Chapitre 4

La méthode des éléments finis

4.1 La méthode de Ritz-Galerkine

La méthode de Ritz-Galerkine est construite sur la formulation variationnelle d'une équation aux dérivées partielles. Par exemple, le problème modèle

$$-\Delta u = f \quad \text{dans } \Omega, \tag{4.1}$$
$$u = 0 \quad \text{sur } \Gamma, \tag{4.2}$$

est associé à la forme variationnelle
Trouver $u \in H_0^1(\Omega)$ *tel que*

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \quad \forall v \in H_0^1(\Omega).$$

Considérons le problème variationnel abstrait
Trouver $u \in V$ *tel que*

$$a(u, v) = l(v) \quad \forall v \in V, \tag{4.3}$$

où V est un espace de Hilbert, a est une forme bilinéaire symétrique, continue et coercive et l une forme linéaire continue de sorte que les hypothèses du lemme de Lax-Milgram soient vérifiées. Il existe alors une unique solution u de ce problème variationnel.

La méthode de Ritz-Galerkine consiste à chercher une solution approchée u_h dans un sous-espace de dimension finie de V. Pour l'étude de la convergence il faut considérer une suite de sous-espaces de V de dimensions de plus en plus grandes s'approchant de V. On définit ainsi une suite de problèmes approchés paramétrés par h qui s'écrivent :
Trouver $u_h \in V_h$ *tel que*

$$a(u_h, v_h) = l(v_h) \quad \forall v_h \in V_h, \tag{4.4}$$

où $V_h \subset V$ est un sous-espace vectoriel de dimension N. Soit $(\varphi_1, \ldots, \varphi_N)$ une base de V_h. Un élément $u_h \in V_h$ peut alors s'écrire $u_h(x) = \sum_{j=1}^N u_j \varphi_j(x)$. Et en prenant $v_h = \varphi_i$ l'équation (4.4) peut s'écrire en utilisant la linéarité de a par rapport à sa première composante

$$\sum_{j=1}^N u_j a(\varphi_j, \varphi_i) = l(\varphi_i).$$

Ainsi en utilisant la symétrie de a, on remarque que la formulation variationnelle discrète (4.4) est équivalente au système linéaire

$$AU_h = L, \tag{4.5}$$

où $A = (a(\varphi_i, \varphi_j))_{1 \leq i,j \leq N}$, L est le vecteur colonne de composantes $l(\varphi_i)$ et U est le vecteur colonne contenant les inconnues u_i qui sont les coefficients de u_h dans la base $(\varphi_1, \ldots, \varphi_N)$.

Théorème 20 *On suppose que a est une forme bilinéaire symétrique, continue et coercive sur un espace de Hilbert V et l une forme bilinéaire continue sur V. Alors le système (4.5) équivaut à la formulation variationnelle discrète (4.4) et admet une solution unique.*

Preuve. Pour $v_h \in V_h$, on note $\tilde{V}$ le vecteur de ses composantes dans la base $(\varphi_1, \ldots, \varphi_N)$.

– Grâce à la linéarité de a et l la relation (4.4) s'écrit de manière équivalente :

$$ {}^t\tilde{V} A U_h = {}^t\tilde{V} L \quad \forall \tilde{V} \in \mathbb{R}^N, \tag{4.6}$$

ce qui signifie que le vecteur $AU_h - L \in \mathbb{R}^N$ est orthogonal à tous les vecteurs de $\mathbb{R}^N$, c'est donc le vecteur nul. Réciproquement, il est clair que (4.5) entraîne (4.6) et donc (4.4).

– Soit $v_h \in V_h$. Alors, comme a est coercive, il existe $\alpha > 0$ tel que

$$ {}^t\tilde{V} A \tilde{V} = a(v_h, v_h) \geq \alpha \|v_h\|^2 \geq 0,$$

et ${}^t\tilde{V} A \tilde{V} = 0 = a(v_h, v_h) \Rightarrow \|v_h\| = 0$, ce qui entraîne que $v_h = 0$ et donc $\tilde{V} = 0$. Donc A est symétrique définie positive et donc inversible.

■

Remarque 11 *On peut sous certaines conditions prendre des espaces approchés V_h qui ne sont pas inclus dans l'espace de départ V. Nous ne considérerons pas ce cas dans ce cours.*

Définition 16 *Dans le cas où les espaces discrets V_h sont inclus dans l'espace V dans lequel est définie la solution exacte, on dit que l'approximation est **conforme**. Dans le cas contraire on dit que l'approximation est **non conforme**.*

L'étude de la convergence de la méthode de Ritz-Galerkine dans le cas d'une approximation conforme est basée sur le lemme fondamental suivant :

Lemme 2 (Céa) *Soit $u \in V$ la solution de (4.3) et $u_h \in V_h$ la solution de (4.4), avec $V_h \subset V$. Alors*

$$\|u - u_h\| \leq C \inf_{v \in V_h} \|u - v\|.$$

Preuve. On a

$$\begin{aligned} a(u, v) &= l(v) \quad \forall v \in V, \\ a(u_h, v_h) &= l(v_h) \quad \forall v_h \in V_h, \end{aligned}$$

comme $V_h \subset V$, on peut prendre $v = v_h$ dans la première égalité et faire la différence, ce qui nous donne

$$a(u - u_h, v_h) = 0 \quad \forall v_h \in V_h.$$

Il en résulte que $a(u - u_h, u - u_h) = a(u - u_h, u - v_h + v_h - u_h) = a(u - u_h, u - v_h)$, car $v_h - u_h \in V_h$ et donc $a(u - u_h, v_h - u_h) = 0$. Il existe alors $\alpha > 0$ et β tels que

$$\begin{aligned} \alpha \|u - u_h\|^2 &\leq a(u - u_h, u - u_h) && \text{car } a \text{ est coercive,} \\ &\leq a(u - u_h, u - v_h) \quad \forall v_h \in V_h, \\ &\leq \beta \|u - u_h\| \|u - v_h\| && \text{car } a \text{ est continue.} \end{aligned}$$

D'où finalement $\|u - u_h\| \leq \frac{\beta}{\alpha} \|u - v_h\|$ pour tout $v_h \in V_h$. On obtient le résultat désiré en passant à l'inf sur V_h. ■

4.2 Approximation polynômiale en dimension 1

Pour utiliser la méthode de Ritz-Galerkine en pratique, il faut trouver une « bonne » suite d'espaces de discrétisation $(V_h)_h$. Une des idées essentielles pour la construction d'une approximation de Galerkine efficace numériquement est d'avoir une matrice A la plus creuse possible, i.e. avec $a(\varphi_i, \varphi_j) = 0$ pour beaucoup de couples (i, j), ce qui permettra de diminuer le nombre d'opérations à effectuer pour la résolution du système linéaire associé. On essaye donc de prendre des fonctions de base à support petit. Il faut évidemment également veiller à ce que l'espace de dimension finie ainsi construit soit bien inclus dans l'espace de départ où est définie la solution exacte.

Premier exemple : on revient au problème modèle du Laplacien en dimension 1

$$\begin{aligned} -u'' &= f \text{ dans }]0,1[, \\ u(0) &= u(1) = 0, \end{aligned}$$

dont la formulation variationnelle s'écrit
Trouver $u \in H_0^1(]0,1[)$ *tel que*

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \quad \forall v \in H_0^1(]0,1[).$$

On cherche un sous-espace V_h de $V = H_0^1(]0,1[$. Pour cela on commence par construire le maillage $(x_0, \ldots, x_{N+1})$ de $]0,1[$ tel que $x_i = ih$, avec $h = \frac{1}{N+1}$.

On construit alors pour $i = 1, \ldots, N$,

$$\varphi_i(x) = \left| \begin{array}{ll} \frac{x - x_{i-1}}{x_i - x_{i-1}} & x \in [x_{i-1}, x_i], \\ \frac{x - x_{i+1}}{x_i - x_{i+1}} & x \in [x_i, x_{i+1}], \\ 0 & \text{sinon.} \end{array} \right.$$

On définit alors $V_h = \mathrm{Vect}(\varphi_1, \ldots, \varphi_N)$, l'espace vectoriel engendré par $(\varphi_1, \ldots, \varphi_N)$.

Proposition 9 *L'espace V_h est un sous-espace vectoriel de $H_0^1(]0,1[$ de dimension N.*

Preuve.

- Les (φ_i) forment un système libre de N vecteurs. En effet supposons qu'il existe des constantes $(c_1, \ldots, c_N)$ tel que $c_1\varphi_1(x) + \cdots + \varphi_N(x) = 0$ pour tout $x \in]0,1[$. Alors en particulier pour $x = x_i$ on a $\varphi_j(x_i) = 0$ pour $j \neq i$ et $\varphi_i(x_i) = 1$, donc $c_i = 0$. Il en résulte que l'espace V_h engendré par les φ_i est bien de dimension N.
- Soit $v_h \in V_h$ quelconque. Montrons que $v_h \in H_0^1(]0,1[$.
  1) On vérifie que $v_h \in L^2(]0,1[)$. La fonction en tant que combinaison linéaire de fonctions continues sur $[0,1]$ est continue sur $[0,1]$, elle y atteint donc son maximum et on a
  $$\int_0^1 v_h^2 \, dx \leq \max_{x \in [0,1]} |v_h(x)|^2 < \infty.$$
  2) On vérifie que $v_h' \in L^2(]0,1[)$. La fonction v_h linéaire par morceaux, n'est pas dérivable aux points x_i, il faut donc calculer v_h' au sens des distributions et le théorème 4 de

recollement, qui se généralise facilement au cas de N sous-domaines, permet d'affirmer que $v_h' \in L^2(]0,1[)$.
3) Pour tout $j = 1, \dots N$, $\varphi_j(0) = \varphi_j(1) = 0$. Donc pour tout $v_h \in V_h$, on a $v_h(0) = v_h(1) = 0$.
On en déduit que $V_h \subset H_0^1(]0,1[$.

■

On note

$$V_h^k = \{v_h \in C^0([a,b]),\ v_{h|[x_i,x_{i+1}]} \text{ est un polynôme de degré } k\},$$

et

$$V_{h,0}^k = \{v_h \in V_h^k \mid v_h(a) = v_h(b) = 0.\}$$

L'espace d'approximation V_h que nous avons construit précédemment correspond a $V_{h,0}^1$.

Remarque 12 *Pour tout $k \geq 1$, on a $V_h^k \subset C^0(]a,b[)$ et $V_h^k \subset H^1(]a,b[)$, mais si l'on considérait des fonctions constantes par morceaux au lieu d'être au moins affine par morceaux, cela ne serait pas le cas. L'espace V_h^0 ne permet pas de définir une approximation conforme dans $H^1(]a,b[)$.*

Remarque 13 *L'élément fini $\mathbb{P}_0$ existe également, les fonctions de base associées sont constantes par maille et on a*

$$\mathbb{P}_0 \subset L^2(]a,b[),\ \mathbb{P}_0 \not\subset H^1(]a,b[),\ \mathbb{P}_0 \not\subset C^0(]a,b[).$$

On ne peut donc pas l'utiliser pour obtenir une approximation conforme de notre problème modèle dont la solution est dans $H^1(]0,1[$.

Pour $v \in C^0(]a,b[)$ on définit le projeté πv de v sur l'espace V_h par

$$(\pi v)(x) = \sum_{i=1}^{N} v(x_i)\varphi_i(x).$$

Théorème 21 *Si la solution exacte de (4.3) est dans $H^2(]a,b[)$, alors on a l'estimation d'erreur $\|u - u_h\|_{H^1} \leq Ch$, où $u_h \in V_h$ est la solution du problème approché (4.4) et $h = \max_i |x_{i+1} - x_i|$.*

Preuve. On a

$$\|u - u_h\|_{H^1}^2 = \|u - u_h\|_{L^2}^2 + \|u' - u_h'\|_{L^2}^2.$$

En utilisant le lemme de Céa, on obtient

$$\|u - u_h\|_{H^1}^2 \leq C \inf_{v_h \in V_h} \|u - v_h\|_{H^1} \leq C\|u - \pi u\|_{H^1}.$$

Il suffit donc d'estimer $\|u - \pi u\|_{H^1}$. Soit donc $x \in [x_i, x_{i+1}]$. Alors

$$\begin{aligned}(\pi u)(x) &= u(x_i)\varphi_i(x) + u(x_{i+1})\varphi_{i+1}(x),\\ &= u(x_i)\frac{x - x_{i+1}}{x_i - x_{i+1}} + u(x_{i+1})\frac{x - x_i}{x_{i+1} - x_i}.\end{aligned}$$

Or, on a $(u - \pi u)(x_i) = 0$, donc

$$(u - \pi u)(x) = \int_{x_i}^{x} (u' - (\pi u)')(y)\, dy, \tag{4.7}$$

$$= \int_{x_i}^{x} \big(u'(y) - \frac{u(x_{i+1}) - u(x_i)}{x_{i+1} - x_i}\big)\, dy, \tag{4.8}$$

or $u \in H^2(]a, b[) \Rightarrow u \in C^1(]a, b[)$, donc, d'après la formule des accroissements finis il existe $\xi \in [x_i, x_{i+1}]$ tel que $u'(\xi) = \frac{u(x_{i+1}) - u(x_i)}{x_{i+1} - x_i}$. Puis pour $y \in [x_i, x]$

$$u'(y) - \frac{u(x_{i+1}) - u(x_i)}{x_{i+1} - x_i} = \int_{\xi}^{y} u''(z)\, dz \leq (\int_{x_i}^{x_{i+1}} |u''|^2)^{\frac{1}{2}} |x_{i+1} - x_i|^{\frac{1}{2}}.$$

Alors

$$|(u' - (\pi u)'(x)|^2 \leq h \int_{x_i}^{x_{i+1}} |u''|^2\, dx,$$

puis

$$\int_{x_i}^{x_{i+1}} |u' - (\pi u)'(x)|^2\, dx \leq h|x_{i+1} - x_i| \int_{x_i}^{x_{i+1}} |u''|^2\, dx \leq h^2 \int_{x_i}^{x_{i+1}} |u''|^2\, dx,$$

et en sommant de 0 à N

$$\int_0^1 |u' - (\pi u)'(x)|^2\, dx \leq h^2 \int_0^1 \|u''\|^2\, dx,$$

soit $\|u - \pi u\|_{L^2} \leq h\|u''\|_{L^2}$. Il en résulte également à l'aide de Cauchy-Schwarz pour (4.7) que

$$|(u - \pi u)(x)| \leq \int_{x_i}^{x} (\int_{x_i}^{x_{i+1}} |u''|^2)^{\frac{1}{2}} |x_{i+1} - x_i|^{\frac{1}{2}}\, dx,$$

$$\leq \big(\int_{x_i}^{x_{i+1}} |u''|^2\, dx\big)^{\frac{1}{2}} |x_{i+1} - x_i|^{\frac{3}{2}},$$

puis

$$\int_{x_i}^{x_{i+1}} |(u - \pi u)(x)|^2\, dx \leq \int_{x_i}^{x_{i+1}} |u''|^2\, dx |x_{i+1} - x_i|^3\, |x_{i+1} - x_i| \leq h^4 \int_{x_i}^{x_{i+1}} |u''|^2\, dx,$$

et en sommant de 0 à N

$$\int_0^1 |(u - \pi u)(x)|^2\, dx \leq h^4 \int_0^1 (u'')^2\, dx,$$

d'où $\|u - \pi u\|_{L^2} \leq h^2\|u''\|_{L^2}$. ■

Remarque 14 *Pour notre problème modèle, il est clair que $f \in L^2$ implique que $u'' \in L^2$.*

4.3 Définition d'un élément fini

On considère un triplet (K, P, Σ) où
(i) K est un sous-ensemble fermé de $\mathbb{R}^n$ d'intérieur non vide,
(ii) P est un espace vectoriel de dimension finie de fonctions définies sur K,
(iii) Σ est un ensemble de formes linéaires sur P de cardinal fini N, $\Sigma = \{\sigma_1, \ldots, \sigma_N\}$.

Définition 17 *On dit que Σ est P-**unisolvant** si pour tout N-uplet $(\alpha_1, \ldots, \alpha_N)$, il existe un unique élément $p \in P$ tel que $\sigma_i(p) = \alpha_i$ pour $i = 1, \ldots, N$.*

Définition 18 *Le triplet (K, P, Σ) de $\mathbb{R}^n$ est appelé **élément fini** de $\mathbb{R}^n$ s'il satisfait (i), (ii) et (iii) et si Σ est P-unisolvant.*

Exemple 1 : Soit $a, b \in \mathbb{R}$, $a < b$. Soit $K = [a, b]$, $P = \mathbb{P}_1$ l'ensemble des fonctions affines sur $[a, b]$, $\Sigma = \{\sigma_a, \sigma_b\}$, où

$$\begin{array}{rcl} \sigma_a : P & \to & \mathbb{R}, \\ p & \mapsto & p(a), \end{array} \qquad \begin{array}{rcl} \sigma_b : P & \to & \mathbb{R}, \\ p & \mapsto & p(b). \end{array}$$

De plus Σ est P-unisolvant : soit $(\alpha, \beta) \in \mathbb{R}^2$. On cherche $p \in P$ tel que $p(a) = \alpha$ et $p(b) = \beta$. L'unique solution est

$$p(x) = (\beta - \alpha)\frac{x - a}{b - a} + \alpha.$$

Exemple 2 : Soit $a, b, c \in \mathbb{R}$, $a < c < b$. Soit $K = [a, b]$, $P = \mathbb{P}_2$ l'ensemble des polynômes quadratiques sur $[a, b]$, $\Sigma = \{\sigma_a, \sigma_b, \sigma_c\}$, où

$$\begin{array}{rcl} \sigma_a : P & \to & \mathbb{R}, \\ p & \mapsto & p(a), \end{array} \qquad \begin{array}{rcl} \sigma_b : P & \to & \mathbb{R}, \\ p & \mapsto & p(b), \end{array} \qquad \begin{array}{rcl} \sigma_c : P & \to & \mathbb{R}, \\ p & \mapsto & p(c). \end{array}$$

Il reste à vérifier que Σ est P-unisolvant : soit $(\alpha, \beta, \gamma) \in \mathbb{R}^3$. On cherche $p \in \mathbb{P}_2$ tel que $p(a) = \alpha$, $p(c) = \gamma$ et $p(b) = \beta$. La forme générique d'un élément $p \in \mathbb{P}_2$ est $p(x) = \lambda x^2 + \mu x + \nu$. Donc

$$\left\{ \begin{array}{l} p(a) = \alpha \\ p(b) = \beta \\ p(c) = \gamma \end{array} \right\} \Leftrightarrow \left\{ \begin{array}{l} \lambda a^2 + \mu a + \nu = \alpha \\ \lambda b^2 + \mu b + \nu = \beta \\ \lambda c^2 + \mu c + \nu = \gamma \end{array} \right\}$$

On a donc un système de 3 équations à 3 inconnues (λ, μ, ν) qui admet une unique solution si

$$\det \begin{pmatrix} a^2 & a & 1 \\ b^2 & b & 1 \\ c^2 & c & 1 \end{pmatrix} = (b - a)(b - c)(c - a) \neq 0,$$

ce qui est le cas car $a < c < b$. Donc Σ est P-unisolvant.

Exemple 3 : Soit $K \subset \mathbb{R}^2$ le triangle non dégénéré de sommets (a, b, c), $P = \mathbb{P}_1(K) = \{\alpha x+\beta y+\gamma, (x, y) \in K\}$ i.e. c'est l'ensemble des fonctions affines sur K et $\Sigma = \{\sigma_a, \sigma_b, \sigma_c\}$, où

$$\begin{array}{rcl} \sigma_a : P & \to & \mathbb{R}, \\ p & \mapsto & p(a), \end{array} \qquad \begin{array}{rcl} \sigma_b : P & \to & \mathbb{R}, \\ p & \mapsto & p(b), \end{array} \qquad \begin{array}{rcl} \sigma_c : P & \to & \mathbb{R}, \\ p & \mapsto & p(c). \end{array}$$

Là encore, Σ est P-unisolvant, comme dans l'exemple précédant on se ramène à un système de 3 équations à 3 inconnues qui admet une solution unique si les points a, b et c sont distincts, ce qui est le cas ici car le triangle est non dégénéré.

Remarque 15 *Les éléments finis que nous avons vu dans les exemples 1, 2 et 3 sont des éléments finis de* **Lagrange**, *pour lesquels les formes linéaires de Σ sont les valeurs des polynômes de P en N points judicieusement choisis (N est la dimension de P et le cardinal de Σ).*

Nous allons maintenant démontrer deux lemmes permettant de caractériser la propriété d'unisolvance, qui sont très utiles dans des cas un peu plus complexes que ceux des exemples précédents.

Lemme 3 *L'ensemble Σ est P-unisolvant si et seulement si les deux propriétés suivantes sont vérifiées*
(i) Il existe N fonctions $p_i \in P$ linéairement indépendantes telles que $\sigma_j(p_i) = \delta_{ij}$ pour $1 \leq i, j \leq N$,
(ii) dim(P) = card(Σ).

Preuve. Soit

$$\begin{array}{rcl} \phi : P & \to & \mathbb{R}^N \\ p & \mapsto & \begin{pmatrix} \sigma_1(p) \\ \vdots \\ \sigma_N(p) \end{pmatrix} \end{array}$$

On voit aisément que Σ est P-unisolvant si et seulement si ϕ est bijective. On va donc utiliser l'application ϕ pour démontrer le lemme.

1) On suppose que Σ est P-unisolvant.
Alors ϕ est une application linéaire bijective, donc un isomorphisme. L'image réciproque par ϕ de la base canonique de $\mathbb{R}^N$ est donc une base de P qui correspond aux (p_i) de (i). De plus la dimension de P est $N = card(\Sigma)$.

2) Réciproquement, on suppose les propriétés (i) et (ii) vérifiées. Montrons alors que ϕ est bijective. D'après (ii), $dim(P) = card(\Sigma) = N = dim(\mathbb{R}^N)$, il suffit donc de montrer que ϕ est surjective. Pour cela, on prend $(\alpha_1, \ldots, \alpha_N) \in \mathbb{R}^N$ quelconque, alors $p = \alpha_1 p_1 + \cdots + \alpha_N p_N$ vérifie que $\sigma_i(p) = \alpha_i$ d'après la propriété (i), d'où la surjectivité. ∎

Exemple 4 : Soit $K \subset \mathbb{R}^2$ le triangle non dégénéré de sommets (a_1, a_2, a_3), $P = \mathbb{P}_2(K)$, l'ensemble des polynômes de degré 2 sur K et $\Sigma = \{\sigma_{a_1}, \sigma_{a_2}, \sigma_{a_3}, \sigma_{a_{12}}, \sigma_{a_{13}}, \sigma_{a_{23}}\}$, où a_{12}, a_{13} et a_{23} sont les milieux des côtés $[a_1, a_2]$, $[a_1, a_3]$,et $[a_2, a_3]$. Pour a désignant l'un des 6 points $(a_1, a_2, a_3, a_{12}, a_{13}a_{23})$, σ_a est l'application

$$\begin{array}{rcl} \sigma_a : P & \to & \mathbb{R}, \\ p & \mapsto & p(a). \end{array}$$

Pour démontrer que Σ est P-unisolvant, nous allons utiliser le lemme 3. Commençons par le (ii). L'ensemble Σ contient 6 éléments et la dimension de $\mathbb{P}_2$ est également 6 (une base de $\mathbb{P}_2$

est donnée par $(1, x, y, x^2, xy, y^2)$). Il reste maintenant à exhiber une base $(p_i)_i$ de P telle que $\sigma_j(p_i) = \delta_{ij}$. Pour cela, il est pratique d'utiliser les coordonnées barycentriques du triangle, qui sont les fonctions affines $(\lambda_1, \lambda_2, \lambda_3)$ qui valent respectivement 1 aux points a_1, a_2, a_3 et qui s'annulent aux deux autres points. Une propriété des coordonnées barycentriques est de prendre une valeur constante sur les droites parallèles au côté opposé du sommet où elles valent 1. En particulier, elles s'annulent sur le coté opposé et valent $\frac{1}{2}$ sur la droite passant par les milieux des segments contenant le sommet où elles valent 1. De ces propriétés, on déduit aisément, avec des notations évidentes que

$$p_1 = \lambda_1(2\lambda_1 - 1), \quad p_2 = \lambda_2(2\lambda_2 - 1), \quad p_3 = \lambda_3(2\lambda_3 - 1),$$

$$p_{12} = 4\lambda_1\lambda_2, \quad p_{13} = 4\lambda_1\lambda_3, \quad p_{23} = 4\lambda_2\lambda_3.$$

D'où l'unisolvance.

Lemme 4 *L'ensemble Σ est P-unisolvant si et seulement si les deux propriétés suivantes sont vérifiées*
(i) $\sigma_j(p) = 0$ pour tout $j = 1, \dots N \Rightarrow p = 0$,
(ii) dim(P) = card(Σ).

Preuve. Même principe que dans la démonstration du lemme précédent en utilisant l'injectivité de ϕ. ■

Exemple 5 : Soit $a, b \in \mathbb{R}$, $a < b$. Soit $K = [a, b]$, $P = \mathbb{P}_3$ l'ensemble des polynômes cubiques sur $[a, b]$, $\Sigma = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$, où

$$\begin{array}{rcl} \sigma_1 : P & \to & \mathbb{R}, \\ p & \mapsto & p(a), \end{array} \qquad \begin{array}{rcl} \sigma_2 : P & \to & \mathbb{R}, \\ p & \mapsto & p(b), \end{array} \qquad \begin{array}{rcl} \sigma_3 : P & \to & \mathbb{R}, \\ p & \mapsto & p'(a), \end{array} \qquad \begin{array}{rcl} \sigma_4 : P & \to & \mathbb{R}, \\ p & \mapsto & p'(b). \end{array}$$

On va utiliser le lemme 4 pour vérifier que Σ est P-unisolvant. Soit $p \in \mathbb{P}_3$ tel que $\sigma_j(p) = 0$ pour tout $j = 1, \dots 4$. Alors en particulier p' est un polynôme de degré 2 qui s'annule en a et en b, donc $p'(x) = \lambda(x - a)(x - b)$ avec $\lambda \in \mathbb{R}$. Le polynôme p s'annule également en a et en b, c'est donc en particulier la primitive de p' qui s'annule en a. Par une intégration par partie on trouve son expression en b

$$p(b) = -\lambda \frac{(b - a)^3}{6}.$$

Et donc comme $a \neq b$, $p(b) = 0$ entraîne que $\lambda = 0$ et donc p est le polynôme nul. Et bien sûr, la dimension de $\mathbb{P}_3$ est 4, ce qui correspond au nombre d'éléments de Σ, d'où l'unisolvance.

Remarque 16 *L'élément fini de l'exemple 5 où les formes linéaires sont donnés par les valeurs des éléments de P et de leurs dérivées en des points est un élément fini de* **Hermite**.

Exemple 6 : Soit $K \subset \mathbb{R}^2$ le triangle non dégénéré de sommets (a_1, a_2, a_3),

$$P = (\mathbb{P}_0(K))^2 \oplus \mathbb{P}_0(K) \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \alpha + \gamma x \\ \beta + \gamma y \end{pmatrix},$$

où α, β et γ sont des réels. L'espace P est clairement un espace vectoriel de dimension 3 de fonctions définies sur K à valeurs dans $\mathbb{R}^2$. En notant n_{ij} la normale au côté $[a_i, a_j]$, on définit les formes linéaires de $\Sigma = \{\sigma_1, \sigma_2, \sigma_3\}$, comme suit

$$\begin{array}{rcl} \sigma_i : P & \to & \mathbb{R}, \\ p & \mapsto & \int_{[a_j, a_k]} p(s) \cdot n_{jk}\, ds. \end{array}$$

Pour démontrer l'unisolvance on va utiliser le lemme 4, en notant tout d'abord que la dimension de P et le cardinal de Σ sont égaux. Ensuite, soit $p \in P$ tel que $\sigma_i(p) = 0$ pour $i = 1, 2, 3$. Commençons par vérifier que dans ce cas $p \cdot n$ est constant sur chacun des cotés du triangle. Considérons le côté $[a_i, a_j]$ et $a \in [a_i, a_j]$ alors

$$(p(a) - p(a_i)) \cdot n_{ij} = \gamma(a - a_i) \cdot n_{ij} = 0.$$

Maintenant grâce à la formule de Green (2.3), on a

$$\int_K \operatorname{div} p \, dx \, dy = \int_{a_1}^{a_2} p(s) \cdot n_{12} ds + \int_{a_2}^{a_3} p(s) \cdot n_{23} ds + \int_{a_3}^{a_1} p(s) \cdot n_{13} ds.$$

Or les termes du second membre correspondent aux $\sigma_i(p)$, ils sont donc nuls par hypothèse et d'autre part $\operatorname{div} p = 2\gamma$. Donc $\gamma = 0$. Il en résulte que p est un vecteur constant. Alors comme $p \cdot n_{12} = p \cdot n_{13} = 0$ et que n_{12} et n_{13} ne sont pas colinéaires, $p = 0$. D'où l'unisolvance.

Définition 19 *Soit (K, P, Σ) un élément fini. Soit C un espace vectoriel contenant P (par exemple $C^0(K), C^1(K), C^2(K), ...$). Pour $v \in C$ on appelle P-interpolé de v l'unique élément $\pi v \in P$ tel que $\sigma_i(\pi v) = \sigma_i(v)$ $\forall i = 1, \ldots, N$.*

Remarque 17 *L'existence et l'unicité de π sont garanties parce que Σ est P-unisolvant.*

Remarque 18 *Pour construire π, on utilise en général la base de P associée à Σ par le lemme 3 :*

$$\pi v = \sum_{j=1}^{N} \sigma_j(v) p_j.$$

Remarque 19 *Pour un élément fini de Lagrange πv n'est autre que l'interpolé de Lagrange.*

4.4 Familles d'éléments finis

Définition 20 *Deux élément finis (K, P, Σ) et $(\hat{K}, \hat{P}, \hat{\Sigma})$ sont dits* **affinement équivalents** *si il existe une application affine inversible $F : \mathbb{R}^n \to \mathbb{R}^n$ telle que*
(i) $K = F(\hat{K})$,
(ii) $P = \{\hat{p} \circ F^{-1}, \ \forall \hat{p} \in \hat{P}\}$,
(iii) Les ensembles $\Sigma = \{(\sigma_i)_{i=1,\ldots,N}\}$ et $\hat{\Sigma} = \{(\hat{\sigma}_i)_{i=1,\ldots,N}\}$ vérifient $\sigma_i(p) = \hat{\sigma}_i(p \circ F)$ $\forall p \in P, i = 1, \ldots N$.

Exemple 1 : Soit $K = [a_1, a_2]$ et $\hat{K} = [\hat{a}_1, \hat{a}_2]$ deux intervalles non vides, $P = \mathbb{P}_1(K)$, $\Sigma = \{\sigma_1, \sigma_2\}$ tels que $\sigma_1(p) = p(a_1)$ et $\sigma_2(p) = p(a_2)$. On considère l'application affine

$$F : \begin{array}{ccc} \hat{K} & \to & K \\ \hat{x} & \mapsto & \alpha\hat{x} + \beta. \end{array}$$

On cherche (α, β) tel que $a_1 = \alpha\hat{a}_1 + \beta$ et $a_2 = \alpha\hat{a}_2 + \beta$. Pour cela, on prend

$$\alpha = \frac{a_2 - a_1}{\hat{a}_2 - \hat{a}_1} \qquad \beta = \frac{\hat{a}_2 a_1 - \hat{a}_1 a_2}{\hat{a}_2 - \hat{a}_1},$$

ce qui nous donne le (i) de la définition. Le (ii) provient du fait que la composée de deux polynômes du premier degré est encore un polynôme du premier degré. Il reste encore à vérifier (iii) : d'une part $\hat{\sigma}_i(\hat{p}) = \hat{p}(\hat{a}_i)$, d'autre part $\sigma_i(p) = p(a_i) = p \circ F(\hat{a}_i) = \hat{p}(\hat{a}_i)$, ce qui nous donne l'égalité souhaitée.

Remarque 20 *Si on choisit l'application affine qui transforme a_1 en $\hat{a}_2$ et a_2 en $\hat{a}_1$, il faut changer $\hat{\Sigma}$ pour avoir l'équivalence affine des deux éléments finis.*

Exemple 2 : On considère les éléments finis de Hermite $\mathbb{P}_3$ basés sur $K = [a_1, a_2]$ et $\hat{K} = [\hat{a}_1, \hat{a}_2]$.
(i) On prend l'application affine F de l'exemple 1 telle que $K = F(\hat{K})$,
(ii) $\mathbb{P}_3(K) = \{\hat{p} \circ F^{-1} \quad \forall \hat{p} \in \mathbb{P}_3(\hat{K})\}$ est la composée d'un polynôme de degré 3 avec une application affine, c'est donc encore un polynôme de degré 3 défini maintenant sur K,
(iii) On a d'abord σ_1 telle que $\sigma_1(p) = p(a_1)$ et σ_2 telle que $\sigma_2(p) = p(a_2)$ pour lesquelles on peut procéder comme dans l'exemple 1. Ensuite on a σ_3 telle que $\sigma_3(p) = p'(a_1) = p'(\alpha\hat{a}_1 + \beta)$ qui est associée à $\hat{\sigma}_3$ telle que $\hat{\sigma}_3(\hat{p}) = \hat{p}'(\hat{a}_1) = (p \circ F)'(\hat{a}_1) = F'(\hat{a}_1)p'(F(\hat{a}_1)) = F'(\hat{a}_1)p'(a_1)$. On a donc l'égalité si et seulement si $F'(\hat{a}_1) = 1$ ce qui équivaut à $\alpha = 1$ ou encore $a_2 - a_1 = \hat{a}_2 - \hat{a}_1$, c'est-à-dire que la longueur de l'intervalle doit être conservé. Le même calcul s'applique également à σ_4.

On en conclut que ces éléments de Hermite sont équivalents si et seulement si K et $\hat{K}$ ont la même longueur.

Exemple 3 : On considère dans $\mathbb{R}^2$ les triangles non dégénérés K de sommets (a_1, a_2, a_3) et $\hat{K}$ de sommets $(\hat{a}_1, \hat{a}_2, \hat{a}_3)$ avec $\hat{a}_1 = (0, 0)$, $\hat{a}_2 = (1, 0)$ et $\hat{a}_3 = (0, 1)$. Et les éléments finis de Lagrange $\mathbb{P}_1(K)$ et $\mathbb{P}_1(\hat{K})$ construits sur ces triangles, i.e. les formes linéaires associées sont les valeurs des polynômes aux sommets des triangles. L'application linéaire F_K telle que

$$F_K \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix} = \begin{pmatrix} a_{2x} - a_{1x} & a_{3x} - a_{1x} \\ a_{2y} - a_{1y} & a_{3y} - a_{1y} \end{pmatrix} \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix} + \begin{pmatrix} a_{1x} \\ a_{2x} \end{pmatrix},$$

où (a_{ix}, a_{iy}) sont les composantes de a_i, fait transforme $\hat{K}$ en K. De plus la propriété (ii) est vérifiée parce que la composée d'un polynôme de degré 1 et d'une application affine reste un polynôme de degré 1 et on a la propriété (iii) car $p(F_K(\hat{a}_i)) = p(a_i)$.

Remarque 21 *L'élément fini basé sur le triangle $\hat{K}$ défini dans l'exemple 3 est appelé élément de référence. En raison de des coordonnées simples de ces sommets on l'utilise souvent pour calculer les $a(\hat{p}_i, \hat{p}_j)$, pour une forme bilinéaire a donnée, les $a(p_i, p_j)$ pour les autres éléments finis en étant ensuite déduits par transformation affine.*

Notation : Soient (K, P, Σ) et $(\hat{K}, \hat{P}, \hat{\Sigma})$ deux éléments finis affinement équivalents pour une application affine F. On utilise la notation suivante : si la fonction v est définie sur K, alors $\hat{v}$ est la fonction définie sur $\hat{K}$ par $\hat{v} = v \circ F$.

Théorème 22 *Soient (K, P, Σ) et $(\hat{K}, \hat{P}, \hat{\Sigma})$ deux éléments finis affinement équivalents pour une application affine F. L'ensemble $\{p_i,\ i = 1 \ldots, N\}$ est la base de P associée à l'élément fini (K, P, Σ) si et seulement si l'ensemble $\{\hat{p}_i,\ i = 1 \ldots, N\}$ est la base de $\hat{P}$ associée à l'élément fini $(\hat{K}, \hat{P}, \hat{\Sigma})$.*
Si de plus on a

$$\sigma_i(v) = \hat{\sigma}_i(\hat{v}), \quad \forall = 1 \ldots, N,\ \forall v \in C,$$

alors

$$\widehat{\pi v} = \hat{\pi}\hat{v} \quad \forall v \in C.$$

Preuve. La première assertion vient du fait que $\sigma_i(p_j) = \hat{\sigma}_i(\hat{p}_j)$. D'autre part

$$\begin{aligned}\hat{\pi}\hat{v} &= \sum_{i=1}^{N} \hat{\sigma}_i(\hat{v})\hat{p}_i, \\ &= \sum_{i=1}^{N} \sigma_i(v)\hat{p}_i = (\sum_{i=1}^{N} \sigma_i(v)p_i) \circ F \\ &= \widehat{\pi v}.\end{aligned}$$

■

4.5 Construction d'espaces d'approximation

La notion d'élément fini que nous avons introduite dans les sections précédentes vont maintenant être utilisées pour construire des espaces d'approximation pour la méthode de Ritz-Galerkine avec des fonctions de base de support petit. Typiquement, l'espace discret associé à un élément fini (K, P, Σ) va être

$$V_h = \{v \in C \text{ (vérifiant les C.L.)} \quad v_{|K_i} \in P\},$$

où C est l'espace de continuité naturel imposé par l'élément fini, i.e. $C = C^0(\Omega)$ pour les éléments finis de Lagrange, $C = C^1(\Omega)$ pour les éléments finis de Hermite.

Concrètement, étant donné un ouvert Ω que l'on suppose polygonal en 2D ou polyédral en 3D pour simplifier (sinon on l'approxime par un ouvert de ce type), on construit une partition de Ω telle que $\bar{\Omega} = \cup K_i$ où les (K_i, P_i, Σ_i) définissent des éléments finis de même type. On suppose que dans cette partition les faces de deux K_i adjacents se correspondent exactement et que les degrés de libertés, i.e. les éléments de Σ, partagés par deux éléments finis adjacents prennent la même valeur. Par exemple dans le cas des éléments finis de Lagrange, si (K_1, P_1, Σ_1) et (K_2, P_2, Σ_2) vérifient $K_1 \cap K_2 \neq \varnothing$ et si un point a définissant un degré de liberté appartient à l'intersection, alors pour tout $p_1 \in P_1$ et pour tout $p_2 \in P_2$, on a $p_1(a) = p_2(a)$. Les propriétés de régularité globale (continuité, dérivabilité,..) de l'espace V_h construit à partir d'un élément fini vont se déduire des contraintes imposées par l'égalité des degrés de liberté communs à deux éléments finis adjacents.

Exemple 1 : On va construire l'espace V_h basé sur l'élément fini de Lagrange $\mathbb{P}_1(K)$ en une dimension. Notons qu'on peut parler sans confusion d'élément fini de Lagrange $\mathbb{P}_1(K)$, car les Σ associés sont univoques. C'est ce que l'on fait couramment.

On construit le maillage $(x_0, \ldots, x_{N+1})$ de $[0, 1]$ de sorte que $x_i = ih$ avec $h = \frac{1}{N+1}$. On définit $K_i = [x_i, x_{i+1}]$ et on considère les éléments finis de Lagrange $(K_i, \mathbb{P}_1(K_i))$. Comme les fonctions de V_h doivent être définie de manière unique aux points x_i, et qu'elles sont affines dans chacun des intervalles $[x_i, x_{i+1}]$, elles sont continues. Par contre elles n'ont aucune raison d'être dérivables aux points x_i et elle ne le sont pas en général. L'espace $C^0(]0, 1[)$ est donc bien l'espace naturel dans lequel sont définies les fonctions basées sur l'élément fini $\mathbb{P}_1$. On définit alors le sous-espace V_h de $H^1(]0, 1[)$ par

$$V_h = \{v \in C^0(]0, 1[) \mid v_{K_i} \in \mathbb{P}_1(K_i)\}.$$

Notons que $C^0(]0, 1[)$ est également l'espace naturel pour les élément finis de Lagrange $\mathbb{P}_k$ avec $k > 1$. La seule contrainte est l'égalité des fonctions aux noeuds du maillage, le fait de

prendre des polynômes de degré plus élevé n'apporte pas de régularité supplémentaire, mais bien évidemment plus de précision.

Exemple 2 : On va construire l'espace V_h basé sur l'élément fini de Lagrange $\mathbb{P}_1(K)$ en dimension deux. Soit Ω un ouvert de $\mathbb{R}^2$ à bord polygonal avec une triangulation $\mathcal{T}_h$ par des triangles K_h de diamètre inférieur à h.

Théorème 23 *L'espace V_h construit à partir de l'élément fini de Lagrange $\mathbb{P}_1(K)$ est un sous-espace de $C^0(\bar{(}\Omega))$ et de $H^1(\Omega)$.*

Preuve. 1) Continuité. Soient K_1 et K_2 deux triangles adjacents et notons a et b leurs sommets communs. On a alors que pour tout $p_1 \in P_1$ et pour tout $p_2 \in P_2$, $p_1(a) = p_2(a)$ et $p_1(b) = p_2(b)$. En supposant, sans restriction de généralité, que le coté $[a, b]$ est parallèle à l'axe (Ox), $p_1 - p_2$ est un polynôme de degré 1 en x qui s'annule en deux points distincts. C'est donc le polynôme nul. Donc $p_1 = p_2$ sur le coté $[a, b]$, d'où la continuité.

2) Soit $u \in V_h$. Alors $u_{|K_i} \in \mathbb{P}_1(K) \subset H^1(K)$. De plus nous venons de voir que u est continu, donc d'après le théorème 4 de recollement $u \in H^1(\Omega)$. ■

On peut donc définir

$$V_h = \{v \in C^0(\bar{(}\Omega)) \mid v_{|K_h} \in \mathbb{P}_1(K_h), \quad \forall K_h \in \mathcal{T}_h\}.$$

On a La dimension de V_h est le nombre de points de la triangulation. On désigne par a_i les sommets de la triangulation et on construit les fonctions φ_i telles que $\varphi_i(a_j) = \delta_{ij}$ et $\varphi_{i|K_h} \in \mathbb{P}_1(K_h)$. Les fonctions $(\varphi_i)_{i=1,\dots,N}$ forment alors une base de V_h.

Remarque 22 *Dans le cas où u s'annule sur tout ou une partie du bord, cette condition vient se rajouter dans l'espace V_h. Il en résulte que les fonctions de base qu'on avait précédemment et qui valent 1 en un point du bord où u s'annule ne font plus partie de V_h. Il en résulte que la dimension de V_h correspond dans ce cas au nombres de sommets de la triangulation moins ceux qui se trouvent sur une partie du bord où u s'annule.*

Exemple 3 : On va construire l'espace V_h basé sur l'élément fini vectoriel construit dans l'exemple 6 de la section 4.3. La triangulation est la même que dans l'exemple précédent. L'égalité des degrés de liberté sur les triangles adjacents entraîne, comme $p \cdot n$ est constant sur les faces que $p \cdot n$ est identique sur la face commune pour les polynômes de deux triangles adjacents, ce qui entraîne, comme nous allons le démontrer dans la proposition suivante que $V_h \subset H(div, \Omega)$. Par contre les fonctions de V_h ne sont pas continues à la traversée des éléments, seule leur composante normale l'est. On a donc

$$V_h = \{v \in H(div, \Omega) \mid v_{|K_h} \in P(K_h), \quad \forall K_h \in \mathcal{T}_h\}.$$

Proposition 10 *Soit v_h une fonction définie sur Ω telle que $v_{h|K} \in P$ et $v_h \cdot n$ est continue à l'interface entre deux triangles adjacents. Alors $v_h \in H(div, \Omega)$.*

Preuve. 1) On a $v_h \in (L^2(\Omega))^2$ car $v_h \in (L^2(K))^2$ en tant que polynôme sur un compact K et

$$\int_\Omega v_h^2 \, dx = \sum_{K \in \mathcal{T}_h} \int_K v_h^2 \, dx < \infty$$

comme somme finie de termes bornés.
2) $\operatorname{div} v_h \in L^2(\Omega)$: Comme $v_{h|K}$ est un polynôme, on a $\operatorname{div} v_h \in L^2(K)$. Par contre v_h n'est

a priori pas dérivable aux interfaces entre les éléments, on va donc calculer $\operatorname{div} v_h$ au sens des distributions. Soit $\varphi \in \mathcal{D}(\Omega)$. Alors

$$\begin{aligned}\langle \operatorname{div} v_h, \varphi\rangle &= -\int_\Omega v_h \cdot \nabla\varphi \, dx \\ &= -\sum_{K\in\mathcal{T}_h} \int_K v_h \cdot \nabla\varphi \, dx \\ &= -\sum_{K\in\mathcal{T}_h} \Big(\int_{\partial K} v_h \cdot n \, \varphi \, ds - \int_K \operatorname{div} v_h \, \varphi \, dx\Big).\end{aligned}$$

On traite les termes sur les bords des éléments K de distinguant deux cas.
Premier cas : Les termes qui sont sur $\partial\Omega$ disparaissent car φ est à support compact.
Deuxième cas : Les termes qui ne sont pas sur $\partial\Omega$ appartiennent exactement à deux triangles K_1 et K_2 et donc en sommant les termes sur l'interface provenant des deux triangles on obtient

$$\int_{\partial K_1\cap\partial K_2} v_h \cdot n_1 \, \varphi \, ds + \int_{\partial K_1\cap\partial K_2} v_h \cdot n_2 \, \varphi \, ds = \int_{\partial K_1\cap\partial K_2} \varphi(v_h \cdot n_1 - v_h \cdot n_1) \, ds = 0,$$

car les normales sortantes de K_1 et K_2 sur cette face vérifient $n_1 = -n_2$ et $v_h \cdot n$ est identique des deux côtés par hypothèse.

On obtient donc

$$\langle \operatorname{div} v_h, \varphi\rangle = \sum_{K\in\mathcal{T}_h} \int_K \operatorname{div} v_h \, \varphi \, dx = \int_\Omega \operatorname{div} v_h \, \varphi \, dx,$$

ce qui signifie que la divergence au sens des distributions de v_h est égale presque partout à $\operatorname{div} v_h$ calculé dans chaque élément. Il en résulte en particulier que $\operatorname{div} v_h \in L^2(\Omega)$ avec

$$\int_\Omega |\operatorname{div} v_h|^2 \, dx = \sum_{K\in\mathcal{T}_h} \int_K |\operatorname{div} v_h|^2 \, dx.$$

■

4.6 Mise en oeuvre de la méthode des élément finis

Le principe de la méthode des éléments finis est d'approximer un problème défini par une formulation variationnelle
Trouver $u \in V$ tel que

$$a(u, v) = l(v) \quad \forall v \in V, \tag{4.9}$$

où a et l vérifient les hypothèses du théorème de Lax-Milgram avec a symétrique, par un problème posé dans un espace de dimension fini V_h construit à partir d'une famille d'éléments finis suivant le principe énoncé dans la section précédente et qui devient "très proche" de V pour h petit :
Trouver $u_h \in V_h$ tel que

$$a(u_h, v_h) = l(v_h) \quad \forall v_h \in V_h. \tag{4.10}$$

En notant $(\varphi_i)_{i=1,\dots,N}$ une base de V_h, on peut décomposer un élément u_h de V_h sur cette base en écrivant

$$u_h(x) = \sum_{j=1}^{N} u_j \varphi_j(x).$$

La formulation variationnelle discrète (4.10) peut alors s'écrire

$$\sum_{j=1}^{N} u_j a(\varphi_j, \varphi_i) = l(\varphi_i) \quad \text{pour } i = 1, \dots N.$$

Alors, en notant

$$A = ((a(\varphi_i, \varphi_j)))_{1 \le i,j \le N}, \quad U = \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix}, \quad b = \begin{pmatrix} l(\varphi_1) \\ \vdots \\ l(\varphi_N) \end{pmatrix},$$

on se ramène à la résolution du système linéaire

$$AU = b,$$

qui admet une solution unique à cause de la coercivité de a.

Il s'agit maintenant d'assembler la matrice A et le second membre b. Pour le second membre on utilise en général une formule de quadrature numérique du même ordre que l'erreur commise dans la méthode d'éléments finis. Par exemple pour une méthode d'éléments finis $\mathbb{P}_1$ en dimension 2, si $l(v) = \int_\Omega fv\, dx\, dy$, on peut utiliser la formule de quadrature

$$\int_K f\varphi_i \, dx \approx \frac{|K|}{3} f(a_i),$$

où a_i est le sommet de K où φ_i vaut 1. Une autre option est d'approcher f par sa projection πf sur l'espace d'éléments finis. On a

$$\pi f = \sum_{j=1}^{N} \sigma_j(f)\varphi_j,$$

et donc

$$\int_\Omega \pi f \varphi_i \, dx = \sum_{j=1}^{N} \sigma_j(f) \int_\Omega \varphi_i \varphi_j \, dx,$$

ce qui peut s'écrire sous forme matricielle $b_h = Ms$. où

$$b_h = (\int_\Omega \pi f \varphi_i \, dx)_i, \quad M = ((\int_\Omega \varphi_i \varphi_j \, dx))_{i,j}, \quad s = (\sigma_j(f))_j.$$

Il faut alors également assembler la matrice de masse M.

Comme l'espace V_h et en particulier ses fonctions de base φ_i sont construites à partir des valeurs par éléments, il est en général beaucoup plus pratique d'assembler les matrices à partir des matrices élémentaires qui contiennent les contributions d'un élément à la matrice globale en décomposant les intégrales sur les éléments

$$\int_\Omega f(x) \, dx = \sum_K \int_K f(x), dx.$$

4.6.1 Assemblage des matrices élémentaires

Les matrices élémentaires sont des matrices pleines ayant comme nombre de lignes et de colonnes la dimension de P. Il est souvent pratique pour l'assemblage des matrices élémentaires de faire un changement de variable permettant de ce ramener à un élément de référence ayant des sommets avec des coordonnées plus simples. Ainsi pour des éléments finis construits sur des triangles, on considère un élément quelconque K de sommets $a_i, i = 1, 2, 3$ ayant pour coordonnées (x_i, y_i) et on prend pour élément de référence $\hat{K}$ ayant pour sommets $\hat{a}_1, \hat{a}_2, \hat{a}_3$ de coordonnées respectives $(0, 0)$, $(1, 0)$ et $(0, 1)$, L'application affine

$$F(x, y) = A \begin{pmatrix} x \\ y \end{pmatrix} + b,$$

avec

$$A = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix}, \quad b = \begin{pmatrix} x_1 \\ y1 \end{pmatrix},$$

transforme $\hat{K}$ en K. On a donc en particulier, en faisant le changement de variable $(\hat{x}, \hat{y}) = F(x, y)$

$$\int_K \varphi_i(x, y)\varphi_j(x, y)\, dx\, dy = \int_{\hat{K}} \varphi_i(F(\hat{x}, \hat{y}))\varphi_j(F(\hat{x}, \hat{y}))(\det A)\, d\hat{x}\, d\hat{y}$$
$$= (\det A) \int_{\hat{K}} \hat{\varphi}_i(\hat{x}, \hat{y})\hat{\varphi}_j(\hat{x}, \hat{y})\, d\hat{x}\, d\hat{y}.$$

De même en utilisant la loi de composition des dérivées on obtient que

$$\hat{\nabla}\hat{\varphi}_i(\hat{x}, \hat{y}) =^t A \nabla\varphi(x, y),$$

et donc

$$\int_K \nabla\varphi_i \, \nabla\varphi_j\, dx\, dy = (\det A) \int_{\hat{K}} ({}^tA)^{-1}\hat{\nabla}\hat{\varphi}_i(\hat{x}, \hat{y}) \cdot ({}^tA)^{-1}\hat{\varphi}_j(\hat{x}, \hat{y})\, d\hat{x}\, d\hat{y}.$$

La matrice $({}^tA)^{-1}$ se calcule à partir de A et à pour expression

$$({}^tA)^{-1} = \frac{1}{\det A} \begin{pmatrix} y_3 - y_1 & y_1 - y_2 \\ x_1 - x_3 & x_2 - x_1 \end{pmatrix}.$$

Les calculs effectifs des fonctions de base sur l'élément de référence peuvent alors se faire à l'aide des coordonnées barycentriques associées à ce triangle et qui ont pour expressions

$$\lambda_1(x, y) = 1 - x - y, \quad \lambda_2(x, y) = x, \quad \lambda_3(x, y) = y.$$

Exemple 1 : Elément fini $\mathbb{P}_1$ Les trois fonctions de base de l'élément de référence sont définies par $\hat{p}_\alpha = \lambda_\alpha, \alpha = 1, 2, 3$. On a alors p.ex.

$$\int_K p_1(x, y)p_2\, dx\, dy = (\det A) \int_{\hat{K}} (1 - x - y)x\, dx\, dy,$$

avec $\det A = (x_2 - x_1)(y_3 - y_1) - (y_2 - y_1)(x_3 - x_1)$.

D'autre part p.ex.

$$\partial_x p_1 = ({}^tA)^{-1}\partial_x\lambda_1 = -\frac{y_3 - y_2}{\det A}.$$

Exemple 2 : Elément fini $\mathbb{P}_2$ Les six fonctions de base de l'élément de référence sont données par

$$p_1 = \lambda_1(2\lambda_1 - 1), \quad p_2 = \lambda_2(2\lambda_2 - 1), \quad p_3 = \lambda_3(2\lambda_3 - 1),$$
$$p_{12} = \frac{1}{4}\lambda_1\lambda_2, \quad p_{13} = \frac{1}{4}\lambda_1\lambda_3, \quad p_{23} = \frac{1}{4}\lambda_2\lambda_3.$$

On en déduit comme dans l'exemple 1 la contribution à la matrice de masse. Et on a par exemple

$$\partial_x p_1 = ({}^tA)^{-1}\partial_x(\lambda_1(2\lambda_2 - 1)) = (y_3 - y_2)(4(x + y) - 3)/(\det A).$$

4.6.2 Assemblage des matrices globales

Pour assembler les matrices globales à partir des matrices élémentaires, on introduit la fonction ϕ qui à un couple (K, α) correspondant à un élément et à une fonction de base $p\alpha$ ($1 \leq \alpha \leq dim(P)$) associe le numéro i ($1 \leq i \leq dim(V_h)$) de la fonction de base φ_i de V_h telle que $(\varphi_i)_{|K} = p_\alpha$. On a alors que la matrice A est donnée par l'algorithme :
A=0
Pour $K = 1, nbelements$,
Pour $\alpha, \beta = 1, dim(P)$

$$M(\phi(K, \alpha), \phi(K, \beta)) = M(\phi(K, \alpha), \phi(K, \beta)) + M_K(\alpha, \beta),$$

où M_K est la matrice élémentaire associée à l'élément K.

4.7 Estimations d'erreur

4.7.1 Erreur d'interpolation locale

Corollaire 2 *Soit Ω un ouvert borné à bord lipschitzien. Soit $V = H^m(\Omega)/\mathbb{P}_{m-1}(\Omega)$. On note $\dot{u}$ l'élément de V représentant de la classe d'équivalence de $u + p$ où $p \in \mathbb{P}_{m-1}(\Omega)$.L'espace V est muni d'une norme définie par*

$$\|\dot{u}\|_m = \inf_{p \in \mathbb{P}_{m-1}(\Omega)} \|u + p\|_m,$$

et d'une semi-norme

$$|\dot{u}|_m = |u|_m.$$

Alors la norme $\|.\|_m$ et la semi-norme $|.|_m$ sont équivalentes sur V.

Preuve. Par définitions des normes ont a $|\dot{u}|_m \leq \|\dot{u}\|_m$. L'autre inégalité découle du théorème 13 en remarquant que par définition de l'espace quotient

$$V \cap \mathbb{P}_{m-1}(\Omega) = \{\dot{0}\}.$$

■

Théorème 24 *Soit (K, P, Σ) un élément fini, $k, m \in \mathbb{N}^*$ vérifiant $m \leq k + 1$. Soit π l'opérateur d'interpolation associé à (K, P, Σ). On suppose que $\mathbb{P}_k(K) \subset P \subset H^m(K)$. Alors il existe une constante $C > 0$ dépendante seulement de k et de K telle que*

$$\|v - \pi v\|_{m,K} \leq C|v|_{k+1,K}.$$

Preuve. 1) L'opérateur $I - \pi$ est linéaire continu pour des éléments finis de Lagrange. L'opérateur d'interpolation π est défini par

$$\pi v = \sum_{i=1}^{N} v(a_i) p_i.$$

Alors

$$\begin{aligned} \|\pi v\|_m &\leq \sum_{i=1}^{N} |v(a_i)| \|p_i\|_m, \\ &\leq (\sum_{i=1}^{N} \|p_i\|_m) \max_{x \in K} |v(x)|, \\ &\leq C \|v\|_{k+1,K}, \end{aligned}$$

car $\sum_{i=1}^{N} \|p_i\|_m$ est constante et grâce au théorème 12 l'injection de $H^{k+1}(\hat{K})$ dans $C^0(\hat{K})$ est continue, i.e.

$$\max_{x \in K} |v(x)| \leq C \|v\|_{k+1,K},$$

car $k + 1 > \frac{n}{2}$.

2) Pour $p \in \mathbb{P}_k$, on a $\pi p = p$, donc

$$v - \pi v = v - p + \pi p - \pi v = (I - \pi)(v - p),$$

donc

$$\begin{aligned} \|v - \pi v\|_m &\leq \|I - \pi\|_{\mathcal{L}(H^{k+1}, H^m)} \inf_p \|v - p\|_{k+1}, \\ &\leq C |v|_{k+1,K}. \end{aligned}$$

■

On considère maintenant deux éléments finis (K, P, Σ) et $(\hat{K}, \hat{P}, \hat{\Sigma})$ affinement équivalents pour une application affine $F : \hat{x} \mapsto A\hat{x} + b$.

Lemme 5 *Pour $m \geq 0$, l'application $v \mapsto \hat{v} = v \circ F$ est un isomorphisme de $H^m(K)$ dans $H^m(\hat{K})$. De plus il existe deux constantes c_1 et c_2 ne dépendant que de m et de la dimension n telles que*

$$\begin{aligned} |\hat{v}|_{m,\hat{K}} &\leq c_1 \|A\|^m |\det A|^{-\frac{1}{2}} |v|_{m,K} \\ |v|_{m,K} &\leq c_2 \|A^{-1}\|^m |\det A|^{\frac{1}{2}} |\hat{v}|_{m,\hat{K}}, \end{aligned}$$

pour tout $v \in H^m(K)$.

Preuve. Notons $A = ((a_{ij}))$. On a

$$\frac{\partial \hat{v}}{\partial \hat{x}_i} = \sum_{j=1}^{N} a_{ji} \frac{\partial v}{\partial x_j} \circ F,$$

donc

$$\frac{\partial^m \hat{v}}{\partial \hat{x}_{i1} \dots \partial \hat{x}_{im}} = \sum_{j_1=1}^{N} \dots \sum_{j_m=1}^{N} a_{j_1 i_1} \dots a_{j_m i_m} \frac{\partial^m v}{\partial x_{j1} \dots \partial x_{jm}} \circ F,$$

donc

$$\Big|\frac{\partial^m \hat{v}}{\partial \hat{x}_{i1}\dots\partial \hat{x}_{im}}\Big| \le \|A\|^m \sum_{j_1=1}^{N}\dots\sum_{j_m=1}^{N}\Big|\frac{\partial^m v}{\partial x_{j1}\dots\partial x_{jm}}\circ F\Big|.$$

Ainsi

$$\int_{\hat{K}}\Big|\frac{\partial^m \hat{v}}{\partial \hat{x}_{i1}\dots\partial \hat{x}_{im}}\Big|^2 d\hat{x} \le c_1\|A\|^{2m}\sum_{j_1=1}^{N}\dots\sum_{j_m=1}^{N}\int_{\hat{K}}\Big|\frac{\partial^m v}{\partial x_{j1}\dots\partial x_{jm}}\circ F\Big|^2 d\hat{x},$$

puis en faisant le changement de variable $x = F(\hat{x})$, $dx = |\det A| d\hat{x}$ dans le membre de droite, il vient

$$\int_{\hat{K}}\Big|\frac{\partial^m \hat{v}}{\partial \hat{x}_{i1}\dots\partial \hat{x}_{im}}\Big|^2 d\hat{x} \le c_1\|A\|^{2m}\sum_{j_1=1}^{N}\dots\sum_{j_m=1}^{N}\int_{K}\Big|\frac{\partial^m v}{\partial x_{j1}\dots\partial x_{jm}}\Big|^2 |\det A|^{-1} dx.$$

Ce qui donne la première inégalité. La démonstration de la deuxième inégalité est identique. ■

Théorème 25 *Soit $(\hat{K},\hat{P},\hat{\Sigma})$ un élément fini, $k,m\in\mathbb{N}^*$ tel que $m\le k+1$. On suppose que $\hat{\pi}\in\mathcal{L}(H^{k+1}(\hat{K}),\hat{P})$ et que $\mathbb{P}_k(\hat{K})\subset\hat{P}\subset H^m(\hat{K})$. Soit ensuite (K,P,Σ) un élément fini affinement équivalent à $(\hat{K},\hat{P},\hat{\Sigma})$ par l'application $F:\hat{x}\mapsto A\hat{x}+b$. Alors, il existe une constante $\hat{c}>0$ telle que*

$$|v-\pi v|_{m,K} \le \hat{c}\|A\|^{k+1}\|A^{-1}\|^m |v|_{k+1,K}.$$

Preuve. On utilise le lemme précédent :

$$|v-\pi v|_{m,K} \le c_2\|A^{-1}\|^m |\det A|^{\frac{1}{2}} |\hat{v}-\hat{\pi}\hat{v}|_{m,\hat{K}},$$

car $\widehat{\pi v}=\hat{\pi}\hat{v}$, puis d'après le théorème 2

$$|\hat{v}-\hat{\pi}\hat{v}|_{m,\hat{K}} \le C|\hat{v}|_{k+1,\hat{K}},$$

puis

$$|\hat{v}|_{k+1,\hat{K}} \le c_1\|A\|^{k+1}|\det A|^{-\frac{1}{2}}|v|_{k+1,K},$$

d'où

$$|v-\pi v|_{m,K} \le c_2 C c_1\|A^{-1}\|^m\|A\|^{k+1}||v|_{k+1,K}.$$

■

Lemme 6 *En notant $h_K, h_{\hat{K}}$ les diamètres de K et $\hat{K}$ et $\rho_K, \rho_{\hat{K}}$ les diamètres des plus grandes sphères contenues dans K et $\hat{K}$, on a*

$$\|A\| \le \frac{h_K}{\rho_{\hat{K}}}, \qquad \|A^{-1}\| \le \frac{h_{\hat{K}}}{\rho_K}.$$

Preuve. Par définition

$$\|A\| = \sup_{\|\xi\|=1}\|A\xi\| = \frac{1}{\rho_{\hat{K}}}\sup_{\|\xi\|=\rho_{\hat{K}}}\|A\xi\|.$$

Or pour ξ tel que $\|\xi\|=\rho_{\hat{K}}$, il existe $\hat{x}_1$, $\hat{x}_2$ dans $\hat{K}$ tels que $\xi=\hat{x}_1-\hat{x}_2$. Alors $A\xi = F(\hat{x}_1)-F(\hat{x}_2)$ et $F(\hat{x}_1)$ et $F(\hat{x}_2)$ sont dans K, donc $\|F(\hat{x}_1)-F(\hat{x}_2)\|\le h_{\hat{K}}$, d'où $\|A\|\le\frac{h_K}{\rho_{\hat{K}}}$. L'autre inégalité se démontre de la même manière. ■

Théorème 26 *Sous les hypothèses du théorème précédent, il existe $\hat{c}$ tel que*

$$|v - \pi v|_{m,K} \leq \hat{c} \frac{h_K^{k+1}}{\rho_K^m} |v|_{k+1,K} \quad \forall v \in H^{k+1}(K). \tag{4.11}$$

Preuve. C'est une conséquence directe des théorèmes et des lemmes précédents. ■

Notons $\sigma_K = \frac{h_K}{\rho_K}$, le paramètre caractérisant la forme du triangle. Plus σ_K est grand, plus le triangle est aplati et σ_K est minimal pour un triangle équilatéral. Dans les démonstrations de convergence, on exige d'une triangulation $\mathcal{T}_h$ que ses triangles ne soient pas trop aplatis, i.e. qu'il existe σ tel que $\sigma_K \leq \sigma \quad \forall K \in \mathcal{T}_h$. La relation (4.11) devient alors

$$|v - \pi v|_{m,K} \leq \hat{c} \sigma^m h_K^{k+1-m} |v|_{k+1,K} \quad \forall v \in H^{k+1}(K).$$

4.7.2 Erreur d'interpolation globale

On va pouvoir passer de l'erreur d'interpolation locale à l'erreur d'interpolation globale grâce à la relation

$$\int_\Omega |\nabla(v - \pi v)|^2 \, dx = \sum_{K \in \mathcal{T}_h} \int_K |\nabla(v - \pi v)|^2 \, dx,$$

qui peut s'écrire aussi $|v - \pi v|_{1,\Omega}^2 = \sum_{K \in \mathcal{T}_h} |v - \pi v|_{1,K}^2$. Cette relation est vérifié à condition que $v, \pi v \in H^1(\Omega)$.

Notons qu'il se peut très bien que $v \in H^1(K)$ pour chaque K de $\mathcal{T}_h$ sans que $v \in H^1(\Omega)$, car il peut y avoir des défauts de régularité apparaissant aux interfaces entre les éléments.

Définition 21 *On dit qu'une famille d'éléments finis $(K, P, \Sigma) \in \mathcal{T}_h$ est de classe C^0 si*

(i) *$P \subset C^0(K)$ pour tout $K \in \mathcal{T}_h$.*

(ii) *Si K_1 et K_2 sont deux éléments adjacents et v_h appartient à l'espace d'approximation construit sur la famille d'éléments finis alors $v_{h|K_1}$ et $v_{h|K_2}$ coïncident sur $K_1 \cap K_2$.*

Pour les estimations d'erreurs globales, on fait les hypothèses suivantes sur la triangulation $\mathcal{T}_h$:

(H1) On suppose que la famille de triangulations est régulière dans les sens suivant :

(i) Il existe une constante σ telle que

$$\forall K \in \cup_h \mathcal{T}_h \quad \frac{h_K}{\rho_K} \leq \sigma.$$

(ii) La quantité $h = \max_{K \in \mathcal{T}_h} h_K$ tend vers 0.

(H2) Tous les éléments finis (K, P, Σ), $K \in \cup_h \mathcal{T}_h$ sont affinement équivalents à un unique élément de référence $(\hat{K}, \hat{P}, \hat{\Sigma})$.

(H3) Tous les éléments finis (K, P, Σ) , $K \in \cup_h \mathcal{T}_h$ sont de classe C^0.

Théorème 27 *On suppose que les hypothèses (H1), (H2) et (H3) sont vérifiées. En outre on suppose qu'il existe un entier $k \geq 1$ tel que*

$$\mathbb{P}_k \subset \hat{P} \subset H^1(\hat{K}),$$

$$K^{k+1}(\hat{K}) \subset C^0(\hat{K}) \qquad (\textit{vrai si } k+1 > \frac{n}{2}).$$

Alors il existe une constante C indépendante de h telle que, pour toute fonction $v \in H^{k+1}(\Omega)$ on a

$$\|v - \pi_h v\|_{k+1} \leq C h^k |v|_{k+1,\Omega}.$$

Preuve. Sous les hypothèses données on a vu que que pour $m = 0$ et $m = 1$ on a

$$\begin{aligned} |v - \pi_h v|_{m,K} &\leq \hat{c}\sigma^m h_K^{k+1-m} |v|_{k+1,K}, \\ &\leq C h^{k+1-m} |v|_{k+1,K}. \end{aligned}$$

Donc en particulier pour $m = 1$

$$|v - \pi_h v|_{1,K} \leq C h^k |v|_{k+1,K}.$$

En prenant le carré et en sommant sur $K\mathcal{T}_h$ on obtient

$$\|v - \pi_h v\|_{1,\Omega}^2 = \sum_{K \in \mathcal{T}_h} \|v - \pi_h v\|_{1,K}^2 \leq C^2 h^{2k} |v|_{k+1,K}^2,$$

car $v \in H^1(\Omega)$ par hypothèse et $\pi v \in H^1(\Omega)$ parce que $\pi v \in H^1(K)$ et $\pi v \in C^0(\Omega)$. ∎

4.7.3 Estimation d'erreur pour les éléments finis

On considère un problème variationnel posé dans $V \subset H^1(\Omega)$.

Théorème 28 *On suppose (H1), (H2) et (H3) vérifiées. En outre on suppose qu'il existe un entier $k \geq 1$ tel que $k+1 > \frac{n}{2}$ avec $\mathbb{P}_k(\hat{K}) \subset P \subset H^1(\hat{K})$ et que la solution exacte du problème variationnel est dans $H^{k+1}(\Omega)$, alors*

$$\|u - u_h\|_{1,\Omega} \leq C h^k |u|_{k+1,\Omega},$$

où $u_h \in V_h$ est la solution discrète.

Preuve. On a d'après le théorème précédent que

$$\|u - \pi_h u\|_{1,\Omega} \leq C h^k |u|_{k+1,\Omega}.$$

D'autre part d'après le lemme de Céa

$$\|u - u_h\|_{1,\Omega} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega} \leq C \|u - \pi_h u\|_{1,\Omega},$$

d'où le résultat. ∎

On vient de montrer une estimation en norme H^1. Nous allons voir maintenant qu'on peut avoir une estimation meilleure dans la norme L^2, grâce à une technique de dualité introduite par Aubin et Nitsche.

On considère le problème variationnel associé à

$$\begin{aligned} -\Delta u &= f \text{ dans } \Omega, \\ u &= 0 \text{ sur } \partial\Omega, \end{aligned}$$

qui s'écrit

Trouver $u \in H_0^1(\Omega)$ telle que

$$(a(u,v) =) \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx (= (f,v)).$$

Théorème 29 *On suppose que (H1), (H2) et (H3) sont vérifiées et qu'il existe un entier $k \geq 1$ tel que $k+1 > \frac{n}{2}$ avec $\mathbb{P}_k(\hat{K}) \subset P \subset H^1(\hat{K})$ et que la solution exacte du problème variationnel est dans $H^{k+1}(\Omega)$. Alors il existe une constante C indépendante de h telle que*

$$|u - u_h|_{0,\Omega} \leq C h^{k+1} |u|_{k+1,\Omega}.$$

Preuve. On va utiliser l'expression suivante de la norme L^2, basée sur la représentation duale :

$$|u|_{0,\Omega} = \sup_{\substack{g \in L^2 \\ g \neq 0}} \frac{\int_\Omega ug \, dx}{|g|_{0,\Omega}}.$$

Notons que cette expression correspond bien à la norme L^2 usuelle. En effet, d'une part par Cauchy-Schwartz on a $\int_\Omega ug \, dx \leq |u|_{0,\Omega}|g|_{0,\Omega}$, donc pour $g \neq 0$ $\frac{\int_\Omega ug \, dx}{|g|_{0,\Omega}} \leq |u|_0$ et d'autre part en prenant $g = u$, on a

$$|u|_0 = \frac{\int u^2 \, dx}{|u|_0} \leq \sup_{\substack{g \in L^2 \\ g \neq 0}} \frac{\int_\Omega ug \, dx}{|g|_{0,\Omega}}.$$

Soit alors $g \in L^2(\Omega)$ quelconque, puis soit φ l'unique solution de

$$a(\varphi, v) = (g, v) \quad \forall v \in V. \tag{4.12}$$

Alors $\varphi \in H^2(\Omega)$ et $\|\varphi\|_{2,\Omega} \leq C|g|_{0,\Omega}$. Et en prenant dans (4.12) $v = u - u_h$ on obtient $a(\varphi, u - u_h) = (g, u - u_h)$, or $\pi_h \varphi \in V_h$, donc d'après la relation d'orthogonalité de Galerkine $a(\pi_h \varphi, u - u_h) = 0$. Il en résulte, en utilisant le théorème précédent que,

$$\begin{aligned} (g, u - u_h) &\leq a(\varphi - \pi_h\varphi, u - u_h), \\ &\leq C\|\varphi - \pi_h\varphi\|\|u - u_h\|, \\ &\leq Ch|\varphi|_{2,\Omega} h^k |u|_{k+1,\Omega}, \\ &\leq Ch^{k+1}|g|_{0,\Omega}|u|_{k+1,\Omega}, \end{aligned}$$

d'où $|u - u_h|_0 \leq Ch^{k+1}|u|_{k+1,\Omega}$. ∎

4.7.4 Estimation d'erreur a posteriori

Les estimations d'erreurs obtenues précédemment donnent une estimation de l'erreur globale en fonction de la solution exacte. Or on sait que cette solution existe et on connaît sa régularité, donc ce type d'erreur permet de prouver la convergence du schéma et donne l'ordre de convergence. Néanmoins elle ne permet pas, lorsque l'on a calculé une solution approchée u_h de connaître localement l'erreur que l'on a commise. Les estimations a posteriori que l'on va dériver dans cette section permettent de déterminer, une fois calculée une solution approchée u_h, une erreur locale en fonction de u_h et des données du problème. Ce qui va permettre de raffiner le maillage aux endroits où c'est nécessaire pour avoir une meilleure approximation de la solution exacte.

Exemple : Dans le cas d'une approximation $\mathbb{P}_1$ en dimension 1, on aura une très bonne approximation de la solution avec très peu de points aux endroits où elle est affine, ou de courbure très faible. Par contre là où la courbure est grande, on aura besoin de plus de points.

Approche pour écrire un code d'éléments finis avec raffinement de maillage : On se donne une tolérance α pour l'erreur locale. On part alors avec un maillage assez grossier et itérativement, on calcule la solution sur le maillage courant, puis on rajoute des points aux endroits où l'erreur locale est supérieure à α jusqu'à ce que l'erreur locale soit partout inférieure à α.

Estimations a posteriori en 1d

Pour calculer les erreurs globales dans la section précédente, nous étions passé par les erreurs locales (par élément) de projection. Nous avions vu que

$$\|(u - \pi_h u)'\|_{L^2(x_i,x_{i+1})} \leq |x_{i+1} - x_i| \|u''\|_{L^2(x_i,x_{i+1})}.$$

L'idée pour avoir une erreur uniformément répartie est d'avoir $|x_{i+1} - x_i| \|u''\|_{L^2(x_i,x_{i+1})}$ à peu près constante. C'est-à-dire que nous allons choisir $|x_{i+1} - x_i|$ petit là où $\|u''\|_{L^2(x_i,x_{i+1})}$ est grand et inversement. On en déduit l'algorithme d'éléments finis adaptatif suivant pour le problème

$$\begin{aligned} -u'' &= f \\ u(0) &= u(1) = 0. \end{aligned}$$

Algorithme :

(i) Initialisation : Partir d'un maillage uniforme (relativement grossier)

$$x_j^0 = \frac{j}{N_0 + 1}.$$

(ii) Étape itérative : Construire le maillage

$$0 = x_0^l < x_1^l < \cdots < x_{N_l}^l < x_{N_l+1}^l,$$

en rajoutant des points là où l'erreur de la solution calculée sur le maillage $(x_i^{l-1})_i$ est trop grande.

Concrètement pour passer du maillage l au maillage $l+1$, on peut utiliser l'indicateur d'erreur suivant :

$$|x_{i+1}^l - x_i^l|^2 \|u''\|^2_{L^2(x_i,x_{i+1})} = |x_{i+1}^l - x_i^l|^2 \|f\|^2_{L^2(x_i,x_{i+1})} \approx |x_{i+1}^l - x_i^l|^3 |f(\frac{x_i^l + x_{i+1}^l}{2})|^2,$$

car $\|f\|^2_{L^2(x_i,x_{i+1})} \approx |x_{i+1}^l - x_i^l||f(\frac{x_i^l+x_{i+1}^l}{2})|^2$. On définit finalement

$$I_i^l = |x_{i+1}^l - x_i^l|^3 |f(\frac{x_i^l + x_{i+1}^l}{2})|^2,$$

qui va servir à estimer l'erreur. Nous pouvons estimer le gain obtenu grâce à l'indicateur d'erreur I_i^l en divisant une maille par deux. Posons $x_{k+1/2}^l = \frac{x_k^l+x_{k+1}^l}{2}$. On a alors

$$|x_{k+1/2}^l - x_k^l|^3 = \frac{1}{8}|x_{k+1}^l - x_k^l|^3,$$

donc

$$|x_{k+1/2}^l - x_k^l|^3 \|f\|^2_{L^2(x_i,x_{i+1})} \leq \frac{1}{8} I_k^l,$$

et de même

$$|x_{k+1}^l - x_{k+1/2}^l|^3 \|f\|^2_{L^2(x_i,x_{i+1})} \leq \frac{1}{8} I_k^l.$$

On divise donc l'indicateur local par 8 en divisant une maille par 2.

Si on veut que l'erreur finale soit inférieure à $\varepsilon > 0$, pour passer du maillage l au maillage $l+1$, on va décider de découper les mailles telles que $I_k^l \geq \frac{\varepsilon^2}{(N_l+1)^2}$ jusqu'à ce qu'on ait pour toutes les mailles $I_k^l < \frac{\varepsilon}{N_l+1}$. Alors, l'erreur globale étant la somme des erreurs locales, elle sera inférieure à ε.

Estimations a posteriori en 2d

On considère le problème de Dirichlet

$$\begin{aligned} -\Delta u &= f \text{ dans } \Omega, \\ u &= 0, \text{ sur } \partial\Omega, \end{aligned}$$

qui admet une unique solution dans $H_0^1(\Omega)$. Nous considérons une approximation de ce problème par éléments finis de Lagrange $\mathbb{P}_1$. On a l'estimation locale de l'erreur de troncature

$$|u - \pi_h u|_{1,K} \leq C \frac{h_K^2}{\rho_K} |u|_{2,K},$$

où h_K est le diamètre de l'élément K et ρ_K le diamètre du cercle inscrit. En supposant qu'on ait un bon maillage, i.e. qu'il existe σ indépendant de l'élément K du maillage tel que $\frac{h_K}{\rho_K} \leq \sigma$, un indicateur de l'erreur locale va être $h_K|u|_{2,K}$. C'est identique au cas 1d, mais ici on ne peut pas directement remplacer $|u|_{2,K}$ par $|f|_{0,K}$, il va donc falloir l'estimer en fonction de u_h. Mais comme $u_h \in \mathbb{P}_1$, $D^2 u_h$ est une distribution de $H^{-1}(\Omega)$ qui s'annule à l'intérieur des triangles. On ne peut donc pas l'utiliser directement. On va construire une approximation numérique en utilisant les triangles voisins. On a d'abord dans K

$$u_h = \sum_{i=1}^{3} u_h(a_i)\varphi_i^K,$$