un esempio di una equazione alle derivate parziali a coefficienti variabili. La formulazione variazionale della precedente richiede l'aggiunta di condizioni al contorno:

$$\begin{aligned} u &= 0 && in \Gamma_D \\ -q \cdot n &= g && in \Gamma_N \end{aligned}$$

con $\partial\Omega = \Gamma = \Gamma_D \cup \Gamma_N$.

Sia $V = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D\}$. Moltiplicando l'equazione di conservazione dell'energia per $v \in V$ e usando il lemma di Green otteniamo:

$$\int_\Omega fv \, dx = \int_\omega v \operatorname{div} q \, dx = \int_\Gamma vq \cdot n \, ds - \int_\Omega q \cdot \nabla v \, dx = \sum_{i=1}^{3} \int_\Omega k_i(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, dx - \int_{\Gamma_N} gv \, ds,$$

e quindi il seguente problema variazionale (alla Galerkin): Trovare $u \in V$ tale che:

$$a(u, v) = f(v) \qquad \forall v \in V,$$

dove:

$$a(v,w) = \sum_{i=1}^{3} \int_{\Omega} k_i(x) \frac{\partial u}{\partial x_i}$$
$$f(v) = \int_{\Omega} fv \, dx + \int_{\Gamma_N} gv \, ds.$$

Si verifica facilmente che la forma bilineare è simmetrica, continua e coerciva, e che la forma lineare e' continua, se esistono due costanti c e C tali per cui:

$$c \leq k_i(x) \leq C, \qquad \forall x \in \Omega; i = 1, 2, 3,$$

e inoltre $f \in L^2(\Omega)$ e $g \in L^2(\Gamma_N)$, e l'area di Γ_D è positiva (non nulla).

3.11 Spazi degli elementi finiti

Bisogna ora specializzare gli spazi V_h. Tali spazi saranno formati da polinomi continui a tratti definiti su suddivisioni del dominio $\Omega \in \mathbb{R}^d$, chiamate triangolazioni. Una triangolazione $\mathcal{T}_h = \{K\}$ è formata quindi dall'unione di elementi T (o suddivisioni) che ricoprono Ω senza sovrapposizioni. Gli spazi che cerchiamo sono sottospazi finito-dimensionali di $H^1(\Omega)$ (o di $H^2(\Omega)$ per PDE del quart'ordine). Essendo formati da polinomi continui a trarri, s avrà:

$$V_h \subset H^1(\Omega) \Leftrightarrow V_h \subset C^0(\bar{\Omega})$$
$$V_h \subset H^2(\Omega) \Leftrightarrow V_h \subset C^1(\bar{\Omega})$$

dove $\bar{\Omega} = \Omega \cup \Gamma$.

3.11.1 Caso bi-dimensionale ($d = 2$)

Sia $\Omega \in \mathbb{R}^2$ con contorno Γ poligonale. Sia $\mathcal{T}_h = \{T\}$ una triangolazione con elementi triangolari T. Sia $\mathcal{P}_r(T)$ il polinomio di grado r in T:

$$\mathcal{P}_r(T) = \{v : v \text{ polinomio di grado } \leq r \text{ in } K\}.$$

Il polinomio in $\mathcal{P}_1(T)$ può essere quindi scritto come:

$$v(x) = a_{00} + a_{10}x_1 + a_{01}x_2, \qquad x \in T, \tag{58}$$

con $a_{ij} \in \mathbb{R}$. Si vede immediatamente che $\psi_1(x) = 1$, $\psi_2(x) = x_1$, $\psi_3(x) = x_2$ formano una base per $\mathcal{P}_1(T)$.

Nel caso quadratico abbiamo:

$$v(x) = a_{00} + a_{10}x_1 + a_{01}x_2 + a_{20}x_1^2 + a_{11}x_1x_2 + a_{02}x_2^2, \qquad x \in T,$$

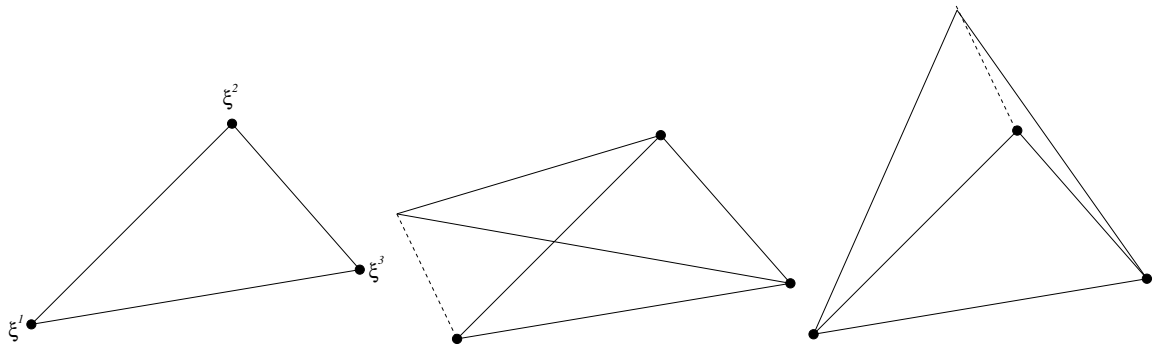

Figura 15: A sinistra: triangolo con gradi di libertà che definiscono una funzione lineare. A destra: esempio di funzioni di base lineari.

con $a_{ij} \in \mathbb{R}$ e una base data da: $\{1, x_1, x_2, x_1^2, x_1x_2, x_2^2\}$. In generale possiamo scrivere:

$$\mathcal{P}_r(T) = \left\{ v : v(x) = \sum_{0 \le i+j \le r} a_{ij} x_1^i x_2^j \text{ for } x \in T \right\} \qquad \dim \mathcal{P}_r(T) = \frac{(r+1)(r+2)}{2}.$$

Esempio 3.52. Polinomi affini su triangoli (Fig.15:

$$V_h = \left\{ v \in C^0(\bar{\Omega}) : v|_T \in \mathcal{P}_1(T), \forall T \in \mathcal{T}_h \right\}.$$

Lo spazio V_h è quindi formato dalle funzioni continue e lineari a tratti. Per descrivere queste funzioni usiamo i "gradi di libertà" che in questo caso sono i nodi di $\mathcal{T}_h$. Quindi ogni funzione $v(x)$ in $V_h(T)$ è univocamente determinata dai valori ai vertici di T. Indicando con ξ^i, $i = 1, 2, 3$, le coordinate di tali vertici, abbiamo per $\alpha_i \in \mathbb{R}$ il seguente:

Teorema 3.9. *Sia $T \in \mathcal{T}_h$ un triangolo i cui vertici hanno coordinate ξ^i, $i = 1, 2, 3$. Una funzione $v(x) \in \mathcal{P}_1(T)$ è determinata univocamente dai suoi valori ai vertici. Cioè, dati i valori $\alpha_i \in \mathbb{R}$, $i = 1, 2, 3$, $v(x) \in \mathcal{P}_1(T)$ è univocamente determinata da:*

$$v(\xi^i) = \alpha_i \qquad i = 1, 2, 3 \tag{59}$$

Dimostrazione. La generica funzione $v(x)$ può essere scritta come in (58). Quindi il sistema lineare (59) ha soluzione unica se la matrice:

$$A = \begin{bmatrix} \xi_1^1 & \xi_2^1 & 1 \\ \xi_1^2 & \xi_2^2 & 1 \\ \xi_1^3 & \xi_2^3 & 1 \end{bmatrix}$$

è non singolare. Ma questo è vero perchè il nucleo di A è vuoto. Infatti, se esistessa un vettore non nullo $a = (a_1, a_2, a_3)$ tale per cui $Aa = 0$, allora si avrebbe l'assurdo che un polinomio in $\mathbb{R}^2$ di grado 1 ammette 3 radici. □

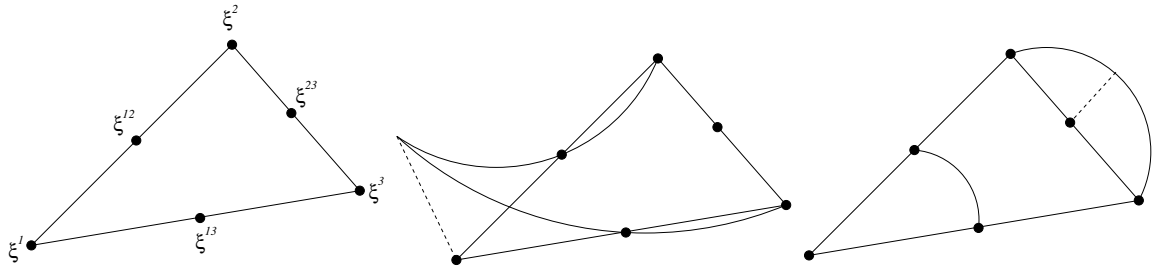

Figura 16: A sinistra: triangolo con gradi di libertà che definiscono una funzione quadratica. A destra: esempio di funzioni di base quadratiche.

Per determinare le funzioni di base basta quindi scegliere i valori di α_i opportunamente. Scegliamo i valori $(1,0,0)$, $(0,1,0)$ e $(0,0,1)$ in accordo con la tecnica dell'interpolazione lagrangiana. E' facile verificare che le funzioni di base $\phi_i(x)$ così determinate sono continue su $\bar{\Omega}$ e hanno gradiente costante a tratti. In ogni elemento T si ha:

$$v(x)|_T = \sum_{i=1}^{3} v(\xi^i)\phi_i(x) \qquad \nabla v(x)|_T = \sum_{i=1}^{3} v(\xi^i)\nabla\phi_i(x).$$

Esempio 3.53. Funzioni di base quadratiche. Lo spazio V_h è dato da:

$$V_h = \left\{ v \in C^0(\bar{\Omega}) : v|_T \in \mathcal{P}_2(T), \forall T \in \mathcal{T}_h \right\}.$$

Per descrivere queste funzioni abbiamo bisogno di 6 "gradi di libertà" per ogni $T \in \mathcal{T}_h$. Scegliamo quindi i valori ai vertici di K e ai punti medi di ogni lato (Fig. 16). Abbiamo il seguente:

Teorema 3.10. *Sia $T \in \mathcal{T}_h$ un triangolo i cui vertici hanno coordinate ξ^i, $i = 1,2,3$. Indichiamo con ξ^{ij} le coordinate dei punti medi del lato del triangolo compreso tra i nodi i e j. Una funzione $v(x) \in \mathcal{P}_2(T)$ è determinata univocamente da:*

$$v(\xi^i) = \alpha_i \qquad i = 1,2,3 \qquad\qquad v(\xi^{ij}) = \alpha_{ij} \qquad i < j, \;\; i,j = 1,2,3.$$

Dimostrazione. Basta verificare che le condizioni $v(\xi^i) = 0$ e $v(\xi^{ij}) = 0$ $i < j, i, j = 1,2,3$ implicano $v = 0$ su tutto T. Consideriamo un lato alla volta. Prendiamo ad esempio il lato tra i nodi 2 e 3 (Fig. 16). La funzione (quadratica) ristretta a questo lato è univocamente determinata dai 3 punti ξ^2, ξ^{23} e ξ^3. Se v è nulla su questi tre nodi, è identicamente nulla sul lato 2-3. Questo implica che si può fattorizzare una funzione $\phi_1(x)$ (il polinomio di grado 1 dell'esempio precedente):

$$v(x) = \phi_1(x) w_1(x).$$

Lo stesso succede ad esempio nel lato tra i nodi 1 e 3, per cui:

$$v(x) = \phi_1(x)\phi_2(x)w_0,$$

dove ora w_0 è una costante. Ora, prendiamo $v(\xi^{12}) = 0$. Abbiamo:

$$0 = v(\xi^{12}) = \phi_1(\xi^{12})\phi_2(\xi^{12})w_0 = \frac{1}{2}\frac{1}{2}w_0,$$

per cui $w_0 = 0$, e quindi il risultato. □

Le funzioni di base quadratiche possono scriversi in termini delle funzioni di base lineari come:

$$v(x)|_K = \sum_{i=1}^{3} v(\xi^i)\phi_i(x)(2\phi_i(x) - 1) + \sum_{\substack{i,j=1 \\ i<j}}^{3} v(\xi^{ij})4\phi_i(x)\phi_j(x).$$

Un esempio di tali funzioni sono mostrate in Fig. 16.

3.12 Stime dell'errore per problemi ellittici

Per equazioni ellittiche la cui forma bilineare associata alla formulazione debole sia coerciva con costante α e continua con costante γ, il lemma di Céa visto nel paragrafo precedente ci assicura che:

$$\|u - u_h\|_V \leq \frac{\gamma}{\alpha} \|u - v\|_V \qquad \forall v \in V_h.$$

Possiamo quindi prendere al posto della funzione v il polinomio interpolatore di u, $\Pi_h u$, così che la stima dell'errore si riconduce alla stima dell'errore di interpolazione $\|u - \Pi_h u\|_V$. Se prendiamo i gradi di libertà definiti nel paragrafo precedente per determinare $\Pi_h u$, basta stimare $\|u - \Pi_h u\|_V$ individualmente su ogni elemento $T \in \mathcal{T}_h$.

Errore di interpolazione. Consideriamo un dominio bidimensionale. L'estensione al caso generale è immediata. Dato $\Omega \in \mathbb{R}^2$ con sufficientemente liscio, sia $\mathcal{T}_h(\Omega)$ una suo partizionamento in triangoli non sovrapposti. In altre parole, indicando con T_j il j-esimo triangolo di $\mathcal{T}_h(\Omega)$ si avrà:

$$\mathcal{T}_h(\Omega) = \bigcup_{j=1}^{M} T_j$$

$$T_j \cap T_i = \begin{cases} \emptyset \\ \sigma_{ij} \end{cases}$$

dove σ_{ij} indica il lato tra i nodi i e j. Per ogni $T \in \mathcal{T}_h$ definiamo:

$$h_T = \text{ il diametro di } T = \text{ lato di lunghezza massima di } T;$$
$$\rho_T = \text{ diametro del cerchio inscritto in } T;$$

La triangolazione $\mathcal{T}_h$ è caratterizzata da un unico parametro di griglia, h, definito da:

$$h = \max_{T \in \mathcal{T}_h} h_T.$$

Si consideri quindi una famiglia di triangolazioni $\{\mathcal{T}_h\}$ di Ω e una corrispondente famiglia di spazi funzionali $V_h = \{v \in H^1(\Omega) : v|T \in \mathcal{P}_1(T)\}$, indicizzati dal parametro h. Assumiamo inoltre che la triangolazione sia "regolare", che esista cioè una costante $\beta > 0$ indipendente da h e dalla particolare triangolazione $\mathcal{T}_h \in \{\mathcal{T}_h\}$ tale che:

$$\frac{\rho_T}{h_T} \geq \beta \qquad \forall T \in \mathcal{T}_h.$$

Il valore di β è una misura del minimo angolo tra tutti i triangoli T, e la regolarità della triangolazione ci assicura che nel processo di limite $h \to 0$ non esistono angoli che tendono a zero. Enunciamo quindi il seguente teorema classico dell'interpolazione polinomiale a tratti, la cui dimostrazione si può trovare per esempio in [7]:

Teorema 3.11. *Sia $T \in \mathcal{T}_h$ un triangolo con vertici ξ^i, $i = 1, 2, 3$. Sia $v(x) \in H^{r+1}(T)$ e sia $\Pi_h v \in \mathcal{P}_r(T)$ la sua interpolante lagrangiana. Allora, per ogni triangolo T si ha:*

$$\|v - \Pi_h v\|_{L^2(T)} \leq C h_T^{r+1} \left\|\partial^{r+1} v\right\|_{L^2(T)},$$
$$|\, v - \Pi_h v \,|_{H^1(T)} \leq C \frac{h_T^{r+1}}{\rho_T} \left\|\partial^{r+1} v\right\|_{L^2(T)}.$$

Osservazione 3.54. Notiamo che nella seconda disuguaglianza abbiamo ora la presenza del paramtero di griglia ρ_T. Questo parametro entra in gioco non appena andiamo a stimare le derivate di v e di $\Pi_h v$ sul T, in quanto si verifica facilmente che la norma del gradiente delle funzioni $v \in H^{r+1}(T)$ è maggiorata da $1/\rho_T$.

Passando alla famiglia di triangolazioni, abbiamo il seguente:

Corollario 3.55. *Se la famiglia di trangolazioni $\{\mathcal{T}_h\}$ è regolare, allora esistono due costanti C_1 e C_2 indipendenti da h e da $v \in H^{r+1}(\Omega)$ tali che:*

$$\|v - \Pi_h v\|_{L^2(\Omega)} \leq C_1 h^{r+1} \left\|\partial^{r+1} v\right\|_{L^2(\Omega)}, \tag{60}$$
$$|\, v - \Pi_h v \,|_{H^1(\Omega)} \leq C_2 h^{r} \left\|\partial^{r+1} v\right\|_{L^2(\Omega)}. \tag{61}$$

Dimostrazione. Dimostriamo il corollaio per $r = 1$ (interpolazione lineare). In questo caso le due stime del teorema (3.11) si possono specificare in:

$$\|v - \Pi_h v\|_{L^2(T)} \leq C h_T^2 \left\|\partial^2 v\right\|_{L^2(T)},$$
$$|\, v - \Pi_h v \,|_{H^1(T)} \leq C \frac{h_T^2}{\rho_T} \left\|\partial^2 v\right\|_{L^2(T)}.$$

Sommando su ogni triangolo $T \in \mathcal{T}_h$ si ottiene:

$$\|v - \Pi_h v\|^2_{L^2(\Omega)} = \sum_{T \in \mathcal{T}_h} \|v - \Pi_h v\|^2_{L^2(T)} \leq \sum_{T \in \mathcal{T}_h} C^2 h_T^4 \left\|\partial^2 v\right\|^2_{L^2(T)} \leq C^2 h^4 \left\|\partial^2 v\right\|^2_{L^2(\Omega)},$$

mentre per la seconda, ricordando che $h_T/\rho_T \leq 1/beta$, abbiamo:

$$|\, v - \Pi_h v \,|^2_{H^1(\Omega)} = \sum_{T \in \mathcal{T}_h} |\, v - \Pi_h v \,|^2_{H^1(T)} \leq \sum_{T \in \mathcal{T}_h} C^2 \frac{h_T^4}{\rho_T^2} \left\|\partial^2 v\right\|^2_{L^2(T)} \leq \frac{C^2}{\beta} h^2 \left\|\partial^2 v\right\|^2_{L^2(\Omega)}.$$

□

Si vede subito che la regolarità della soluzione determina, assieme all'ordine dell'interpolazione, l'accuratezza dello schema. Si avrà in generale, per $1 \leq s \leq r + 1$:

$$\|v - \Pi_h v\|_{L^2(T)} \leq C h_T^s \|\partial^s v\|_{L^2(T)},$$
$$|\, v - \Pi_h v \,|_{H^1(T)} \leq C h_T^{s-1} \|\partial^s v\|_{L^2(T)}.$$

Stima dell'errore FEM per problemi ellittici e regolarità della soluzione. Dal lemma di Céa si ottiene immediatamente la stima dell'errore del metodo FEM sostituendo a v l'interpolante di u:

$$\|u - u_h\|_V \leq \frac{\gamma}{\alpha} \|u - \Pi_h u\|_V,$$

e usando le stime dell'errore di interpolazione possiamo specificare le stime di errore per i diversi problemi FEM.

Ad esempio, consideriamo l'equazione di Poisson:

$$-\Delta u = f \qquad \text{in } \Omega$$
$$u = 0 \text{ in } \Gamma = \partial\Omega$$

Sia quindi $V = H_0^1(\Omega)$ e $V_h = \{v \in V : v|_T \in \mathcal{P}_r(T) \forall T \in \mathcal{T}_h\}$. Allora, per il problema di Poisson con condizioni al contorno di Dirichlet abbiamo:

$$\|u - u_h\|_{H^!(\Omega)} \leq C h^r |\, u \,|_{H^{r+1}(\Omega)}.$$

La teoria delle equazioni ellittiche ci dice che se Γ è liscio, e in particolare è una cruva senza ha cuspidi o angoli, si ha la seguente stima:

$$\|u\|_{H^{s+2}(\Omega)} \leq C \|f\|_{H^s(\Omega)}, \tag{62}$$

o, con parole approssimative, la soluzione è più regolare della forzante f di "due derivate".

Se Γ non è liscio, tale stima potrebbe non essere vera (neanche per $s = 0$!). Per esempio, se Ω non è convesso ed ha un punto angoloso, la soluzione avrà una singolarita in tale punto anche se f è liscia. Possiamo pensare di approssimare in tale punto la soluzione u come (usiamo qui coordinate polari centrate nel punto angoloso):

$$u(r, \theta) = r^\gamma \alpha(\theta) + \beta(r, \theta) \qquad \gamma = \frac{\pi}{\omega}, \tag{63}$$

dove ω è l'angolo formato dal contorno. Si può dimostrare che la stima (62) vale con $s = 0$ se $\omega < \pi$, caso di dominio convesso con contorno poligonale. Se invece $\omega > \pi$, una funzione della forma (63) non appartiene a $H^2(\Omega)$ se $\alpha \neq 0$. In particolare, si vede subito che:

$$\int_\Omega | \partial^s u |^2 \, dx \approx C \int_0^R [r^{\gamma - s}]^2 r \, dr.$$

Tale integrale esiste finito, e quindi $u \in H^s(\Omega)$, se $s < \gamma + 1$. Per un problema FEM per la soluzione dell'equazione di Poisson con dominio a contorno polinomiale non convesso, possiam quindi scrivere per ogni $\epsilon > 0$:

$$\|u - u_h\|_{H^1(\Omega)} \leq C h^{\gamma - \epsilon} \|u\|_{H^{\gamma - \epsilon + 1}(\Omega)} = C h^{\gamma - \epsilon},$$

dove $\gamma = \pi/\omega$, e ω è l'angolo massimo dei punti angolosi di Γ. Per esempio, se $\gamma = 2/3$, corrispondente ad un angolo concavo $\omega = 3\pi/2$, non si ha convergenza piena ($O(h)$) del metodo FEM:

$$\|u - u_h\|_{H^1(\Omega)} \leq C h^{\frac{2}{3} - \epsilon}.$$

Esistono metodi cosidetti adattativi che cercano di diminuire la dimensione caratteristica della griglia h_T nei punti dove si prevede convergenza non ottimale. Non si discuterà di tali metodi qui. Basti sapere che sono molto usati anche se la complicazione nel costruire la griglia in maniera adattativa ne rende difficoltosa l'applicazione concreta.

Stima dell'errore in L^2 per l'equazione di Poisson. Dall'analisi fatta fino a ora, si riconosce che si ha una discrepanza tra ordine di convergenza dello schema FEM dovuta al lemma di Céa e ordine di convergenza dell'errore di interpolazione. Infatti, tutto quello che possiamo dire al momento per l'errore del FEM è:

$$\|u - u_h\|_{H^1(\Omega)} \leq C h | u |_{H^2(\Omega)},$$

mentre per funzioni di base lineari a tratti l'errore di interpolazione è:

$$\|u - \Pi_h u\|_{L^2(\Omega)} \leq Ch^2 |\, u \,|_{H^2(\Omega)},$$

e si vede che non è imediato avere on $O(h^2)$ nell'errore FEM. In realtà la convergenza ottimale è quadratica. Il procedimento per dimostrarlo è chiamato "trucco di Aubin-Nitsche", e porta al seguente:

Teorema 3.12. *Sia Ω un cominio poligonale convesso e u_h la soluzione FEM dell'equazione di Poisson con funzioni di base lineari a tratti. Allora esiste una costante C indipendente da h e u tale che:*

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^2 |\, u \,|_{H^2(\Omega)}.$$

Dimostrazione. Per brevità definiamo la funzione errore $e = u - u_h$. Dalla consistenza dello schema possiamo scrivere immediatamente:

$$a\,(e, v) = 0 \qquad \forall v \in V_h, \tag{64}$$

Vogliamo quindi stimare ora la norma L^2 dell'errore, $\|e\|_{L^2(\Omega)} = (e, e)^{1/2}$. Sia φ soluzione del problema ausiliario:

$$\begin{aligned} -\Delta\varphi &= e \qquad \text{in } \Omega \\ \varphi &= 0 \text{ in } \Gamma. \end{aligned}$$

Siccome Ω è convesso, vale la (62) con $s = 0$:

$$\|\varphi\|_{H^2(\Omega)} \leq C \|e\|_{L^2(\Omega)}\,. \tag{65}$$

Usando la formula di Green e il fatto che $e = 0$ in Γ, osservando dalla (64) che $a\,(e, \Pi_h\varphi) = 0$, otteniamo:

$$(e, e) = (e, -\Delta\varphi) = a\,(e, \varphi) = a\,(e, \varphi - \Pi_h\varphi)\,.$$

Usando il lemma di Green e sfruttando il fatto che $\varphi = 0$ al bordo si ottiene:

$$\|e\|^2_{L^2(\Omega)} \leq \int_\Omega \nabla e \nabla(\varphi - \Pi_h\varphi \leq \|\nabla e\|_{L^2(\Omega)} \|\nabla(\varphi - \Pi_h\varphi)\|_{L^2(\Omega)}$$

$$\|e\|_{H^1(\Omega)} \|\varphi - \Pi_h\varphi\|_{H^1(\Omega)}\,.$$

Possiamo usare ora la stima di interpolazione (61) con $r = 1$:

$$\|e\|^2_{L^2(\Omega)} \leq C \|e\|_{H^1(\Omega)} h |\, \varphi \,|_{H^2(\Omega)};$$

la stima (65) relativa al problema di Poisson ausiliario ci fornisce quindi:

$$\|e\|^2_{L^2(\Omega)} \leq Ch \|e\|_{H^1(\Omega)} h \|e\|_{L^2(\Omega)},$$

da cui, dividendo per $\|e\|_{L^2(\Omega)}$, si ottiene:

$$\|e\|_{L^2(\Omega)} \leq Ch \|e\|_{H^1(\Omega)};$$

ricordando il Lemma di Céa, si ha infine:

$$\|u - u_h\|_{L^2(\Omega)} \leq ch^2 |\, u \,|_{H^2(\Omega)}.$$

□

3.13 Stima del condizionamento della matrice di rigidezza

Siamo in grado ora di provare l'affermazione riportata nell'osservazione 3.12. Prendiamo come esempio il caso dell'equazione di Poisson discretizzata con elementi finiti lineari su una triangolazione regolare $\mathcal{T}_h$ con parametro di mesh h (cfr. paragrafo 3.12). In questo caso, la matrice di rigidezza è data da:

$$A = \{a_{ij}\} \qquad a_{ij} = a(\phi_i, \phi_j) \qquad a(\phi_i, \phi_j) = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j,$$

con $\phi_j \in \mathcal{P}_1(\mathcal{T}_h)$. Dimostriamo quindi il seguente:

Teorema 3.13. *Il numero di condizionamento della matrice di rigidezza A può essere stimato come:*

$$\kappa(A) = \mathcal{O}\left(h^{-2}\right).$$

In particolare, l'autovalore massimo di $\lambda_1 A = \mathcal{O}(1)$ e quello minimo vale $\lambda_n A = \mathcal{O}(h^2)$.

Prima di procedere alla dimostrazione, dimostriamo il seguente risultato, noto con il nome di "stima inversa" perchè ci permette di stimare il gradiente della soluzione con la soluzione stessa, con la conseguente comparsa di un fattore $1/h$.

Lemma 3.56 (Stima inversa)**.** *Esistono due costanti c e C dipendenti solo dalle costanti di regolarià della triangolazione $\mathcal{T}_h$ tali che per ogni $v = \sum_{i=1}^N \alpha_i\phi_i \in V_h$:*

$$ch^2 |\, \alpha \,|^2 \leq \|v\|^2_{L^2(\Omega)} \leq Ch^2 |\, \alpha \,|^2; \tag{66}$$

$$a(v, v) = \int_\Omega |\, \nabla v \,|^2 \, dx \leq Ch^{-2} \|v\|^2_{L^2(\Omega)}. \tag{67}$$

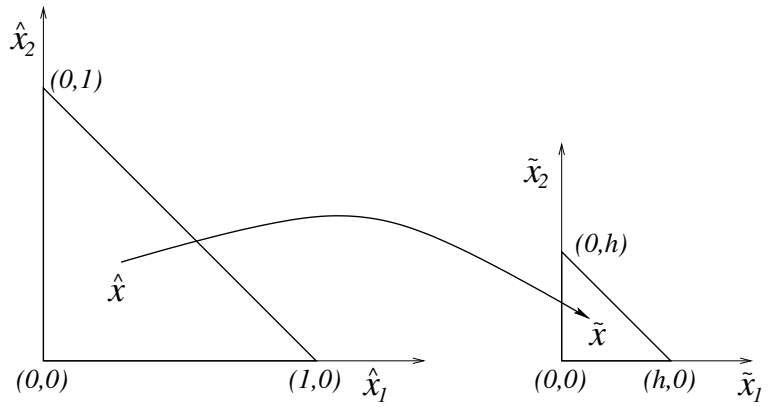

Figura 17: Trasformazione di coordinate per passare dal triangolo di riferimento $\hat{T}$ al triangolo scalato $\tilde{T}$.

Dimostrazione. Bisogna dimostrare che per ogni triangolo $T \in \mathcal{T}_h$ di vertici $\xi^{(i)}$, $i = 1, 2, 3$, e ogni $v \in \mathcal{P}_1(T)$, abbiamo:

$$ch_T^2 \sum_{i=1}^{3} \mid v(\xi^{(i)} \mid^2 \leq \|v\|^2_{L^2(T)} \leq Ch_T^2 \sum_{i=1}^{3} \mid v(\xi^{(i)}) \mid^2, \tag{68}$$

$$\int_T \mid \nabla v \mid^2 dx \leq Ch_T^{-2} \int_T \mid v \mid^2 dx. \tag{69}$$

La dimostrazione poi segue sommando su tutti i triangoli.

La strategia per la dimostrazione delle disuguaglianze precedenti è quella di mostrare che valgono per il triangolo di riferimento $\hat{T}$ avente coordinate nodali $\hat{\xi}^{(1)} = (0, 0)$, $\hat{\xi}^{(2)} = (1, 0)$, e $\hat{\xi}^{(3)} = (0, 1)$ e poi usare una trasformazione affine per passare da $\hat{T}$ nel piano di coordinate $(\hat{x_1}, \hat{x_2})$ ad un triangolo qualsiasi T nel piano (x_1, x_2).

Consideriamo quindi il triangolo di riferimento $\hat{T}$. Sia $\hat{\phi}_i(x)$ la classica funzione di base $\mathcal{P}_1(\hat{T})$ e sia:

$$\hat{v}(\hat{x}) = \sum_{i=1}^{3} \gamma_i \hat{\phi}_i(\hat{x}), \qquad \forall \hat{x} \in \hat{T}.$$

Sia $\gamma = (\gamma_i, \gamma_2, \gamma_3)$ e dimostriamo che la funzione $f : \mathbb{R}^3 \to \Omega$ definita da:

$$f(\gamma) = \frac{\int_{\hat{T}} \mid \nabla \hat{v} \mid^2 d\hat{x}}{\int_{\hat{T}} \hat{v}^2 \, d\hat{x}}, \gamma \neq 0,$$

soddisfa:

$$f(\gamma) \leq C \qquad \forall \gamma \in \mathbb{R}^3, \gamma \neq 0. \tag{70}$$

Da questa ne discende la (69) con $T = \hat{T}$ e $h_T = \sqrt{2}$. Si noti che $f(\gamma)$ è una funzione omogenea di grado zero ($f(\alpha\gamma) = f(\gamma)\ \forall\alpha \in \mathbb{R}$, $\alpha \neq 0$. Quindi, per dimostrare la (70) dimostriamo che la $f(\gamma)$ è continua e limitata in una palla $B = \{\gamma \in \mathbb{R}^3 : \|\gamma\| = 1\}$. Per prima cosa notiamo che $f(\gamma) \neq 0$ per $\gamma \in B$, e si vede facilmente che è continua. Ma siccome B è chiuso e limitato in $\mathbb{R}^3$, allora f raggiunge il massimo in B in $\hat{T}$.

Prima di passare ad un triangolo generico T, passiamo ad un triangolo $\tilde{T}$ simile a $\hat{T}$ ma scalato con h (cfr. figura 17), quindi isoscele con lati pari a h e ipotenusa $h_{\tilde{T}} = \sqrt{h}$. La trasformazione $F : \hat{T} \to \tilde{T}$ è definita da:

$$\tilde{x} = F(\hat{x}) = (h\hat{x}_1, h\hat{x}_2)\,, \hat{x} \in \hat{T}.$$

Per ogni funzione $v \in \mathcal{P}_1(\tilde{T})$ abbiamo:

$$\hat{v}(\hat{x}) = \tilde{v}\left(F(\hat{x})\right), \qquad \hat{x} \in \hat{T},$$

e possiamo calcolare le componenti dello Jacobiano della trasformazione:

$$\frac{\partial \hat{v}}{\partial \hat{x}_i} = \frac{\partial \tilde{v}}{\partial \tilde{x}_1}\frac{\partial \tilde{x}_1}{\partial \hat{x}_i} + \frac{\partial \tilde{v}}{\partial \tilde{x}_2}\frac{\partial \tilde{x}_2}{\partial \hat{x}_i} = \frac{\partial \tilde{v}}{\partial \hat{x}_i}h.$$

Nello stesso modo possiamo dire che $\nabla\hat{v} = \nabla\tilde{v}$ e ovviamente $d\tilde{x} = h^2 d\hat{x}$. Per cui si ottiene:

$$\int_{\tilde{T}|} h^{-2}|\ \nabla\tilde{v}\ |^2\, d\tilde{x} = \int_{\hat{T}} |\ \nabla\hat{v}\ |^2 \leq d\hat{x} \leq C\int_{\hat{T}} \hat{v}^2\, dx = Ch^{-2}\int_{\tilde{T}} v^2\, d\tilde{x}.$$

Ora, per passare dal triangolo $\hat{T}$ ad un triangolo qualsiasi T possiamo ragionare in maniera del tutto analoga. Costruiamo la trasformazione:

$$x = F(\hat{x}) = \xi^{(1)} + \left(\xi^{(2)} - \xi^{(1)}\right)\hat{x}_1 + \left(\xi^{(3)} - \xi^{(1)}\right)\hat{x}_2,$$

usare il fatto che $|\ \xi^{(i)} - \xi^{(1)}\ | \leq Ch_T$, $i = 1, 2, 3$ e $dx = Ch_t^2 d\hat{x}$ per le proprietà di regolarità della triangolazione $\mathcal{T}_h$. □

Dimostrazione del teorema 3.13. Una generica funzione $v \in V_h$ può essere scritta come combinazione lineare delle funzioni di base:

$$v(x) = \sum_{i=1}^{N} \beta_i \phi_i(x),$$

per cui:

$$a\,(v, v) = \beta \cdot A\beta,$$

con $\beta = \{\beta_i\}$. Quindi, per le stime inverse equazioni (66) e (67) del Lemma 3.56, abbiamo:

$$\frac{\beta \cdot A\beta}{\|\beta\|^2} = \frac{a(v,v)}{\|v\|^2} \leq Ch^{-2}\frac{\|v\|^2_{L^2(\Omega)}}{\|\beta\|^2} \leq C^2 \qquad \forall \beta \in \mathbb{R}^N.$$

D'altro canto la coercività della forma bilineare $a(\cdot,\cdot)$ e la (66) si ha ($\|v\|_{H^1(\Omega} \geq \|v\|_{L^2(\Omega)}$):

$$\frac{\beta \cdot A\beta}{\|\beta\|^2} = \frac{a(v,v)}{\|v\|^2} \geq \alpha\frac{\|v\|^2_{L^2(\Omega)}}{\|\beta\|^2} \geq C\alpha h^2 \qquad \forall \beta \in \mathbb{R}^N,$$

da cui, evidentemente, esistono costanti c e C indipendenti da h tali che:

$$\lambda_{\max} \leq C, \qquad \lambda_{\min} \geq ch^2,$$

e quindi $\kappa(A) = \lambda_{\max}/\lambda_{\min} \leq Ch^{-2}$. □

Osservazione 3.57. La matrice di rigidezza A è spesso scalata con una costante dell'ordine di $\mathcal{O}(h^2)$, e.g., ogni riga viene moltiplicata per l'area afferente al nodo relativo. In questo caso la matrice ha autovalori minimi e massimi tali che $\lambda_{\min} = \mathcal{O}(1)$ e $\lambda_{\max} = \mathcal{O}(h^{-2})$. Si può dimostrare questi autovalori tendono agli autovalori dell'operatore laplaciano continuo, che giaciono in un intervallo illimitato (Λ, ∞), $\Lambda > 0$.

Osservazione 3.58. Si ricorda che il metodo del gradiente coniugato precondizionato (PCG) converge con un numero di iterazioni che è proporzionale alla radice dell'indice di condizionamento spettrale. Quindi, quando si usa il PCG per risolvere i sistemi lineari che scaturiscono dalla discretizzazione di un operatore di diffusione coercivo su una sequenza di mesh raffinate in modo da dimezzare ogni volta il parametro di griglia h (ad esempio, si può ottenere una triangolazione raffinata uniformemente con h dimezzato congiungendo tutti i punti medi di ogni triangolo), il numero di iterazioni del PCG per arrivare alla soluzione don una fissata tolleranza raddoppia ad ogni raffinamento. Questo è un fenomeno tipico di tutti i problemi "ellittici".

4 Equazioni in forma mista

Cominciamo questo capitolo partendo da due esempio significativi. Il primo riguarda l'equazione di Stokes, il secondo l'equazione di Darcy, ambedue modelli molto importanti nelle applicazioni della fluidodinamica computazionale.

Esempio 4.1 (Equazione di Stokes)**.** Le equazioni di Stokes stazionarie per un fluido incomprimibile e Newtoniano sono un'approssimazione linearizzata delle equazioni di Navier-Stokes valide al limite per il numero di Reynolds che tende a zero. Esse si scrivono come:

$$\begin{aligned} -\mu\Delta u + \nabla p &= f && \text{in } \Omega, \\ \operatorname{div} u &= 0 && \text{in } \Omega, \\ u &= 0 && \text{in } \Gamma, \end{aligned} \tag{71}$$

dove μ è la viscosità dinamica del fluido, $u \in \mathbb{R}^3$ è il vettore veolcità del fluido, e gli operatori di div e Δ sono intesi per componenti.

Una formulazione variazionale può essere ricavata come segue. Prendiamo funzioni test vettoriali $v \in [H_0^1(\Omega)]^3$ che soddisfino all'ulteriore condizione di avere divergenza nulla, $\div v = 0$. Moltiplicando la i-esima equazione per la componente v_i, integrando e applicando il lemma di Green, otteniamo (usiamo notazioni tensoriali, per cui la presenza di due stessi indici indica la somma dei termini corrispondenti):

$$\int_\Omega f_i v_i \, dx = \mu \int_\Omega \nabla u_i \cdot \nabla v_i \, dx - \int_\Gamma u \cdot n v_i \, ds + \int_\Gamma p n_i v_i \, ds - \int_\Omega p v_{i,i} \, dx;$$

equazione che può essere scritta come:

$$\int_\Omega f_i v_i \, dx = \mu \int_\Omega \nabla u_i \cdot \nabla v_i \, dx - \int_\Gamma u \cdot n v_i \, ds + \int_\Gamma p n_i v_i \, ds - \int_\Omega p v_{i,i} \, dx = \mu \int_\Omega \nabla u_i \cdot \nabla v_i \, dx,$$

poichè $v_i = 0$ in Γ e $\operatorname{div} v = v_{i,i} = 0$ in Ω. Si noti che usando il "double dot product" o prodotto tensoriale, e definendo lo spazio $V(\Omega) = \left\{ v \in [H_0^1(\Omega)]^3 : \operatorname{div} v = 0 \right\}$, si può scrivere:

$$\mu \int_\Omega \nabla u : \nabla v \, dx = \int_\Omega f \cdot v \, dx \qquad \forall v \in V(\Omega).$$

Si può dimostrare che ogni funzione u che soddisfa la precedente soddisfa anche l'equazione di Stokes. La formulazione variazionale quindi diventa: trovare $u \in V$ tale che:

$$a(u, v) = L(v) \qquad \forall v \in V, \tag{72}$$

dove:

$$a(v, w) = \mu \int_\Omega \nabla u : \nabla v \, dx, \qquad L(v) = \int_\Omega f \cdot v \, dx.$$

Si noti però che non c'è un'equazione variazionale per la pressione, per cui qualcosa sembra mancare in questa formulazione variazionale, per via del fatto che lavoriamo su uno spazio in cui le funzioni test hanno divergenza nulla.

La formulazione variazionale semplicemente consiste nel rimpiazzare V con un sottoinsieme finito-dimensionale. Assumendo di essere in un dominio bidimensionale, per semplicità, lo spazio V possiamo riscriverlo come:

$$V(\Omega) = \left\{ v = (v_1, v_2) \in \left[H_0^1(\Omega) \right]^2 : \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} v = 0 \text{ in } \Omega \right\},$$

con $\Omega \subset \mathbb{R}^2$. Se Ω è un insiem semplicemente connesso, allora $\operatorname{div} v = 0$ se e solo se esiste una funzione "di corrente" $\varphi \in H_0^2(\Omega)$ tale che

$$v = \operatorname{rot} \varphi = \left(\frac{\partial \varphi}{\partial x_2}, -\frac{\partial \varphi}{\partial x_1} \right).$$

Le funzioni φ sono quindi di classe $C^1(\Omega)$. Sia quindi $W_h \subset H_0^2(\Omega)$ un sottospazio finito dimensionale. Le funzioni di base per W_h saranno polinomi di quinto grado $\varphi \in \mathcal{P}_5(T)$, determinati univocamente dalle seguenti condizioni (denotiamo con $\xi^{(i)}$ le coordinate dei 3 vertici del triangolo T e con $\xi^{(ij)}$ la coordinata del punto medo del lato compreso tra i vertici i e j):

$$D^\alpha \varphi(\xi^{(i)}), \qquad i = 1, 2, 3; \mid \alpha \mid \leq 2$$
$$\frac{\partial \varphi}{\partial n}(\xi^{(ij)}, i, j = 1, 2, 3; i < j.$$

Possiamo quindi costruire il nostro spazio FEM come:

$$V_h = \{v : v = \operatorname{rot} \varphi, \varphi \in W_h\},$$

e formulare il metodo FEM sostituendo V con V_h in (72), a cui corrisopnderà intuitivamente una stima dell'errore:

$$\|u - u_h\|^{H^1(\Omega)} \leq Ch^4 \mid u \mid_{H^5(\Omega)}.$$

Come si evince dall'esempio precedente, trovare funzioni di base per V_h che soddisfino alla condizione di incomprimibilità ($\operatorname{div} v = 0$) non è assolutamente facile, e in tre dimensioni le cose si complicano ulteriormente. E' allora conveniente lavorare in forma "mista", utilizzando come incognite esplicite del problema (71) sia la velocità u che la pressione p. Si noti che la pressione è però definita a meno di una costante. Per rendere la soluzione unica, bisogna aggiungere una condizione per esempio di media nulla della pressione:

$$\int_\Omega p \, dx = 0.$$

4.1 Formulazione mista per equazioni ellittiche

Riformuliamo in forma mista l'equazione ellittica scalare del secondo ordine (equazione di diffusione):

$$-\operatorname{div}(a(x)\nabla p) = f \qquad \text{in } \Omega \tag{73}$$
$$p = 0 \qquad \text{in } \Gamma = \partial\Omega, \tag{74}$$

dove il coefficiente di diffusione $a(x)$ è limitato superiormente e inferiormente da costanti positive, e $\Omega \subset \mathbb{R}^d$, $d = 2, 3$. Questa equazione esprime, ad esempio, la conservazione della quantità di moto di un fluido in moto laminare. Quindi p rappresenta la pressione del fluido e $u = -a(x)\nabla p$ la velocità del fluido. Lo scopo è di approssimare simultanemante sia p che u sperando di ottenere migliori proprietà della soluzione (p, u) rispetto alla classica soluzione che usa p come unica incognita e calcola la velocità per derivazione. Il problema si trasforma quindi

in un sistema di equazioni di primo grado (abbiamo posto $\mu = 1/a(x)$) dove le somiglianze con il problema di Stokes sono evidenti:

$$\begin{aligned} \mu u + \nabla p &= 0 && \text{in } \Omega, \\ \operatorname{div} u &= f && \text{in } \Omega, \\ p &= 0 && \text{in } \Gamma, \end{aligned}$$

Moltiplicando la prima equazione per funzioni test vettoriali e la seconda per funzioni test scalari otteniamo la formulazione variazionale:

$$\begin{aligned} \int_\Omega \mu u \cdot v \, dx - \int_\Omega p \operatorname{div} v \, dx &= 0 && \forall v \in H(\operatorname{div}, \Omega), \\ \int_\Omega q \operatorname{div} u \, dx &= \int_\Omega f q \, dx && \forall q \in L^2(\Omega), \end{aligned}$$

dove lo spazio $H(\operatorname{div}, \Omega)$ è uno spazio di Hilbert formato da:

$$H(\operatorname{div}, \Omega) = \left\{ v \in \left[L^2\right]^d : \operatorname{div} v \in L^2(\Omega) \right\},$$

con norma definita da:

$$\|v\|_{H(\operatorname{div},\Omega)} = \|v\|_{L^2(\Omega)} + \|\operatorname{div} v\|_{L^2(\Omega)}.$$

Si osservi che questa formulazione prevede la divergenza della soluzione e delle funzioni test, ma non include derivate prime arbitrarie. La conseguenza è che funzioni di base vettoriali che sono polinomiali a tratti devono avere solo la componente normale continua.

Il nostro problema può ora essere scritto usando una forma bilineare simmetrica definita da:

$$C\left((u,p),(v,q)\right) = \int_\Omega \mu u \cdot v \, dx - \int_\Omega p \operatorname{div} v \, dx - \int_\Omega q \operatorname{div} u \, dx.$$

e una forma lineare definita da:

$$L\left((v,q)\right) = -\int_\Omega f q \, dx,$$

e imponendo[6]

$$C\left((u,p),(v,q)\right) = L\left((v,q)\right) \qquad \forall (v,q) \in H(\operatorname{div}, \Omega) \times L^2(\Omega).$$

La forma $C(\cdot,\cdot)$ non è coerciva, ma si può dimostrare che soddisfa alla condizione inf-sup (49) e quindi per la simmetria anche alla (50).

[6]Le due equazioni separate sono ottenute usando $(v, 0)$ e $(0, q)$ nel secondo argomento della forma bilineare.

4.2 Elementi finiti misti

Sia $\mathcal{T}_h(\Omega)$ una trangolazione regolare di Ω con parametro di griglia h. Dobbiamo costruire gli spazi FEM:

$$V_h \subset H(\mathrm{div}, \Omega) \text{ e } W_h \subset L^2(\Omega).$$

Si può dimostrare che La condizione inf-sup per garantire la stabilità dello schema impone una relazione tra V_h e Q_h, nel senso che V_h deve essere "sufficientemente più ricco" di Q_h. La formulazione agli elementi finiti misti possiamo scriverla direttamente come: trovare $(u_h, p_h) \in V_h \times Q_h$ tali che:

$$\begin{aligned} a(u_h, v) - b(p_h, v) &= 0 \qquad \forall v \in V_h, \\ b(q, u_h) &= f(q) \qquad \forall q \in Q_h, \end{aligned} \tag{75}$$

dove:

$$a(v, w) = \int_\Omega \mu v \cdot w \, dx \qquad b(v, q) = \int_\Omega q \, \mathrm{div}\, v \qquad f(q) = \int_\Omega f q \, dx.$$

Per ricavare le stime di convergenza assumiamo alcune proprietà degli spazi V_h e Q_h e poi andremo a definirli nel paragrafo sccessivo. In particolare assumiamo:

$$\mathrm{div}\, V_h = Q_h, \tag{76}$$

e che esista un operatore (di proiezione) $\Pi_h : [H^1(\Omega)]^d \to V_h$ tale che:

$$\int_\Omega \mathrm{div}(u - \Pi_h u) q = 0 \qquad \forall u \in \left[H^1(\Omega)\right]^d, \forall q \in Q_h. \tag{77}$$

Sotto queste ipotesi, la soluzione è unica. Prendiamo infatti $f = 0$ e dimostriamo che questo implica $(u_h, p_h) = (0, 0)$. Per $f = 0$ il sistema si scrive:

$$\begin{aligned} \int_\Omega \mu u_h \cdot v \, dx - \int_\Omega p_h \, \mathrm{div}\, v \, dx &= 0 \qquad \forall v \in V_h \\ \int_\Omega q \, \mathrm{div}\, u_h \, dx &= 0 \qquad \forall q \in Q_h \end{aligned}$$

Siccome $\mathrm{div}\, V_h \subset Q_h$, prendiamo $q = \mathrm{div}\, u_h$ nella seconda equazione, e otteniamo $\mathrm{div}\, u_h = 0$. Prendendo $v = u_h$ nella prima, otteniamo subito $u_h = 0$. Ma siccome $\mathrm{div}\, V_h \supset Q_h$ e $p_h \in Q_h$, scegliamo $v \in V_h$ tale che $\mathrm{div}\, v = p_h$, da cui risulta $p_h = 0$.

La stima dell'errore è data dal seguente:

Teorema 4.1. *Siano V_h e Q_h spazi FEM, aventi la proprietà* (76) *e sia Pi_h il proiettore definito in* (77)*. Allora esiste una costante C indipendente da h tale che:*

$$\|u - u_h\|_{L^2(\Omega)} \leq C \left\{ \|u - \Pi_h u\|_{L^2(\Omega)} \right\}.$$

Dimostrazione. Per sottrazione si ottiene l'equazione per l'errore:

$$a\left((u-u_h),v\right)-b\left((p-p_h),v\right)=0 \qquad \forall v\in V_h, \tag{78}$$

$$b\left(q,(u-u_h)\right)=0 \qquad \forall q\in Q_h. \tag{79}$$

L'ultima delle precedenti, osservando che vale (84), si può scrivere come:

$$b\left(q,(\Pi_h u-u_h)\right)=0 \qquad \forall q\in Q_h.$$

Prendendo $q=\mathrm{div}(\Pi_h u-u_h)$ si arriva subito a:

$$\mathrm{div}(\Pi_H u-u_h)=0.$$

Prendendo adesso $v=\Pi_h u-u_h$ nella (78) otteniamo:

$$a\left((u-u_h),\Pi_h u-u_h)\right)=\int_\Omega \mu(u-u_h)\cdot(\Pi_h u-u_h)\,dx=0. \tag{80}$$

Siccome:

$$\|\Pi_h u-u_h\|^2_{L^2(\Omega)}\le \|a\|_\infty a\left((\Pi_h u-u_h),\Pi_h u-u_h\right)=\|a\|_\infty\int_\Omega \mu(\Pi_h u-u_h)^2\,dx.$$

Sottraendo la (80) e raccogliendo, si ha:

$$\begin{aligned}\|\Pi_h u-u_h\|^2_{L^2(\Omega)} &\le \|a\|_\infty\int_\Omega \mu(\Pi_h u-u_h)^2\,dx\\ &\le \|a\|_\infty\int_\Omega \mu\left[(\Pi_h u-u_h)(\Pi_h u-u_h)-(u-u_h)(\Pi_h u-u_h)\right]\,dx\\ &\le \|a\|_\infty\int_\Omega \mu(\Pi_h u-u_h)(\Pi_h u-u)\,dx\\ &\le \|a\|_\infty\|\mu\|_\infty\|\Pi_h u-u_h\|_{L^2(\Omega)}\|\Pi_h u-u\|_{L^2(\Omega)}.\end{aligned}$$

Il risultato segue dividendo per $\|\Pi_h u-u_h\|_{L^2(\Omega)}$. □

La stima dell'errore del metodo agli elementi finiti misti non può prescindere dalle stime di interpolazione per funzioni scalari e vettoriali e dalla classica stima di regolarità della soluzione in funzione dei dati. Assumiamo quindi valide le seguenti stime:

$$\|p\|_{H^2(\Omega)}\le C\|f\|_{L^2(\Omega)} \tag{81}$$

$$\|q-P_h q\|_{L^2(\Omega)}\le Ch\|q\|_{H^1(\Omega)} \qquad \forall q\in H^1(\Omega); \tag{82}$$

$$\|v-\Pi_h v\|_{L^2(\Omega)}\le Ch\|\|_{H^1(\Omega)} \qquad \forall v\in\left[H^1(\Omega)\right]^n; \tag{83}$$

$$\|\Pi_h v\|_{L^2(\Omega)}\le C\|v\|_{H^1(\Omega)}. \tag{84}$$

Quindi possiamo ora dimostrare il seguente:

Teorema 4.2. *Siano V_h e Q_h spazi FEM misti che soddisfano* (76) *e 77, e inoltre il proiettore* Π_h *soddisfi* (84)*, allora esiste una costante C indidpendente da h tale che:*

$$\|p - p_h\|_{L^2(\Omega)} \leq C \left[\|p - Php\|_{L^2(\Omega)} + \|u - \Pi_h u\|_{L^2(\Omega)} \right].$$

Dimostrazione. Si noti che (77) insieme a (84) implica che per ogni $q \in Q_h$ esiste $v \in V_h$ tale che $\operatorname{div} v = q$ e $\|v\|_{L^2(\Omega)} \leq C \|q\|_{L^2(\Omega)}$. Infatti, prendiamo il problema ausiliario:

$$\begin{aligned} \Delta\varphi &= q && \in \Omega, \\ \varphi &= 0 && \in \partial\Omega, \end{aligned}$$

e definiamo $w = \nabla\varphi$. Dalla (81), otteniamo $\|w\|_{H^1(\Omega)} \leq C \|f\|_{L^2(\Omega)}$. Allora si può dimostrare utilizzando la (76) e (84), che la funzione $v = \Pi_h w$ soddisfa le condizioni richieste.

Dalle equazioni dell'errore, notando che siccome per ogni $v \in V_h$ dalla (76) si deduce che $(P_h p, q) = (p, q)$ per ogni $q \in Q_h$ e $\operatorname{div} v \in Q_h$, si ricava:

$$\int_\Omega (p - p_h) \operatorname{div} v \, dx = \int_\Omega (P_h p - p_h) \operatorname{div} v \, dx = \int_\Omega (u - u_h) \cdot v \, dx;$$

Prendendo $v \in V_h$, tale che $\operatorname{div} v = (P_h p - p_h)$ e sia valida la seguente:

$$\|v\|_{L^2(\Omega)} \leq C \|P_h p - p_h\|_{L^2(\Omega)},$$

si ottiene:

$$\|P_h p - p_h\|^2_{L^2(\Omega)} \leq C \|u - u_h\|_{L^2(\Omega)} \|P_h p - p_h\|_{L^2(\Omega)}.$$

La dimostrazione si conclude usando il Teorema 4.1 e la disuguaglianza triangolare. □

Questo teorema, insieme alle (82) e (83) ci dice che lo schema gli elementi finiti misti converge linearmente in p_h e u_h se la soluzione è sufficientemente regolare. In realtà esistono teoremi di "superconvergenza" per la p_h che in punti opportuni può convergere con ordine superiore a uno, in tutti quei casi in cui la (82) può essere scritta con un esponente di h maggiore di uno.

Quello che ci manca adesso è di costruire spazi V_h e Q_h che soddisfino a tutte le proprietà richieste.

4.2.1 Spazi di Raviart-Thomas $\mathbf{RT}_k$

Si ricorda che lo spazio V_h è un sottospazio di $H(\operatorname{div}, \Omega)$, e che bisognerà individuare anche il proiettore Π_h opportuno. Consideriamo per semplicità il caso di dominio bi-dimensionale e triangulazione regolare $\mathcal{T}_h = \{T\}$ definita nel capitolo precedente. Consideriamo qui solamente gli spazi di Raviart-Thomas di grado k ($\mathcal{RT}_k$), rimandando per esempio a [2] per approfondimenti. Definiamo quindi la famiglia di spazi $\mathcal{RT}$ sul triangolo $T \in \mathcal{T}_h$:

$$\mathcal{RT}_k(T) = [\mathcal{P}_k]^2 \oplus x\mathcal{P}_k$$

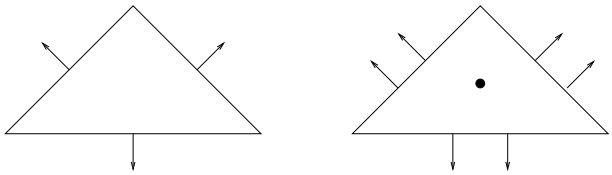

Figura 18: Posizione dei "gradi di libertà" per spazi $\mathcal{RT}_0$ (destra) e $\mathcal{RT}_1$ (sinistra). In corrispondenza delle frecce va specificata la componente normale $v \cdot n$, mentre nel punto si utilizza il valore della funzione v.

dove in $x \in \mathbb{R}^n$. Dunque lo spazio V_h è dato da:

$$V_h = \{v \in H(\mathrm{div}, \Omega) : v|_T \in \mathcal{RT}_k(T) \ \ \forall T \in \mathcal{T}_h\}.$$

Indichiamo nel prosieguo con ξ_i il nodo i-esimo di T e la sua coordinata, e con σ_i il lato i-esimo (faccia in 3D) e con ν_i il corrisopondente versore normale, $i = 1, 2, 3$, secondo numerazioni indipendenti locali. Si possono dimostrare le seguenti proprietà degli spazi $\mathcal{RT}_k$:

Lemma 4.2. *1.* $\dim \mathcal{RT}_k(T) = (k+1)(k+3)$*;*

2. se $v \in \mathcal{RT}_k(T)$*, allora* $v \cdot \nu_i \in \mathcal{P}_k(\sigma_i)$*;*

3. se $\mathrm{div}\, v = 0$*,* $v \in \mathcal{RT}_k(T)$*, allora* $v \in [\mathcal{P}_k(T))^n$*.*

Per quanto riguarda lo spazio $Q_h \subset L^2(\Omega)$, non ci sono particolari condizioni di regolarità da imporre, per cui abbiamo:

$$Q_h = \left\{q \in L^2(Omega) : q|_T \in \mathcal{P}_k(T) \ \ \forall T \in \mathcal{T}_h\right\}.$$

Ci manca quindi solo la definizione dell'operatore Π_h. Per costruire tale operatore, osserviamo che una funzione vettoriale le cui componenti siano polinomi continui a tratti appartiene a $H(\mathrm{div}, \Omega)$ solo se la sua componente normale alle facce di T è continua (basta pplicare il teorema della divergenza). Quindi possiamo scegliere come "gradi di libertà" per la definizione delle funzioni di base $k+1$ punti so ogni lato di T. Per esempio, per $k = 0$ abbiamo un punto (prenderemo il punto medio) per faccia dove imporre la continuità della componente normale, mentre per $k = 1$ avremo due punti per faccia per la componente normale più un punto centrale per la funzione v (due gradi di libertà) (si veda la Figura 4.2.1). Abbiamo quindi il seguente:

Lemma 4.3 (operatore Π_h)**.** *Dato il triangolo* $T \in \mathcal{T}_h$ *e una funzione vettoriale* $v \in [H^1(T)]^2$*, esiste un unico* $\Pi_T v \in \mathcal{RT}_k(T)$ *tale che:*

$$\int_{\sigma_i} \Pi_T v \cdot \nu_i p_k \, d\sigma = \int_{\sigma_i} v \cdot \nu_i p_k \, d\sigma \qquad \forall p_k \in \mathcal{P}_k(T), i = 1, 2, 3,$$

e

$$\int_T \Pi_T v \cdot p_{k-1} \, dx = \int_T v \cdot p_{k-1} \, dx \qquad \forall p_{k-1} \in [\mathcal{P}_{k-1}(T)]^2.$$

A questo punto la convergenza si può dimostrare con argomenti simili al caso degli elementi finiti normali (caso lagrangiano), a patto che tutti gli operatori che si usano conservino i polinomi quando si trasforma il triangolo generico nel triangolo di riferimento di coordinate $\xi_1 = (0,0)$, $\xi_2 = (1,0)$, $\xi_3 = (0,1)$. Per questo scopo si usa la trasformata di Piola, definita nel modo seguente. Data la mappa (affine) F che trasforma il triangolo $\tilde{T}$ nel triangolo T, definiamo $\tilde{v} \in \left[L^2(\tilde{T})\right]^2$ come:

$$v(x) = \frac{1}{|\det J(\tilde{x})|} J(\tilde{x})\tilde{v}(\tilde{x}),$$

dove $x = F(\tilde{x})$ e $J(\tilde{x})$ è la matrice Jacobiana di F. Abbiamo quindi:

Lemma 4.4. *Esiste una costante $C > 0$ tale che per ogni $v \in [H^m(T)]$ con $1 \leq m \leq k+1$, tale che:*

$$\|v - \Pi_T v\|_{L^2(T)} \leq C h_T^m \|v\|_{H^m(T)}.$$

Da questo lemma, usando le stime di interpolazione, si ottiene:

Teorema 4.3. *Sia $\{\mathcal{T}_h\}$ una famiglia di triangolazioni regolari e sia $u \in \left[H^{k+1}(\Omega)\right]^2$ e $p \in H^{k+1}(\Omega)$, allora la soluzione numerica $(u_h, p_h) \in V_h \times Q_h$ ottenuta con il metodo agli elementi finiti misti soddisfa:*

$$\|u - u_h\|_{L^2(\Omega)} \leq C h^{k+1} \|u\|_{H^{k_1}(\Omega)},$$

e

$$\|p - p_h\|_{L^2(\Omega)} \leq C h^{k+1} \left[\|u\|_{H^{k_1}(\Omega)} + \|p\|_{H^{k_1}(\Omega)}\right].$$

4.2.2 Uno sguardo alla condizione inf-sup

Riscriviamo il sistema (75), dove abbiamo però cambiato segno alla definizione della forma bilineare $b(p,q)$ e alla forma lineare $f(q)$:

$$\begin{aligned} a(u_h, v) + b(p_h, v) &= 0 \qquad \forall v \in V_h, \\ +b(q, u_h) &= f(q) \qquad \forall q \in Q_h, \end{aligned}$$

dove:

$$a(v,w) = \int_\Omega \mu v \cdot w \, dx \qquad b(v,q) = -\int_\Omega q \operatorname{div} v \qquad f(q) = -\int_\Omega f q \, dx.$$

Usando per esempio le funzioni di base $\mathcal{RT}_0$-P_0 definite in una triangolazione $\mathcal{T}_h$ formata da N_T triangoli e N_f lati (facce) possiamo esprimere u_h e p_h tramite le funzioni di base $v_k \in V_h$, $k = 1, \ldots, N_f$ e $q_t \in Q_h$, $t = 1, \ldots, N_T$:

$$u_h = \sum_{i=1}^{N_f} u_i v_i, \tag{85}$$

$$p_h = \sum_{m=1}^{N_T} p_m q_m. \tag{86}$$

Sostituendo, otteniamo il seguente sistema lineare:

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} \tag{87}$$

dove le matrici A, B hanno dimensioni $N_f \times N_f$ e $N_T \times N_f$, rispettivamente, e sono date da:

$$\begin{aligned} A = \{a_{ij}\} \;\; a_{ij} &= a(v_i, v_j), \qquad i, j = 1, \ldots, N_f \\ B = \{b_{lm}\} \;\; b_{lm} &= b(q_l, v_m), \qquad l = 1, \ldots, N_T \;, m = 1, \ldots, N_f, \end{aligned}$$

i vettori $u \in \mathbb{R}^{N_f}$ e $p \in \mathbb{R}^{N_T}$ contengono i coefficienti delle combinazioni lineari (85) e (86), e i vettori $f \in \mathbb{R}^{N_f}$ e $g \in \mathbb{R}^{N_T}$ sono termini noti. Si noti che nel nostro caso $f = 0$, ma se ci fossero condizioni al contorno di Neumann non omogenee sarebbe $f \neq 0$.

Guardiamo il sistema (87) dal punto di vista algebrico. Indicheremo con $\mathcal{A}$ la matrice del sistema:

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix},$$

mentre il sistema lineare completo sarà indicato con $\mathcal{A}x = b$, con ovvio significato dei termini.

Osserviamo che A è simmetrica e definita positiva, il sistema completo è simmetrico e corrisponde al classico problema di "punto sella": la soluzione $x = (u, p) \in \mathbb{R}^{N_f \cdot N_t}$ di tale sistema lineare risolve il seguente problema di minimo vincolato:

$$\min_{u \in \mathbb{R}^{N_f}} \frac{1}{2} u^T A u - f^T v \tag{88}$$

$$\text{soggetto a } Bp = g, \tag{89}$$

dove ora a variabile p ha il ruolo di moltiplicatore di Lagrange. Ogni soluzione (u^*, p^*) è un punto di sella per la Lagrangiana:

$$\mathcal{L}(u, p) = \frac{1}{2} u^T A u - f^T u + (bu - g)^T p,$$

nel senso che la coppia (u, p) deve soddisfare:

$$\mathcal{L}(u^*, p) \leq \mathcal{L}(u^*, p^*) \leq \mathcal{L}(u, p^*), \qquad \forall u \in \mathbb{R}^{N_f} \qquad \forall p \in \mathbb{R}^{N_T},$$

ovvero, in maniera equivalente:

$$\min_u \max_p \mathcal{L}(u, p) = \mathcal{L}(u^*, p^*) = \max_p \min_u \mathcal{L}(u, p).$$

La matrice $\mathcal{A}$ può essere fattorizzata a blocchi nei seguenti modi:

$$\begin{aligned}\mathcal{A} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} &= \begin{bmatrix} I & 0 \\ A^{-1}B & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} I & A^{-1}B^T \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} A & 0 \\ B & S \end{bmatrix} \begin{bmatrix} I & A^{-1}B^T \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ BA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B^T \\ 0 & S \end{bmatrix}\end{aligned}$$

dove il *complemento di Schur* S è dato da $S = -(BA^{-1}B^T)$. E' chiaro dalle precedenti espressioni che se A è non-singolare la condizione per l'esistenza dell'inversa di $\mathcal{A}$ è che B abbia rango massimo ($\text{rank}\,(B) = N_T$). Abbiamo infatti il seguente teorema che stabilisce le condizioni di non-singolarità della matrice $\mathcal{A}$ [1]:

Teorema 4.4. *Sia A simmetrica e semidefinita positiva, e sia B di rango massimo. Allora la matrice $\mathcal{A}$ è non-singolare se solo se* $\ker(A) \cap \ker(B) = \{0\}$.

Dimostrazione. Condizione sufficiente. Sia $x = (u, p)^T$ tale che $\mathcal{A}x = 0$. Quindi avremo $Au + B^T p = 0$ e $Bu = 0$. Quindi $u^T Au = -u^T B^T p = -(Bu)^T p = 0$. Per ipotesi, A è simmetrica e demipositiva definita, per cui $u^T Au = 0$ implica $Au = 0$, e quindi $u \in \ker(A) \cap \ker(B)$, da cui $u = 0$. Inoltre, $B^T p = 0$ e il fatto che B ha rango massimo implica $p = 0$.

Condizione necessaria. Assumiamo ora che $\ker(A) \cap \ker(B) \neq \{0\}$, e sia $u \in \ker(A) \cap \ker(B)$, con $u \neq 0$. Per $x = (u, 0)^T$ si ha $\mathcal{A}x = 0$, e quindi $\mathcal{A}$ è singolare, e la condizione è anche necessaria. □

Si può dimostrare che u^* è la proiezione A-ortogonale (cioè proiezione rispetto al prodotto scalare $(v, w)_A = v^T Aw$ sullo spazio dei vincoli $\mathcal{C} = \{p \in \mathbb{R}^{N_T} : Bp = g\}$.

Sia quindi B di rango massimo, e sia $\beta^2 = \sigma_m in(B)$ il minimo valore singolare di B che richiediamo essere strettamente maggiore di zero così come α, l'autovalore minimo di A. L'inversa della matrice $\mathcal{A}$ si può scrivere esplicitamente:

$$\mathcal{A}^{-1} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1}(I - B^T S^{-1} BA^{-1}) & A^{-1}B^T S^{-1} \\ S^{-1}BA^{-1} & S^{-1} \end{bmatrix},$$

e si hanno immediatamente le seguenti stime:

$$\|u\|_A \leq \|f\|_{A^{-1}} \leq \frac{1}{\alpha}\|f\| \qquad \|p\| \leq \frac{1}{\beta}\|f\|_{A^{-1}} \leq \frac{1}{\alpha\beta}\|f\|$$

da cui si vede che il problema è ben posto se esistono delle costanti β^* e α^* indipendenti da h tali che $\beta > \beta^* > 0$ e $\alpha > \alpha^* > 0$.

Lemma 4.5. *La condizione $\beta = \sigma_{min}(B) > 0$ è equivalente alla condizione di inf-sup per il problema di punto sella:*

$$\inf_{q\in\mathbb{R}^{N_T}} \sup_{v\in\mathbb{R}^{N_f}} \frac{q^T B v}{\|q\|\,\|v\|} > \beta^2 > 0 \qquad \forall q \neq 0,\ e\ \forall v \neq 0,$$

o, equivalentemente:

$$\max_{v\in\mathbb{R}^{N_f}} \frac{q^T B v}{\|v\|} > \beta^2 \|q\| \qquad \forall q \in \mathbb{R}^{N_T}, q \neq 0.$$

Dimostrazione. Sia $B = U\Sigma V^T$ la decomposizione in valori singolari della matrice B. Prendiamo:

$$q = u_i, i = 1, \ldots, N_T \text{ e } v = \sum_{j=1}^{N_f} \gamma_j v_j.$$

Dall'ortogonalità della matrice V si ha che $\|v\|^2 = \sum \gamma_i^2$. Sostituendo otteniamo:

$$\frac{q^T B v}{\|v\|} = \frac{e_i^T \Sigma V^T v}{\sqrt{\sum_j \gamma_j^2}} = \frac{\sigma_i \gamma_i}{\sqrt{\sum_j \gamma_j^2}} \geq \sigma_i \geq \beta^2.$$

D'altro canto, prendendo $q = \sum_{j=1}^{N_T} \xi_j u_j$ e indicando con $\gamma = \{\gamma_i\}$ il vettore dei coefficienti γ_i, otteniamo:

$$\max_{v\in\mathbb{R}^{N_f}} \frac{q^T B v}{\|v\|} = \max_{\gamma\neq 0} \sum_{i=1}^{N_f} \xi_i \frac{\sigma_i \gamma_i}{\sqrt{\sum_j \gamma_j^2}} \geq \beta^2 \sum_{i=1}^{N_T} \frac{\xi_i^2}{\sqrt{\sum_j \xi_j^2}} = \beta^2 \|q\|.$$

□

In maniera del tutto equivalente, che non si riporta qui, si può estendere queste considerazioni al problema continuo.

La convergenza di MFEM richiede che la matrice $\mathcal{A}$ sia invertibile per ogni $\mathcal{T}_h$ uniformemente per $h \to 0$. Sia quindi $\mathcal{A}_h$ la matrice relativa a $\mathcal{T}_h$. Al variare di h abbiamo una successione di problemi di punto sella del tipo:

$$\begin{bmatrix} A_h & B_h^T \\ B_h & 0 \end{bmatrix} \begin{bmatrix} u_h \\ p_h \end{bmatrix} = \begin{bmatrix} f_h \\ g_h \end{bmatrix}.$$

Ogni sistema è risolubile se $\sigma_{min}(B_h) = \beta_h^2 \geq 0$, ovvero:

$$\inf_{q \in W_h} \sup_{v \in V_h} \frac{q^T B_h v}{\|q\| \, \|v\|} > \beta_h^2 > 0 \qquad \forall q \neq 0, \text{ e } \forall v \neq 0,$$

Quindi gli spazi V_h e W_h devono soddisfare la condizione inf-sup. Se la soddisfano, allora si ha unicità, esistenza, cdipendenza continua dai dati, e si riesce a dimostrare la convergenza di MFEM. La costante dell'errore nella stima di convergenza dipende ovviamente da $1/\alpha$ e $1/\beta$. Se β decresce per $h \to 0$, la velocità di convergenza non sarà quella ottimale.

Se invece gli spazi V_h e W_h non soddisfano la condizione inf-sup, la convergenza non è assicurata. Si possono verificare diverse condizioni:

- lo spazio delle funzioni v per cui $b(p, v) = g$ è vuoto. Questo potrebbe succedere per esempio se $N_T > N_f$ (ci sono più vincoli che equazioni). Un tipico esempio sono la coppia di funzioni di base $\mathrm{P}_1/\mathrm{P}_0$ usate per discretizzare l'equazione di Stokes.

- la matrice del sistema di punto sella diventa singolare o "quasi" singolare. Un tipico sintomo di questo problema è la comparsa di oscillazioni spurie nella soluzione p.

- la condizione che B sia di rango massimo è soddisfatta ma il valore singolare minimo tende a zero con h $\beta_h = \mathcal{O}\left(h^k\right)$. In questo caso si potrebbe avere convergenza fino ad un certo valore di h, e poi la velocità di convergenza degenera fino a scomparire.

Osserviamo che il modo più facile per curare la mancanza della condizione inf-sup è di garantire che lo spazio V_h sia sufficientemente più ricco dello spazio W_h, in modo tale da non avere vincoli troppo stringenti o addirittura sovrabbondanti.

4.2.3 Sulla soluzione del sistema lineare

Un modo semplice per risolvere il sistema lineare 87 nasce dall'osservazione che la matrice A è invertibile (la forma $a(u, v)$ è coerciva. Si può quindi semplificare il sistema invertendo A e esplicitando formalmente u:

$$u = A^{-1}(f - B^T p),$$

per cui ci si riduce al sistema simmetrico e definito positivo:

$$BA^{-1}B^T p = BA^{-1}f - g.$$

Questo modo di procedere è computazionalmente impraticabile per via della necessità di invertire la matrice A. Infatti, A^{-1} è piena e quindi il costo per calcolare l'inversa è improponibile.

Una tecnica più semplice è quella di procedere alla cosidetta "ibridizzazione", che da' luogo al metodo degli elementi finiti misti ibridi. Si tratta di rendere la matrice A facilmente invertibile, e poi procedere come sopra. Per fare questo, si rilassa l'ipotesi che $V_h \subset H(\mathrm{div}, \Omega)$, e si impone solamente che $V_h \subset H(\mathrm{div}, T)$, per ogni $T \in \mathcal{T}_h$. Le funzioni di V_h non hanno più la componente normale continua, per cui possono essere definiti indipendentemente elemento per elemento. La continuità dei flussi normali sui bordi degli elementi viene poi imposta esplicitamente tramite la tecnica dei moltiplicatori di Lagrange.

Per semplificare, utilizziamo i classici elementi $\mathcal{RT}_0 - P_0$. Introduciamo il seguente spazio (lo spazio $\mathcal{RT}_0 - P_0$ discontinuo):

$$\widetilde{V}_h = \left\{ v \in \left[L^2(\Omega) \right]^2 : v|_T \in \mathcal{RT}_0 \ \forall T \in \mathcal{T}_h \right\}.$$

Si noti che $V_h \subset \widetilde{V}_h$ e che $v \in V_h$ se e solo se $v \in H(\mathrm{div}, \Omega)$. Mettiamo insieme tutti i lati σ dei triangoli di $\mathcal{T}_h$ in Γ_h:

$$\Gamma_h = \bigcup_{T \in \mathcal{T}_h} \partial T = \bigcup_{i=1}^{N_f} \sigma_i,$$

e introduciamo lo spazio dei moltiplicatori λ sull'insieme Γ_h:

$$\Lambda_h = \left\{ \mu \in L^2(\Gamma_h) : \mu|_\sigma \in \mathcal{P}_0(\sigma) \ \forall \sigma \in \Gamma_h \right\},$$

e la funzione bilineare $d(v, \mu)$ definita da:

$$d(v, \mu) = -\sum_{T \in \mathcal{T}_h} \int_{\partial T} \mu \, v|_T \cdot \nu \, dx = \sum_{\sigma \in \Gamma_h} \int_\sigma \mu \left[v \cdot \nu \right] \, ds,$$

per ogni $v \in \widetilde{V}_h$ e ogni $\mu \in \Lambda_h$, dove ν_T è la normale esterna a ∂T e $[v \cdot \nu]$ è il "salto" della componente normale di v attraverso il lato σ. Si noti che $d(v, \mu) = 0$ in ogni $\sigma \in \Gamma_h$ se e solo se $v \in H(\mathrm{div}, \Omega)$ (ovvero $v \in V_h$). Siccome $b(q, v)$ non è definita in $\widetilde{V}_h$, definiamo la forma bilineare "di griglia" $b_h(q, v)$ come:

$$b_h(q, v) = \sum_{T \in \mathcal{T}_h} \int_T q \, \mathrm{div} \, v \, dx.$$

per ogni $q \in W_h$ e ogni $v \in \widetilde{V}_h$. Consideriamo il seguente problema FEM: Trovare $(\tilde{u}_h, \tilde{p}_h, \lambda_h) \in \widetilde{V}_h \times W_h \times \Lambda_h$ tali che:

$$\begin{aligned} a(\tilde{u}_h, v) + b_h(\tilde{p}_h, v) + d(v, \lambda_h) &= f \qquad \forall v \in \widetilde{V}_h \\ b_h(q, \tilde{u}_h) &= g \qquad \forall q \in W_h \\ d(\tilde{u}_h, \mu) &= 0 \qquad \forall \mu \in \Lambda_h. \end{aligned}$$

Il corrispondente sistema algebrico diventa quindi:

$$\begin{cases} \tilde{A}u + B_h^T p + C\lambda &= f \\ B_h u &= g \\ Cu &= 0 \end{cases},$$

con ovvia espressione per le matrici. Ora, la matrice $\tilde{A}$ è diagonale a blocchi con blocchi 3×3, corrispondenti alle tre facce dei triangoli, e quindi è facilmente invertibile senza sforzo computazionale importante. Procedendo all'eliminazione a blocchi, si ottiene quindi nei diversi passi::

$$u = \tilde{A}^{-1}\left(f - B_h^T p - C\lambda\right),$$

e sostituendo:

$$\begin{cases} B_h\tilde{A}^{-1}B_h^T p + B_h C\lambda = B_h\tilde{A}^{-1}f - g \\ C\tilde{A}^{-1}B_h^T p + C\tilde{A}^{-1}C\lambda = -C\tilde{A}^{-1}f \end{cases}.$$

Notando che la matrice $H = B_h\tilde{A}^{-1}B_h^T$ è diagonale e facilmente invertibile e denotando $S = \tilde{A}^{-1}B_h^T$, otteniamo:

$$p = H^{-1}\left[f - S^T g\right],$$

da cui, indicando con M la matrice diagonale a blocchi $M = \tilde{A}^{-1} - SH^{-1}S^T$, si ottiene il sistema finale di dimensioni $N_f \times N_f$ per i moltiplicaroti di Lagrange λ su ciascuna faccia:

$$C^T M C\lambda = C^T\left[Mg - SH^{-1}f\right].$$

Si può dimostrare che questo sistema è simmetrico e definito positivo, e quindi può essere risolto con il metodo del gradiente coniugato precondizionato. Inoltre, il numero di elementi non nulli per riga è al massimo pari a 5, e quindi la matrice è considerevolmente più sparsa rispetto ad esempio al metodo di Galerkin $\mathcal{P}_1$. Di contro in media una triangolazione bidimensionale ha un numero di lati che è circa 3 volte il numero di nodi, mentre tale fattore diventa 7 in triangolazioni tridimensionali (tetraedriche).

4.2.4 Confronto sperimentale tra Galerkin P_1 e FEM misti $\mathcal{RT}_0 - P_0$

Per apprezzare meglio le motivazioni più pratiche che fanno dell'approccio misto un metodo assai utile, ancorchè numericamente meno efficiente, riportiamo qui alcune considerazioni che scaturiscono da esempi test semplificati. Consideriamo quindi il dominio e la mesh di Figura 19 dove andiamo a risolvere l'equazione (73) con le condizioni di Dirichlet e Neumann illustrate in figura. Queste condizioni sono tali per cui si crea un flusso di materia, dato da $u(x) =$

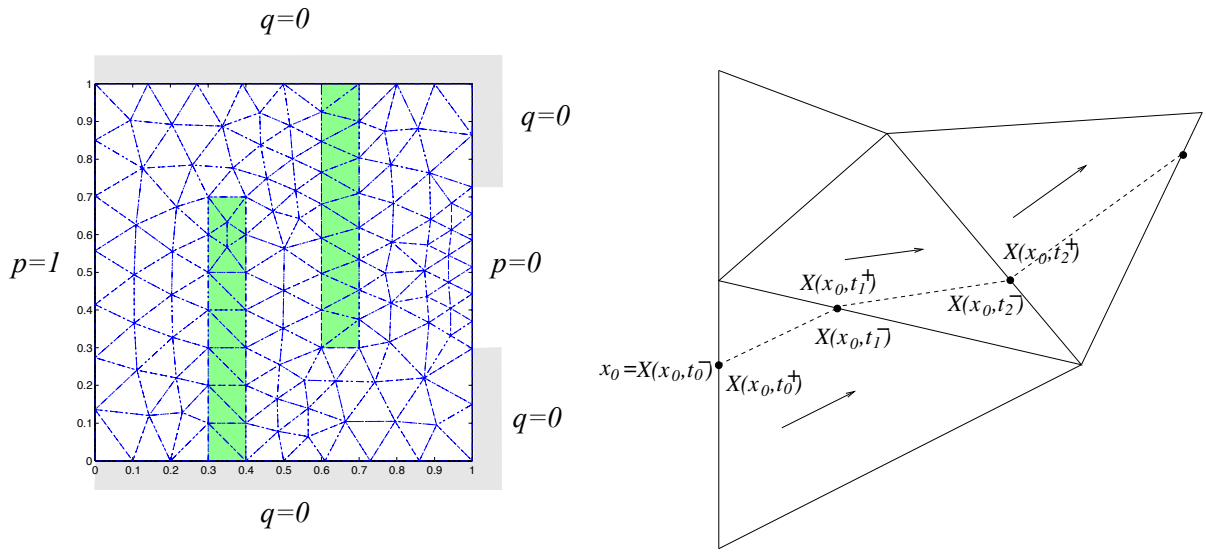

Figura 19: A destra: dominio e condizioni al contorno per la soluzione dell'equazione (73). Il flusso normale al bordo è identificato con $q = u \cdot n$. A sinistra: esemplificazione grafica e notazioni usate nella procedura di "particle tracking" (sinistra).

$-a(x)\nabla p(x)$, dal contorno "di entrata", coincidente con tutto il lato di sinistra del quadrato, verso l'uscita localizzata sulla porzione centrale del lato destro. Si noti nel dominio la presenza dei due "pilastri" segnati in verde in figura 19 ove si è specificato un coefficiente di diffusione $a(x) = 10^{-12}$, mentre nel resto del dominio $a(x) = 1$. Nei due pilastri il flusso di massa è quindi praticamente impedito, e le traiettorie del moto (a potenziale) devono andare dal bordo di sinistra verso il bordo di destra "circumnavigano" i pilastri a basso coefficiente di diffusione.

Per evidenziare le differenze tra i due campi di moto (Galerkin P_1 e Misti $\mathcal{RT}_0 - P_0$) si mostrano le traiettorie di 100 particelle inizialmente uniformemente distribuite sul contorno "di entrata". Le traiettorie sono calcolate valutando numericamente il seguente integrale:

$$X(x_0, t) = \int_{t_0}^{t} u(X(x_0, \tau))\, d\tau$$

dove $X(x_0, t)$ è la posizione al tempo t della particella rilasciata al tempo $t = 0$ nel punto x_0. Tale integrale è valutato sfruttando il fatto che il vettore velocità u è costante su ogni triangolo secondo il seguente algoritmo, semplificato graficamente in Figura 19. Si discretizza il tempo t in maniera tale che $t_0 = 0$, $t_k = t_{k-1} + h_{k-1}$. Si parte a $k = 0$ dal punto $X(x_0, 0) = x_0$. Tale punto giace per ipotesi nel lato $\sigma_{ij} \in T_r$, lato del bordo di Dirichlet $p = 1$ appartenente al triangolo $T_r \in \mathcal{T}_h$. Il punto successivo $X(x_0, t_1^-)$ sarà il punto che giace nel contorno di T_r costruito a partire da $X(x_0, 0)$ muovendosi lungo la direzione u_r, dove u_r è la velocità (costante) in T_r. Si continua così individuando il nuovo triangolo T_s avente in comune con T_r il segmento contentne il punto $X(x_0, t_1^-)$. Chiamiamo $X(x_0, t_1^+)$ lo stesso punto $X(x_0, t_1^-)$,

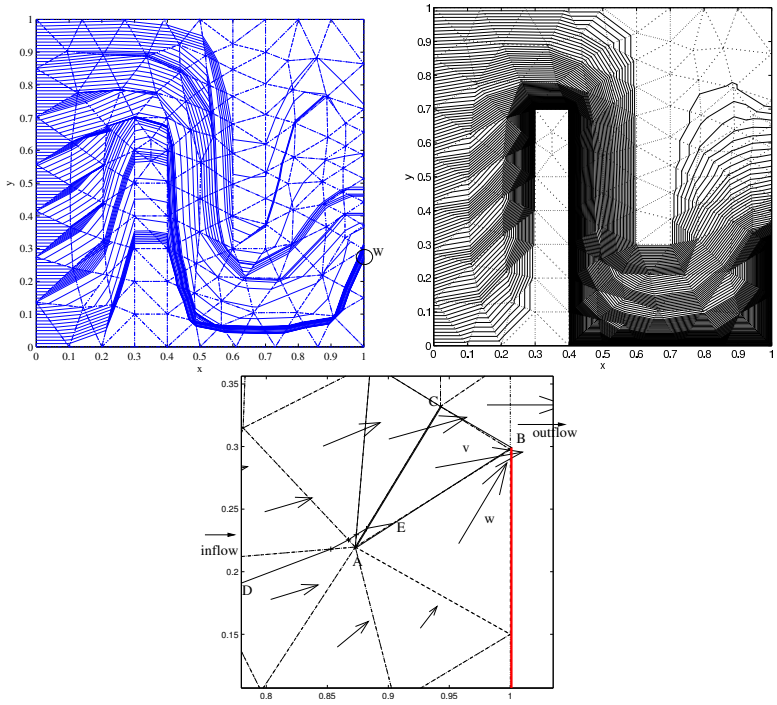

Figura 20: Traiettorie calcolate dal campo di velocità numerico ottenuto con Galerkin P_1 (in alto a sinistra) e con il metodo agli elementi finiti misti $\mathcal{RT}_0 - P_0$ (in alto a destra). In basso si mostra un dettaglio del campo di velocità (costante a tratti) calcolato con Galerkin P_1, Si noti che ci sono diversi triangoli adiacenti in cui le velocità hanno componente normale sul lato in comune con segni discordi. Questo corrisponde ad un errore nel bilancio di massa che evidenzia l'esistenza di una sorgente (o una emissione) di massa non fisiche (spurie) localizzate sul lato in questione.

pensato però come appartenente T_s, e procediamo come nel passo precedente. Indicando con $u(x_{b_k})$ il vettore velocità calcolato nel triangolo contenente $X(x_0, t_k^+)$, lo schema si può scrivere come:

$$X(X(x_0, t_k), t_{k+1}^-) = X(x_0, t_k^+) + \lambda u(x_{b_k}),$$

dove λ è il coefficiente che individua la lunghezza del segmento $X(x_0, t_{k+1}^+) - X(x_0, t_k^+)$. La procedura è graficamente illustrata in Figura 19.

La Figura 20 (in alto) mostra le traiettorie calcolate con l'algoritmo precedente a partire dal campo di moto P_1 (figura di sinistra) e dal campo di moto $\mathcal{RT}_0 - P_0$ (figura di destra) partento da 100 punti x_0 distribuiti uniformemente nel contorno "di entrata". La differenza tra le due figure salta agli occhi. Per prima cosa si vede che, al contrario del metoto $\mathcal{RT}_0 - P_0$, le traiettorie P_1 non sono uniformemente spaziate all'interno del dominio, convergendo in diversi cluster. La seconda osservazione è che alcune delle traiettorie P_1 escono dal dominio attraverso

facce di bordo dove si è imposta una condizione di flusso nullo. E' evidente quindi che il campo di moto P_1 non è conservativo. Tutto ciò invece non succede con il campo di moto $\mathcal{RT}_0 - P_0$, che risulta sempre conservativo. Il riquadro inferioro in Figura 20 mostra un dettaglio riportante alcuni triangoli della mesh con i relativi vettori velocità P_1 e una traiettoria calcolata. A partire dal punto D entrante dal bordo di sinistra, la traiettoria prosegue parallelamente al vettore velocità fino all'intersezione con il bordo del triangolo, e vi così fino al punto E. Qui, la traiettoria viene direzionata dal vettore w verso l'interno del triangolo di uscita, con evidente contraddizione. Infatti le componenti dei vettori v e w normali al lato AB sono di segno opposto, evidenziando la loro discontinuità. Fisicamente, componenti normali che puntano una contro l'altra corrispondono ad un punto assorbente, non presente però nell'equazione, evidenziano così la proprietà di non conservatività del campo di moto P_1. Tutto questo non succede nel caso degli MFEM $\mathcal{RT}_0 - P_0$, dando ragione alle differenze così marcate nelle traiettorie riportate nel pannello in alto di Figura 20.

Due osservazioni sono necessarie per chiarire in maniera più approfondita le conseguenze di tali errori. La prima riguarda la convergenza degli schemi di Galerkin. Come abbiamo visto in precedenza tali schemi sono convergenti al tendere a zero del parametro di mesh h. A convergenza gli errori di bilancio sono nulli, o a tutti gli effetti trascurabili. Nella pratica, però, si lavora a mesh fissata, con parametro h dettato dalla memoria finita dei computer in uso, talchè l'errore di bilancio è sempre presente, ancorchè tendente a zero per $h \to 0$.

La seconda osservazione riguarda l'errore di bilancio relativo alla soluzione scalare p. Tale errore, infatti, è sempre trascurabile se calcolato sul corretto volume di controllo e non sull'elemento, e non va confuso con l'errore di bilancio che scaturisce dai vettori velocità, quantità calcolate da un processo di derivazione. Questo è il principale motivo per cui si è cominciato a lavorare, nella metà degli anni '70, a schemi che utilizzassero come incognite sia la pressione che la velocità, dando vita quindi al grande filone dei metodi "misti".

Come abbiamo notato anche in precedenza, la complessità computazionale dei metodi misti è di gran lunga più elevata rispetto ai metodi tradizionali. Per questo motivo, di recente la ricerca si è concentrata nello sviluppo di metodi cosidetti di "post-processing" che prendono le velocità e.g. P_1 e ricostruiscono un campo di moto conservativo. Anche se ancora sperimentali, tali schemi sembrano essere molto promettenti per recuperare il gap compuazionale mantenendo al contempo la conservatività del campo di moto.

5 Equazioni paraboliche

Consideriamo l'equazione parabolica di riferimento:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \operatorname{div}\left[a(x)\nabla u\right] + f(x,t) && t \in [0,T];\ \ x \in \Omega \subset \mathbb{R}^d, \\ u(x,0) &= u_0(x) && t = 0,\ x \in \Omega, \\ u(x,t) &= u_D(x,t) && t \in [0,T],\ x \in \Gamma_D \subset \partial\Omega, \\ -a(x)\nabla u(x,t)\cdot\nu &= q_N(x,t) && t \in [0,T],\ x \in \Gamma_N \subset \partial\Omega, \end{aligned} \tag{90}$$

che fisicamente rappresenta ad esempio la trasmissione del calore in un mezzo conduttivo con conuttività termica data dalla funzione $a(x)$ che ovviamente assumiamo maggiore di zero sempre per garantire l'ellitticità necessaria. La soluzione di tale problema ora dipende non solo dalle variabili spaziali ma anche dal tempo $u(x,t)$.

5.1 Problema modello mono-dimensionale

Per meglio comprendere come si comporta la soluzione di (90), studiamo un problema semplificato in maniera esplicita. Si consideri quindi il problema mono-dimensionale:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} && x \in (0,\pi),\ \ t > 0; \\ u(x,0) &= u_0(x) && x \in (0,\pi), \\ u(0,t) &= u(\pi,t) = 0 && t > 0. \end{aligned}$$

Si può usare la separazione delle variabili e la tecnica di Fourier per trovare la soluzione analitica di tale equazione. E' infatti immediatamente verificabile che tale soluzione è data da:

$$u(x,t) = \sum_{j=1}^{\infty} u_{0,j}\, \mathrm{e}^{-j^2 t} \sin(jx), \qquad u_{0,j} = \sqrt{2/\pi} \int_0^{\pi} u_0(x) \sin(jx)\, dx, \qquad j = 1,2,\ldots.$$

dove $u_{0,j}$ sono i coefficienti di Fourier del dato iniziale $u_0(x)$ nella base (ortonormale in $L^2((0,\pi))$ $\{\sqrt{2/\pi}\ \sin(jx)\}_{j=1}^{\infty}$. La componente di frequenza j (relativa a $sin(jx)$) ha una scala temporale propria dell'ordine $\mathcal{O}\left(j^{-2}\right)$. Al variare di j si hanno quindi componenti che decadono velocemente nel tempo e la soluzione diventa sempre più regolare all'aumentare di t. Intuitivamente questo è proprio quello che ci aspettiamo da un'equazione di diffusione. Si noti però che a tempi piccoli la soluzione non è necessariamente regolare e potremmo avere che $\|\dot{u}(t)\| = \|\dot{u}(\cdot,t)\|_{L^2((0,\pi))} \to \infty$ per $t \to 0$ in funzione delle condizioni iniziali. Per esempio, se prendiamo come condizioni inziali $u_0(x) = \pi - x$, allora $u_{0,j} = C/j$, e per $t \to 0$ si ha che $\|\dot{u}(t)\| \approx Ct^{-\alpha}$ con $\alpha = 3/4$. e invece prendiamo $u0(x) = \min(x, \pi - x)$, troviamo $u_{0,j} = C/j^2$ e $\|\dot{u}(t)\| \approx Ct^{-\alpha}$ con $\alpha = 1/4$. In generale, se $u_{0,j}$ decade più velocemente di $j^{-2.5}$ per $j \to \infty$, allora $\|\dot{u}(t)\|$ è limitata per $t \to 0$.

La soluzione avrà in generale un transitorio iniziale dove alcune derivate potranno avere bassa regolarità, ma a tempi sufficientemente grandi la soluzione sarà liscia. Si noti che la presenza di forzanti $f(t)$ oscillatorie potrebbe generare transitori importanti non solamente per tempi piccoli.

Le stime a priori pricipali che si possono dimostrare facilmente usando i metodi del paragrafo successivo o la formula di Parseval sono:

$$\|u(t)\| \leq \|u_0\|, \qquad t \in (0,T) \tag{91}$$

$$\|\dot{u}(t)\| \leq \frac{C}{t}\|u_0\|, \qquad t \in (0,T). \tag{92}$$

5.2 Formulazione variazionale

Possiamo scrivere la formulazione variazionale di tale equazione immediatamente usando le tecniche sviluppate nel paragrafo precedente. Assumiamo come nel paragrafo precedente di avere condizioni al contorno di Dirichlet omogenee (nulle). Per fare questo, usiamo la separazione delle variabili e moltiplichiamo per una funzione test $v \in H_0^1(\Omega)$, lavorando a t fissato, $t \in I = (0,T]$. Dato $t > 0$, cerchiamo una funzione $u(t) : \Omega \to \mathbb{R}$ tale che:

$$\left(\frac{du}{dt}, v\right) + a(u,v) = (f,v) \qquad \forall v \in V;\ t \in I.$$

con $u(0) = u_0$. Studiamo ora alcune stime a priori per la formulazione variazionale precedente. Prendiamo quindi $v = u(t)$ ottenendo:

$$\left(\frac{du(t)}{dt}, u(t)\right) + a(u(t), u(t)) = (f(t), u(t)).$$

Il primo termine diventa:

$$\left(\frac{du(t)}{dt}, u(t)\right) = \int_\Omega \frac{du(t)}{dt} u(t)\, dx = \frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)}.$$

Gli altri termini possono essere stimati usando la coercività di $a(\cdot,\cdot)$ (eq. (46)), la disuguaglianza di Poincaré (Lemma 3.37) e la disugaglianza di Schwartz per il secondo membro. Si ottiene quindi:

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} + \alpha\|\nabla u(t)\|^2_{L^2(\Omega)} \leq \|f(t)\|_{L^2(\Omega)}\|u(t)\|_{L^2(\Omega)}.$$

Usando la disuguaglianza di Young, valida per ogni scalare α e β reali e ogni $\epsilon > 0$:

$$\alpha\beta \leq \epsilon\alpha^2 + \frac{1}{4\epsilon}\beta^2,$$

e di nuovo la disuguaglianza di Poincaré, otteniamo:

$$\frac{1}{2}\frac{d}{dt}\left\|u(t)\right\|^2_{L^2(\Omega)} + \alpha\left\|\nabla u(t)\right\|^2_{L^2(\Omega)} \leq \frac{C^2_\Omega}{2\alpha}\left\|f(t)\right\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\left\|\nabla u(t)\right\|^2_{L^2(\Omega)}.$$

Integrando nel tempo tra 0 e t, otteniamo:

$$\left\|u(t)\right\|^2_{L^2(\Omega)} + \alpha\int_0^t \left\|\nabla u(\tau)\right\|^2_{L^2(\Omega)}\ d\tau \leq \left\|u(0)\right\|^2_{L^2(\Omega)} + \frac{C^2_\Omega}{\alpha}\int_0^t \left\|f(\tau)\right\|^2_{L^2(\Omega)}\ d\tau.$$

Il termine di sinistra rappresenta l'energia totale del sistema al tempo t, che deve intuitivamente essere minore o uguale all'energia iniziale sommata a quella della forzante. Si noti che per $\alpha = 1$, il termine di sinistra è proprio la norma $H^1(\Omega)$ (al quadrato) di $u(t)$, da cui discende la (91) nel caso $f = 0$.

Un'altra stima a priori si può ottenere osservando che:

$$\frac{1}{2}\frac{d}{dt}\left\|u(t)\right\|^2 = \left\|u(t)\right\|\frac{d}{dt}\left\|u(t)\right\|,$$

da cui, con pochi passaggi (asusmento $\left\|u(t)\right\|_{L^2(\Omega)} \neq 0$:

$$\frac{d}{dt}\left\|u(t)\right\|_{L^2(\Omega)} + \frac{\alpha}{C}\left\|\nabla u(t)\right\|_{L^2(\Omega)} \leq \left\|f(t)\right\|_{L^2(\Omega)},$$

da cui dopo integrazione tra 0 e t:

$$\left\|u(t)\right\|_{L^2(\Omega)} \leq \left\|u(o)\right\|_{L^2(\Omega)} + \int_0^t \left\|f(\tau)\right\|_{L^2(\Omega)}\ d\tau,$$

da cui di nuovo discende la (91) nel caso $f = 0$.

5.3 Formulazione FEM

Consideriamo la formulazione di Galerkin con funzioni di base polinomiali lagrangiane. Quindi per ogni $t > 0$, cerchiamo una funzione $u_h(t) \in V_h$ tale che:

$$\begin{aligned} &\left(\frac{\partial u_h(t)}{\partial t}, v\right) + a\left(u_h, v\right) = \left(f, v\right) && \forall v \in V_h, \\ &\left(u_h(0), v\right) = \left(u_0, v\right). \end{aligned} \tag{93}$$

L'equazione precedente è il risultato dell'applicazione del Metodo delle Linee (MOL, Method of Lines) alla equazione originale [4].

Separando le variabili x e t, scriviamo u_h in funzione della base $\{\phi_i\}$ di V_h:

$$u_h(t,x) = \sum_{i=1}^{N} u_j(t)\phi_j(x),$$

e sostituendo otteniamo:

$$\sum_{j=1}^{N} \frac{du_j(t)}{dt} (\phi_j, \phi_i) + \sum_{j=1}^{N} u_j a (\phi_j, \phi_i) = (f(t), \phi_i) \qquad i = 1, \ldots, N.$$

Questo è un sistema $N \times N$ di ODE che possiamo scrivere in forma matriciale come:

$$P\dot{u} + Au = b, \tag{94}$$

dove il vettore $u = \{u_i(t)\}$ raccoglie i coefficienti di $u_h(x,t)$, la matrice di massa P ha elementi dati da $p_{i,j} = (\phi_i, \phi_j)$, la matrice di rigidezza è la solita $a_{ij} = a(\phi_i, \phi_j)$, il termine noto è dato da $b_i = (f(t), \phi_i)$ e la soluzione iniziale è data da $u_i = (u_0, \phi_i)$. Ambedue le matrici P e A sono simmetriche e definite positive. Possiamo quindi riscrivere la (94) invertendo formalmente la P. Scriviamo quindi $P = E^T E$ e il nuovo sistema diventa:

$$\dot{\eta}(t) + \tilde{A}\eta(t) = g(t), \qquad \eta(0) = \eta_0. \tag{95}$$

La matrice $\tilde{A} = E^{-T} A E^{-1}$ è simmetrica e definita positiva con numero di condizionamento spettrale $\kappa(\tilde{A}) = \mathcal{O}(h^{-1})$. La soluzione di questo sistema è data da:

$$\eta(t) = \mathrm{e}^{-\tilde{A}t}\, \eta_0 + \int_0^t \mathrm{e}^{-\tilde{A}(t-\tau)}\, g(\tau)\, d\tau. \tag{96}$$

Il sistema di ODE (94) è di tipo "stiff" come dimostrato dal grande intervallo di variazione degli ordini di grandezza degli autovalori di $\tilde{A}$, ciè dal fatto che $\kappa(\tilde{A})$ è grande.

Ritorniamo ora al problema semidiscreto e riportiamo stime a priori e stime dell'errore. Le stime a priori per il sistema (93) sono equivalenti, e si dimostrano in maniera equivalente, a quelle viste precedentemente, e si ottiene:

$$\|u_h(t)\|^2_{L^2(\Omega)} + \alpha \int_0^2 \|\nabla u_h(\tau)\|^2_{L^2(\Omega)}\, d\tau \leq \|u_{0,h}\|^2_{L^2(\Omega)} + \frac{C^2}{\alpha} \int_0^t \|f(\tau)\|^2_{L^2(\Omega)}\, d\tau, \tag{97}$$

ovvero:

$$\|u_h(t)\|_{L^2(\Omega)} \leq \|u_{0,h}\|_{L^2(\Omega)} + \int_0^t \|f(\tau)\|_{L^2(\Omega)}\, d\tau, \tag{98}$$

Il problema semi-continuo può essere studiato nel dettaglio. Assumiamo per questo un dominio Ω poligonale e assumiamo di usare elementi $\mathcal{P}_1$. Assumiamo inoltre che $a(x) = 1$ e che $\mathcal{T}_h$ sia una triangolazione regolare di Ω con passo h. Allora abbiamo il seguente:

Teorema 5.1. *Sotto le ipotesi menzionate sopra, esiste una costante C tale per cui:*

$$\max_{t\in(0,T)} \|u(t) - u_h(t)\|_{L^2(\Omega)} \leq C \left(1 + |\log(\tfrac{T}{h^2})|\right) \max_{t\in(0,T)} h^2 \|u(t)\|_{H^2(\Omega)},$$

dove u è la soluzione di (90) *e u_h la souzione di* (93)*.*

Dimostrazione. Definiamo il seguente problema ausiliario: dato $t \in (0,T)$ sia $\varphi_h : (0,t) \to V_h$ una funzione che soddisfa:

$$\begin{aligned} &-(\dot{\varphi}_h(\tau), v) + a(\varphi_h(\tau), v) = 0 \qquad\qquad \forall v \in V_h, \tau \in (0,t), \\ &\varphi_h(t) = e_h(t), \end{aligned} \tag{99}$$

dove $e_h(\tau) = u_h(\tau) - \tilde{u}_h(\tau)$ e $\tilde{u}_h(\tau)$ soddisfa a:

$$a(u(\tau) - \tilde{u}_h(\tau), v) = 0 \qquad \forall v \in V_h, \tau \in (0,T).$$

Prendendo $v = e_h(\tau)$ nella prima delle (99), e ponendo $\epsilon(\tau) = u(\tau) - \tilde{u}_h(\tau)$, otteniamo:

$$\begin{aligned} \|e_h(t)\|^2 &= \int_0^t \left[-(\dot{\varphi}_h(\tau), e_h(\tau)) + a(\varphi_h(\tau), e_h(\tau))\right] \, d\tau + (\varphi_h(t), e_h(t)) \\ &= \int_0^t \left[(\varphi_h(\tau), \dot{e}_h(\tau)) + a(\varphi_h(\tau), e_h(\tau))\right] \, d\tau + (\varphi_h(0), e_h(0)) \\ &= \int_0^t \left[(\varphi_h(\tau), \dot{\epsilon}_h(\tau)) + a(\varphi_h(\tau), \epsilon_h(\tau))\right] \, d\tau + (\varphi_h(0), \epsilon_h(0)) \\ &= -\int_0^t (\dot{\varphi}_h(\tau), \epsilon_h(\tau)) \, d\tau + (\varphi_h(t), \epsilon_h(t)), \end{aligned}$$

da cui si ottiene:

$$\|e_h(t)\|_{L^2(\Omega)} \leq -\int_0^t (\epsilon_h(\tau), \dot{\varphi}(\tau)) \, ds + (\epsilon_h(\tau), \varphi_h(t)).$$

Usando la soluzione esplicita 96 applicata alla soluzione del problema ausiliario 99 e successivamente applicando il Lemma 3.56, si può dimostrare:

$$\begin{aligned} \|\varphi_h(\tau)\|_{L^2(\Omega)} &\leq \|\epsilon_h(t)\|_{L^2(\Omega)}, \qquad 0 \leq \tau \leq t \\ \int_0^t \|\dot{\varphi}_h(\tau)\|_{L^2(\Omega)} \, d\tau &\leq C\left(1 + |\log(\tfrac{t}{h^2})|\right) \|\epsilon_h(t)\|_{L^2(\Omega)}, \end{aligned}$$

da cui facilmente:

$$\|e_h(t)\|_{L^2(\Omega)} \leq C\left(1 + |\log(\tfrac{t}{h^2})|\right) \max_{0 \leq \tau \leq t} \|\epsilon_h(t)\|_{L^2(\Omega)}.$$

La dimostrazione si conclude verificando che $u - u_h = \epsilon_h - e_h$ e usando la stima L^2 del Teorema 3.12 applicata a $\epsilon_h(\tau) = u(\tau) - \tilde{u}_h(\tau)$. □

Questo teorema, in particolare, ci dice che l'accuratezza della soluzione al tempo t non dipende dall'accuratezza con cui abbiamo risolto il transitorio iniziale. Questo è molto importante quando si cercano soluzioni indipendenti da tale transitorio.

5.4 Discertizzazione spazio-temporale

Nel seguito studieremo i più semplici metodi basati sulla discretizzazione temporale di Eulero all'indietro e in avanti. Ovviamente tecniche più avanzate possono (e devono in certi casi) ssere applicate per avere maggior efficienza computazionale e miglior controllo sull'accuratezza della soluzione. Come vedremo, però, la complicazione che deve essere risolta è data dall'interazione non semplice tra la discretizzazione temporale e quella spaziale.

Analizziamo dapprima qualitativamente come i risultati della sezione (5.1) si riflettono nella versione discreta dell'equazione. Riprendiamo l'eq. 95, la cui soluzione si può scrivere in funzione degli autovalori μ_i e corrispondenti autovettori z_j della matrice $\tilde{A}$. Nel caso $g(t) = 0$ si ha la seguente rappresentazione della soluzione:

$$\eta(t) = \sum_{j=1}^{N} (\eta_0, z_j)\, \mathrm{e}^{-\mu_j t}\, z_j.$$

Si può verificare facilmente che gli autovalori della matrice delle masse P hanno tutti ordine di grandezza $\mathcal{O}(1)$. Quindi gli autovalori di $\tilde{A}$ sono quelli di A, per cui (vedi Teorema 3.13) $\mu_1 = \mathcal{O}(1)$ e $\mu_n = \mathcal{O}(h^{-2})$. Gli autovalori più grandi corrispondono quindi ai modi (autovettori) più oscillanti, mentre gli autovalori più piccoli corrispondono agli autovettori più regolari. Le componenti della soluzione $\eta(t)$ sono caratterizzate da scale temporali assai diverse, variabili tra $\mathcal{O}(h^{-2})$ a $\mathcal{O}(1)$, segnale di "stiffness" importante. E' quindi necessario usare metodi impliciti. Nel seguito studieremo l'incondizionata stabilità del metodo di Eulero implicito, e verificheremo come le condizioni di stabilità dello schema di Eulero esplicito ne precludano l'utilizzo efficiente.

5.4.1 Il metodo di Eulero implicito (all'indietro)

Indichiamo con $I = [0, T]$ l'intervallo di integrazione temporale ($T > 0$) e sia $0 = t_0 \leq t_1 \leq \ldots \leq t_M$, una partizione di I dove $t_{n+1} = t_n + k_n$, e $I_n = (t_n, t_{n+1})$.

Partiamo dal problema semidiscreto (93). Sostituiamo al posto di $\partial u_h(t)\partial t$ il suo rapporto incrementale, possiamo scrivere il seguente problema: Trovare $u_h^n \in V_h$ tale che:

$$\begin{aligned} \left(\frac{u_h^{n+1} - u_h^n}{k_n}, v\right) + a\left(u_h^{n+1}, v\right) &= (f(t_{n+1}, v) && \forall v \in V_h \;\; n = 0, 1, \ldots, N-1, \\ \left(u_h^0, v\right) &= (u_0, v) && \forall v \in V_h. \end{aligned} \tag{100}$$

Questo corrisponde all'applicazione del metodo di Eulero all'indietro al sistema di ODE (94):

$$\left(\frac{1}{k_n} P + A\right) u^{n+1} = \frac{1}{k_n} P u^n + b^{n+1}. \tag{101}$$

Si vede quindi che ad ogni passo temporale si deve risolvere un sistema lineare di dimensioni $n \times n$ la cui matrice è data da $M = P/k_n + A$. Ovviamente si può pensare di fattorizzare la

matrice M una volta per tutti e usare la fattorizzazione ogni volta che risolviamo il sistema. Questo non è di solito conveniente perchè, da quello che abbiamo visto prima, è conveniente aumentare k_n ad ogni passo man mano che le componenti a scala temporale "veloce" vengono risolte. Bisogna quindi pensare di fattorizzare P e A per formare M ad ogni passo temporale. Anche questo potrebbe non essere fattibile se le dimensioni della mesh sono grandi, per cui si ricorre spesso a metodi di tipo gradiente coniugato precondizionato.

La stabilità dello schema di Eulero implicito si verifica immediatamente. Infatti, prendendo $v = u_h$ in (100) otteniamo per $f(t) = 0$ [7]:

$$\left\|u_h^{n+1}\right\|^2 - \left(u_h^{n+1}, u_h^n\right) + k_n a\left(u_h^{n+1}, u_h^{n+1}\right) = 0.$$

Usando la disuguaglianza di Young, si ottiene:

$$\frac{1}{2}\left\|u_h^{n+1}\right\|^2 - \frac{1}{2}\left\|u_h^n\right\|^2 + k_n a\left(u_h^{n+1}, u_h^{n+1}\right) \leq 0, \qquad n = 1, \ldots, N$$

Sommando su n si ha:

$$\left\|u_h^{n+1}\right\|^2 + 2\sum_{j=1}^{N} k_n a\left(u_h^j, u_h^j\right) \leq \left\|u_h^0\right\|^2,$$

e usando la coercività di $a(,)$ si ha immediatamente:

$$\left\|u_h^n\right\| \leq \left\|u_h^0\right\| \leq \|u_0\|, \qquad n = 1, \ldots, N, \tag{102}$$

che è evidentemente una stima analoga a quelle di stabilità del sistema semidiscreto (97) e (98).

Un altro modo per analizzare la stabilità del metodo di Eulero implicito parte dal sistema algebrico (101). Assumendo ancora $b = 0$, il sistema diventa:

$$\left(\frac{1}{k_n}P + A\right) u^{n+1} = \frac{1}{k_n} P u^n.$$

Come abbiamo detto in precedenza, la matrice delle masse P è simmetrie e definita positiva, quindi invertibile, e il suo indice di condizionamento è dell'ordine $\mathcal{O}(1)$. Moltiplicando per k_n e per P^{-1} si ottiene formalmente:

$$u^{n+1} = \left(I + k_n P^{-1} A\right)^{-1} u^n.$$

La matrice $I + k_n P^{-1} A$ è simile alla matrice $I + k_n L^{-1} A L^{-T}$, dove $P = LL^T$. Quindi la stabilità di Eulero implicito è garantita perchè

$$\left\|(I + k_n L^{-1} A L^{-T})^{-1}\right\| = \max_{i=1,n}\left[\frac{1}{\lambda_i(I + k_n L^{-1} A L^{-T})}\right] < 1$$

[7] Omettiamo il pedice nel simbolo di norma e usiamo la convenzione che sono tutte norme in $L^2(\Omega)$.

poichè gli autovalori di $P^{-1}A$ sono tutti positivi e $k_n > 0$, da cui si ricava di nuovo la (102).

Ovviamente usando il metodo di Eulero il cui errore di troncamento è $\mathcal{O}(k_n)$ la convergenza ottimale si ottiene se $k_n = \mathcal{O}(h^2)$, una restrizione troppo forte per gi usi pratici. Per questo motivo, spesso si cerca di usare schemi di discretizzazione temporale con ordine di accuratezza più elevato, come il metodo di Crank-Nicolson, che studiamo di seguito.

5.4.2 Il metodo di Crank-Nicolson

Il metodo di Crank-Nicolson (o metodo dei trapezi) si ottiene sempre usando un semplice rapporto incrementale per l'approssimazione della derivata temporale ma usando una media degli altri termini del sistema tra il passo t_{n+1} e il passo t_n:

$$\begin{aligned}\left(\frac{u_h^{n+1}-u_h^n}{k_n}, v\right) + \frac{1}{2}\left[a\left(u_h^{n+1}, v\right) + a\left(u_h^n, v\right)\right] &= \frac{1}{2}\left[(f(t_{n+1}), v) + (f(t_n), v)\right] \\ &\qquad \forall v \in V_h \;\; n = 0, 1, \ldots, N-1, \\ \left(u_h^0, v\right) &= (u_0, v) \qquad \forall v \in V_h.\end{aligned} \tag{103}$$

ovvero in termini algebrici:

$$\left(\frac{1}{k_n}P + \frac{1}{2}A\right)u^{n+1} = \left(\frac{1}{k_n}P - \frac{1}{2}A\right)u^n + \frac{1}{2}\left(b^{n+1} + b^n\right). \tag{104}$$

Questo schema ha un errore di troncamento dell'ordine $\mathcal{O}(k_n^2)$, e quindi più accurato rispetto a Eulero implicito. Lo schema è incondizionatamente stabile come Eulero implicito. Infatti:

$$\left\|(I + \frac{1}{2}k_n L^{-1}AL^{-T})^{-1}(I - \frac{1}{2}k_n L^{-1}AL^{-T})\right\| = \max_{i=1,n}\left|\frac{2I - k_n\lambda_i(L^{-1}AL^{-T})}{2I + k_n\lambda_i(L^{-1}AL^{-T})}\right| < 1.$$

5.4.3 Il metodo di Eulero esplicito (o in avanti)

Il metodo di Eulero esplicito si scrive nel modo seguente:

$$\begin{aligned}\left(\frac{u_h^{n+1}-u_h^n}{k_n}, v\right) + a\left(u_h^n, v\right) &= \left(f(t_n), v\right) \\ &\qquad \forall v \in V_h \;\; n = 0, 1, \ldots, N-1, \\ \left(u_h^0, v\right) &= (u_0, v) \qquad \forall v \in V_h.\end{aligned} \tag{105}$$

ovvero in termini algebrici:

$$\left(\frac{1}{k_n}P\right)u^{n+1} = \left(\frac{1}{k_n}P - A\right)u^n + b^n. \tag{106}$$

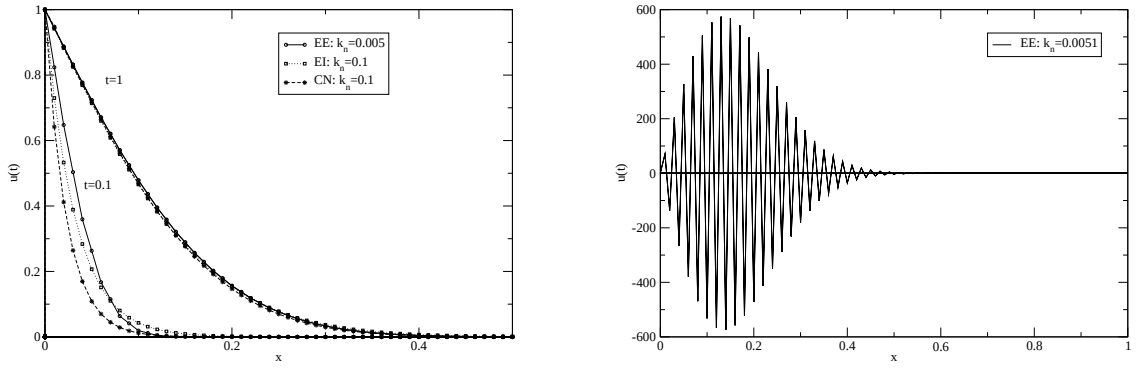

Figura 21: Risultati numerici per Eulero esplicito (EE), Eulero implicito (EI) e Crank-Nicolson per la soluzione del problema mono-dimensionale con elementi finiti P1. A destra si riporta il caso stabile, a sinistra il caso instabile.

Questo schema ha un errore di troncamento dell'ordine $\mathcal{O}(k_n)$, e quindi ha la stessa accuratezza di Eulero implicito. Lo schema è però stabile sotto condizione. Infatti:

$$\left\|(I - k_n L^{-1} A L^{-T})\right\| = \max_{i=1,n} \left|I - k_n \lambda_i (L^{-1} A L^{-T})\right| \leq 1,$$

relazione soddisfatta se $k_n \leq \frac{2}{\lambda_1(P^{-1}A)} = \mathcal{O}(h^2)$, poichè, come da Teorema 3.13, $\lambda_{\min}(A) = \mathcal{O}(h^2)$. Quindi lo schema di Eulero esplicito è stabile sotto la condizione:

$$k_n \leq Ch^2 \qquad \text{ovvero } \frac{\sqrt{Ck_n}}{h} \leq 1. \tag{107}$$

La costante C dipende dal coefficiente di diffusione $a(x)$, e la precedente può essere interpretata fisicamente come una condizione simile alla condizione CFL (Courant-Friedrichs-Lewy), che dice che la risoluzione spaziale e quella temporale devono essere tali per cui la variazione della soluzione in un passo di tempo k_n (di dimensione $\sqrt{Dk_n}$) deve rimanere all'interno di una cella (di dimensioni h). Cioè lo schema numerico deve essere in grado di risolvere correttamente le componenti più "rapide" della soluzione durante il transitorio iniziale.

Nella pratica, non si usano mai metodi espliciti dato che la condizione di stabilità richiede passi temporali troppo piccoli (variano con il quadrato di h). Conviene quindi ricorrere a metodi impliciti: anche se si è costretti a risolvere un sistema lineare ad ogni passo temporale, il fatto che, soprattutto dopo il primo transitorio, si possa aumentare il k_n fa sì che il costo computazionale sia di gran lunga inferiore a quello di un metodo esplicito.

Verifiche numeriche su caso monodimensionale. In questo paragrafo riportiamo una veloce verifica numerica dei metodi di Eulero Implicito e di Eulero Esplicito in condizioni stabili

e non. In prticolare ci riferiamo alla soluzione agli elementi finiti P_1 del seguente problema:

$$\begin{aligned}
\frac{\partial u}{\partial t} &= D\frac{\partial^2 u}{\partial x^2} && x \in (0,\infty)\,; \\
u(0,t) &= 1 && \forall t; \\
u(\infty,t) &= 0 && \forall t; \\
u(x,0) &= \begin{cases} u_0, & \text{se } x = 0, \\ 0, & \text{se } x \in (0,\infty). \end{cases} .
\end{aligned}$$

Tale equazione ammette soluzione esplicita che vale:

$$u(x,t) = u_0 \text{erfc}\left(\frac{x}{2\sqrt{Dt}}\right).$$

Figura 21 (pannello di destra) riporta le soluzioni ottenute per $D = 10^{-2}$ ai tempi $t = 0.1$ e $t = 1.0$, utilizzando FEM-P1 assieme a Eulero Esplicito (EE), Eulero implicito (EI), Crank-Nicolson (CN), utilizzando $h = 1/100$ e $k_n = 0.1$ per EI e CN, mentre $k_n = 0.005$ per EI (tale valore garantisce la stabilità). Nel pannello si sinistra viene riportato il risultato per EI con $k_n = 0.0051$. Si vedono le classiche oscillazioni, il cui valore aumenta all'aumentare di k_n.

A Appendice A: Discretizzazione alle differenze finite dell'equazione di convezione e diffusione.

Il metodo alle differenze finite per la discretizzazione dell'equazione della convezione e diffusione nel caso monodimensionale procede nel modo. Si consideri una discretizzazione uniforme di passo h dell'intervallo $[0, 1]$, per cui ciascun sotto intervallo è caratterizzato dai suoi estremi indicati con x_i e $x_{i+1} = x_i + h$. Sia u_i l'approssimazione numerica della soluzione $u(x)$ nel punto x_i: $u_i \approx u(x_i)$. Usando in maniera sistematica lo sviluppo in serie di Taylor di u_i con passo h si ottiene:

$$u_{i+1} = u_i + hu_i' + \frac{h^2}{2}u_i'' + \dots \tag{108}$$

$$u_{i-1} = u_i - hu_i' + \frac{h^2}{2}u_i'' + \dots. \tag{109}$$

Sommando membro a membro le due equazioni, e trascurando i termini dello sviluppo di ordine superiore, si ottiene la seguente approssimazione della derivata seconda nel nodo i-esimo:

$$u_i'' \approx \frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2},$$

mentre la derivata prima sempre nel punto x_i può essere approssimata da:

$$u_i' \approx \frac{u_{i+1} - u_{i-1}}{2h},$$

che si ottiene sottraendo le due espansioni di Taylor sopra riportate. Inserendo le due approssimazioni nella equazione differenziale si ottiene la seguente equazione alle differenze:

$$\frac{D}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \frac{b}{2h}(u_{i+1} - u_{i-1}) = 0,$$

che può essere scritta per ogni nodo i della griglia computazionale. Tale equazione, che coincide con la (39), è caratterizzata dal secondo ordine di approssimazione ($O(h^2)$).

Per la derivata prima si può ricavare un'approssimazione in avanti o una all'indietro esplicitando la (108) o la (109) in maniera opportuna, ottenendo la discretizzazione "upwind" o "downwind":

$$\frac{D}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \frac{b}{2h}(u_{i+1} - u_{i-1}) = 0,$$
$$\frac{D}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \frac{b}{2h}(u_{i+1} - u_{i-1}) = 0.$$

Riferimenti bibliografici

[1] M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 2005.

[2] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer-Verlag, Berlin, 1991.

[3] G. Gambolati. *Lezioni di metodi numerici per ingegneria e scienze applicate*. Cortina, Padova, Italy, 2 edition, 2002. 619 pp.

[4] C. T. Kelley. *Iterative Methods for Linear and Nonlinear Equations*. SIAM, 1995.

[5] R. Peyret and T. D. Taylor. *Computational Methods for Fluid Flow*. Springer-Verlag, New York, 1983.

[6] A. Quarteroni. *Matematica Numerica per Problemi Differenziali*. Springer-Verlag Italia, Milano, 2003.

[7] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*, volume 23 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1994.

[8] A. Quarteroni, R. Sacco, and F. Saleri. *Matematica Numerica*. Springer-Verlag Italia, Milano, 2008.