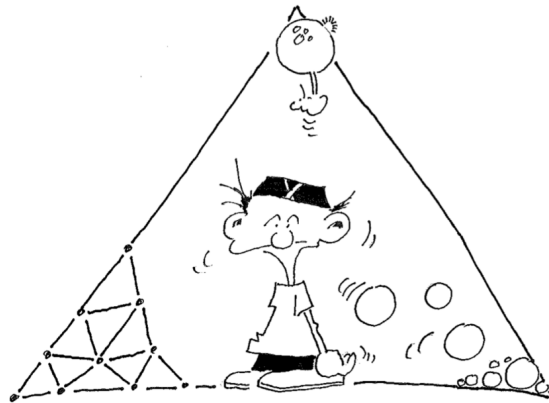




Ecole Polytechnique de Louvain



**INTRODUCTION
AUX
ELEMENTS FINIS**
*...et au génie logiciel numérique
en langage C*

V. Legat

Notes provisoires pour le cours LEPL1110
Année académique 2020-21 (version 2.0 1-2-2021)

*Ce document est une oeuvre originale protégée par le droit d'auteur.
Copyright V. Legat, mars 2021*

Ce texte est toujours une version provisoire. Malgré tout le soin apporté à sa rédaction, il est possible que quelques erreurs soient toujours présentes dans le texte. Tout commentaire, critique ou suggestion de votre part, est évidemment le bienvenu. Il vous est possible de m'envoyer vos commentaires directement par courrier électronique à l'adresse suivante : vincent.legat@uclouvain.be

Les éventuels errata du texte seront disponibles sur le site Web du cours. J'adresse toute ma reconnaissance à Joachim Van Verdegheem qui a éliminé un joli paquet de fautes d'orthographe entre la version de janvier 2013 et celle de janvier 2014.

Table des matières

1	Interpolation sur un maillage non structuré	1
1.1	Eléments finis unidimensionnels	2
1.2	Eléments finis bidimensionnels	7
1.2.1	Eléments triangulaires	9
1.2.2	Eléments quadrilatères	12
2	Eléments finis pour des problèmes elliptiques	19
2.1	Formulations forte, faible et discrète	19
2.2	Exemple unidimensionnel	29
2.2.1	Formulation faible	30
2.2.2	Formulation discrète	32
2.2.3	Construction du système algébrique	33
2.3	Eléments finis bidimensionnels	42
2.3.1	Construction du système algébrique	42
2.3.2	Triangles de Turner pour l'équation de Poisson	46
2.3.3	Résolution du système linéaire par élimination de Gauss	53
2.3.4	Traitement de conditions essentielles inhomogènes	57
3	Techniques de résolution des systèmes linéaires creux	61
3.1	Méthodes directes	62

3.1.1	Factorisation de matrices creuses	64
3.1.2	Solveur bande	67
3.1.3	Solveur frontal	67
3.2	Méthodes itératives	71
3.2.1	Méthode de la plus grande pente	72
3.2.2	Méthode des gradients conjugués	75
3.2.3	Préconditionnement	79
4	Théorie de la meilleure approximation	83
4.1	Espaces d'Hilbert	84
4.2	Espaces de Sobolev	86
4.3	Estimations d'erreur a priori	90
4.3.1	Erreur de l'interpolation : lemme de Bramble-Hilbert	91
4.3.2	Théorème de la meilleure approximation	93
4.3.3	Lemme de Cea	95
4.4	Estimation d'erreur a posteriori	97
4.4.1	Stratégies adaptatives	99
5	Méthodes d'éléments finis pour des problèmes d'advection-diffusion	103
5.1	Equation scalaire de transport	104
5.1.1	A propos de la méthode des caractéristiques	105
5.1.2	Méthode de Galerkin	106
5.1.3	Cas unidimensionnel	108
5.2	Equation scalaire d'advection-diffusion	113
5.2.1	Méthodes de Petrov-Galerkin	115
5.2.2	Cas unidimensionnel	116

Fun, frustrations and tricks of the trade
(Titre du premier chapitre du livre de B.M. Irons)

Chapitre 1

Interpolation sur un maillage non structuré

Supposons que nous disposions d'une fonction connue $u \in \mathcal{U}$. La méthode des éléments finis consiste à écrire cette interpolation sous la forme suivante

$$u^h(\mathbf{x}) = \sum_{j=1}^n U_j \tau_j(\mathbf{x}) \quad (1.1)$$

où U_j sont de *valeurs nodales* inconnues, tandis que τ_j sont des *fonctions de forme* spécifiées a priori et appartenant à l'espace \mathcal{U} . Les fonctions de forme sont choisies afin qu'aucune d'entre elles ne puisse être obtenue par combinaison linéaire des autres.

Il y a donc n degrés de liberté pour définir un élément particulier du sous-espace discret $\mathcal{U}^h \subset \mathcal{U}$. La dimension \mathcal{U}^h est clairement n et les fonctions de forme sont une base de cet espace dont tous les éléments sont obtenus par une combinaison linéaire unique des éléments de cette base. La plupart des fonctions de forme utilisées sont associées à un point particulier de l'espace \mathbf{X}_i et satisfont la propriété suivante :

$$\tau_j(\mathbf{X}_i) = \delta_{ij} \quad (1.2)$$

où δ_{ij} est le symbole de Kronecker ($\delta_{ij} = 1$ si $i = j$ et 0 autrement).

On peut directement en déduire que les valeurs nodales sont les valeurs de u^h en \mathbf{X}_i .

$$\begin{aligned}
 u^h(\mathbf{X}_i) &= \sum_{j=1}^n U_j \tau_j(\mathbf{X}_i), \\
 &\downarrow \\
 u^h(\mathbf{X}_i) &= \sum_{j=1}^n U_j \delta_{ij} \\
 &\downarrow \\
 u^h(\mathbf{X}_i) &= U_i.
 \end{aligned}$$

1.1 Eléments finis unidimensionnels

Considérons, par exemple, $u(x) = 2(1 - x^3)x$. Nous désirons obtenir une interpolation $u^h(x)$ au moyen de polynômes par morceaux. Sur la figure 1.1, nous considérons un cas très simple. Le domaine Ω est divisé en deux segments, qui sont appelés *éléments finis*. Leur longueur est respectivement 0.6 et 0.4. Sur chaque élément, nous désirons remplacer la fonction u par une fonction polynomiale.

Le cas le plus simple est celui où l'approximation est un polynôme d'ordre zéro, c'est-à-dire une constante sur chaque élément. On a ici choisi la valeur de u au centre de l'élément comme constante. L'approximation u^h est continue par morceaux.

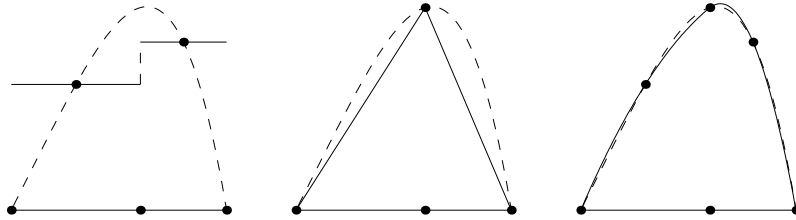


Figure 1.1: Interpolation constante, linéaire et quadratique par morceaux de la fonction $u(x) = 2(1 - x^3)x$ sur deux éléments de longueur 0.6 et 0.4 respectivement.

Nous pouvons ensuite introduire une interpolation linéaire par morceaux au moyen de polynômes du premier degré. L'approximation est maintenant continue. Dans la troisième partie de la figure, on introduit un polynôme du second degré dans chaque élément. Les coefficients du polynôme sont déterminés par l'imposition de la valeur de u aux extrémités et au milieu de l'élément.

Après avoir présenté le problème au moyen d'un exemple simple, définissons en termes plus généraux le processus d'approximation de u sur le domaine Ω en un nombre donné d'*éléments* $\Omega_e =]X_e, X_{e+1}[$, segments ouverts successifs qui ne se recouvrent pas. On définit ainsi une partition de Ω appelée *maillage*. Les points X_e aux extrémités des éléments sont appelés les *sommets* du maillage. De manière plus formelle, un maillage d'un domaine Ω doit être tel que :

$$\overline{\Omega} = \bigcup_{e=1}^{N_1} \{\overline{\Omega}_e\}, \quad \Omega_e \cap \Omega_f = \emptyset, \quad \text{si } e \neq f.$$

La taille d'un tel maillage est caractérisée par l'un des deux nombres :

- N_0 le nombre de sommets,
- $N_1 = N_0 - 1$ le nombre d'éléments.

Elément parent

Pour construire une interpolation polynômiale sur ces éléments qui peuvent être de longueurs diverses, il est opportun d'établir un isomorphisme entre chaque point x de chaque élément Ω_e et un point ξ de l'intervalle ouvert $\widehat{\Omega} =]-1, +1[$,

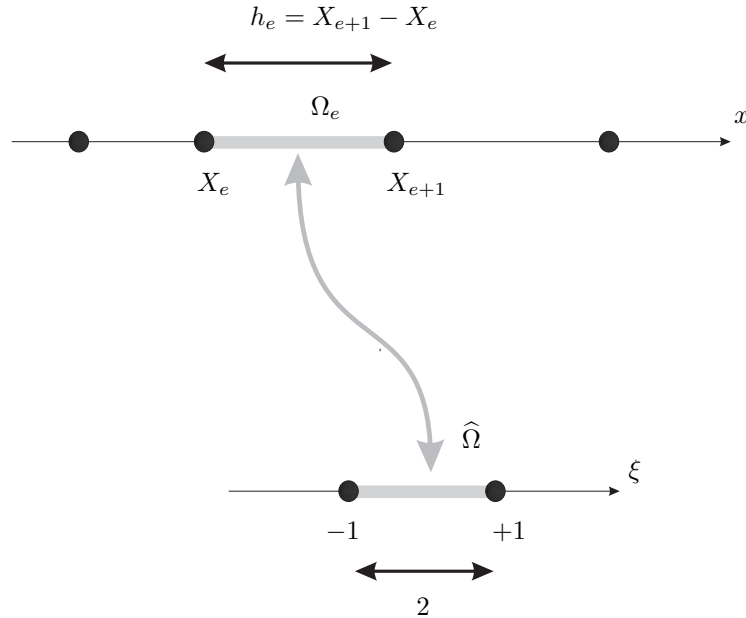


Figure 1.2: Isomorphisme entre un élément fini quelconque Ω_e et l'élément parent $\widehat{\Omega}$.

$$\begin{aligned}
x(\xi) &= \xi \frac{(X_{e+1} - X_e)}{2} + \frac{(X_{e+1} + X_e)}{2}, \\
\xi(x) &= \frac{2x - (X_{e+1} + X_e)}{(X_{e+1} - X_e)}.
\end{aligned} \tag{1.3}$$

L'isomorphisme est illustré à la figure 1.2. L'intervalle $\hat{\Omega}$ est appelé *l'élément parent*. Les coordonnées $x \in \Omega_e$ et $\xi \in \hat{\Omega}$ sont reliées entre elles de manière univoque par les relations (1.3).

Valeurs nodales et fonctions de forme locales

Définissons sur $\hat{\Omega}$ une base de polynômes ϕ_i de degré p de la manière suivante. Nous sélectionnons un nombre $p + 1$ de noeuds Ξ_i sur l'intervalle parent. A chaque noeud est associé une fonction de la base, c'est-à-dire un polynôme de degré p valant l'unité pour le noeud auquel elle est associée et s'annulant aux autres noeuds.

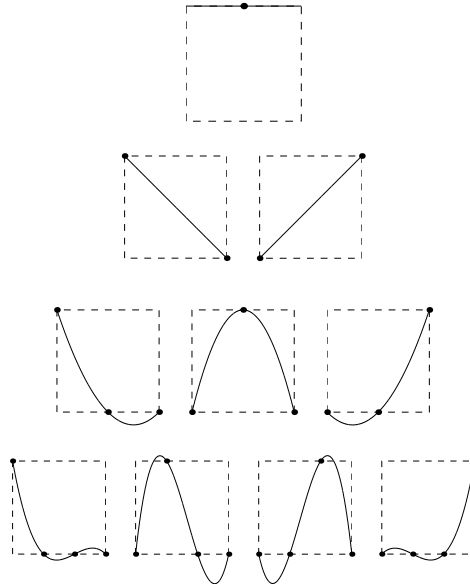


Figure 1.3: Fonctions de forme constante, linéaire, quadratique et cubique.

Il est dès lors possible de construire des bases de fonctions de forme de degrés divers :

- Pour les polynômes de degré 0, on a besoin d'une seule fonction de base,

$$\phi_1(\xi) = 1. \tag{1.4}$$

- Pour les polynômes du premier degré, nous choisissons les deux noeuds aux extrémités de l'élément parent, et obtenons une base des fonctions linéaires,

$$\begin{aligned}\phi_1(\xi) &= (1 - \xi)/2, \\ \phi_2(\xi) &= (1 + \xi)/2.\end{aligned}\tag{1.5}$$

- Pour les polynômes du deuxième degré, nous pouvons identifier deux noeuds aux extrémités et un noeud au centre de l'élément parent. Nous obtenons ainsi une base des fonctions quadratiques,

$$\begin{aligned}\phi_1(\xi) &= -\xi(1 - \xi)/2, \\ \phi_2(\xi) &= \xi(1 + \xi)/2, \\ \phi_3(\xi) &= (1 - \xi)(1 + \xi).\end{aligned}\tag{1.6}$$

La figure 1.3 montre clairement l'allure et la signification de ces fonctions. Il est clair que d'autres choix sont possibles et nous aborderons ce problème plus tard.

Une interpolation u^h de degré p sur l'élément Ω_e passant par des valeurs nodales U_i^e s'obtient ensuite par combinaison linéaire des fonctions de base

$$u^h(x) = \sum_{i=1}^{p+1} U_i^e \phi_i^e(x), \quad x \in \Omega_e,\tag{1.7}$$

où les U_i^e sont les *valeurs nodales locales* inconnues a priori, tandis que les fonctions de base $\phi_i^e(x) = \phi_i(\xi(x))$ connues a priori sont appelées les *fonctions de forme locales*. De façon plus synthétique, on parle d'*élément unidimensionnel constant discontinu*, d'*élément unidimensionnel linéaire continu* ou *élément unidimensionnel quadratique continu* pour caractériser le choix d'un type d'élément et d'une base de fonctions de forme sur cet élément.

Valeurs nodales et fonctions de forme globales

Nous avons montré qu'une interpolation peut être construite en associant les valeurs nodales et les fonctions de forme dans chaque élément. Par la suite, il sera utile d'écrire une expression sur le domaine Ω plutôt que sur les éléments séparés.

Dans cette optique, nous sélectionnons un nombre N de *noeuds* dont on donne les positions X_i^{node} au sein du maillage. Ces noeuds peuvent être à nouveau les extrémités d'éléments, ou des points à l'intérieur des éléments. Il est fondamental de distinguer, dès à présent, les concepts de noeud et de sommet, comme nous l'illustrons sur la figure 1.4. Un sommet est un point à l'intersection de plusieurs éléments finis, tandis qu'un noeud est un point où est définie une valeur discrète ou nodale de l'interpolation.

L'interpolation globale u^h peut alors s'écrire,

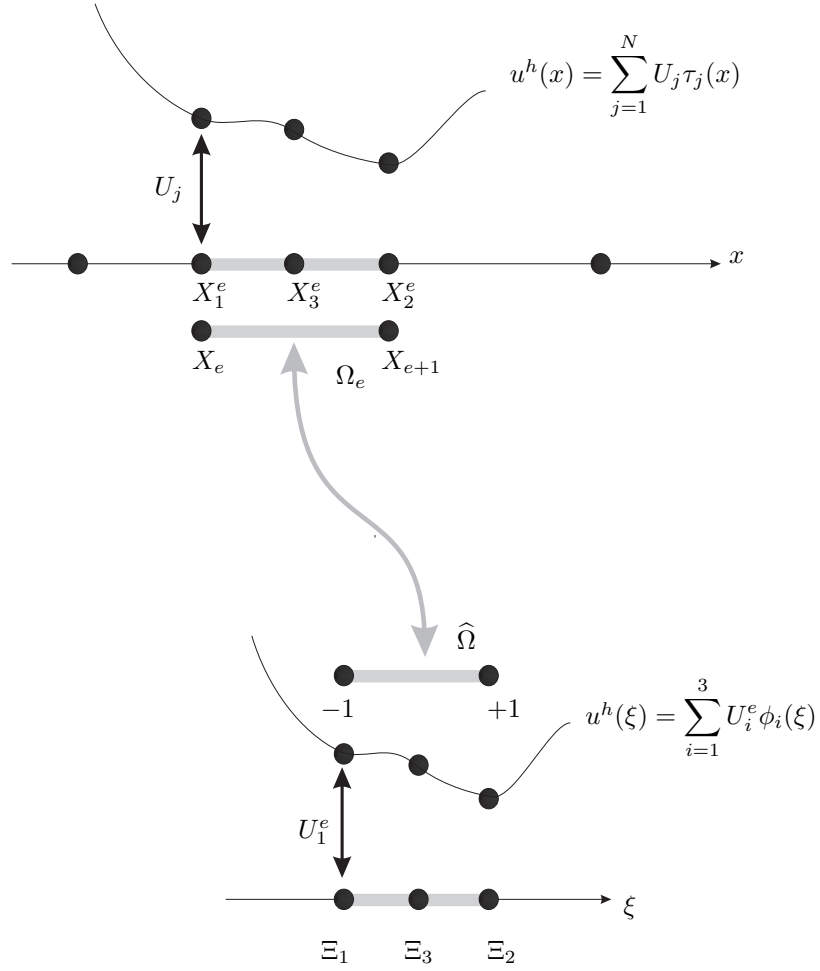


Figure 1.4: Concept de noeuds, de valeurs nodales et de sommets d'un élément.

$$u^h(x) = \sum_{j=1}^N U_j \tau_j(x) \quad (1.8)$$

où les U_j sont les *valeurs nodales globales*, tandis que les $\tau_j(x)$ sont les *fonctions de forme globales*. Cette équation montre clairement que l'approximation u^h dépend d'un nombre fini de valeurs nodales. Appelons \mathcal{U} l'espace fonctionnel dans lequel est incluse la solution exacte u . On constate que les fonctions de forme globales forment la base d'un sous-espace \mathcal{U}^h de dimension N , appelé un *sous-espace discret*.

Afin de construire la forme globale (1.8) à partir de la forme locale (1.7), il est nécessaire de définir la correspondance entre noeuds locaux et noeuds globaux de la manière suivante

	Indice local de noeud i
Indice d'éléments e	Indice global de noeud j

Ces correspondances peuvent être établies très facilement dans le cas unidimensionnel.

- Nous pouvons construire ainsi toutes les fonctions de forme globales linéaires sur un élément Ω_e .

$$\begin{aligned}
\tau_e(x) &= \phi_1^e(x), & x \in \Omega_e, \\
\tau_{e+1}(x) &= \phi_2^e(x), & x \in \Omega_e, \\
\tau_f(x) &= 0, & f \notin \{e, e+1\} \text{ et } x \in \Omega_e.
\end{aligned} \tag{1.9}$$

- Il est facile de suivre le même raisonnement pour une interpolation quadratique et d'obtenir les fonctions de forme globales correspondantes sur un élément Ω_e .

$$\begin{aligned}
\tau_{2e-1}(x) &= \phi_1^e(x), & x \in \Omega_e, \\
\tau_{2e}(x) &= \phi_3^e(x), & x \in \Omega_e, \\
\tau_{2e+1}(x) &= \phi_2^e(x), & x \in \Omega_e, \\
\tau_f(x) &= 0, & f \notin \{2e-1, 2e, 2e+1\} \text{ et } x \in \Omega_e.
\end{aligned} \tag{1.10}$$

Le fait que les fonctions τ_i soient des fonctions polynômiales par morceaux qui s'annulent en dehors d'un support compact est une des caractéristiques principales de la méthode des éléments finis. Nous verrons ultérieurement que c'est aussi une des clés du succès et de la popularité de la méthode.

1.2 Eléments finis bidimensionnels

Nous allons maintenant présenter quelques espaces discrets \mathcal{U}^h habituellement utilisés dans les applications bidimensionnelles. Typiquement, les espaces d'éléments finis sont définis à partir de

- la description de la division du domaine Ω en éléments,
- la définition de la forme de la restriction de tout élément de \mathcal{U}^h sur un élément : en général, il s'agira d'un polynôme, mais on pourrait imaginer autre chose. Typiquement, on a donc un nombre fini de degré de liberté pour déterminer cette restriction,

- la définition de contraintes sur chaque élément afin que cette restriction puisse être fixée de manière unique. Typiquement, il s'agit des valeurs nodales, mais il pourrait aussi s'agir d'autre chose.

En d'autres mots, on voit bien qu'un élément précis de l'espace discret sera uniquement déterminé par l'ensemble des valeurs nodales. Toutefois, cet élément reste une fonction, dans le cas qui nous concerne, une interpolation par ses valeurs nodales.

Tout d'abord, considérons quelques exemples de maillages d'éléments finis. Le choix du maillage dépend d'abord de la géométrie du problème. La taille des éléments dépend principalement de l'allure de la fonction cherchée. Pratiquement, on utilise des éléments petits par rapport à la taille du domaine là où la fonction représentée par éléments finis varie très rapidement. La plupart des éléments que nous utiliserons ci-dessous sont capables de représenter exactement des fonctions qui dépendent linéairement des coordonnées. Pour une telle variation linéaire, il n'est pas nécessaire d'utiliser des petits éléments.

Le découpage en éléments devra toujours respecter la règle suivante : deux éléments distincts ne peuvent avoir en commun qu'un sommet ou un côté entier. Il est toutefois possible de déroger à cette règle moyennant certains développements...

Un maillage bidimensionnel d'éléments finis est identifié par :

- le tableau des coordonnées de chaque sommet,
- le tableau d'appartenance des sommets aux éléments, c'est-à-dire, pour chaque élément, la liste ordonnée des sommets qui en font partie,
- éventuellement, le tableau d'appartenance des côtés aux éléments, c'est-à-dire, pour chaque élément, la liste ordonnée des côtés qui en font partie.

Chaque élément est numéroté de manière unique dans un ordre qui n'est en général *pas arbitraire*. On fait de même pour tous les côtés et tous les sommets. Nous verrons plus loin qu'il existe plusieurs manières de numéroter les éléments, les côtés et les sommets dans un maillage d'éléments finis.

La taille d'un maillage est donc caractérisée par trois nombres :

- N_0 le nombre de sommets,
- N_1 le nombre de côtés,
- N_2 le nombre d'éléments.

Il est important de toujours distinguer clairement le concept de noeud et de sommet. Un sommet est un point à l'intersection de plusieurs arêtes communes d'éléments finis, tandis qu'un noeud est un point où est définie une valeur discrète ou nodale de l'approximation. A partir de la taille d'un maillage, il est en général facile de déduire n le nombre de valeurs nodales ou de noeuds d'un type d'approximation.

Exemple de maillage

Nous donnons ci-dessous un exemple complet de données qui pourraient être associées au réseau de la figure 1.5.

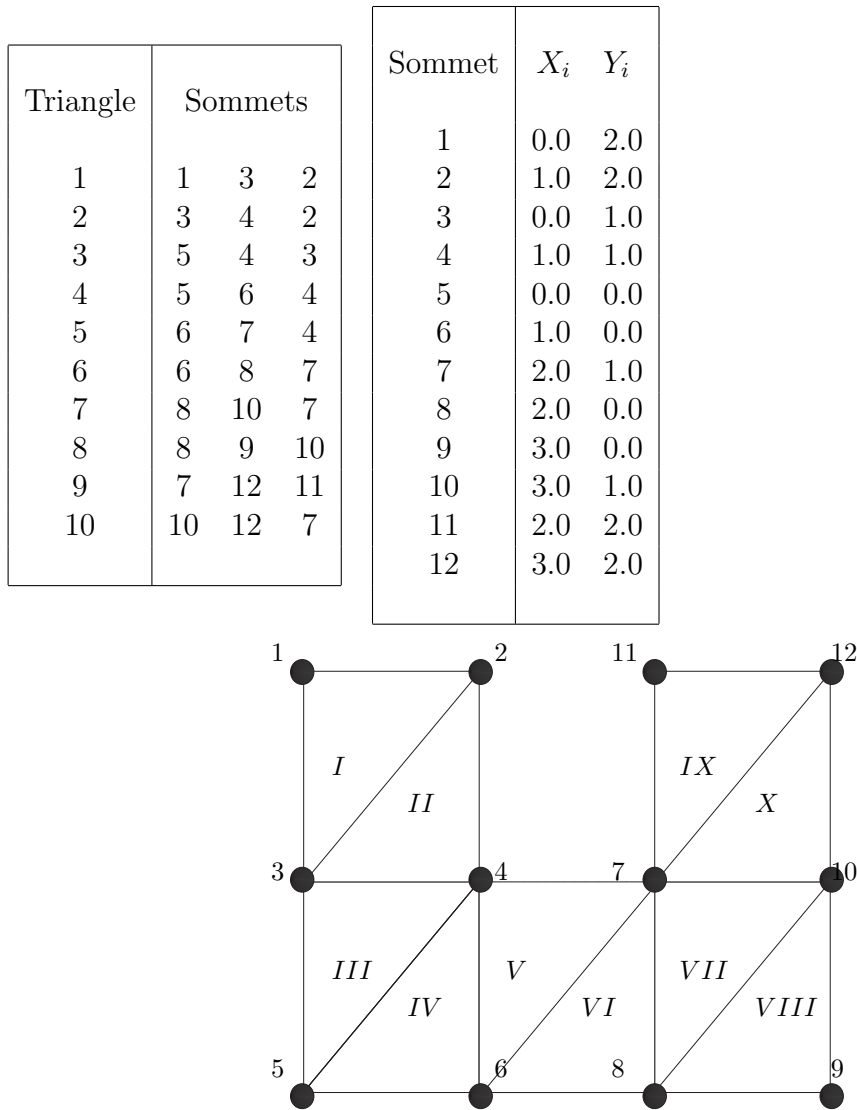


Figure 1.5: Numérotation du maillage.

1.2.1 Éléments triangulaires

Élément parent

Pour obtenir facilement les fonctions de forme, considérons un isomorphisme entre l'élément triangulaire Ω_e et l'élément parent triangulaire $\hat{\Omega}$ défini dans le plan $\boldsymbol{\xi} = (\xi, \eta)$.

Le triangle parent généralement utilisé dans la littérature¹ est défini par les coordonnées de ses trois sommets qui sont respectivement $\Xi_1 = (0, 0)$, $\Xi_2 = (1, 0)$, $\Xi_3 = (0, 1)$. Considérons maintenant les trois sommets d'un élément quelconque Ω_e , dont les coordonnées respectives sont $\mathbf{X}_1^e = (X_1^e, Y_1^e)$, $\mathbf{X}_2^e = (X_2^e, Y_2^e)$, $\mathbf{X}_3^e = (X_3^e, Y_3^e)$. La correspondance entre $\hat{\Omega}$ et Ω_e est donnée par

$$\mathbf{x}(\xi) = (1 - \xi - \eta)\mathbf{X}_1^e + \xi\mathbf{X}_2^e + \eta\mathbf{X}_3^e. \quad (1.11)$$

On observe immédiatement que $\mathbf{x}(\Xi_i) = \mathbf{X}_i$. Il est évidemment possible d'obtenir la relation inverse $\xi(\mathbf{x})$, puisque l'isomorphisme est ici une relation linéaire. Mais, nous n'effectuerons pas ce calcul, car nous n'en aurons jamais besoin.

Valeurs nodales et fonctions de forme locales

Définissons sur $\hat{\Omega}$ une base de polynômes ϕ_i de degré p de la manière suivante. Nous sélectionnons un nombre adéquat de noeuds Ξ_i sur le triangle parent. A chaque noeud est associé une fonction de la base, c'est-à-dire, un polynôme de degré p valant l'unité pour le noeud auquel elle est associée et s'annulant aux autres noeuds. Il suffit ensuite de construire des bases de fonctions de forme de degrés divers :

- Pour les polynômes de degré 0, on a besoin d'une seule fonction de base,

$$\phi_1(\xi, \eta) = 1. \quad (1.12)$$

- Pour les polynômes du premier degré, nous choisissons les trois sommets de l'élément parent, et obtenons une base des fonctions linéaires,

$$\begin{aligned} \phi_1(\xi, \eta) &= (1 - \xi - \eta), \\ \phi_2(\xi, \eta) &= \xi, \\ \phi_3(\xi, \eta) &= \eta. \end{aligned} \quad (1.13)$$

Un polynôme du premier degré dans un espace à deux dimensions est bien caractérisé par trois coefficients, ce qui signifie que définir trois fonctions de forme associées aux sommets constitue un choix naturel. Il est utile d'examiner la continuité qu'engendre cet élément. Considérons deux éléments voisins avec deux noeuds communs et une vue de la fonction u^h . Il est clair que la valeur de u^h entre deux noeuds dépend seulement de ses valeurs à ces deux noeuds, et que u^h est continue à l'interface de ces éléments.

¹Ce n'est peut-être pas le choix le plus élégant, mais le résultat du poids des traditions.

- Pour les polynômes du deuxième degré, il y a lieu de définir six fonctions de base distinctes. On choisit de les associer aux trois sommets et à trois noeuds milieux des segments. Nous obtenons ainsi une base des fonctions quadratiques,

$$\begin{aligned}
\phi_1(\xi, \eta) &= 1 - 3(\xi + \eta) + 2(\xi + \eta)^2, \\
\phi_2(\xi, \eta) &= \xi(2\xi - 1), \\
\phi_3(\xi, \eta) &= \eta(2\eta - 1), \\
\phi_4(\xi, \eta) &= 4\xi(1 - \xi - \eta), \\
\phi_5(\xi, \eta) &= 4\xi\eta, \\
\phi_6(\xi, \eta) &= 4\eta(1 - \xi - \eta).
\end{aligned} \tag{1.14}$$

Nous observons à nouveau que la valeur de u^h le long d'un côté du triangle est entièrement déterminée par ses valeurs nodales le long de ce côté, et nous concluons que u^h est une fonction continue.

- Il convient de noter que nous avons totale liberté pour définir d'autres fonctions de forme locales. On pourrait envisager un autre choix pour les polynômes du premier degré. Choisissons maintenant non plus les trois sommets de l'élément parent, mais les 3 noeuds milieux des segments. Nous obtenons une nouvelle base des fonctions linéaires,

$$\begin{aligned}
\phi_1(\xi, \eta) &= 1 - 2\eta, \\
\phi_2(\xi, \eta) &= -1 + 2(\xi + \eta), \\
\phi_3(\xi, \eta) &= 1 - 2\xi.
\end{aligned} \tag{1.15}$$

Chacune de ces fonctions prend la valeur unité au centre d'un segment et s'annule bien au milieu des autres. Il y a toutefois une différence fondamentale entre les approximations du premier degré définies par les fonctions (1.13) et par les fonctions (1.15), respectivement. Sur la frontière commune de deux triangles adjacents, on observe dans le premier un raccord continu qui fait défaut dans le second choix des fonctions de forme. En effet, dans le cas d'une interpolation linéaire entre les milieux des côtés, la fonction u^h peut être discontinue le long d'un côté car elle ne dépend pas que des valeurs nodales aux noeuds situées le long de ce côté.

Un élément totalement quelconque u^h de degré p sur l'élément Ω_e , s'obtient ensuite par combinaison linéaire des fonctions de base

$$u^h(\mathbf{x}) = \sum_{i=1}^{p+1} U_i^e \phi_i^e(\mathbf{x}), \quad \mathbf{x} \in \Omega_e, \tag{1.16}$$

où les U_i^e sont les *valeurs nodales locales* inconnues à priori, tandis que les fonctions de base $\phi_i^e(\mathbf{x}) = \phi_i(\boldsymbol{\xi}(\mathbf{x}))$ connues a priori sont appelées les *fonctions de forme locales*. On parle aussi de façon synthétique, d'*élément triangulaire constant discontinu*, d'*élément*

triangulaire linéaire continu, d'*élément triangulaire quadratique continu* ou d'*élément triangulaire linéaire discontinu* pour caractériser nos quatre exemples de choix d'élément et de fonctions de bases locales. La définition de fonctions de forme globales suit alors exactement la même logique que dans le cas unidimensionnel.

Approximations polynômiales complètes

Une des caractéristiques des éléments triangulaires est la possibilité de construire un ensemble de fonctions de forme *complètes sur le triangle parent* avec un nombre optimal de degrés de liberté. Un ensemble de fonctions de forme $\phi_i(\xi, \eta)$ sera dit *complet pour l'ordre p* , si une combinaison linéaire appropriée des $\phi_i(\xi, \eta)$ permet de reproduire n'importe quel polynôme d'ordre p .

Une telle propriété découle immédiatement du triangle de Pascal (1.17) qui fournit l'ensemble des termes d'un polynôme à deux variables (ξ, η) . Par exemple, un polynôme d'ordre 2 requiert six termes qui sont exactement le nombre de valeurs nodales des fonctions de forme quadratiques sur le triangle.

$$\begin{array}{ccccccc}
 & & & & 1 & & 0 \\
 & & & \xi & & \eta & 1 \\
 & \xi^2 & & \xi\eta & & \eta^2 & 2 \\
 \xi^3 & & \xi^2\eta & & \xi\eta^2 & & \eta^3 & 3 \\
 \dots & & & & & & &
 \end{array} \tag{1.17}$$

1.2.2 Éléments quadrilatères

Élément parent

Introduisons à présent un isomorphisme entre un élément quadrilatère quelconque Ω_e et l'élément parent carré $\hat{\Omega}$ défini dans le plan $\boldsymbol{\xi} = (\xi, \eta)$.

Le carré parent généralement utilisé dans la littérature ² est défini par les coordonnées de ses quatre sommets qui sont respectivement $\boldsymbol{\Xi}_1 = (1, 1)$, $\boldsymbol{\Xi}_2 = (-1, 1)$, $\boldsymbol{\Xi}_3 = (-1, -1)$ et $\boldsymbol{\Xi}_4 = (1, -1)$. Considérons maintenant les quatre sommets d'un élément quelconque Ω_e , dont les coordonnées respectives sont $\mathbf{X}_i^e = (X_i^e, Y_i^e)$ avec $i = 1 \dots 4$. La correspondance entre $\hat{\Omega}$ et Ω_e est donnée par

²A nouveau, ceci n'est peut-être pas le choix le plus élégant, puisque la superficie du carré parent est 4... alors que l'unité aurait été pratique. En plus, si le segment et le carré parents sont cohérents entre eux, le triangle parent ne rentre pas dans la même logique. Toutefois, ceci ne doit pas vous troubler, car les éléments parents peuvent être choisis de manière totalement arbitraire.

$$\begin{aligned}
\mathbf{x}(\xi, \eta) = & \frac{(1+\xi)(1+\eta)}{4} \mathbf{X}_1^e + \frac{(1-\xi)(1+\eta)}{4} \mathbf{X}_2^e \\
& + \frac{(1-\xi)(1-\eta)}{4} \mathbf{X}_3^e + \frac{(1+\xi)(1-\eta)}{4} \mathbf{X}_4^e.
\end{aligned} \tag{1.18}$$

On observe à nouveau que $\mathbf{x}(\Xi_i) = \mathbf{X}_i$. Il est important d'observer par contre que cette correspondance n'est plus une relation linéaire. Il ne sera pas automatiquement possible d'obtenir la relation inverse $\xi(\mathbf{x})$. Toutefois, nous obtiendrons une transformation bijective entre l'ensemble des points du carré parent et ceux de l'élément quadrilatère si ce dernier est convexe. En d'autres mots, il suffit d'exiger que les angles intérieurs du quadrilatère de la figure soient inférieurs à π pour définir une transformation licite. Il est alors possible, au moins formellement, d'inverser (1.18), et d'obtenir $\xi(\mathbf{x})$. Mais, nous n'effectuons pas ce calcul, car nous n'en aurons jamais besoin.

Valeurs nodales et fonctions de forme locales

Définissons sur $\hat{\Omega}$ une base de polynômes ϕ_i de degré p de la manière suivante. Nous sélectionnons un nombre adéquat de noeuds Ξ_i sur le carré parent. A chaque noeud est associé une fonction de la base, c'est-à-dire, un polynôme de degré p valant l'unité pour le noeud auquel elle est associée et s'annulant aux autres noeuds. Il suffit ensuite de construire des bases de fonctions de forme de degrés divers :

- Pour les polynômes de degré 0, on a besoin d'une seule fonction de base,

$$\phi_1(\xi, \eta) = 1. \tag{1.19}$$

- Choisissons maintenant les quatre sommets de l'élément parent, et obtenons une base des fonctions bilinéaires,

$$\begin{aligned}
\phi_1(\xi, \eta) &= (1+\xi)(1+\eta)/4, \\
\phi_2(\xi, \eta) &= (1-\xi)(1+\eta)/4, \\
\phi_3(\xi, \eta) &= (1-\xi)(1-\eta)/4, \\
\phi_4(\xi, \eta) &= (1+\xi)(1-\eta)/4.
\end{aligned} \tag{1.20}$$

Ces fonctions de forme sont obtenues par le produit terme à terme des fonctions de forme unidimensionnelles linéaires. Ces fonctions varient linéairement le long des côtés et une fonction u^h construite comme combinaison linéaire de telles fonctions de forme bilinéaires sera continue sur les segments frontières entre éléments.

La génération de fonctions de forme bidimensionnelles par produit terme à terme des fonctions unidimensionnelles est une des caractéristiques de la *famille des éléments de Lagrange*. Les fonctions de forme bilinéaires sont aussi appelées des fonctions de forme de Lagrange d'ordre 1.

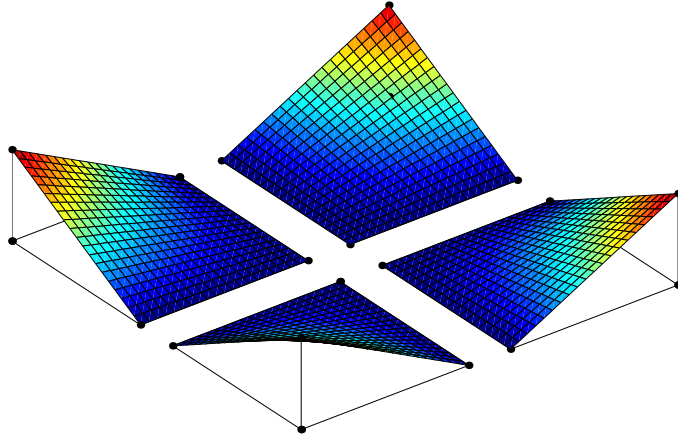


Figure 1.6: Fonctions de forme de Lagrange continues d'ordre un sur un quadrilatère ou fonctions de forme bilinéaires.

Par contre, il faut observer que ces fonctions ne sont plus linéaires, mais le produit de fonctions linéaires : d'où, le nom de fonctions de forme bilinéaires. Si la valeur de u^h est donnée aux quatre sommets, il est impossible de construire un polynôme du premier degré qui ne contient par définition que trois degrés de liberté. Il n'est donc pas possible de construire une interpolation linéaire continue sur un élément quadrilatère.

- Les 9 fonctions de forme de Lagrange d'ordre deux ou fonctions de forme biquadratiques sont obtenues également par les multiplications respectives des fonctions unidimensionnelles quadratiques suivant les directions ξ et η .

$$\begin{aligned}
\phi_1(\xi, \eta) &= \xi(1 + \xi)\eta(1 + \eta)/4, \\
\phi_2(\xi, \eta) &= -\xi(1 - \xi)\eta(1 + \eta)/4, \\
\phi_3(\xi, \eta) &= \xi(1 - \xi)\eta(1 - \eta)/4, \\
\phi_4(\xi, \eta) &= -\xi(1 + \xi)\eta(1 - \eta)/4, \\
\phi_5(\xi, \eta) &= (1 + \xi)(1 - \xi)\eta(1 + \eta)/2, \\
\phi_6(\xi, \eta) &= -\xi(1 - \xi)(1 - \eta)(1 + \eta)/2, \\
\phi_7(\xi, \eta) &= -(1 - \xi)(1 + \xi)\eta(1 - \eta)/2, \\
\phi_8(\xi, \eta) &= \xi(1 + \xi)(1 - \eta)(1 + \eta)/2, \\
\phi_9(\xi, \eta) &= (1 - \xi)(1 + \xi)(1 - \eta)(1 + \eta).
\end{aligned} \tag{1.21}$$

Ces fonctions sont associées aux quatres sommets, aux quatres points milieux du segment et au centre du carré. Sur un côté de l'élément, une fonction u^h est entièrement caractérisée par les valeurs nodales de ce côté et est donc continue.

Il est possible d'engendrer avec ces fonctions de forme un polynôme arbitraire du second degré mais pas du troisième degré malgré la présence de termes en $\xi^2\eta$, $\xi\eta^2$

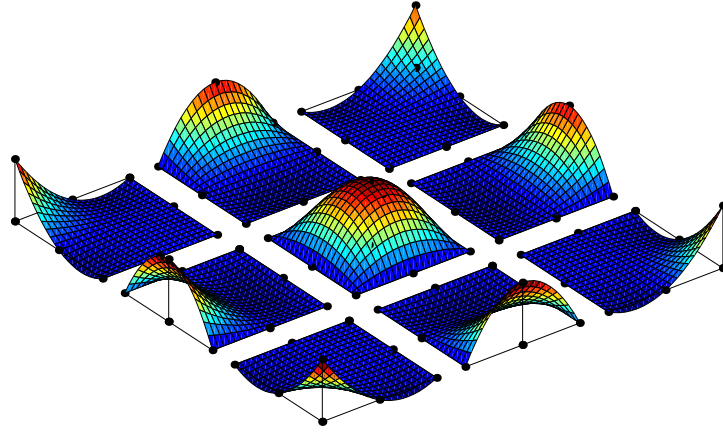


Figure 1.7: Fonctions de forme de Lagrange continues d'ordre deux sur un quadrilatère ou fonctions de forme biquadratiques.

et $\xi^2\eta^2$. Il n'est pas possible de produire, par exemple, ξ^3 par une combinaison linéaire des ϕ_i . En d'autres mots, ces fonctions de forme sont complètes à l'ordre deux, mais pas à l'ordre 3, car seule une partie des termes cubiques est représentable par une combinaison linéaire des fonctions de forme biquadratiques. Elles ne sont ni vraiment cubiques, ni vraiment quadratiques... elles ont même un terme d'ordre quatre. Schématiquement, les termes présents peuvent être représentés comme suit :

$$\begin{array}{ccccccc}
 & & & 1 & & & 0 \\
 & \xi & & & \eta & & 1 \\
 \xi^2 & & \xi\eta & & \eta^2 & & 2 \\
 & \xi^2\eta & & \xi\eta^2 & & & 3 \\
 & & \xi^2\eta^2 & & & & 4
 \end{array} \quad (1.22)$$

- Les 8 fonctions de forme de Serendip d'ordre deux sont définies en n'associant des fonctions de forme qu'aux sommets et milieux des côtés. Dans la famille d'éléments dits de Serendip, on ne retient que les noeuds situés sur la frontière de l'élément de Lagrange correspondant. On impose aussi que la valeur des fonctions de forme sur la frontière de l'élément soit identique à celle des fonctions lagrangiennes correspondantes.

On obtient facilement³ les huit fonctions de forme de Serendip d'ordre deux :

$$\begin{aligned}
\phi_1(\xi, \eta) &= (1 + \xi)(1 + \eta)(\eta + \xi - 1)/4, \\
\phi_2(\xi, \eta) &= (1 - \xi)(1 + \eta)(\eta - \xi - 1)/4, \\
\phi_3(\xi, \eta) &= (1 - \xi)(1 - \eta)(-\eta - \xi - 1)/4, \\
\phi_4(\xi, \eta) &= (1 + \xi)(1 - \eta)(-\eta + \xi - 1)/4, \\
\phi_5(\xi, \eta) &= (1 - \xi)(1 + \xi)(1 + \eta)/2, \\
\phi_6(\xi, \eta) &= (1 - \xi)(1 + \eta)(1 - \eta)/2, \\
\phi_7(\xi, \eta) &= (1 + \xi)(1 - \xi)(1 - \eta)/2, \\
\phi_8(\xi, \eta) &= (1 + \xi)(1 + \eta)(1 - \eta)/2.
\end{aligned} \tag{1.23}$$

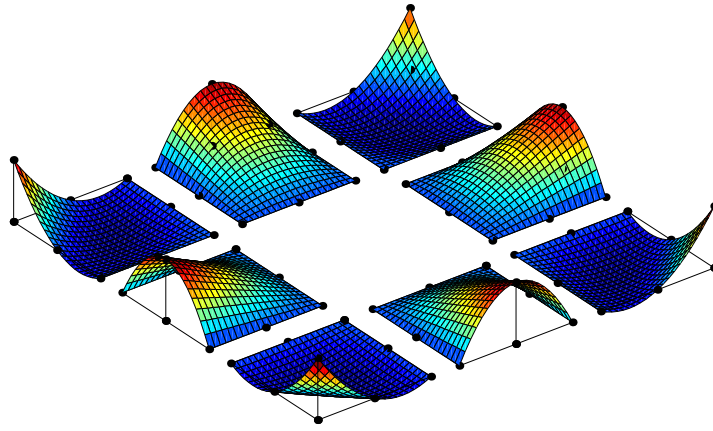


Figure 1.8: Fonctions de forme de Serendip continues d'ordre deux sur un quadrilatère.

³Comment peut-on trouver facilement les équations des fonctions de forme ? Oui, c'est facile, mais il y a une petite astuce... Considérons par exemple l'élément de Serendip à 8 noeuds. Nous désirons trouver une famille de fonctions de forme qui prennent la valeur unité à un noeud, et zéro aux autres noeuds. On peut facilement obtenir de telles fonctions au moyen de polynômes du troisième degré obtenus en multipliant les membres de gauche des équations des droites passant par les autres points. Pour le point 1, par exemple, les trois droites ont pour équation,

$$\begin{aligned}
1 + \xi &= 0, \\
1 + \eta &= 0, \\
1 - \xi - \eta &= 0.
\end{aligned}$$

La fonction $(1 + \xi)(1 + \eta)(1 - \xi - \eta)$ vaut zéro en chaque noeud, sauf au noeud 1, où sa valeur est -4. La fonction $(1 + \xi)(1 + \eta)(\xi + \eta - 1)/4$ a donc toutes les propriétés recherchées. De même, pour le noeud 5, les trois droites ont pour équation

$$\begin{aligned}
1 + \xi &= 0, \\
1 - \xi &= 0, \\
1 + \eta &= 0.
\end{aligned}$$

Ce qui donne une fonction de forme $(1 + \xi)(1 - \xi)(1 + \eta)/2$.

Schématiquement, les termes présents peuvent être représentés comme suit :

$$\begin{array}{ccccccc}
 & & & 1 & & & 0 \\
 & & \xi & & \eta & & 1 \\
 \xi^2 & & & \xi\eta & & \eta^2 & 2 \\
 & \xi^2\eta & & & \xi\eta^2 & & 3
 \end{array} \quad (1.24)$$

On voit donc que seul le terme du quatrième ordre a disparu. Ces fonctions de forme sont donc toujours complètes à l'ordre deux. En d'autres mots, il est toujours possible d'engendrer au moyen de ces fonctions de forme un polynôme arbitraire du second degré

$$a_1 + a_2\xi + a_3\eta + a_4\xi\eta + a_5\xi^2 + a_6\eta^2, \quad (1.25)$$

défini sur $\widehat{\Omega}$. L'élément de Serendip à 8 noeuds est complet jusqu'à l'ordre deux sur *l'élément parent*, tout comme l'élément à 9 noeuds.

Tout l'intérêt de la neuvième fonction de forme biquadratique apparaît lorsqu'on regarde si les éléments sont complets après application de l'isomorphisme. En d'autres mots, est-ce qu'une combinaison appropriée des fonctions de forme peut reproduire n'importe quel polynôme du second ordre

$$b_1 + b_2x(\xi, \eta) + b_3y(\xi, \eta) + b_4x(\xi, \eta)y(\xi, \eta) + b_5x^2(\xi, \eta) + b_6y^2(\xi, \eta), \quad (1.26)$$

défini sur Ω_e . Après application de la transformation (1.18) qui contient des termes en ξ , η et $\xi\eta$, l'expression (1.26) peut être réduite sous la forme

$$c_1 + c_2\xi + c_3\eta + c_4\xi\eta + c_5\xi^2 + c_6\eta^2 + c_7\xi\eta^2 + c_8\eta\xi^2 + c_9\eta^2\xi^2. \quad (1.27)$$

On voit clairement qu'une telle propriété sera satisfaite par l'élément lagrangien à 9 noeuds et non par l'élément de Serendip à 8 noeuds. Ceci explique partiellement les modestes performances numériques des fonctions de formes de Serendip dans le cas d'un maillage d'éléments non rectangulaires.

Chapitre 2

Eléments finis pour des problèmes elliptiques

De nombreux problèmes de physique/mathématique se ramènent à la résolution d'une équation de Poisson sur un domaine ouvert Ω du plan (x, y) . Celle-ci s'écrit avec des conditions aux frontières de deux types : des conditions de Dirichlet sur une partie de la frontière Γ_D et des conditions de Neumann le long de la partie Γ_N de la frontière ($\partial\Omega = \Gamma_D \cup \Gamma_N$ et $\Gamma_D \cap \Gamma_N = \emptyset$).

Dans ce chapitre, nous appliquons la méthode des éléments finis pour l'équation de Poisson qui est le paradigme des problèmes elliptiques. Nous définissons également le concept de principes variationnels, de formulations forte, faible et discrète. Pour des problèmes elliptiques, nous montrons qu'il est possible de construire une fonctionnelle scalaire qui atteint un minimum absolu pour la solution exacte du problème.

2.1 Formulations forte, faible et discrète

Il existe deux manières distinctes d'écrire ce problème : la formulation forte usuelle en termes d'équation aux dérivées partielles et la formulation faible *quasiment (!)* équivalente. Cette formulation faible ou variationnelle est strictement équivalente à un problème de minimisation. Intuitivement, pour la plupart des ingénieurs et des physiciens, écrire un problème aux limites (*boundary value problem*) sous une formulation variationnelle qui revient à minimiser ou à maximiser une fonctionnelle donne l'impression que le problème de l'existence et de l'unicité est automatiquement satisfait. En toute rigueur, ce n'est pas aussi simple. C'est exact à condition de recourir à des espaces de fonctions adéquats. Finalement, montrer l'équivalence entre une formulation différentielle et une formulation variationnelle dans le cas à plusieurs dimensions n'est pas simple. A plusieurs dimensions, il n'est généralement plus possible d'établir l'existence d'une solution de la formulation forte.

Formulation forte

La *formulation forte* d'un problème elliptique aux conditions aux limites s'écrit dès lors comme suit :

Trouver $u(\mathbf{x}) \in \mathcal{U}_s$ tel que

$$\begin{aligned} \nabla \cdot (a \nabla u) + f &= 0, & \forall \mathbf{x} \in \Omega, \\ \mathbf{n} \cdot (a \nabla u) &= g, & \forall \mathbf{x} \in \Gamma_N, \\ u &= t, & \forall \mathbf{x} \in \Gamma_D, \end{aligned}$$

(2.1)

où $\mathbf{n} = (n_x, n_y)$ représente la normale sortante de la courbe Γ et a est un paramètre constant¹. Sur Γ_D , la valeur de u est imposée à t , tandis que sur Γ_N , la valeur du flux $\mathbf{n} \cdot (a \nabla u)$ est imposée à la valeur de g . Nous n'avons pas encore précisé quel était cet espace \mathcal{U}_s dans lequel nous recherchons la fonction u : il s'agit évidemment d'une question capitale que nous allons laisser ouverte pour le moment, pour mieux y revenir plus tard².

Ce problème aux conditions frontières peut modéliser la conduction de la chaleur, où a est la conductivité thermique, f une source de chaleur et u la température. Il peut aussi s'agir de transfert de masse, où a est la diffusivité et u la concentration, de membrane élastique tendue où u est la flèche, a la tension dans la membrane et f la force latérale.

Formulation faible

Tout d'abord, introduisons quelques notations

$$\begin{aligned} \langle f g \rangle &= \int_{\Omega} f g \, d\Omega, \\ \ll f g \gg &= \int_{\partial\Omega} f g \, ds, \\ \ll f g \gg_N &= \int_{\Gamma_N} f g \, ds, \\ \ll f g \gg_D &= \int_{\Gamma_D} f g \, ds. \end{aligned}$$
(2.2)

¹Notons que nous supposons que a soit constant uniquement afin d'alléger les notations. Nous pourrions ainsi envisager de traiter le problème $\nabla \cdot (a(\mathbf{x}) \nabla u(\mathbf{x})) + f(\mathbf{x}) = 0$ de la même manière que celle que nous présenterons dans ce chapitre.

²Toutefois, vous pouvez déjà savoir que l'indice s est utilisé pour préciser qu'il s'agit de l'espace de la formulation forte (*strong formulation*)

Définissons de manière très approximative un espace $\widehat{\mathcal{U}}$ ne contenant que des fonctions qui s'annulent sur Γ_D et un espace \mathcal{U} ne contenant que des fonctions qui satisfont les conditions de Dirichlet. Armés de toutes ces superbes notations, nous pouvons maintenant effectuer un peu d'algèbre à partir de (2.1)

$$\begin{aligned}
& \langle \widehat{u} (\nabla \cdot (a \nabla u) + f) \rangle = 0, \quad \forall \widehat{u} \in \widehat{\mathcal{U}}, \\
& \quad \downarrow \\
& \text{En effectuant une intégration par parties,} \\
& \langle \nabla \cdot (\widehat{u} a \nabla u) \rangle - \langle (\nabla \widehat{u}) \cdot (a \nabla u) \rangle + \langle \widehat{u} f \rangle = 0, \quad \forall \widehat{u} \in \widehat{\mathcal{U}}, \\
& \quad \downarrow \\
& \text{En vertu du théorème de la divergence,} \\
& \ll \mathbf{n} \cdot (\widehat{u} a \nabla u) \gg - \langle (\nabla \widehat{u}) \cdot (a \nabla u) \rangle + \langle \widehat{u} f \rangle = 0, \quad \forall \widehat{u} \in \widehat{\mathcal{U}}, \\
& \quad \downarrow \\
& \text{En vertu de la définition de } \widehat{\mathcal{U}}, \\
& \ll \widehat{u} \underbrace{\mathbf{n} \cdot (a \nabla u)}_g \gg_N - \langle (\nabla \widehat{u}) \cdot (a \nabla u) \rangle + \langle \widehat{u} f \rangle = 0, \quad \forall \widehat{u} \in \widehat{\mathcal{U}}.
\end{aligned}$$

La *formulation faible* ou variationnelle est alors immédiatement déduite comme suit :

Trouver $u(\mathbf{x}) \in \mathcal{U}$ tel que

$$\underbrace{\langle (\nabla \widehat{u}) \cdot (a \nabla u) \rangle}_{a(\widehat{u}, u)} = \underbrace{\langle \widehat{u} f \rangle + \ll \widehat{u} g \gg_N}_{b(\widehat{u})}, \quad \forall \widehat{u} \in \widehat{\mathcal{U}},$$

(2.3)

Pour obtenir ce résultat, on tient compte du fait que \widehat{u} s'annule sur Γ_D et que sur Γ_N , le flux a été prescrit comme égal à g . Les conditions de Dirichlet et de Neumann sont également appelées conditions essentielles et naturelles. L'adjectif “naturel” indique ici que la condition peut être incluse naturellement dans l'écriture du principe variationnel, par opposition à la condition essentielle imposée a priori dans l'espace fonctionnel.

Ici, on pourrait croire que la formulation différentielle ou forte et la formulation variationnelle ou faible sont parfaitement équivalentes. En fait, c'est nettement moins évident

car cela dépend de la définition des espaces \mathcal{U}_s , \mathcal{U} et $\widehat{\mathcal{U}}$. Une définition judicieuse de ces espaces est un ingrédient essentiel pour obtenir des résultats théoriques et nous y reviendrons dans les sections ultérieures. Si les espaces \mathcal{U} et $\widehat{\mathcal{U}}$ sont simplement définis comme l'ensemble des fonctions de \mathcal{U}_s s'annulant ou valant t sur Γ_D , on obtient évidemment une stricte équivalence³.

Mais, lorsqu'on écrit une formulation variationnelle du type de (2.3), il est naturel et tentant d'un point de vue mathématique de considérer des espaces \mathcal{U} et $\widehat{\mathcal{U}}$ un peu plus grands (c'est-à-dire qui contiennent un peu plus de fonctions et ces quelques nouvelles venues sont justement très intéressantes : pas seulement des curiosités mathématiques, mais également des solutions tout à fait utiles d'un point de vue physique). D'un point de vue mathématique, on souhaitera travailler avec des espaces d'Hilbert pour démontrer, entre autres choses, l'existence et l'unicité de la solution de la formulation faible. Par contre, en déduire ensuite la même propriété pour la formulation forte n'est vraiment pas évident et est même souvent impossible.

Souvent, nous allons nous limiter au cas où les conditions de Dirichlet sont homogènes ($t = 0$). Dans une telle perspective, on observe immédiatement que les deux espaces \mathcal{U} et $\widehat{\mathcal{U}}$ coïncident et seront notés simplement \mathcal{U} . Cela simplifie énormément la présentation des résultats théoriques : le cas spécifique des conditions de Dirichlet non homogènes sera traité ultérieurement, lorsque nous considérerons le cas des éléments finis contraints.

Problème de minimum

L'intérêt majeur de la formulation faible est qu'elle est parfaitement équivalente à problème de minimisation défini comme suit :

Trouver $u(\mathbf{x}) \in \mathcal{U}$ tel que

$$J(u) = \min_{v \in \mathcal{U}} \underbrace{\left(\frac{1}{2} a(v, v) - b(v) \right)}_{J(v)}, \quad (2.4)$$

Considérons une fonctionnelle associée au problème de Poisson et écrivons :

$$J(v) = \frac{1}{2} \langle \nabla v \cdot a \nabla v \rangle - \langle f v \rangle - \ll g v \gg_N. \quad (2.5)$$

Nous allons montrer l'équivalence entre la formulation faible et la recherche de la

³Et encore.... pas vraiment aussi simple : noter par exemple que pour pouvoir appliquer le théorème de la divergence, la frontière $\partial\Omega$ ne peut pas être totalement arbitraire : elle doit être constituée d'un nombre fini de morceaux continûment différentiables. Il n'est donc pas nécessaire que la normale soit définie en chaque point, car le domaine Ω peut être ce qu'on appelle un ouvert à bords lipschitziens. Dans cette hypothèse, la normale peut n'être définie que presque partout. En particulier, il est donc possible de considérer le cas de coins dans la frontière en respectant rigoureusement les mathématiques.

fonction qui minimise la fonctionnelle. Envisageons, donc, la recherche d'une valeur stationnaire $J(u)$ d'une fonctionnelle J . On appelle calcul des variations l'art de trouver une fonction qui minimise ou maximise une fonctionnelle, généralement une intégrale sur un espace de fonctions avec ou sans contrainte.

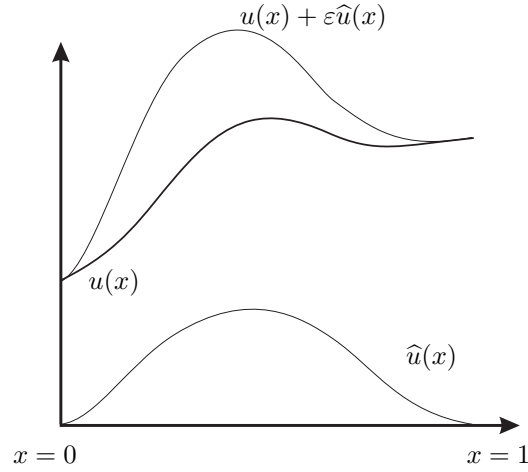


Figure 2.1: Recherche de la fonction u qui rend stationnaire une fonctionnelle J sur le domaine $\Omega =]0, 1[$: concept de variation \hat{u} .

Recherchons une fonction u qui minimise J et dont la valeur sur la totalité de la courbe frontière $\partial\Omega$ soit prescrite à t . Introduisons \hat{u} une fonction arbitraire qui s'annule sur Γ_D . Parmi les fonctions satisfaisant les conditions essentielles, une fonction u rend la fonctionnelle (2.5) stationnaire si la condition suivante est satisfaite.

$$\frac{dJ(u + \varepsilon\hat{u})}{d\varepsilon}\bigg|_{\varepsilon=0} = 0, \quad \forall \hat{u}, \quad (2.6)$$

Il suffit ensuite d'effectuer encore et toujours un peu d'algèbre

$$\begin{aligned} J(u + \varepsilon\hat{u}) &= \frac{1}{2} \langle \nabla u \cdot a \nabla u \rangle + \varepsilon \langle \nabla u \cdot a \nabla \hat{u} \rangle + \frac{\varepsilon^2}{2} \langle \nabla \hat{u} \cdot a \nabla \hat{u} \rangle \\ &- \langle f u \rangle - \varepsilon \langle f \hat{u} \rangle \\ &- \underbrace{\ll g u \gg_N}_{J(u)} - \varepsilon \ll g \hat{u} \gg_N \end{aligned}$$

$$\frac{dJ(u + \varepsilon\hat{u})}{d\varepsilon}\bigg|_{\varepsilon=0} = \langle \nabla u \cdot a \nabla \hat{u} \rangle - \langle f \hat{u} \rangle - \ll g \hat{u} \gg_N$$

et d'observer que la condition (2.6) correspond exactement à la formulation faible (2.3), tandis que la formulation d'Euler-Lagrange du problème de minimisation de la fonctionnelle J correspond exactement à la formulation forte (2.1).

La condition (2.6) est nécessaire pour que la fonctionnelle atteigne un minimum en la solution u , mais n'est pas suffisante. Il faut donc encore vérifier que u correspond bien à un minimum en effectuant à nouveau un petit peu d'algèbre pour une fonction quelconque $v \in \mathcal{U}$ qu'il est possible de décomposer en une combinaison de u et d'une variation $\hat{u} \in \hat{\mathcal{U}}$

$$\begin{aligned}
J(v) &= J(u + \hat{u}) \\
&= \frac{1}{2} \langle a(\nabla u + \nabla \hat{u})^2 \rangle - \langle f(u + \hat{u}) \rangle - \ll g(u + \hat{u}) \gg_N, \\
&= \frac{1}{2} \langle \nabla u \cdot a \nabla u \rangle + \langle \nabla u \cdot a \nabla \hat{u} \rangle + \frac{1}{2} \langle \nabla \hat{u} \cdot a \nabla \hat{u} \rangle \\
&\quad - \langle f u \rangle \quad - \langle f \hat{u} \rangle \\
&\quad - \underbrace{\ll g u \gg_N}_{J(u)} \quad - \underbrace{\ll g \hat{u} \gg_N}_{=0} \quad \underbrace{\quad}_{\geq 0} \\
&\hspace{15em} \text{cfr (2.3)} \hspace{10em} \text{car } a \text{ positif !} \\
&\geq J(u)
\end{aligned}$$

Le dernier terme est toujours du signe de a . La fonctionnelle J atteint donc un minimum en u pour autant que a soit positif. Heureusement, dans tous les problèmes physiques modélisés par l'équation de Poisson, les paramètres de diffusion sont toujours positifs (conductibilité, module de Young, viscosité ..etc) pour que le modèle soit cohérent d'un point de vue physique. Les formulations (2.3) et (2.4) sont donc bien strictement équivalentes.

En outre, si une solution du problème (2.3) ou (2.4) existe, cette solution est unique. Procédons par l'absurde. Considérons deux fonctions u et v solutions du problème et écrivons successivement (2.3) pour les deux solutions :

$$\begin{aligned}
\langle (\nabla \hat{u}) \cdot (a \nabla u) \rangle &= \langle \hat{u} f \rangle + \ll \hat{u} g \gg_N, & \forall \hat{u} \in \hat{\mathcal{U}}. \\
\langle (\nabla \hat{u}) \cdot (a \nabla v) \rangle &= \langle \hat{u} f \rangle + \ll \hat{u} g \gg_N, & \forall \hat{u} \in \hat{\mathcal{U}}. \\
&\downarrow \text{En soustrayant les 2 lignes précédentes,} \\
\langle (\nabla \hat{u}) \cdot (a \nabla u - a \nabla v) \rangle &= 0, & \forall \hat{u} \in \hat{\mathcal{U}}. \\
&\downarrow \text{En choisissant } u - v \text{ comme fonction } \hat{u} \text{ arbitraire,} \\
\langle a(\nabla u - \nabla v)^2 \rangle &= 0. \\
&\downarrow \text{Si } \Gamma_D \text{ n'est pas un ensemble vide,} \\
u - v &= 0.
\end{aligned}$$

L'unicité de la solution est acquise si une condition de Dirichlet est imposée au moins en un point. Cela correspond parfaitement à l'intuition physique qui suggère aussi de toujours imposer une condition de Dirichlet en au moins en point du domaine. Notez aussi qu'il est essentiel que a soit toujours strictement positif (ou négatif) pour déduire ce résultat d'unicité.

Les éléments finis sont une méthode variationnelle

Comme \mathcal{U}^h a été introduit comme un sous-espace de \mathcal{U} , une façon naturelle d'obtenir les U_i est d'exiger que :

Trouver $u^h(\mathbf{x}) \in \mathcal{U}^h$ tel que

$$\underbrace{\langle \nabla \hat{u}^h \cdot (a \nabla u^h) \rangle}_{a(\hat{u}^h, u^h)} = \underbrace{\langle \hat{u}^h f \rangle + \ll \hat{u}^h g \gg_N}_{b(\hat{u}^h)}, \quad \forall \hat{u}^h \in \mathcal{U}^h,$$

(2.7)

ou de manière strictement équivalente :

Trouver $u^h(\mathbf{x}) \in \mathcal{U}^h$ tel que

$$J(u^h) = \min_{v^h \in \mathcal{U}^h} \underbrace{\left(\frac{1}{2} a(v^h, v^h) - b(v^h) \right)}_{J(v^h)}, \quad (2.8)$$

Le problème (2.7) est souvent appelé *méthode de Galerkin*, tandis la formulation en terme de minimisation (2.8) est souvent présentée comme la *méthode de Ritz*. Toutefois, comme c'est équivalent, on pourrait aussi dire exactement le contraire... Nous appellerons le problème (2.7) ou (2.8) la *formulation discrète*.

Cette formulation discrète est un système linéaire de n équations à n inconnues. Pour obtenir les n équations algébriques fournissant les valeurs nodales, il suffit de substituer u^h dans l'expression de la fonctionnelle ⁴

$$\begin{aligned}
J(u^h) &= \frac{1}{2} \int_{\Omega} (\nabla u^h) \cdot (\nabla u^h) \, d\Omega - \int_{\Omega} f u^h \, d\Omega, \\
J(u^h) &= \frac{1}{2} \int_{\Omega} \left(\sum_{i=1}^n U_i \nabla \tau_i \right) \cdot \left(\sum_{j=1}^n U_j \nabla \tau_j \right) \, d\Omega - \int_{\Omega} f \left(\sum_{i=1}^n U_i \tau_i \right) \, d\Omega, \\
J(u^h) &= \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n U_i U_j \int_{\Omega} (\nabla \tau_i) \cdot (\nabla \tau_j) \, d\Omega - \sum_{i=1}^n U_i \int_{\Omega} f \tau_i \, d\Omega, \\
J(u^h) &= \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n U_i U_j A_{ij} - \sum_{i=1}^n U_i B_i,
\end{aligned} \quad (2.9)$$

où la matrice A_{ij} et le vecteur B_i sont définis par

$$\begin{aligned}
A_{ij} &= \int_{\Omega} (\nabla \tau_i) \cdot (\nabla \tau_j) \, d\Omega, \\
B_i &= \int_{\Omega} f \tau_i \, d\Omega.
\end{aligned} \quad (2.10)$$

Le minimum de la fonctionnelle est obtenu lorsque

$$0 = \frac{\partial J(u^h)}{\partial U_i} = \sum_{j=1}^n A_{ij} U_j - B_i, \quad i = 1, \dots, n. \quad (2.11)$$

⁴Afin de simplifier l'algèbre, nous ne tiendrons pas compte de l'intégrale le long de Γ_N , mais l'extension est évidente. Nous supposons aussi que $a = 1$, sans aucune perte de généralité.

Ce sont exactement les équations que nous aurions obtenues en sélectionnant les n fonctions de forme τ_j comme fonction arbitraire \hat{u}^h dans (2.7). Comme ces fonctions de forme forment une base du sous-espace d'approximation, cela revient bien à considérer que les équations sont satisfaites pour n'importe quel élément de \mathcal{U}^h .

Les éléments finis sont une méthode de résidus pondérés

Il existe une autre manière d'obtenir la formulation discrète qui peut paraître plus intuitive, sans devoir explicitement construire une formulation faible. Il s'agit ici de trouver une manière de sélectionner les n valeurs nodales U_j afin que la fonction $u^h(s)$ approxime "au mieux" la fonction $u(x)$, solution du problème original. Il faut évidemment préciser ce que nous entendons par l'expression "au mieux".

Nous avons n paramètres inconnus à déterminer. Dans les méthodes de résidus pondérés, il est possible de distinguer trois types d'approches :

- Les équations différentielles sont satisfaites par tout approximant et les paramètres sont déterminés pour satisfaire les conditions frontières. C'est l'approche frontière.
- Les conditions frontières sont satisfaites par tout approximant et les paramètres sont déterminés afin que les équations soient satisfaites. C'est l'approche intérieure.
- Les approximants ne vérifient a priori ni les équations, ni les conditions frontières. Il faudra déterminer les paramètres afin que l'équation et les conditions frontières soient satisfaites. C'est l'approche mixte.

En général, on adopte la seconde approche pour les conditions essentielles dans le cadre de méthodes d'éléments finis. C'est ce que nous faisons implicitement en définissant les espaces \mathcal{U} et \mathcal{U}^h comme ne contenant que des solutions satisfaisant les conditions de Dirichlet. Par contre, pour les conditions de Neumann, nous construirons une méthode mixte. Comme a priori, les éléments de \mathcal{U}^h (y compris u^h le meilleur d'entre eux, en général) ne satisfont pas l'équation différentielle de (2.1), ils donnent lieu à un résidu

$$r^h = \nabla \cdot (a \nabla u^h) + f. \quad (2.12)$$

La méthode des résidus pondérés consiste alors à évaluer les valeurs nodales afin de minimiser le résidu d'une certaine manière. Parmi les manières de minimiser le résidu, citons :

- La *technique de collocation* qui impose l'annulation du résidu aux noeuds.

$$r^h(\mathbf{X}_i) = 0, \quad i = 1, \dots, n, \quad (2.13)$$

- La *technique de Galerkin* consiste à annuler en moyenne le produit des résidus avec les fonctions de forme.

$$\langle \tau_i r^h \rangle = 0, \quad i = 1, \dots, n, \quad (2.14)$$

Le choix d'une technique influence fortement la précision (et donc la fiabilité) de l'approximation produite et nous expliquerons ultérieurement pourquoi nous sélectionnons ici la technique de Galerkin. On peut effectuer un peu d'algèbre (toujours la même !!) avec les équations (2.14) :

$$\begin{aligned}
& \langle \tau_i r^h \rangle = 0, \quad i = 1, \dots, n, \\
& \quad \downarrow \\
& \text{Par définition de } r^h, \\
& \langle \tau_i (\nabla \cdot (a \nabla u^h)) \rangle + \langle \tau_i f \rangle = 0, \quad i = 1, \dots, n, \\
& \quad \downarrow \\
& \text{En effectuant une intégration par parties,} \\
& - \langle (\nabla \tau_i) \cdot (a \nabla u^h) \rangle + \langle \nabla \cdot (\tau_i a \nabla u^h) \rangle + \langle \tau_i f \rangle = 0, \quad i = 1, \dots, n, \\
& \quad \downarrow \\
& \text{En vertu du théorème de la divergence,} \\
& - \langle (\nabla \tau_i) \cdot (a \nabla u^h) \rangle + \ll \mathbf{n} \cdot (\tau_i a \nabla u^h) \gg + \langle \tau_i f \rangle = 0, \quad i = 1, \dots, n, \\
& \quad \downarrow \\
& \text{En vertu de la définition de } \mathcal{U}^h, \\
& - \langle (\nabla \tau_i) \cdot (a \nabla u^h) \rangle + \ll \tau_i \underbrace{\mathbf{n} \cdot (a \nabla u^h)}_{\approx g} \gg_N + \langle \tau_i f \rangle = 0, \quad i = 1, \dots, n, \\
& \quad \downarrow \\
& \text{Par définition de } u^h, \\
& \sum_{j=1}^n \underbrace{\langle (\nabla \tau_i) \cdot (a \nabla \tau_j) \rangle}_{A_{ij}} U_j - \underbrace{\left(\ll \tau_i g \gg_N + \langle \tau_i f \rangle \right)}_{B_i} = 0, \quad i = 1, \dots, n,
\end{aligned}$$

où la matrice A_{ij} (souvent appelée matrice de raideur) et le vecteur B_i (souvent appelé vecteur des forces nodales) sont définis par,

$$\begin{aligned} A_{ij} &= \langle (\nabla \tau_i) \cdot (\nabla \tau_j) \rangle, \\ B_i &= \langle f \tau_i \rangle + \ll \tau_i g \gg_N. \end{aligned} \quad (2.15)$$

Ceci constitue bien la généralisation des définitions (2.10) pour des conditions de Neumann inhomogènes.

On obtient finalement le système algébrique qui définit la formulation discrète :

Trouver $U_j \in \mathcal{R}^n$ tels que

$$\sum_{j=1}^n A_{ij} U_j = B_i, \quad i = 1, n.$$

(2.16)

On a ainsi obtenu exactement la même formulation discrète qu'en ayant effectué la démarche variationnelle.

2.2 Exemple unidimensionnel

Tout d'abord, considérons le cas unidimensionnel de l'équation de Poisson avec le problème suivant :

Trouver $u(x)$ tel que

$$\frac{d^2 u}{dx^2} + f = 0, \quad \forall x \in \Omega,$$

$$\begin{aligned} u(0) &= 0, \\ u(1) &= 0, \end{aligned}$$

(2.17)

Supposons qu'il n'y ait strictement aucun espoir⁵ d'obtenir la solution analytique de (2.17). Nous désirons calculer une représentation approchée qui sera voisine de la solution exacte dans un sens qui sera défini plus loin. La formulation (2.17) est la *formulation forte* de notre problème.

⁵Ce qui est en fait complètement stupide car la solution analytique est vraiment facile à obtenir !

2.2.1 Formulation faible

Tout d'abord, les notations (2.2) deviennent maintenant

$$\begin{aligned} \langle f \, g \rangle &= \int_0^1 f g \, dx, \\ \ll f \, g \gg &= \left[f g \right]_0^1. \end{aligned} \tag{2.18}$$

Notons que comme nous n'avons introduit que des conditions essentielles homogènes, il n'est pas nécessaire de distinguer l'espace des solutions \mathcal{U} et l'espace des variations $\widehat{\mathcal{U}}$ dans notre formulation faible. Notons aussi que nous n'avons pas encore vraiment défini nos espaces : nous le ferons plus tard !

Introduisons la fonctionnelle correspondant à notre problème

$$J(v) = \underbrace{\frac{1}{2} \int_{\Omega} \frac{dv}{dx} \frac{dv}{dx} \, dx}_{a(v, v)} - \underbrace{\int_{\Omega} f v \, dx}_{b(v)}, \tag{2.19}$$

Et finalement la formulation peut s'écrire soit sous la forme (2.3), soit sous la forme d'un problème de minimum (2.4).

Trouver $u(x) \in \mathcal{U}$ tel que

$$\underbrace{\langle \frac{d\widehat{u}}{dx} \frac{du}{dx} \rangle}_{a(\widehat{u}, u)} = \underbrace{\langle \widehat{u} f \rangle}_{b(\widehat{u})}, \quad \forall \widehat{u} \in \mathcal{U},$$

(2.20)

Trouver $u(x) \in \mathcal{U}$ tel que

$$J(u) = \min_{v \in \mathcal{U}} \underbrace{\left(\frac{1}{2} a(v, v) - b(v) \right)}_{J(v)},$$

(2.21)

On observe ici clairement que la solution du problème formulé sous la forme faible (2.20) ou (2.21) ne sera solution du problème fort que si sa dérivée seconde existe et est continue. Le problème faible est faible dans le sens qu'il peut avoir une solution qui est inacceptable pour le problème fort : c'est donc, en termes imagés, une formulation laxiste...

Interprétation physique : problème de la corde tendue

Afin de donner un sens physique aux formulations (2.17), (2.20) ou (2.21), considérons un problème physique qu'ils pourraient modéliser. En l'occurrence, considérons une corde de longueur $L = 1$ tendue par une traction d'intensité $T = 1$, et soumise à une charge latérale répartie $f(x)$. Nous désirons construire la fonctionnelle d'énergie potentielle $J(u)$, dépendant de la flèche latérale $u(x)$. Nous examinons une théorie de petits déplacements tels que

$$\left(\frac{du}{dx}\right)^2 \ll 1. \quad (2.22)$$

L'hypothèse des petits déplacements permet de remplacer le sinus de l'angle de la pente de la corde en tout point par la dérivée de la déformée puisque le cosinus est proche de zéro. Exprimer l'équilibre vertical des forces pour un intervalle quelconque $]a, b[$ consiste donc à écrire :

$$\begin{aligned} \int_a^b f \, dx &= T \frac{du}{dx}(a) - T \frac{du}{dx}(b) && \forall a, b \\ \int_a^b f \, dx &= -T \left[\frac{du}{dx} \right]_a^b && \forall a, b \\ &\downarrow \text{Si la fonction } \frac{du}{dx} \text{ est continue !} \\ \int_a^b f + T \frac{d^2u}{dx^2} \, dx &= 0 && \forall a, b \\ f + T \frac{d^2u}{dx^2} &= 0 \end{aligned}$$

Lors du déplacement latéral de la corde, son allongement produit des tensions complémentaires que nous supposons petites par rapport à T . Dans ces conditions, si l est la longueur finale après l'application de la charge latérale, l'énergie potentielle est donnée par

$$J(u) = T(l - L) - \int_0^L f u \, dx$$

où nous distinguons l'énergie élastique accumulée par l'allongement de la corde sous la traction T et le travail des forces appliquées obtenu par le produit de celles-ci par les déplacements correspondants.

En tenant compte de l'hypothèse des petits déplacements, il est possible de calculer la longueur finale l à partir du déplacement latéral u .

$$l = \int_0^L dl = \int_0^L \sqrt{dx^2 + du^2} = \int_0^L \sqrt{1 + \left(\frac{du}{dx}\right)^2} dx \approx \int_0^L \left(1 + \frac{1}{2} \left(\frac{du}{dx}\right)^2\right) dx$$

on en déduit l'expression suivante de l'énergie potentielle

$$J(u) = \frac{1}{2} \int_0^L T \left(\frac{du}{dx}\right)^2 dx - \int_0^L f u dx$$

Le problème fort (2.17) correspond à une équation locale de conservation de la quantité de mouvement, le problème faible (2.20) correspond à l'application du principe des travaux virtuels, tandis que le problème de minimum (2.21) correspond au principe de minimisation de l'énergie potentielle du problème (2.17). Il est important ici de se rappeler que les formes locales des principes de conservation ont été déduits à partir de forme intégrale sur des volumes de contrôle ou des volumes matériels. En d'autres mots, la formulation originale en termes physiques est la formulation faible et il est donc naturel de s'attacher à la résolution de celle-ci lors de la construction d'une méthode numérique telle que celle des éléments finis.

2.2.2 Formulation discrète

Comme l'espace \mathcal{U}^h ne doit contenir que des fonctions respectant les conditions frontières, les deux valeurs nodales aux extrémités sont toujours contraintes à la valeur nulle. La dimension de cet espace est clairement $N - 2$, lorsque nous considérons notre exemple (2.17).⁶ Nous imposons d'abord

$$\begin{aligned} U_1 &= 0, \\ U_N &= 0, \end{aligned} \tag{2.23}$$

et ensuite il reste les $N - 2$ inconnues, solutions du système discret de $N - 2$ équations à $N - 2$ inconnues :

$$\sum_{j=2}^{N-1} A_{ij} U_j = B_i, \quad i = 2, \dots, N - 1. \tag{2.24}$$

où la matrice A_{ij} et le vecteur B_i sont donnés par :

⁶Il convient ici d'observer que la valeur $N - 2$ que nous obtenons ici correspond exactement à la valeur n de la section précédente. Il y aussi un décalage dans la numérotation, puisque les inconnues sont numérotées de U_2 à U_{N-1} .

$$\begin{aligned}
A_{ij} &= \int_{\Omega} \frac{d\tau_i}{dx}(x) \frac{d\tau_j}{dx}(x) dx, \\
B_i &= \int_{\Omega} f(x) \tau_i(x) dx,
\end{aligned} \tag{2.25}$$

2.2.3 Construction du système algébrique

Après avoir expliqué les idées fondamentales de la méthode des éléments finis, détaillons brièvement les étapes principales qui doivent être accomplies pour calculer les coefficients du système algébrique. A nouveau, nous prendrons (2.17) et sa forme discrétisée (2.24) comme exemple.

Connaissant les fonctions de forme globales τ_i et la fonction f , nous devons calculer les composantes A_{ij} de la matrice de raideur et B_i du vecteur de forces nodales, données par (2.25). Nous avons vu plus haut qu'une fonction de forme τ_i est associée au noeud i , et s'annule hors des éléments auxquels appartient le noeud i . Soient Ω_e et Ω_{e+1} les deux éléments contenant le noeud. Au lieu de calculer A_{ij} sur le domaine Ω , il suffit d'effectuer l'intégration sur ces deux seuls éléments et donc,

$$\begin{aligned}
A_{ij} &= \int_{\Omega_e} \tau_{j,x}(x) \tau_{i,x}(x) dx + \int_{\Omega_{e+1}} \tau_{j,x}(x) \tau_{i,x}(x) dx, \\
B_i &= \int_{\Omega_e} f(x) \tau_i(x) dx + \int_{\Omega_{e+1}} f(x) \tau_i(x) dx.
\end{aligned} \tag{2.26}$$

Les seuls coefficients non nuls de A_{ij} sont ceux qui sont associés à des noeuds j et i qui apparaissent simultanément dans l'élément Ω_e ou Ω_{e+1} . Dans la plupart des problèmes d'éléments finis, la matrice d'éléments finis est donc creuse. C'est le choix de fonctions de base avec un support limité à deux éléments qui est la cause de cette caractéristique fondamentale pour la résolution du système algébrique.

A titre d'exemple, considérons une fonction de forme globale linéaire et le tableau de correspondance entre noeuds globaux et locaux qui y est associé.

Eléments	Noeuds
\dots e $e + 1$ \dots	$i - 1 \quad i$ $i \quad i + 1$

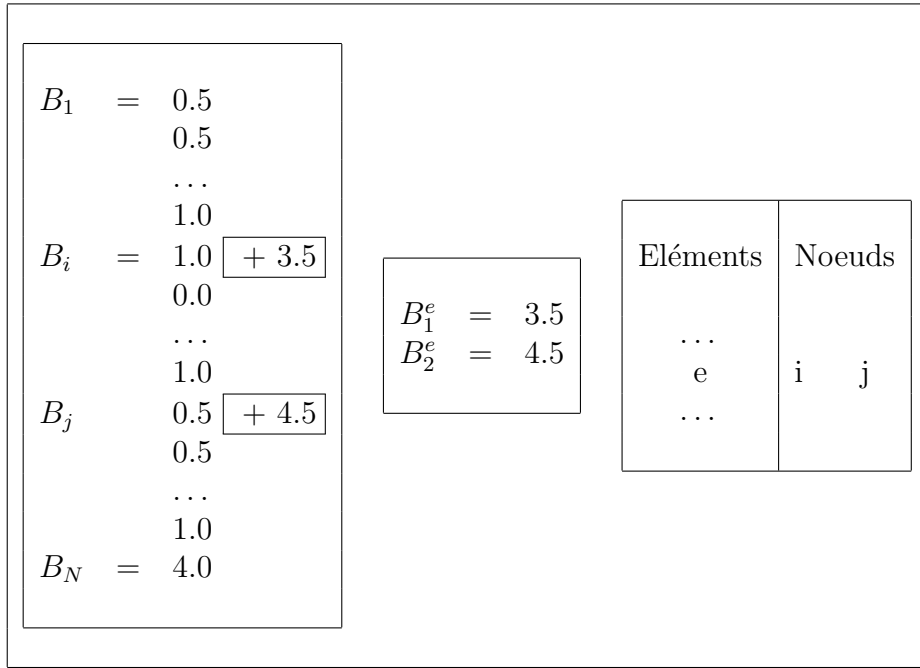
Nous obtenons pour un i donné (qui n'est pas sur la frontière), la composante B_i et trois uniques composantes non nulles de A_{ij} en vertu de l'isomorphisme entre l'élément parent et un élément quelconque.

$$\begin{aligned}
A_{i \ i-1} &= \int_{\Omega_e} \phi_{2,x}^e(x) \phi_{1,x}^e(x) dx, \\
A_{ii} &= \int_{\Omega_e} \phi_{2,x}^e(x) \phi_{2,x}^e(x) dx + \int_{\Omega_{e+1}} \phi_{1,x}^{e+1}(x) \phi_{1,x}^{e+1}(x) dx, \\
A_{i \ i+1} &= \int_{\Omega_{e+1}} \phi_{1,x}^{e+1}(x) \phi_{2,x}^{e+1}(x) dx, \\
B_i &= \int_{\Omega_e} \phi_2^e(x) f(x) dx + \int_{\Omega_{e+1}} \phi_1^{e+1}(x) f(x) dx.
\end{aligned} \tag{2.27}$$

Les équations (2.27) suggèrent un moyen facile de calculer les composantes A_{ij} et B_i . Au lieu de les calculer sur tout le domaine, nous pouvons considérer séparément chaque élément. Calculons des *matrices de raideur locales* et des *vecteurs de forces nodales locaux* définis par

$$\begin{aligned}
A_{ij}^e &= \int_{\Omega_e} \phi_{i,x}^e(x) \phi_{j,x}^e(x) dx, \\
B_i^e &= \int_{\Omega_e} f(x) \phi_i^e(x) dx.
\end{aligned} \tag{2.28}$$

De la relation (2.27), nous voyons que A_{11}^e est une partie de $A_{i-1 \ i-1}$, A_{22}^e est une partie de A_{ii} , etc. En d'autres mots, lorsqu'on construit une matrice de raideur locale pour un élément, on construit une partie de la matrice globale qui doit être assemblée sur la base de la table de correspondance entre noeuds. La matrice et le vecteur globaux peuvent donc être obtenus par ajouts successifs des matrices et des vecteurs locaux. Le procédé d'assemblage d'un vecteur local au sein du vecteur global est illustré par le schéma suivant :



Nous obtenons ainsi pour un indice i fixé la matrice et le vecteur globaux de notre exemple,

$$\begin{aligned}
 A_{i \ i-1} &= A_{21}^e, \\
 A_{ii} &= A_{22}^e + A_{11}^{e+1}, \\
 A_{i \ i+1} &= A_{12}^{e+1}, \\
 B_i &= B_2^e + B_1^{e+1}.
 \end{aligned} \tag{2.29}$$

En dehors des opérations logiques liées à l'assemblage de la matrice et du vecteur globaux, le calcul se réduit à l'évaluation d'intégrales de produits de fonctions connues sur les éléments.

Une procédure habituelle est de réduire le domaine d'intégration à l'élément parent $\hat{\Omega}$ et d'utiliser la transformation entre un élément quelconque et l'élément parent. Nous obtenons ainsi l'ensemble des matrices locales et des vecteurs locaux,

$$\begin{aligned}
A_{ij}^e &= \int_{\Omega_e} \phi_{i,x}^e(x) \phi_{j,x}^e(x) dx, \\
&= \int_{-1}^1 \left(\phi_{i,\xi}(\xi) \frac{d\xi}{dx} \right) \left(\phi_{j,\xi}(\xi) \frac{d\xi}{dx} \right) \left(\frac{dx}{d\xi} d\xi \right), \\
&= \int_{-1}^1 \phi_{i,\xi}(\xi) \phi_{j,\xi}(\xi) \frac{d\xi}{dx} d\xi, \\
&= \frac{2}{(X_{e+1} - X_e)} \int_{-1}^1 \phi_{i,\xi}(\xi) \phi_{j,\xi}(\xi) d\xi, \\
B_i^e &= \int_{\Omega_e} \phi_i^e(x) f(x) dx, \\
&= \frac{(X_{e+1} - X_e)}{2} \int_{-1}^1 \phi_i(\xi) f(x(\xi)) d\xi.
\end{aligned} \tag{2.30}$$

Exemple numérique

Considérons le cas où $f(x) = x$. On obtient immédiatement la solution analytique,

$$u(x) = \frac{x(1 - x^2)}{6}. \tag{2.31}$$

Pour obtenir la solution du problème par éléments finis, divisons le domaine Ω en N_1 éléments de longueur égale. Les sommets sont numérotés de 1 à $N_0 = N_1 + 1$. La longueur h de chaque élément et la position X_i de chaque sommet sont données par

$$\begin{aligned}
h &= \frac{1}{N_1}, \\
X_i &= (i - 1)h.
\end{aligned} \tag{2.32}$$

Pour la simplicité, adoptons l'élément linéaire continu. Les noeuds sont les sommets du maillage et le nombre de valeurs nodales est $N = N_0$. Les fonctions de forme sur l'élément parent ainsi que leurs dérivées sont :

$$\begin{aligned}
\phi_1(\xi) &= (1 - \xi)/2, & \phi_{1,\xi}(\xi) &= -1/2, \\
\phi_2(\xi) &= (1 + \xi)/2, & \phi_{2,\xi}(\xi) &= 1/2.
\end{aligned} \tag{2.33}$$

En vertu de (2.30) nous obtenons pour l'élément Ω_e la matrice locale,

$$A_{ij}^e = \frac{1}{h} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix},$$

Afin de calculer aisément les vecteurs locaux, nous écrivons la fonction source $f(x(\xi)) = x(\xi)$ sur l'élément Ω_e sous la forme :

$$\begin{aligned} x(\xi) &= \xi \frac{(X_{e+1} - X_e)}{2} + \frac{(X_{e+1} + X_e)}{2}, \\ &= X_e \frac{1 - \xi}{2} + X_{e+1} \frac{1 + \xi}{2}, \\ &= X_e \phi_1(\xi) + X_{e+1} \phi_2(\xi). \end{aligned}$$

Ce qui permet d'écrire pour l'ensemble des éléments

$$B_i^e = \frac{h}{2} \begin{bmatrix} X_e \int_{-1}^1 \phi_1(\xi) \phi_1(\xi) d\xi + X_{e+1} \int_{-1}^1 \phi_1(\xi) \phi_2(\xi) d\xi \\ X_e \int_{-1}^1 \phi_2(\xi) \phi_1(\xi) d\xi + X_{e+1} \int_{-1}^1 \phi_2(\xi) \phi_2(\xi) d\xi \end{bmatrix},$$

ou finalement en termes de coordonnées des sommets

$$B_i^e = \frac{h}{6} \begin{bmatrix} 2X_e + X_{e+1} \\ X_e + 2X_{e+1} \end{bmatrix}.$$

Examinons d'abord le cas $N = 3$ où $h = 0.5$. Les matrices et vecteurs locaux des deux éléments sont donnés par

$$\begin{aligned} A_{ij}^1 &= \frac{1}{h} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, & B_i^1 &= \frac{h}{6} \begin{bmatrix} 1/2 \\ 1 \end{bmatrix}, \\ A_{ij}^2 &= \frac{1}{h} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, & B_i^2 &= \frac{h}{6} \begin{bmatrix} 2 \\ 5/2 \end{bmatrix}. \end{aligned}$$

Le système global avant l'imposition des conditions essentielles est ensuite assemblé

$$\frac{1}{h} \begin{bmatrix} 1 & -1 & \\ -1 & 2 & -1 \\ & -1 & 1 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix} = \frac{h}{6} \begin{bmatrix} 1/2 \\ 6/2 \\ 5/2 \end{bmatrix}. \quad (2.34)$$

La valeur de u^h étant fixée au noeud 1 et au noeud 3, seule la seconde équation est conservée. On obtient ainsi la valeur $U_2 = h^2/4 = 1/16$, qui est égale à la valeur analytique

au milieu de la corde. Ce résultat est inattendu, car les fonctions de forme sont une base des polynômes de degré 1 alors que la solution est de degré 3. En fait, pour les problèmes unidimensionnels, on peut prouver que les valeurs nodales exactes sont obtenues lorsque les fonctions de forme satisfont l'équation différentielle homogène, ce qui est bien le cas ici. Toutefois, l'interpolation entre les valeurs nodales est linéaire et s'écarte sensiblement de la solution analytique.

Si le nombre de noeuds passe à 5, on obtient $h = 0.25$ et le système non contraint est :

$$\frac{1}{h} \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \\ U_5 \end{bmatrix} = \frac{h}{6} \begin{bmatrix} 1/4 \\ 6/4 \\ 12/4 \\ 18/4 \\ 11/4 \end{bmatrix}. \quad (2.35)$$

Après l'imposition des conditions aux frontières essentielles, on obtient le système

$$\frac{1}{h} \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & \end{bmatrix} \begin{bmatrix} U_2 \\ U_3 \\ U_4 \end{bmatrix} = \frac{h}{6} \begin{bmatrix} 6/4 \\ 12/4 \\ 18/4 \end{bmatrix},$$

dont la solution nous donne

$$\begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \\ U_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 5/128 \\ 8/128 \\ 7/128 \\ 0 \end{bmatrix}.$$

Les solutions analytique et approchée sont dessinées sur la figure 2.2. De manière surprenante, la solution numérique et la solution analytique coïncident aux noeuds : cette propriété qui n'existe que pour les problèmes unidimensionnels est une sorte de miracle numérique qui ne se reproduira plus dans les problèmes bidimensionnels.

Exemple numérique : condition naturelle

Supposons que nous souhaitions imposer une condition naturelle sur la dérivée première $u_{,x}(1) = G$, au lieu d'imposer la valeur de l'inconnue à l'extrémité droite du domaine. Ce nouveau problème a comme solution analytique

$$u(x) = \frac{x(3 - x^2)}{6}. \quad (2.36)$$

Il ne faut désormais fixer qu'une seule valeur nodale sur base de l'unique condition essentielle $u(0) = 0$

$$U_1 = 0, \quad (2.37)$$

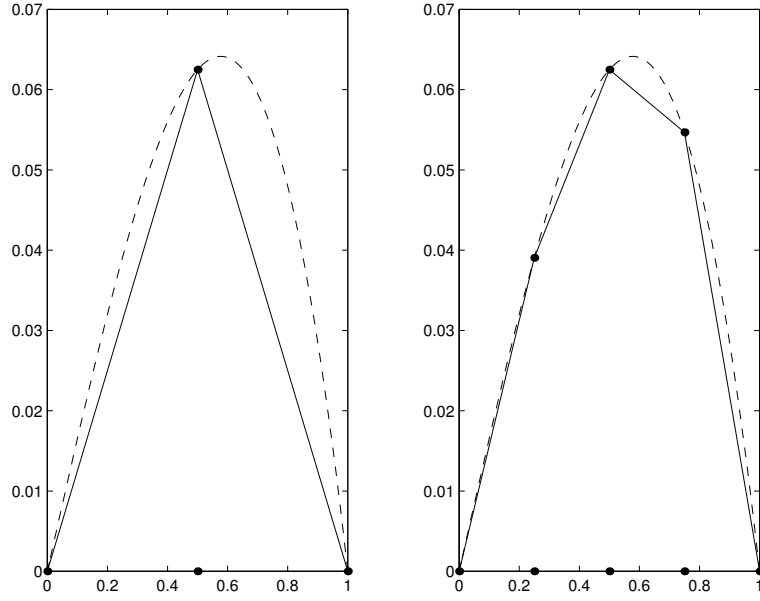


Figure 2.2: Superposition de la solution analytique du problème $u,_{xx} + x = 0$ avec $u(0) = u(1) = 0$ et des solutions discrètes avec 2 et 4 éléments finis linéaires continus respectivement. Les valeurs nodales des solutions discrètes sont exactes dans ce cas particulier.

et obtenons le système algébrique de $N - 1$ inconnues qui est presque identique à (2.24). L'unique différence est la présence d'une équation supplémentaire pour l'inconnue U_N . Cette dernière équation est un tout petit peu différente des autres...

$$\begin{aligned}
\int_{\Omega} \tau_N r^h dx &= \int_{\Omega} \tau_N \frac{d^2 u^h}{dx^2} dx + \int_{\Omega} \tau_N f dx, \\
&= - \int_{\Omega} \frac{d\tau_N}{dx} \frac{du^h}{dx} dx + \int_{\Omega} \tau_N f dx + \left[\tau_N \frac{du^h}{dx} \right]_0^1, \\
&= - \sum_{j=2}^N \left(\int_{\Omega} \frac{d\tau_N}{dx} \frac{d\tau_j}{dx} dx \right) U_j + \int_{\Omega} \tau_N f dx + \tau_N|_{x=1} \frac{du^h}{dx} \Big|_{x=1}, \\
&= - \sum_{j=2}^N A_{Nj} U_j + B_N + G.
\end{aligned}$$

On constate que la dérivée première en $x = 1$ apparaît “naturellement” dans l'équation supplémentaire. Pour imposer la condition naturelle, il a suffit d'y substituer la valeur requise. Pour simplifier encore, supposons maintenant que $G = 0$.

Lorsque le domaine est divisé en deux éléments, nous retournons aux équations (2.34)

et conservons comme inconnues U_2 et U_3 . Le système devient

$$\frac{1}{h} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} U_2 \\ U_3 \end{bmatrix} = \frac{h}{6} \begin{bmatrix} 6/2 \\ 5/2 \end{bmatrix},$$

dont la solution se compose de $U_2 = 11/48$ et $U_3 = 1/3$ qui sont précisément les valeurs analytiques. La solution par éléments finis est représentée à la figure 2.3. Bien que la solution numérique aux noeuds soit identique à la solution analytique, la condition naturelle est loin d'être satisfaite. Lorsque le domaine est divisé en 4 éléments on obtient

$$\frac{1}{h} \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \\ & & -1 & 1 \end{bmatrix} \begin{bmatrix} U_2 \\ U_3 \\ U_4 \\ U_5 \end{bmatrix} = \frac{h}{6} \begin{bmatrix} 6/4 \\ 12/4 \\ 18/4 \\ 11/4 \end{bmatrix},$$

dont la solution est

$$\begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \\ U_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 47/384 \\ 11/48 \\ 39/128 \\ 1/3 \end{bmatrix}.$$

Nous retrouvons à nouveau la solution analytique. On voit également que la condition naturelle est mieux satisfaite.

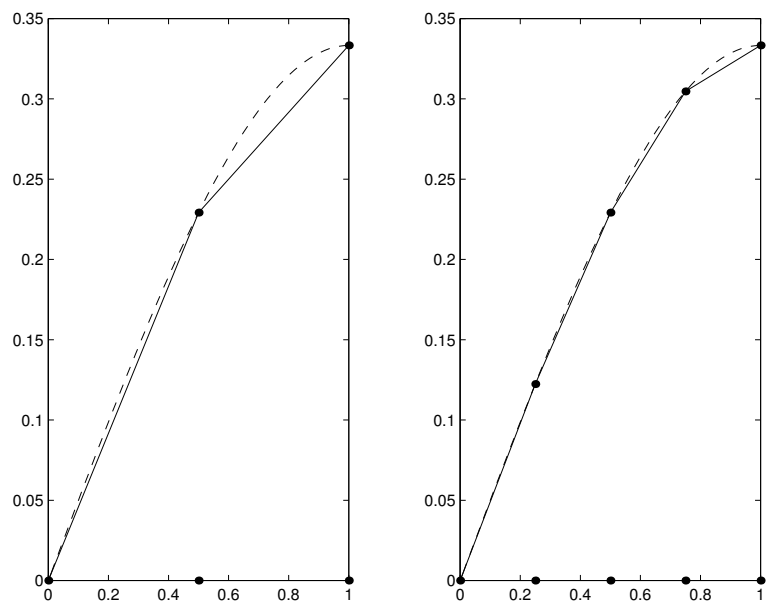


Figure 2.3: Superposition de la solution analytique du problème $u_{,xx} + x = 0$ avec $u(0) = u_{,x}(1) = 0$ et des solutions discrètes avec 2 et 4 éléments finis linéaires continus respectivement. Les valeurs nodales des solutions discrètes sont aussi exactes dans ce cas particulier.

2.3 Eléments finis bidimensionnels

Pour la résolution numérique du système discret (2.16) obtenu par une méthode d'éléments finis, il est nécessaire d'évaluer les coefficients de la matrice de raideur locale et du vecteur local des forces nodales. Dans cette section, nous montrons comment ces intégrations sont effectuées en pratique. Pour des éléments quelconques, il ne sera pas toujours possible de calculer ces intégrales de manière analytique et il faudra parfois recourir à une intégration numérique.

2.3.1 Construction du système algébrique

Il s'agit donc de calculer les coefficients de la matrice de raideur locale associée à l'équation de Poisson sur un élément quelconque Ω_e .

$$A_{ij}^e = \int_{\Omega_e} \nabla \phi_i^e \cdot \nabla \phi_j^e d\Omega \quad (2.38)$$

Nous allons omettre les suffixes e dans les fonctions de forme pour éviter d'alourdir les notations. Il y a a priori deux aspects gênants dans le calcul des coefficients de ce type de matrice :

- Il faut effectuer l'intégration de surface sur un élément aux frontières quelconques (éventuellement courbes).
- On ne dispose pas d'une expression du gradient des fonctions de forme par rapport aux coordonnées.

Pour effectuer avec plus de facilité l'intégration et contourner la première difficulté, il suffit d'observer qu'il vaut mieux effectuer cette intégration sur l'élément parent $\hat{\Omega}$, plutôt que sur l'élément quelconque Ω_e . Il faudra toutefois tenir compte de ce que l'on appelle le jacobien de la transformation. Considérons donc la définition de la transformation entre l'élément parent $\hat{\Omega}$ et l'élément quelconque Ω_e .

$$\begin{array}{c} \mathbf{x}(\boldsymbol{\xi}) = \sum_{i=1}^n \mathbf{X}_i^e \phi_i^x(\boldsymbol{\xi}), \\ \downarrow \\ d\mathbf{x}(\boldsymbol{\xi}) = \underbrace{\sum_{i=1}^n \mathbf{X}_i^e \frac{\partial \phi_i^x}{\partial \boldsymbol{\xi}}(\boldsymbol{\xi})}_{\frac{\partial \mathbf{x}}{\partial \boldsymbol{\xi}}(\boldsymbol{\xi})} \cdot d\boldsymbol{\xi}, \end{array}$$

où le gradient de la transformation est défini par :

$$\frac{\partial \mathbf{x}}{\partial \boldsymbol{\xi}}(\boldsymbol{\xi}) = \begin{bmatrix} \frac{\partial x}{\partial \xi}(\xi, \eta) & \frac{\partial x}{\partial \eta}(\xi, \eta) \\ \frac{\partial y}{\partial \xi}(\xi, \eta) & \frac{\partial y}{\partial \eta}(\xi, \eta) \end{bmatrix} \quad (2.39)$$

Il est important ici de bien observer la manière dont le gradient d'un vecteur est défini dans les équations précédentes⁷. Le déterminant J_e du gradient de la transformation de l'élément parent vers un élément Ω_e est appelé le jacobien de la transformation :

$$J_e(\boldsymbol{\xi}) = \det \left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\xi}}(\boldsymbol{\xi}) \right) = \det \begin{bmatrix} \frac{\partial x}{\partial \xi}(\xi, \eta) & \frac{\partial x}{\partial \eta}(\xi, \eta) \\ \frac{\partial y}{\partial \xi}(\xi, \eta) & \frac{\partial y}{\partial \eta}(\xi, \eta) \end{bmatrix} \quad (2.40)$$

Il est important de noter qu'en toute généralité, la transformation n'est pas affine et que le jacobien n'est pas constant.

Rapport des aires élémentaires

Pour pouvoir remplacer une intégration sur l'élément Ω_e par une intégration sur l'élément parent, il faut connaître le rapport entre des aires élémentaires.

⁷C'est ici que vous découvrez tout le problème de notation que pose le gradient d'un vecteur... Ceci est une petite note spécialement écrite pour Fabrice Loix que ce problème intéresse tout particulièrement. Ecrivons la relation

$$d\mathbf{v} = \frac{\partial \mathbf{v}}{\partial \mathbf{x}} \cdot d\mathbf{x}, \quad \text{ou en notation indicée...} \quad dv_i = \frac{\partial v_i}{\partial x_j} dx_j.$$

Le premier indice est alors associé naturellement à la composante du vecteur que l'on dérive. Mais, on peut aussi se représenter le gradient d'un vecteur comme le produit tensoriel de ∇ avec \mathbf{v}

$$\frac{\partial \mathbf{v}}{\partial \mathbf{x}} = \nabla \mathbf{v} = \frac{\partial v_j}{\partial x_i},$$

on est alors tenté d'établir la convention inverse. Ce qui est parfaitement légitime à partir du moment où celui qui écrit et celui qui lit l'équation sont d'accord entre eux...

La situation se corse pour l'exemple suivant : que signifient les termes $\mathbf{v} \cdot (\nabla \mathbf{v})$ ou $(\nabla \mathbf{v}) \cdot \mathbf{v}$? Des scientifiques utilisent ces deux notations pour désigner exactement le terme d'inertie des équations de Navier-Stokes. En fonction de la convention choisie, on optera pour l'une ou l'autre option pour écrire le terme adéquat.

$$d\Omega = \text{norme} \left(\left(\frac{\partial \mathbf{x}}{\partial \xi} d\xi \right) \times \left(\frac{\partial \mathbf{x}}{\partial \eta} d\eta \right) \right),$$

\downarrow Par définition du produit vectoriel,

$$d\Omega = \underbrace{\det \left(\frac{\partial \mathbf{x}}{\partial \xi} \right)}_{J_e} \underbrace{d\xi d\eta}_{d\hat{\Omega}},$$

Calcul du gradient des fonctions de forme

Le procédé de calcul du gradient des fonctions de forme sur un élément quelconque s'obtient en observant que :

$$\begin{aligned}
 \frac{\partial \phi_i}{\partial \mathbf{x}} &= \frac{\partial \phi_i}{\partial \xi} \cdot \frac{\partial \xi}{\partial \mathbf{x}} \\
 &= \frac{\partial \phi_i}{\partial \xi} \cdot \left(\frac{\partial \mathbf{x}}{\partial \xi} \right)^{-1} \\
 &= \frac{\partial \phi_i}{\partial \xi} \cdot \left(\sum_{i=1}^n \mathbf{x}_i^e \frac{\partial \phi_i^x}{\partial \xi}(\xi) \right)^{-1}
 \end{aligned}$$

On peut en déduire immédiatement une procédure de calcul, il suffit d'utiliser successivement les relations suivantes

- Calcul du gradient de la transformation :

$$\begin{aligned}
 \frac{\partial \mathbf{x}}{\partial \xi} &= \sum_{i=1}^n \mathbf{x}_i^e \frac{\partial \phi_i^x}{\partial \xi} \\
 &\downarrow
 \end{aligned}$$

$$\begin{bmatrix} \frac{\partial x}{\partial \xi}(\xi, \eta) & \frac{\partial x}{\partial \eta}(\xi, \eta) \\ \frac{\partial y}{\partial \xi}(\xi, \eta) & \frac{\partial y}{\partial \eta}(\xi, \eta) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n X_i^e \frac{\partial \phi_i^x}{\partial \xi}(\xi, \eta) & \sum_{i=1}^n X_i^e \frac{\partial \phi_i^x}{\partial \eta}(\xi, \eta) \\ \sum_{i=1}^n Y_i^e \frac{\partial \phi_i^x}{\partial \xi}(\xi, \eta) & \sum_{i=1}^n Y_i^e \frac{\partial \phi_i^x}{\partial \eta}(\xi, \eta) \end{bmatrix}$$

- Calcul du jacobien de la transformation

$$\begin{aligned}
 J_e &= \det \left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\xi}} \right), \\
 &\downarrow \\
 J_e(\xi, \eta) &= \frac{\partial x}{\partial \xi}(\xi, \eta) \frac{\partial y}{\partial \eta}(\xi, \eta) - \frac{\partial x}{\partial \eta}(\xi, \eta) \frac{\partial y}{\partial \xi}(\xi, \eta).
 \end{aligned}$$

- Calcul de l'inverse du gradient de la transformation

$$\begin{aligned}
 \frac{\partial \boldsymbol{\xi}}{\partial \mathbf{x}} &= \left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\xi}} \right)^{-1} \\
 &\downarrow \\
 \begin{bmatrix} \frac{\partial \xi}{\partial x}(\xi, \eta) & \frac{\partial \xi}{\partial y}(\xi, \eta) \\ \frac{\partial \eta}{\partial x}(\xi, \eta) & \frac{\partial \eta}{\partial y}(\xi, \eta) \end{bmatrix} &= \frac{1}{J_e(\xi, \eta)} \begin{bmatrix} \frac{\partial y}{\partial \eta}(\xi, \eta) & -\frac{\partial x}{\partial \eta}(\xi, \eta) \\ -\frac{\partial y}{\partial \xi}(\xi, \eta) & \frac{\partial x}{\partial \xi}(\xi, \eta) \end{bmatrix}
 \end{aligned}$$

- Calcul du gradient des fonctions de forme

$$\begin{aligned}
 \frac{\partial \phi_i}{\partial \mathbf{x}} &= \frac{\partial \phi_i}{\partial \boldsymbol{\xi}} \cdot \frac{\partial \boldsymbol{\xi}}{\partial \mathbf{x}} \\
 &\downarrow \\
 \begin{bmatrix} \frac{\partial \phi_i}{\partial x}(\xi, \eta) \\ \frac{\partial \phi_i}{\partial y}(\xi, \eta) \end{bmatrix}^T &= \begin{bmatrix} \frac{\partial \phi_i}{\partial \xi}(\xi, \eta) \\ \frac{\partial \phi_i}{\partial \eta}(\xi, \eta) \end{bmatrix}^T \cdot \begin{bmatrix} \frac{\partial \xi}{\partial x}(\xi, \eta) & \frac{\partial \xi}{\partial y}(\xi, \eta) \\ \frac{\partial \eta}{\partial x}(\xi, \eta) & \frac{\partial \eta}{\partial y}(\xi, \eta) \end{bmatrix}
 \end{aligned}$$

On observe en particulier qu'il n'a pas été nécessaire de construire la transformation inverse $\boldsymbol{\xi}(\mathbf{x})$. Si la transformation entre (x, y) et (ξ, η) est supposée bijective, il n'existe pas en général d'expression analytique de $(\xi(x, y), \eta(x, y))$. Au passage, on observe qu'il est indispensable que le jacobien ne s'annule en aucun point pour que la transformation soit bijective. Il est facile de montrer que c'est le cas pour des triangles et des quadrilatères convexes.

Finalement, le calcul des coefficients de la matrice (2.38) se réduit à intégrer une

fonction un peu plus compliquée sur le triangle parent.

$$\begin{aligned}
A_{ij}^e &= \int_{\Omega_e} \nabla \phi_i^e \cdot \nabla \phi_j^e d\Omega \\
&= \int_{\hat{\Omega}} \underbrace{\left(\frac{\partial \phi_i}{\partial x}(\xi, \eta) \frac{\partial \phi_j}{\partial x}(\xi, \eta) + \frac{\partial \phi_i}{\partial y}(\xi, \eta) \frac{\partial \phi_j}{\partial y}(\xi, \eta) \right)}_{F(\xi, \eta)} J_e(\xi, \eta) d\hat{\Omega}.
\end{aligned} \tag{2.41}$$

Pour ce type d'intégration, il faut recourir aux techniques de l'intégration numérique de Gauss-Legendre et de Hammer sur le carré ou le triangle parent respectivement. Nous rappelons brièvement ces techniques ci-après.

2.3.2 Triangles de Turner pour l'équation de Poisson

Considérons un maillage d'éléments triangulaires linéaires continus (aussi appelés éléments de Turner). Nous souhaitons résoudre l'équation de Poisson sur celui-ci. Soit u la fonction que nous désirons approcher par la solution discrète u^h de la méthode d'éléments finis. Nous cherchons ici les valeurs nodales U_i aux sommets des éléments triangulaires, et nous admettons que l'interpolation linéaire entre ces valeurs nodales constitue le type d'approximation que nous désirons utiliser. La force est décrite par une approximation constante F_i sur une partie des éléments, et est supposée nulle sur les autres éléments.

Les conditions frontières sont décrites comme suit. Sur une partie des noeuds, les valeurs T_i sont données et sur une partie des segments, le flux est imposé à une valeur constante G_i . Sur les segments frontières où on ne prescrit ni une condition essentielle, ni une condition naturelle non-homogène (non nulle), une condition naturelle homogène (nulle) sera appliquée par défaut. Typiquement, les données de ce problème peuvent être écrites sous la forme suivante

- Description des triangles : tableau d'appartenance des sommets aux éléments. Les sommets sont donnés suivant un parcours anti-horlogique de la frontière de l'élément.
- Description des sommets : coordonnées des sommets.
- Description de f^h : liste des éléments où F_i est non nul.
- Description de g^h : liste des segments où le flux approximé par la constante G_i est non nul. Les segments sont identifiés par le couple de leurs deux sommets, dont l'ordre est à nouveau fixé par un parcours anti-horlogique de la frontière du domaine.
- Description de t_h : liste des sommets où la valeur nodale U_i est contrainte à la valeur de T_i .

La construction des matrices locales et des vecteurs de forces généralisées se fait en appliquant la procédure que nous venons de décrire...

- Calcul des fonctions de forme sur l'élément parent

$$\begin{aligned}
\phi_1(\xi, \eta) &= -\xi - \eta + 1, & \frac{\partial \phi_1}{\partial \xi}(\xi, \eta) &= -1, & \frac{\partial \phi_1}{\partial \eta}(\xi, \eta) &= -1, \\
\phi_2(\xi, \eta) &= \xi, & \frac{\partial \phi_2}{\partial \xi}(\xi, \eta) &= 1, & \frac{\partial \phi_2}{\partial \eta}(\xi, \eta) &= 0, \\
\phi_3(\xi, \eta) &= \eta, & \frac{\partial \phi_3}{\partial \xi}(\xi, \eta) &= 0, & \frac{\partial \phi_3}{\partial \eta}(\xi, \eta) &= 1,
\end{aligned}$$

- Calcul du gradient de la transformation

$$\begin{aligned}
\frac{\partial \mathbf{x}}{\partial \boldsymbol{\xi}} &= \sum_{i=1}^n \mathbf{X}_i^e \frac{\partial \phi_i^x}{\partial \boldsymbol{\xi}} \\
&\downarrow \\
\begin{bmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} \end{bmatrix} &= \begin{bmatrix} X_2^e - X_1^e & X_3^e - X_1^e \\ Y_2^e - Y_1^e & Y_3^e - Y_1^e \end{bmatrix}
\end{aligned}$$

- Calcul du jacobien de la transformation

$$\begin{aligned}
J_e &= \det \left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\xi}} \right), \\
&\downarrow \\
J_e &= (X_2^e - X_1^e)(Y_3^e - Y_1^e) - (Y_2^e - Y_1^e)(X_3^e - X_1^e).
\end{aligned}$$

- Calcul de l'inverse du gradient de la transformation

$$\begin{aligned}
\frac{\partial \boldsymbol{\xi}}{\partial \mathbf{x}} &= \left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\xi}} \right)^{-1} \\
&\downarrow \\
\begin{bmatrix} \frac{\partial \xi}{\partial x} & \frac{\partial \xi}{\partial y} \\ \frac{\partial \eta}{\partial x} & \frac{\partial \eta}{\partial y} \end{bmatrix} &= \frac{1}{J_e} \begin{bmatrix} (Y_3^e - Y_1^e) & -(X_3^e - X_1^e) \\ -(Y_2^e - Y_1^e) & (X_2^e - X_1^e) \end{bmatrix}
\end{aligned}$$

- Calcul du gradient des fonctions de forme

$$\begin{aligned} \frac{\partial \phi_i}{\partial \mathbf{x}} &= \frac{\partial \phi_i}{\partial \boldsymbol{\xi}} \cdot \frac{\partial \boldsymbol{\xi}}{\partial \mathbf{x}} \\ &\downarrow \\ \begin{bmatrix} \frac{\partial \phi_1}{\partial x} & \frac{\partial \phi_1}{\partial y} \\ \frac{\partial \phi_2}{\partial x} & \frac{\partial \phi_2}{\partial y} \\ \frac{\partial \phi_3}{\partial x} & \frac{\partial \phi_3}{\partial y} \end{bmatrix} &= \begin{bmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \frac{1}{J_e} \begin{bmatrix} (Y_3^e - Y_1^e) & -(X_3^e - X_1^e) \\ -(Y_2^e - Y_1^e) & (X_2^e - X_1^e) \end{bmatrix} \end{aligned}$$

Il est dès lors facile d'obtenir les coefficients de la matrice locale :

$$\begin{aligned} A_{ij}^e &= \int_{\Omega_e} \frac{\partial \phi_i}{\partial x} \frac{\partial \phi_j}{\partial x} + \frac{\partial \phi_i}{\partial y} \frac{\partial \phi_j}{\partial y} d\Omega, \\ B_i^e &= \underbrace{F_i^e \int_{\Omega_e} \phi_i^e d\Omega}_{C_i^e} + \sum_{f=1}^3 G^f \underbrace{\int_{\Gamma_f} \phi_i^f(x, y) d\Gamma}_{D_i^f}, \end{aligned}$$

où F_i^e est l'approximation locale par une constante de la fonction f , tandis qu'on approxime par une constante G^f le flux imposé au travers d'un des trois côtés Γ_f de l'élément considéré. Ces intégrales n'interviendront que pour des éléments dont un ou plusieurs côtés est inclus dans la partie Γ_N de la frontière. On observe que les forces généralisées contiennent une contribution C_i^e de la source répartie f par unité d'aire et une contribution D_i^e du flux g le long de la frontière.

Comme les gradients des fonctions ϕ_i^e sont constants sur l'élément et que J_e vaut le double de la surface de l'élément, on obtient immédiatement

$$A_{ij}^e = \left(\frac{\partial \phi_i}{\partial x} \frac{\partial \phi_j}{\partial x} + \frac{\partial \phi_i}{\partial y} \frac{\partial \phi_j}{\partial y} \right) \frac{J_e}{2},$$

où les gradients des fonctions de forme sont donnés par

$$\begin{aligned} \phi_{1,x}^e &= (Y_2^e - Y_3^e)/J_e, & \phi_{1,y}^e &= (X_3^e - X_2^e)/J_e, \\ \phi_{2,x}^e &= (Y_3^e - Y_1^e)/J_e, & \phi_{2,y}^e &= (X_1^e - X_3^e)/J_e, \\ \phi_{3,x}^e &= (Y_1^e - Y_2^e)/J_e, & \phi_{3,y}^e &= (X_2^e - X_1^e)/J_e. \end{aligned} \tag{2.42}$$

Comme les fonctions ϕ_i^e forment des tétraèdres de hauteur unitaire et dont la base est l'élément, on obtient tout aussi facilement

$$C_i^e = \frac{F_i^e J_e}{6}.$$

Calculons ensuite la contribution du terme de flux. Considérons un segment Γ_f au travers duquel un flux constant G^f est imposé entre les noeuds de coordonnées (X_i^f, Y_i^f) avec $i = 1, 2$. La contribution au vecteur des forces est donnée par

$$D_i^f = G^f \int_{\Gamma_f} \phi_i^f(x, y) d\Gamma, \quad i = 1, 2.$$

où la constante G^f est une approximation locale du flux g sur le segment Γ_f . On obtient toujours aisément

$$D_i^f = G^f \frac{1}{2} \underbrace{\sqrt{(X_2^f - X_1^f)^2 + (Y_2^f - Y_1^f)^2}}_{J_f},$$

où le jacobien constant J_f peut être associé au segment Γ_f . Il s'agit bien du rapport entre sa longueur et celle du segment parent $]0, 1[$.

Il suffit ensuite de procéder à l'assemblage, à l'imposition des conditions essentielles et à la résolution du système. Ces étapes du calcul vont maintenant être illustrées par un exemple encore plus simple.

Exemple numérique vraiment élémentaire

Étudions la résolution du problème de l'équation de Poisson soumise à des conditions frontières de Dirichlet homogènes

$$\begin{aligned} \nabla^2 u(x, y) - 1 &= 0, & 0 < x < 2, \quad 0 < y < 2, \\ u(0, y) = u(2, y) &= 0, & 0 < y < 2, \\ u(x, 0) = u(x, 2) &= 0, & 0 < x < 2. \end{aligned} \tag{2.43}$$

La solution analytique du problème est donnée par

$$u(x, y) = - \sum_{m, n \text{ impairs}} \frac{64}{\pi^4 (m^2 + n^2) mn} \sin\left(\frac{m\pi x}{2}\right) \sin\left(\frac{n\pi y}{2}\right). \tag{2.44}$$

Ce problème possède d'importantes propriétés de symétrie. En fait, comme le montre la figure 2.4, il suffit de considérer un huitième du domaine d'intégration, avec cette fois, des conditions aux frontières naturelles le long de deux axes de symétrie.

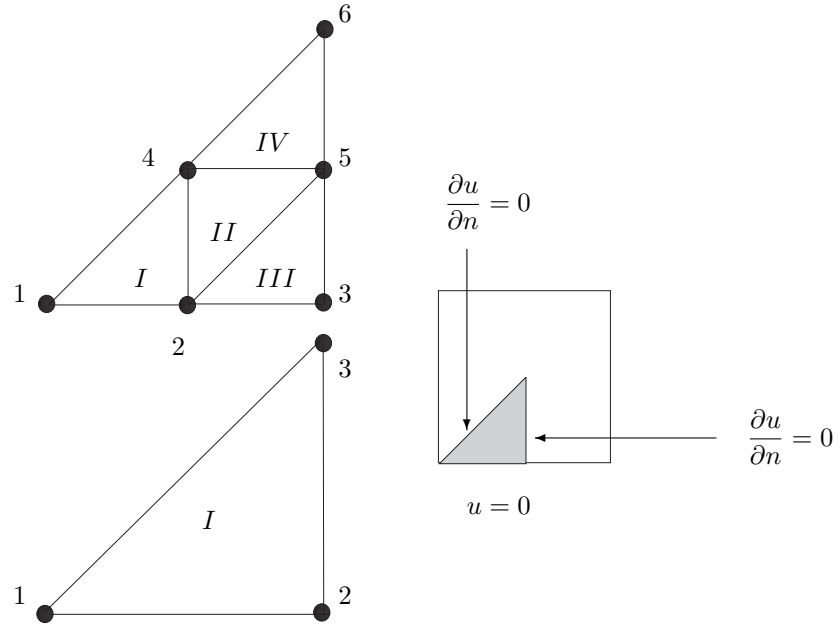


Figure 2.4: Domaine d'intégration et domaine de calcul réduit par symétrie. Deux maillages d'éléments finis sont construits sur le domaine réduit.

Considérons d'abord le cas simple du premier maillage où un seul élément est utilisé. Les données du problèmes sont réduites à

Triangle	Sommets			Sommet	X_i	Y_i	Sommet	T_i
1	1	2	3	1	0.0	0.0	1	0.0
				2	1.0	0.0	2	0.0
				3	1.0	1.0		

On observe que $J = 1$ et on obtient

$$A_{ij} = \begin{bmatrix} 1/2 & -1/2 & 0 \\ -1/2 & 1 & -1/2 \\ 0 & -1/2 & 1/2 \end{bmatrix}, \quad B_i = \begin{bmatrix} -1/6 \\ -1/6 \\ -1/6 \end{bmatrix}.$$

Les conditions essentielles nous imposent ici $U_1 = U_2 = 0$. En vertu de (2.56), la seule équation non triviale est

$$\frac{1}{2}U_3 = -\frac{1}{6}.$$

On trouve ainsi, avec un seul élément, $U_3 = -1/3$. La valeur analytique (2.44) à

l'intersection des diagonales est -0.2946 , soit une erreur relative de 13%.

Considérons ensuite le second maillage plus dense où les noeuds sont numérotés de 1 à 6. La composition des éléments, les coordonnées des noeuds et les conditions essentielles sont données par

Triangle	Sommets			Sommet	X_i	Y_i	Sommet	T_i
1	1	2	4	1	0.0	0.0	1	0.0
2	2	5	4	2	0.5	0.0	2	0.0
3	2	3	5	3	1.0	0.0	3	0.0
4	4	5	6	4	0.5	0.5		
				5	1.0	0.5		
				6	1.0	1.0		

On calcule facilement pour chaque élément $\phi_{i,x}^e$ et $\phi_{i,y}^e$. Tous les jacobiens J_e valent ici $1/4$. Les matrices de raideur locales et les vecteurs de forces nodales locaux sont donnés par

$$A_{ij}^1 = \begin{bmatrix} 1/2 & -1/2 & 0 \\ -1/2 & 1 & -1/2 \\ 0 & -1/2 & 1/2 \end{bmatrix}, \quad B_i^1 = \begin{bmatrix} -1/24 \\ -1/24 \\ -1/24 \end{bmatrix},$$

$$A_{ij}^2 = \begin{bmatrix} 1/2 & 0 & -1/2 \\ 0 & 1/2 & -1/2 \\ -1/2 & -1/2 & 1 \end{bmatrix}, \quad B_i^2 = \begin{bmatrix} -1/24 \\ -1/24 \\ -1/24 \end{bmatrix},$$

$$A_{ij}^3 = \begin{bmatrix} 1/2 & -1/2 & 0 \\ -1/2 & 1 & -1/2 \\ 0 & -1/2 & 1/2 \end{bmatrix}, \quad B_i^3 = \begin{bmatrix} -1/24 \\ -1/24 \\ -1/24 \end{bmatrix},$$

$$A_{ij}^4 = \begin{bmatrix} 1/2 & -1/2 & 0 \\ -1/2 & 1 & -1/2 \\ 0 & -1/2 & 1/2 \end{bmatrix}, \quad B_i^4 = \begin{bmatrix} -1/24 \\ -1/24 \\ -1/24 \end{bmatrix}.$$

Il faut ensuite procéder à l'assemblage dans la matrice globale de dimension 6×6 . Pour chaque matrice A_{ij}^e , on se réfère au tableau d'appartenance pour retrouver le terme correspondant de la matrice globale. On obtient finalement le système algébrique global non contraint

$$\begin{bmatrix} 1/2 & -1/2 & & & & \\ -1/2 & 2 & -1/2 & -1 & & \\ & -1/2 & 1 & & -1/2 & \\ & -1 & & 2 & -1 & \\ & & -1/2 & -1 & 2 & -1/2 \\ & & & & -1/2 & 1/2 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \\ U_5 \\ U_6 \end{bmatrix} = \begin{bmatrix} -1/24 \\ -1/8 \\ -1/24 \\ -1/8 \\ -1/8 \\ -1/24 \end{bmatrix}.$$

En tenant compte des conditions aux frontières essentielles $U_1 = U_2 = U_3 = 0$, il reste alors le système d'équations

$$\begin{bmatrix} 2 & -1 & \\ -1 & 2 & -1/2 \\ & -1/2 & 1/2 \end{bmatrix} \begin{bmatrix} U_4 \\ U_5 \\ U_6 \end{bmatrix} = \begin{bmatrix} -1/8 \\ -1/8 \\ -1/24 \end{bmatrix}.$$

On obtient ainsi $U_6 = -0.3125$, soit une erreur de 6% par rapport à la solution analytique. Lorsque le domaine est divisé en 16 éléments, on obtient $U_6 = -0.3013$ (erreur de 2.3%), et s'il est divisé en 64 éléments, on a $U_6 = -0.2969$ (erreur de 0.8%).

Dans cet exemple, nous avons une approximation du premier ordre pour des triangles à 3 noeuds, et du second ordre pour les triangles à 6 noeuds. En ce qui concerne la norme L^2 de l'erreur, nous devons nous attendre à un facteur h^2 pour les éléments linéaires et h^3 pour les éléments quadratiques. Le tableau ci-dessous montre que l'erreur évolue bien dans ce sens.

Eléments		Triangles linéaires			Triangles quadratiques		
N_1	h	N	$u^h(0)$	$e(0)$	N	$u^h(0)$	$e(0)$
1	1	3	-0.3333	13.1 %	6	-0.3000	1.83 %
4	1/2	6	-0.3125	6.1 %	15	-0.2950	0.14 %
16	1/4	15	-0.3013	2.3 %	45	-0.2947	0.03 %
64	1/8	45	-0.2969	0.8 %			

On voit que, pour un nombre d'éléments fixé, l'erreur par rapport à la solution analytique est plus grande pour les triangles à 3 noeuds, en vertu du plus grand nombre de noeuds disponibles pour les triangles quadratiques. En fait, même pour un nombre de noeuds égal, on constate que l'erreur est plus basse pour le triangle à 6 noeuds que pour le triangle à 3 noeuds. La figure 2.5, présente le graphe du logarithme de l'erreur locale en fonction du logarithme de h pour les éléments linéaires. On voit que la pente tend vers la valeur théorique -2 lorsque h diminue.

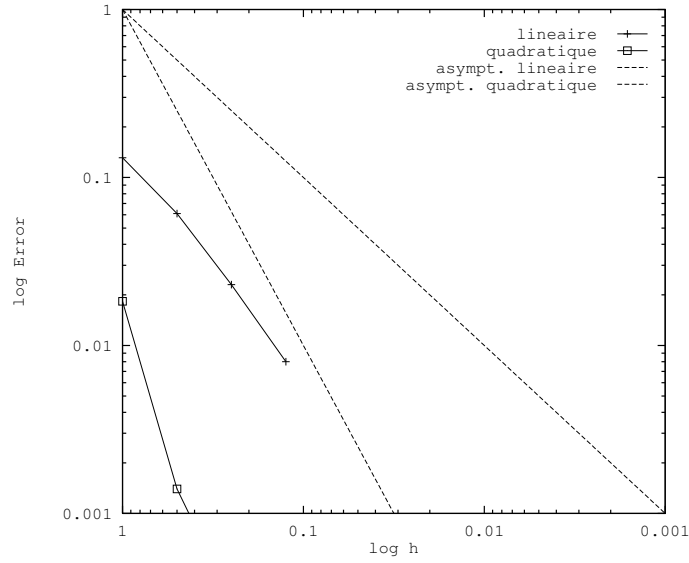


Figure 2.5: Taux de convergence des éléments triangulaires à 3 noeuds.

2.3.3 Résolution du système linéaire par élimination de Gauss

Dans une première étape, nous allons résoudre le système linéaire obtenu en faisant appel à une technique d'élimination gaussienne pour une matrice pleine. Ce n'est toutefois pas la meilleure stratégie car la matrice obtenue est creuse et il est donc plus indiqué d'utiliser un solveur spécialisé pour des matrices creuses.

Nous verrons aussi que des méthodes itératives sont souvent une alternative nettement plus avantageuse qu'une technique d'élimination gaussienne, même si celle-ci a aussi certains avantages.

Systèmes triangulaires

Si la matrice \mathbf{A} est *triangulaire supérieure*, le système $\mathbf{Au} = \mathbf{f}$ a la forme suivante

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ & a_{22} & \dots & a_{2n} \\ & & \dots & \\ & & & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{bmatrix}. \quad (2.45)$$

Si $a_{jj} \neq 0$, pour $j = 1, 2, \dots, n$, la solution du système (2.45) est donnée par le calcul successif de x_n, x_{n-1}, \dots, x_1 au moyen de la relation

$$x_j = (b_j - \sum_{k=j+1}^n a_{jk}x_k) / a_{jj} \quad (2.46)$$

Cette procédure porte le nom de *substitution arrière* (*backward substitution*). De la même façon, si \mathbf{A} est une matrice *triangulaire inférieure*, on peut calculer successivement x_1, x_2, \dots, x_n . Cette procédure est une *substitution directe* (*forward substitution*). Si un élément diagonal est nul, la procédure décrite ne peut pas convenir puisqu'elle implique alors une division par 0. Dans ce cas, le système ne présente pas de solution unique : il est impossible ou indéterminé.

Triangularisation

Deux systèmes linéaires sont dits *équivalents* si ces deux systèmes ont les mêmes solutions. On peut montrer que les opérations suivantes permettent de transformer un système en un autre système équivalent :

- la permutation de deux équations dans un système
- la multiplication d'une équation par un nombre non nul
- le remplacement d'une équation par la somme de cette équation et d'un multiple d'une autre équation

La résolution d'un système linéaire par *élimination de Gauss* est une méthode qui consiste à utiliser ces opérations élémentaires pour obtenir un système linéaire équivalent plus facile à résoudre, par exemple un système triangulaire. On parle dans ce cas de *triangularisation* du système.

Rappelons brièvement comment fonctionne l'élimination gaussienne pour le système suivant

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ & & \dots & \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{bmatrix}. \quad (2.47)$$

L'obtention d'un système triangulaire se fait en effectuant successivement des combinaisons linéaires des équations du système original afin d'obtenir un système triangulaire équivalent noté :

$$\underbrace{\begin{bmatrix} a_{11}^{(n)} & a_{12}^{(n)} & \dots & a_{1n}^{(n)} \\ & a_{22}^{(n)} & \dots & a_{2n}^{(n)} \\ & & \dots & \\ & & & a_{nn}^{(n)} \end{bmatrix}}_{\mathbf{A}^{(n)}\mathbf{x}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}}_{\mathbf{b}^{(n)}} = \begin{bmatrix} b_1^{(n)} \\ b_2^{(n)} \\ \dots \\ b_n^{(n)} \end{bmatrix} \quad (2.48)$$

L'obtention de $\mathbf{A}^{(k+1)}$ et $\mathbf{b}^{(k+1)}$ à partir de $\mathbf{A}^{(k)}$ et $\mathbf{b}^{(k)}$ est réalisé comme suit :

$$\boxed{\begin{aligned} a_{ij}^{(k+1)} &= a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)} & i = k+1, \dots, n, j = k, \dots, n \\ b_i^{(k+1)} &= b_i^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} b_k^{(k)} & i = k+1, \dots, n, \end{aligned}} \quad (2.49)$$

L'élimination gaussienne démarre avec $\mathbf{A}^{(1)} = \mathbf{A}$ et $\mathbf{b}^{(1)} = \mathbf{b}$ et fonctionne seulement si $a_{kk}^{(k)}$ ne s'annule jamais pendant la procédure. Graphiquement, une matrice $\mathbf{A}^{(k)}$ a l'allure suivante

$$\begin{bmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \dots & a_{1k}^{(k)} & \dots & a_{1n}^{(k)} \\ & a_{22}^{(k)} & \dots & a_{2k}^{(k)} & \dots & a_{2n}^{(k)} \\ & & & \dots & & \\ & & & a_{kk}^{(k)} & \dots & a_{kn}^{(k)} \\ & & & & \dots & \\ & & & a_{nk}^{(k)} & \dots & a_{nn}^{(k)} \end{bmatrix}$$

L'étape $k+1$ de l'élimination gaussienne sert à annuler les coefficients de la $k^{\text{ème}}$ colonne situés sous la diagonale. Le système triangulaire supérieur ainsi obtenu peut alors être très facilement résolu par la technique de *backward substitution*.

Implémenter l'élimination gaussienne n'est pas une chose très compliquée. La solution et le membre de droite peuvent être stockés dans le même vecteur et les lignes de code peuvent être écrites comme suit :

```
double *matrixSolve(double **A, double *B, int size)

{
    int    i, j, k;
```

```

/* Gauss elimination */
for (k=0; k < size; k++) {
    if ( A[k][k] == 0 ) Error("zero pivot");
    for (i = k+1 ; i < size; i++) {
        factor = A[i][k] / A[k][k];
        for (j = k+1 ; j < size; j++)
            A[i][j] = A[i][j] - A[k][j] * factor;
        B[i] = B[i] - B[k] * factor; }}

/* Back-substitution */
for (i = (size)-1; i >= 0 ; i--) {
    factor = 0;
    for (j = i+1 ; j < size; j++)
        factor += A[i][j] * B[j];
    B[i] = ( B[i] - factor)/A[i][i]; }
return(B);
}

```

Une telle implémentation est dite *en place* puisqu'on utilise les structures de données du système fourni pour calculer la solution. C'est économique d'un point de vue mémoire, mais il faut garder à l'esprit que les données sont perdues après l'exécution de la procédure.

Et la question du pivotement...

Dans la procédure décrite ci-dessus, on a travaillé de façon systématique, en gardant x_1 dans la première équation, x_2 dans la deuxième, ... On a donc pris les pivots dans l'ordre systématique a_{11} , a_{22} , ... Cet ordre n'est pas nécessairement le meilleur.

Par exemple, il pourrait arriver qu'un pivot soit nul. Dans ce cas, la procédure n'est plus utilisable puisque le multiplicateur m contient l'élément pivot au dénominateur. On peut alors essayer de permuter les équations de façon à obtenir un nouveau pivot non nul pour continuer la procédure. Si l'on n'y parvient pas, c'est que la matrice \mathbf{A} est singulière et le système ne présente pas une solution unique.

Cette méthode n'est évidemment applicable que si le terme A_{11} de la première équation est non nul. Il doit en être de même pour toutes les équations suivantes lors de l'élimination successive. On sait toutefois par un théorème d'algèbre que tous les éléments diagonaux d'une matrice définie positive sont et restent positifs durant l'élimination. La méthode de résolution proposée est dès lors valable lorsque la méthode des éléments finis est basée sur un principe de minimisation d'une forme définie positive. Le théorème est d'ailleurs utilisé en cours d'exécution du programme. Si un élément diagonal s'avère être négatif ou nul, on peut y trouver deux raisons : le programme contient une erreur, ou les conditions essentielles insérées dans le programme sont insuffisantes. Ce dernier cas se produit par exemple si seules des conditions naturelles sont imposées pour résoudre l'équation de Poisson. Une fonction linéaire en x et y est une solution de l'équation homogène et les

conditions aux frontières doivent lever l'indétermination.

Si l'on travaille avec un nombre limité de décimales, l'utilisation d'un pivot trop petit peut faire apparaître des erreurs d'arrondi qui peuvent aller jusqu'à compromettre complètement le résultat obtenu. Pour réduire l'effet de ces erreurs d'arrondi, une règle de bonne pratique consiste à utiliser à chaque étape le plus grand pivot possible (en valeur absolue). Cette stratégie s'appelle le *pivotement partiel*. Comme \mathbf{A} est une matrice définie positive, on peut démontrer qu'il n'est alors ni nécessaire, ni utile d'effectuer un pivotement partiel pour éviter des instabilités numériques. En d'autres mots, il sera toujours possible d'effectuer l'élimination gaussienne dans n'importe quel ordre !

2.3.4 Traitement de conditions essentielles inhomogènes

La solution approchée de notre problème elliptique modèle (2.1) peut être construite sur base de fonctions de forme globales de la manière suivante :

$$u(\mathbf{x}) \approx u^h(\mathbf{x}) = \sum_{i \in \mathcal{I}} U_i \tau_i(\mathbf{x}). \quad (2.50)$$

La liste des indices des N noeuds définit l'ensemble $\mathcal{I} = \{\infty, \epsilon, \dots, \mathcal{N}\}$ au sein duquel nous distinguons deux sous-ensembles :

- $\mathcal{I}_{\mathcal{D}}$: la liste des indices des $N_{\mathcal{D}}$ noeuds dont les valeurs nodales sont prescrites par les conditions essentielles. Ils forment l'équivalent discret de $\Gamma_{\mathcal{D}}$. Les indices des noeuds n'appartenant pas à la liste $\mathcal{I}_{\mathcal{D}}$ forment la liste complémentaire $\overline{\mathcal{I}_{\mathcal{D}}}$.
- $\mathcal{I}_{\mathcal{N}}$: la liste des indices des $N_{\mathcal{N}}$ noeuds dont les valeurs nodales sont sujets à une condition naturelle. Ils sont évidemment associés à $\Gamma_{\mathcal{N}}$. On définit également la liste complémentaire $\overline{\mathcal{I}_{\mathcal{N}}}$.

Le respect des conditions de Dirichlet du problème (2.1) s'obtient en exigeant

$$U_i = T_i \quad i \in \mathcal{I}_{\mathcal{D}} \quad (2.51)$$

où T_i est par exemple la valeur $t(\mathbf{X}_i)$. Cela assure un strict respect des conditions essentielles aux noeuds de l'ensemble $\mathcal{I}_{\mathcal{D}}$. Par contre, observons que la valeur imposée entre les noeuds n'est en général qu'approximativement égale à la fonction t . En d'autres mots, nous avons remplacé la fonction t par l'approximation t^h

$$t(\mathbf{x}) \approx t^h(\mathbf{x}) = \sum_{i \in \mathcal{I}_{\mathcal{D}}} T_i \tau_i(\mathbf{x}), \quad (2.52)$$

où le fait de sélectionner $T_i = t(\mathbf{X}_i)$ n'est pas l'unique possibilité qui nous soit offerte, ni

la meilleure du point de vue de la précision. Toutefois, c'est en pratique la solution la plus souvent utilisée. De manière similaire, si nous ne disposons que de valeurs nodales pour les fonctions f et g , nous utilisons en fait des approximations f^h et g^h , qui seront aussi la cause d'erreurs additionnelles dans la résolution numérique.

Pour déterminer les valeurs nodales restantes, nous souhaitons toujours minimiser la fonctionnelle $J(v)$. Le nombre de degrés de liberté disponibles est maintenant $N - N_D$, et on peut alors obtenir les $N - N_D$ équations algébriques permettant de déterminer ces degrés de liberté en recherchant les conditions de stationnarité de la fonctionnelle J pour $v = u^h(\mathbf{x})$:

$$\begin{aligned}
J(u^h) &= \frac{1}{2} \left\langle \left(\sum_{i \in \mathcal{I}} U_i \nabla \tau_i \right) \cdot a \left(\sum_{j \in \mathcal{I}} U_j \nabla \tau_j \right) \right\rangle & - & \left\langle f \left(\sum_{i \in \mathcal{I}} U_i \tau_i \right) \right\rangle \\
& & - & \ll g \left(\sum_{i \in \mathcal{I}_N} U_i \tau_i \right) \gg, \\
J(u^h) &= \frac{1}{2} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{I}} \left\langle \nabla \tau_i \cdot a(\nabla \tau_j) \right\rangle U_i U_j & - & \sum_{i \in \mathcal{I}} \left\langle f \tau_i \right\rangle U_i \\
& & - & \sum_{i \in \mathcal{I}_N} \ll g \tau_i \gg U_i, \\
& \underbrace{\frac{1}{2} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{I}} A_{ij} U_i U_j}_{\text{matrice}} & \underbrace{\sum_{i \in \mathcal{I}} B_i U_i}_{\text{vecteur}}.
\end{aligned}$$

où la matrice A_{ij} et le vecteur B_i sont définis par

$$\begin{aligned}
A_{ij} &= \left\langle \nabla \tau_i \cdot a(\nabla \tau_j) \right\rangle & i, j \in \mathcal{I}, \\
B_i &= \left\langle f, \tau_i \right\rangle + \ll g, \tau_i \gg & i \in \mathcal{I}_N, \\
B_i &= \left\langle f, \tau_i \right\rangle & i \in \overline{\mathcal{I}_N}.
\end{aligned} \tag{2.53}$$

Ne tenons pas compte pour l'instant des conditions essentielles, et supposons que toutes les N valeurs nodales soient inconnues. Le minimum de la fonctionnelle est alors obtenu lorsque

$$0 = \frac{\partial J(u^h)}{\partial U_i} = \sum_{j \in \mathcal{I}} A_{ij} U_j - B_i, \quad i \in \mathcal{I}. \tag{2.54}$$

Ce qui correspond exactement à l'expression (2.11).

Envisageons, maintenant, le cas de conditions aux frontières essentielles. La liste des noeuds où la valeur de l'inconnue est imposée est donnée par l'ensemble \mathcal{I}_D , tandis que la liste des $N - N_D$ valeurs nodales restantes est $\overline{\mathcal{I}_D}$. La solution discrète se décompose en une partie inconnue et une partie contrainte par les conditions essentielles.

$$u(\mathbf{x}) \approx u^h(\mathbf{x}) = \sum_{i \in \overline{\mathcal{I}_D}} U_i \tau_i(\mathbf{x}) + \sum_{i \in \mathcal{I}_D} T_i \tau_i(\mathbf{x}). \quad (2.55)$$

L'expression de la fonctionnelle devient dès lors :

$$\begin{aligned} J(u^h) &= \frac{1}{2} \sum_{i \in \overline{\mathcal{I}_D}} \sum_{j \in \overline{\mathcal{I}_D}} A_{ij} U_i U_j - \sum_{i \in \overline{\mathcal{I}_D}} B_i U_i \\ &\quad + \sum_{i \in \overline{\mathcal{I}_D}} \sum_{j \in \mathcal{I}_D} A_{ij} U_i T_j \\ &\quad + \frac{1}{2} \sum_{i \in \mathcal{I}_D} \sum_{j \in \mathcal{I}_D} A_{ij} T_i T_j - \sum_{i \in \mathcal{I}_D} B_i T_i, \end{aligned}$$

et les conditions de stationnarité deviennent :

$$0 = \frac{\partial J(u^h)}{\partial U_i} = \sum_{j \in \overline{\mathcal{I}_D}} A_{ij} U_j - B_i + \sum_{k \in \mathcal{I}_D} A_{ik} T_k, \quad i \in \overline{\mathcal{I}_D}. \quad (2.56)$$

L'imposition d'une condition essentielle transforme le système algébrique (2.54) en un système réduit (2.56). Alors que l'obtention du système réduit peut sembler laborieuse, il existe un moyen très simple d'imposer en pratique des conditions aux frontières essentielles sans modifier le nombre de variables ni la position des lignes et des colonnes dans la matrice de raideur. Considérons en effet la matrice de raideur et les forces nodales connues, là où les conditions essentielles ne sont pas imposées, et supposons que la k -ième valeur nodale soit imposée à T_k . On procédera comme suit:

- On calcule la matrice de raideur complète A_{ij} et les forces nodales B_i .
- On multiplie la k -ième colonne de la matrice de raideur par la valeur T_k , et on la soustrait au vecteur des forces nodales.
- La k -ième ligne et la k -ième colonne de la matrice sont remplacées par une ligne et une colonne de zéros.
- Le terme A_{kk} est remplacé par 1.
- La composante B_k est remplacée par T_k .

Cette procédure a l'avantage d'être simple à implémenter dans un programme de calcul. Elle conserve le caractère symétrique et défini positif de la matrice de raideur sans modifier ses dimensions. On la réalise simplement par les lignes de code suivantes


```

void      matrixConstrain(double **A, double *B,
                          int size, int myNode, double myValue)

/* A, b, size      linear system,
   myNode myValue  index of the node to be constrained
                   to the value */
{
    int i

    for (i=0; i < size; i++) {
        B[i] -= myValue * A[i][myNode];
        A[i][myNode] = 0.0; }
    for (i=0; i < size; i++)
        A[myNode][i] = 0.0;
    A[myNode][myNode] = 1.0;
    B[myNode] = myValue;
}

```

Elle peut être appliquée directement sur les matrices locales, avant la procédure d'assemblage.

Chapitre 3

Techniques de résolution des systèmes linéaires creux

Rappelons d'abord que la discrétisation de problèmes elliptiques linéaires est le domaine usuel d'application des éléments finis. Dans un tel cas, nous obtenons une formulation discrète consistant en la recherche de la solution d'un système linéaire d'équations

Trouver $U_j \in \mathbb{R}^n$ tels que

$$\sum_{j=1}^n A_{ij} U_j = F_i, \quad i = 1, n.$$

(3.1)

où A_{ij} est une matrice carrée de taille $n \times n$. Cette matrice est creuse, symétrique et définie positive. L'unique solution du problème (3.1) peut aussi être vue comme la solution du problème de minimisation

Trouver $U_j \in \mathbb{R}^n$ tels que

$$J(U_j) = \min_{V_j \in \mathbb{R}^n} \left(\underbrace{\sum_{i=1}^n \sum_{j=1}^n \frac{1}{2} V_i A_{ij} V_j}_{J(V_j)} - \sum_{i=1}^n V_i F_i \right),$$

(3.2)

Caractériser le vecteur U_j comme solution de (3.1) ou (3.2) est totalement équivalent.

Tout d'abord, nous allons ici utiliser les notations habituelles pour les matrices ¹. Il n'y aura plus de confusion possible avec les notations tensorielles de la mécanique des

¹On observera que le symbole \cdot est désormais strictement réservé au produit scalaire entre deux vecteurs.

meilleux continus, puisqu'on ne considérera que la formulation discrète de nos équations. Nous allons donc écrire (3.1) et (3.2) de la manière :

Trouver $\mathbf{x} \in \mathbb{R}^n$ tel que

$$\mathbf{Ax} = \mathbf{b}$$

(3.3)

Trouver $\mathbf{x} \in \mathbb{R}^n$ tel que

$$J(\mathbf{x}) = \min_{\mathbf{v} \in \mathbb{R}^n} \underbrace{\left(\frac{1}{2} \mathbf{v} \cdot \mathbf{Av} - \mathbf{b} \cdot \mathbf{v} \right)}_{J(\mathbf{v})}$$

(3.4)

3.1 Méthodes directes

Dans cette section, nous allons considérer les techniques ou méthodes de résolution directe de (3.3) basées sur l'élimination gaussienne. Ensuite, nous donnerons une très brève introduction aux méthodes algorithmiques de recherche du minimum de (3.4) ou méthodes de résolution itérative de (3.3).

Il est important de mettre en évidence que, pour des problèmes hyperboliques ou d'advection-diffusion, ainsi que pour des formulations mixtes, on n'obtient pas toujours un système discret avec une matrice définie positive. Dans de tels cas, il n'est pas toujours évident de construire des méthodes itératives efficaces. En pratique, on revient souvent dans de tels cas à une élimination gaussienne. C'est la robustesse de cette approche et son caractère universel qui en fait encore aujourd'hui le solveur le plus populaire dans les codes commerciaux d'éléments finis.

Quelques améliorations sont toutefois faciles à réaliser...

Factorisation LU

On peut développer une autre procédure de résolution des systèmes linéaires en factorisant la matrice \mathbf{A} en produit d'une matrice triangulaire inférieure \mathbf{L} (*lower*) et d'une matrice triangulaire supérieure \mathbf{U} (*upper*). Si une telle factorisation est possible, la résolution du système $\mathbf{Ax} = \mathbf{b}$ peut s'effectuer en trois étapes suivantes :

- Factorisation de \mathbf{A} en \mathbf{LU}

- Résolution de $\mathbf{Lz} = \mathbf{b}$ par substitution directe
- Résolution de $\mathbf{Ux} = \mathbf{z}$ par substitution arrière

Une procédure rapide pour la factorisation \mathbf{LU} est justement l'élimination de Gauss. Lorsque l'on effectue une triangularisation, la matrice triangulaire supérieure se construit en faisant apparaître des "0" en dessous de la diagonale principale. Si, à chaque étape, on remplace les "0" qui apparaissent ainsi par les *multiplicateurs* utilisés, on obtient finalement une matrice \mathbf{F} qui fournit facilement la factorisation cherchée. En effet, on peut montrer que \mathbf{U} est formée de la partie triangulaire supérieure de \mathbf{F} tandis que \mathbf{L} est obtenue en prenant la partie de \mathbf{F} en dessous de la diagonale, complétée d'une diagonale de "1".

$$\begin{array}{ll} l_{ik} = -\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} & i = k+1, \dots, n, k = 1, \dots, n \\ l_{ii} = 1 & i = 1, \dots, n \end{array} \quad (3.5)$$

Lorsque l'on emploie la factorisation \mathbf{LU} avec la méthode utilisant l'élimination de Gauss, le nombre d'opérations nécessaires est donc semblable à celui qui est nécessaire pour la méthode de triangularisation directe avec la matrice augmentée. Cependant, l'avantage de la factorisation \mathbf{LU} apparaît si l'on doit résoudre plusieurs fois un système avec des seconds membres \mathbf{b} différents. Dans ce cas, la factorisation \mathbf{LU} de la matrice \mathbf{A} ne doit être faite qu'une seule fois, chaque nouveau second membre ne demandant que la résolution finale par substitution directe puis inverse, alors que la méthode de triangularisation doit être recalculée complètement pour chaque nouveau deuxième membre. Comme il arrive fréquemment que l'on souhaite analyser successivement plusieurs cas de charges distincts pour une même structure, les solveurs directs implémentés dans les codes d'éléments finis sont presque toujours des factorisations \mathbf{LU} , même lorsqu'on parle d'élimination gaussienne.

Factorisation de Cholesky

Lorsque la matrice \mathbf{A} est symétrique définie-positive, on peut la factoriser sous la forme suivante \mathbf{BB}^T avec $\mathbf{B} = \mathbf{DL}$. La matrice \mathbf{D} est une matrice diagonale dont les éléments sont donnés par

$$d_{kk} = \sqrt{a_{kk}^{(k)}}$$

tandis que les valeurs des éléments de \mathbf{L} et des coefficients $a_{kk}^{(k)}$ sont celles que l'on obtiendrait via une élimination gaussienne.

Il est possible d'obtenir les éléments de la matrice \mathbf{B} en utilisant une technique alternative dite méthode de *Cholesky*

$$\begin{aligned}
b_{11} &= \sqrt{a_{11}} \\
b_{i1} &= \frac{a_{i1}}{b_{11}} & i = 2, \dots, n \\
b_{jj} &= \sqrt{a_{jj} - \sum_{k=1}^{j-1} b_{jk}^2} & j = 2, \dots, n \\
b_{ij} &= \frac{a_{ij} - \sum_{k=1}^{j-1} b_{ik}b_{jk}}{b_{jj}} & j = 2, \dots, n, i = j+1, \dots, n
\end{aligned} \tag{3.6}$$

Pour une matrice symétrique définie positive, il est possible de démontrer que tous les nombres apparaissant sous les racines carrées seront toujours tous positifs. Une analyse du nombre d'opérations montre qu'il y a avantage à utiliser la méthode de Cholesky plutôt que la méthode de Gauss dans le cas d'une matrice symétrique et définie positive.

3.1.1 Factorisation de matrices creuses

Construire la matrice obtenue par une discrétisation par éléments finis risque d'occuper une place importante en mémoire dès le moment où le problème compte un nombre élevé de variables. Par ailleurs, effectuer une élimination gaussienne classique donne lieu à un grand nombre d'opérations inutiles dès le moment où la matrice de raideur contient de nombreux termes nuls. Nous allons voir qu'il est possible de réduire considérablement la place prise par la matrice.

Une première démarche consiste à tenir compte du fait que la matrice est symétrique; on peut donc ne retenir en mémoire que les termes situés sur et en-dessous de la diagonale principale, ce qui donne lieu à un gain de mémoire d'environ 50%. D'autre part, on tiendra compte de ce que la matrice de raideur est creuse, c'est-à-dire que le pourcentage d'éléments nuls y est élevé.

Il y a zéro et zéro

Lorsqu'on travaille avec des matrices creuses, il faut distinguer les zéros logiques, des zéros numériques. Les zéros logiques sont les éléments d'une matrice dont on peut déduire la valeur nulle à partir de la structure topologique du problème. Ils sont totalement indépendants des valeurs numériques des éléments non-nuls de la matrice et leur présence peut être déduite de la structure logique du problème. Au contraire, les zéros numériques

ne doivent leur valeur nulle qu'à une annulation providentielle de plusieurs termes : ils sont quasiment imprévisibles a priori.

Une mise en oeuvre intelligente de la factorisation de matrices creuses peut, en général, tirer relativement facilement profit des zéros logiques, mais beaucoup plus difficilement des zéros numériques. On pourrait évidemment les détecter au vol et ainsi en tenir compte en cours d'exécution de l'algorithme. Mais, en général, on néglige les zéros numériques et on considère que tout élément d'une matrice ou d'un vecteur qui n'est pas un zéro logique est non nul et sera traité comme tel. En outre, il arrive souvent qu'on décide de traiter certains zéros logiques comme de simples zéros numériques, afin de ne pas compliquer de manière excessive la structure de données. Typiquement, la structure du solveur bande est plus simple que celle du solveur frontal. Mais on ne tire pas profit d'autant de zéros logiques dans le premier cas que dans le second.

Problème du fill-in

Même quand la matrice à factoriser comporte beaucoup d'éléments nuls, les parties triangulaire supérieure **U** et inférieure **L** peuvent en avoir nettement moins. En d'autres mots, les matrices triangulaires auront toujours nettement moins de zéros logiques que **A**. C'est le phénomène de remplissage ou de fill-in.

Ci-après, on considère un cas pathologique où **A** est très creux tandis que la triangularisation génère une matrice **U** (tout comme **L**) totalement pleine.

$$\mathbf{A} = \begin{bmatrix} \diamond & \diamond & \diamond & \diamond & \diamond & \diamond & \diamond \\ \diamond & \diamond & & & & & \\ \diamond & & \diamond & & & & \\ \diamond & & & \diamond & & & \\ \diamond & & & & \diamond & & \\ \diamond & & & & & \diamond & \\ \diamond & & & & & & \diamond \end{bmatrix} \quad \mathbf{U} = \begin{bmatrix} \diamond & \diamond & \diamond & \diamond & \diamond & \diamond & \diamond \\ & \diamond & \diamond & \diamond & \diamond & \diamond & \diamond \\ & & \diamond & \diamond & \diamond & \diamond & \diamond \\ & & & \diamond & \diamond & \diamond & \diamond \\ & & & & \diamond & \diamond & \diamond \\ & & & & & \diamond & \diamond \\ & & & & & & \diamond \end{bmatrix}$$

L'effet de remplissage peut être prédit en ne tenant compte que de la structure logique (zéros logiques) de la matrice.

L'effet de remplissage peut être modifié par la numérotation des inconnues et des équations du système, c'est-à-dire par des permutations de lignes et de colonnes de la matrice. Pour pouvoir exploiter un maximum de zéros, les solveurs directs creux essaient de minimiser le remplissage en exploitant une bonne permutation de la matrice avant d'effectuer la factorisation. Par exemple, si on inverse totalement l'ordre des lignes et des colonnes de la matrice, on obtient une triangularisation qui ne souffre d'aucun effet de

remplissage...

$$\mathbf{A} = \begin{bmatrix} \diamond & & & & & & \diamond \\ & \diamond & & & & & \diamond \\ & & \diamond & & & & \diamond \\ & & & \diamond & & & \diamond \\ & & & & \diamond & & \diamond \\ & & & & & \diamond & \diamond \\ \diamond & \diamond & \diamond & \diamond & \diamond & \diamond & \diamond \end{bmatrix} \quad \mathbf{U} = \begin{bmatrix} \diamond & & & & & & \diamond \\ & \diamond & & & & & \diamond \\ & & \diamond & & & & \diamond \\ & & & \diamond & & & \diamond \\ & & & & \diamond & & \diamond \\ & & & & & \diamond & \diamond \\ & & & & & & \diamond \end{bmatrix}$$

Stratégie générale de résolution

Comme l'effet de remplissage peut être prédit avant d'effectuer la factorisation, la résolution d'un système creux se fait par le biais de quatre étapes

- Déterminer la numérotation optimale des inconnues, ou trouver la permutation optimale des lignes et des colonnes, afin de limiter l'effet de remplissage. Comme la matrice a une structure logique symétrique, on évite évidemment de perdre cette propriété, on effectue toujours les permutations de lignes et de colonnes de manière symétrique.
- Allouer des structures de données permettant de stocker les éléments non nuls de la matrice et les éléments additionnels résultant de l'effet de remplissage.
- Effectuer la factorisation de la matrice (en général, "en place", c'est-à-dire en utilisant le même espace mémoire pour les données et le résultat...)
- Résoudre le (ou les) systèmes triangulaires creux pour obtenir le vecteur des inconnues.

La minimisation du remplissage n'est pas le critère absolu. En effet, il faut aussi considérer la facilité avec laquelle on peut représenter la matrice \mathbf{A} et les conséquences d'une structure de données trop lourde. Cette stratégie est évidemment valable dans le cas où il n'est pas nécessaire d'effectuer des pivotages.

La théorie des graphes permet de construire un modèle rigoureux de l'effet de remplissage et de proposer des stratégies de numérotation des inconnues et des algorithmes de factorisation permettant de minimiser cet effet. Néanmoins, la recherche de la permutation d'une matrice telle que l'effet de remplissage soit minimal est un problème très difficile. Il s'agit d'un problème dit *NP* complet pour lequel il n'existe pas d'algorithme de complexité polynomiale. On doit donc utiliser des heuristiques qui minimisent raisonnablement l'effet de remplissage. Un des algorithmes de numérotation les plus populaires est connu sous le nom d'algorithme de *Reverse-Cuthill-MacKee* et est encore à la base de la renumérotation effectuée de manière automatique dans de nombreux codes commerciaux.

3.1.2 Solveur bande

La largeur de bande d'une matrice A_{ij} est la plus grande distance à la diagonale que peut atteindre un élément non nul de la matrice, augmentée d'une unité, ce que nous écrirons

$$\beta(\mathbf{A}) - 1 = \max_{ij} \{|i - j|, \forall (i, j) \text{ tels que } a_{ij} \neq 0\}. \quad (3.7)$$

On voit immédiatement que la largeur de bande d'une matrice diagonale est l'unité, tandis que celle d'une matrice dont tous les éléments sont non nuls est égale à l'ordre de la matrice. On dira qu'une matrice est *bande* lorsque sa largeur de bande est faible par rapport à son ordre.

$$\mathbf{A} = \begin{bmatrix} \diamond & \diamond & \diamond & & & & \\ & \diamond & & \diamond & & & \\ \diamond & & & \diamond & \diamond & & \\ & \diamond & \diamond & \diamond & & \diamond & \\ & & \diamond & & \diamond & \diamond & \diamond \\ & & & \diamond & \diamond & \diamond & \\ & & & & \diamond & \diamond & \diamond \\ & & & & & \diamond & \diamond \\ & & & & & & \diamond \end{bmatrix} \quad \mathbf{U} = \begin{bmatrix} \diamond & \diamond & \diamond & & & & \\ & \diamond & \diamond & \diamond & & & \\ & & \diamond & \diamond & \diamond & & \\ & & & \diamond & \diamond & \diamond & \\ & & & & \diamond & \diamond & \diamond \\ & & & & & \diamond & \diamond \\ & & & & & & \diamond \\ & & & & & & & \diamond \end{bmatrix}$$

La propriété essentielle des matrices bandes est que la largeur de bande d'une matrice reste constante durant le processus d'élimination gaussienne décrit plus haut. Ceci se comprend facilement en analysant la procédure ; si le terme A_{ij} est nul, le processus d'élimination du terme A_{jj} ne modifie aucun terme dans la i -ème ligne, et les zéros sont préservés. Cette propriété nous autorise à ne garder en mémoire que les éléments de la matrice de raideur dont la différence entre l'indice de ligne et l'indice de colonne est strictement inférieure à la largeur de bande. Finalement, en vertu de la symétrie de la matrice de raideur, il suffit de ne retenir que les éléments situés d'un côté de la diagonale, en-dessous par exemple. On comprend que, si la matrice est bande, seule la bande de termes non nuls sera conservée en mémoire. Il faut utiliser une numérotation afin de rendre minimale la largeur de bande.

3.1.3 Solveur frontal

La plupart des grands programmes d'éléments finis utilisent aujourd'hui la méthode frontale qui, à l'origine, fut développée en vue de la résolution de grands systèmes sur des ordinateurs disposant de peu de mémoire centrale (et en l'absence de mémoire virtuelle). L'idée sous-jacente de cette méthode est qu'il n'est pas nécessaire de stocker l'entièreté de la matrice $\mathbf{A}^{(k)}$ durant l'élimination gaussienne dans la mémoire à accès rapide : ce qui peut être difficile. En effet, on peut se contenter d'en stocker une petite partie seulement, appelée *matrice active* dans la mémoire active et de communiquer avec la mémoire secondaire seulement au début et à la fin de chaque étape de l'élimination. Cette idée est schématisée sur l'exemple de la matrice bande :

$$\mathbf{A} = \begin{bmatrix} \diamond & \diamond & \diamond & & & & \\ \diamond & \diamond & \diamond & \diamond & & & \\ \diamond & & \diamond & \diamond & \diamond & & \\ & \diamond & \diamond & \diamond & \diamond & \diamond & \\ & & \diamond & \diamond & \diamond & \diamond & \diamond \\ & & & \diamond & \diamond & \diamond & \diamond \\ & & & & \diamond & \diamond & \diamond \end{bmatrix} \quad \mathbf{A}^{(k)} = \begin{bmatrix} \diamond & \diamond & \diamond & & & & \\ & \diamond & & & & & \\ & & \blacklozenge & \blacklozenge & \blacklozenge & & \\ & & \blacklozenge & \blacklozenge & \blacklozenge & \diamond & \\ & & \blacklozenge & \blacklozenge & \blacklozenge & \diamond & \diamond \\ & & & \diamond & \diamond & \diamond & \diamond \\ & & & & \diamond & \diamond & \diamond \end{bmatrix}$$

Il faut toutefois préciser que la méthode frontale n'exige pas d'avoir une matrice bande et qu'elle peut tirer profit d'éventuels trous dans celle-ci. En ce sens, elle est plus puissante que les solveurs bandes (y compris les solveurs à bande de taille variable, appelés *skyline solvers*).

Pour en expliquer le fonctionnement, considérons un maillage de six éléments rectangulaires. La table d'appartenance du maillage est donnée comme suit,

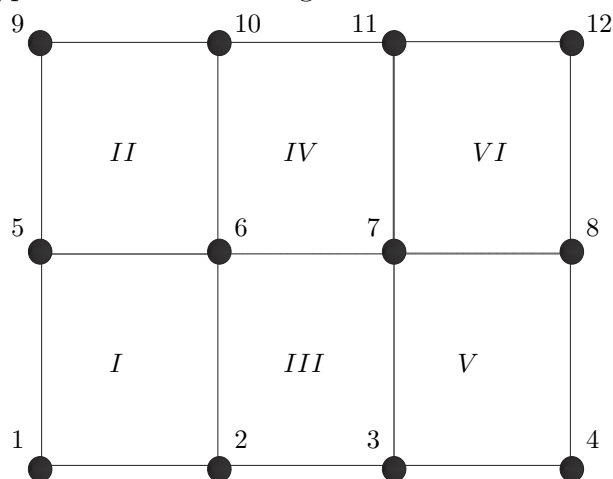


Figure 3.1: Numérotation du maillage.

Elément	Sommets			
1	1	2	6	5
2	5	6	10	9
3	2	3	7	6
4	11	10	6	7
5	3	4	8	7
6	7	8	12	11

A dessein, nous n'avons pas numéroté les noeuds de manière optimale pour la résolution par matrice bande. Les éléments sont numérotés de 1 à 6. Lors de la résolution frontale,

c'est la numérotation des éléments qui gouvernera la taille des matrices de calcul. Lorsqu'un domaine de calcul a une forme allongée, il faut numéroter les éléments dans le sens de la largeur pour minimiser le coût calcul.

Il est alors facile d'assigner à chaque noeud ce qu'on appelle un *élément de disparition*. Si nous comptons les éléments de 1 à 6, on voit par exemple que le noeud 1 disparaît avec l'élément 1, le noeud 2 avec l'élément 3, le noeud 6 avec l'élément 4, ... Pour obtenir l'élément de disparition de chaque noeud, il suffit de parcourir le tableau d'appartenance de 1 à 6. Chaque fois que le noeud i est rencontré dans l'élément Ω_e , on met à jour l'élément de disparition du i -ème noeud à la valeur e . La valeur finale est l'élément de disparition.

Sommet	Elément de disparition					
1	1					
2	1	3				
3		3		5		
4				5		
5	1	2				
6	1	2	3	4		
7			3	4	5	6
8					5	6
9		2				
10		2		4		
11				4		6
12						6

Réalisons ensuite la simulation de l'assemblage de lignes et colonnes pour les variables nodales. Le procédé est appliqué de la manière suivante. Dans la colonne de gauche, nous indiquons les lignes de la matrice globale allouées aux diverses variables dans la matrice frontale, lorsque l'on parcourt la table d'appartenance. Pour l'élément 1, nous plaçons les noeuds 1, 2, 6, 5 dans les premières lignes de la matrice globale 1, 2, 3 et 4. Ce sont les *destinations* des variables au sein de la matrice globale stockée sous la forme de la *matrice active*. Constatant à ce moment que 1 est l'élément de disparition du noeud 1, nous l'encadrons et nous libérons la ligne allouée à ce noeud, avant de passer au second élément.

Elément	Occupation de la matrice active lors de l'assemblage des éléments					
	1	2	3	4	5	6
Destination						
1	1	10	10	10	4	12
2	2	2	2	11	11	11
3	6	6	6	6	8	8
4	5	5	3	3	3	
5		9	7	7	7	7
6						

A l'élément 2, les noeuds 9 et 10 reçoivent une nouvelle destination. En particulier, la ligne 1 est libre et reçoit le noeud 10. Nous constatons que 2 est l'élément de disparition des noeuds 5 et 9 et les encadrons dans les lignes 4 et 5 qui deviennent libres. Le procédé se poursuit pour tous les éléments jusqu'à l'élimination des derniers noeuds lors de l'assemblage simulé de l'élément 6.

Le nombre de destinations ainsi ouvertes est la largeur de front, n_{act} , qui indique le nombre de variables actives. Nous allons montrer que tout le processus d'élimination de Gauss peut être effectué dans une matrice active de taille n_{act}^2 tandis que les lignes éliminées sont stockées (éventuellement sur disque) dans une matrice de taille $n \times n_{act}$ où n est le nombre de variables.

Supposons en effet que nous créons la matrice de raideur élément par élément, et que nous remplissons les lignes et colonnes dans l'ordre de l'apparition des noeuds. L'élément 1 nous donne les variables 1, 2, 6 et 5. Nous créons aussi le vecteur des forces nodales dans le même ordre. Nous constatons à ce moment que le noeud 1 se trouve dans son élément de disparition. La ligne du noeud ne changera plus, et il est dès lors possible de l'éliminer. S'il y a lieu, c'est le moment d'introduire les modifications de la matrice causées par une condition essentielle. Pour l'élimination, on divise la ligne 1 par l'élément diagonal. L'élimination de la variable 1 est alors opérée dans les trois autres lignes. On place alors dans une matrice de stockage la ligne 1 ainsi que la composante de force nodale divisée par le terme diagonal appelé *pivot*. On conserve aussi l'indice de la variable éliminée. L'élément diagonal est réinitialisé à zéro. On poursuit alors la même procédure pour les autres éléments. Dès le moment où une variable disparaît, sa ligne et sa colonne peuvent être réaffectées. On voit facilement qu'en suivant la destination prédite par la simulation de l'assemblage et en procédant à l'élimination dès qu'une variable disparaît, la taille de la matrice correspondant au maillage considéré ne dépassera pas 5^2 , ou n_{act}^2 .

Lorsqu'on arrive au bout de l'élimination dans l'élément 6, on obtient une équation à une inconnue, car la triangulation de la matrice de raideur est achevée. Il est alors facile de procéder à la substitution inverse de la méthode de Gauss et d'obtenir la valeur

de toutes les variables jusqu'à la première. L'originalité et l'efficacité de cette méthode proviennent de la construction et de l'élimination simultanées effectuées sur la matrice de raideur.

Il est facile de voir que le coût du calcul est essentiellement proportionnel à $\mathcal{O}(n n_{act}^2)$ (c'est-à-dire n éliminations dans des matrices carrées n_{act}^2). Il est dès lors important de réduire la valeur de n_{act} . Cette valeur dépend du nombre maximal de noeuds sur la frontière qui sépare à tout moment les éléments éliminés de ceux qui ne le sont pas encore. La valeur de n_{act} dépend uniquement de la numérotation des éléments et est totalement indépendante de la numérotation des noeuds. Lorsqu'un domaine de calcul a une forme allongée, il faut donc numéroter les éléments dans le sens de la largeur pour minimiser la largeur de front.

La méthode frontale est aujourd'hui encore la méthode d'élimination directe la plus populaire dans les grands codes d'éléments finis. La factorisation de matrices à largeur de bande variable se révèle souvent une alternative plus rapide. Finalement, il convient de signaler que le recours à des techniques itératives pour la résolution du système algébrique se justifie pleinement en particulier pour des matrices symétriques définies positives.

3.2 Méthodes itératives

Nous allons maintenant considérer des méthodes itératives pour la solution d'un problème de minimisation écrit sous la forme (3.4)

Trouver $\mathbf{x} \in \mathbb{R}^n$ tel que

$$J(\mathbf{x}) = \min_{\mathbf{v} \in \mathbb{R}^n} \underbrace{\left(\frac{1}{2} \mathbf{v} \cdot \mathbf{A} \mathbf{v} - \mathbf{b} \cdot \mathbf{v} \right)}_{J(\mathbf{v})}$$

où la matrice \mathbf{A} est une matrice carrée creuse, symétrique, définie positive et de taille $n \times n$. Comme nous l'avons déjà mentionné un nombre incalculable de fois, l'application des éléments finis à un problème linéaire elliptique conduit typiquement à un problème discret de la forme (3.4).

Les méthodes itératives pour trouver une solution de (3.4) ou de (3.3) jouent un rôle très important dans de nombreuses applications. Un des intérêts majeurs de telles méthodes est qu'il est possible de ne pas devoir stocker la matrice \mathbf{A} qui, même si elle creuse, peut être de très grande taille. Contrairement à ce qui est mentionné dans de nombreuses références, il n'est pas nécessaire de construire la matrice globale du problème, car on peut construire seulement des résidus locaux et n'assembler que ceux-ci pour obtenir une expression du résidu global. Ce sont donc des méthodes particulièrement utiles, lorsque la taille de la matrice globale rend son stockage impossible. En conclusion, ces méthodes sont intéressantes car elles ne nécessitent qu'un stockage et un nombre

d'opérations proportionnels au nombre d'inconnues du système, si elles convergent dans un nombre raisonnable d'itérations.

Nous allons considérer des méthodes itératives ou des algorithmes de minimisation de (3.4) qui peuvent être décrits dans le schéma suivant

Soit \mathbf{x}^0 une approximation initiale de la solution exacte \mathbf{x} de (3.4),

il s'agit alors construire une série d'approximations $\mathbf{x}^i \rightarrow \mathbf{x}$ de la manière suivante

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{d}^k$$

(3.8)

où \mathbf{d}^k est la direction de recherche et $\alpha_k > 0$ est le pas. Une large famille de méthodes peuvent être construites en sélectionnant, de manière empirique et pragmatique ou parfois de manière plus rigoureuse, la direction de recherche \mathbf{d}^k et le pas α_k . Nous allons ici donner une brève introduction à la méthode du gradient et à la méthode du gradient conjugué, ainsi qu'aux versions préconditionnées de ces méthodes.

3.2.1 Méthode de la plus grande pente

Pour une valeur \mathbf{x}^k , la direction de la plus grande pente est donnée par le gradient $\nabla J(\mathbf{x}^k)$. Il apparaît donc logique de sélectionner $\mathbf{d}^k = -\nabla J(\mathbf{x}^k)$ et d'obtenir la famille des méthodes du gradient.

$$\mathbf{d}^k = - \underbrace{(\mathbf{A}\mathbf{x}^k - \mathbf{b})}_{\mathbf{r}^k}$$

où \mathbf{r}^k est le vecteur du résidu du système linéaire pour \mathbf{x}^k .

Il faut encore sélectionner une valeur pour α_k . Un choix simple est de sélectionner une valeur constante α . Une manière plus compliquée est de calculer le pas optimal α_k en résolvant le problème de minimisation unidimensionnel suivant

Trouver $\alpha_k \geq 0$ tel que

$$J(\mathbf{x}^k + \alpha_k \mathbf{d}^k) = \min_{\alpha \geq 0} \underbrace{\left(\frac{1}{2}(\mathbf{x}^k + \alpha \mathbf{d}^k) \cdot \mathbf{A}(\mathbf{x}^k + \alpha \mathbf{d}^k) - \mathbf{b} \cdot (\mathbf{x}^k + \alpha \mathbf{d}^k) \right)}_{J((\mathbf{x}^k + \alpha \mathbf{d}^k))}$$

(3.9)

La condition de stationnarité de ce problème s'écrit sous la forme

$$\begin{aligned}
\frac{dJ(\mathbf{x}^k + \alpha_k \mathbf{d}^k)}{d\alpha} &= 0 \\
&\downarrow \\
(\mathbf{A}(\mathbf{x}^k + \alpha_k \mathbf{d}^k) - \mathbf{b}) \cdot \mathbf{d}^k &= 0 \\
&\downarrow \\
\underbrace{(\mathbf{A}\mathbf{x}^k - \mathbf{b})}_{-\mathbf{d}^k} \cdot \mathbf{d}^k &= -\alpha_k \mathbf{d}^k \cdot \mathbf{A} \mathbf{d}^k \\
&\downarrow \\
\frac{\mathbf{r}^k \cdot \mathbf{r}^k}{\mathbf{r}^k \cdot \mathbf{A} \mathbf{r}^k} &= \alpha_k
\end{aligned}$$

Il s'agit toutefois d'une optimisation à court terme, car une analyse plus fine montre que la méthode de la plus grande pente est une méthode qui converge lentement.

Analyse de la convergence

Par souci de simplicité, nous allons considérer le cas où le pas est fixé à une constante α^2 .

$$\begin{aligned}
\mathbf{x}^{k+1} &= \mathbf{x}^k - \alpha (\mathbf{A}\mathbf{x}^k - \mathbf{b}) \\
\mathbf{x}^{k+1} - \mathbf{x} &= \mathbf{x}^k - \mathbf{x} - \alpha (\mathbf{A}\mathbf{x}^k - \mathbf{b}) \\
\mathbf{x}^{k+1} - \mathbf{x} &= \mathbf{x}^k - \mathbf{x} - \alpha (\mathbf{A}\mathbf{x}^k - \mathbf{b}) + \alpha (\mathbf{A}\mathbf{x} - \mathbf{b}) \\
\underbrace{\mathbf{x}^{k+1} - \mathbf{x}}_{\mathbf{e}^{k+1}} &= (\boldsymbol{\delta} - \alpha \mathbf{A}) \underbrace{(\mathbf{x}^k - \mathbf{x})}_{\mathbf{e}^k}
\end{aligned}$$

où \mathbf{e}^k est le vecteur d'erreur à l'itération k et $\boldsymbol{\delta}$ est la matrice identité.

Pour finir l'analyse de convergence, il faut utiliser quelques résultats classiques de l'algèbre linéaire et écrire

$$\|\mathbf{e}^{k+1}\| \leq \|\boldsymbol{\delta} - \alpha \mathbf{A}\| \|\mathbf{e}^k\|,$$

²Les propriétés de la méthode de la plus grande pente utilisant un pas optimal sont fort semblables. L'optimisation locale du calcul du pas n'apporte pas une amélioration réelle de la convergence.

où les normes d'un vecteur et d'une matrice sont définies respectivement par :

$$\begin{aligned}\|\mathbf{u}\| &= \sqrt{\mathbf{u} \cdot \mathbf{u}}, \\ \|\mathbf{A}\| &= \max_{\mathbf{u} \in \mathbb{R}_0^n} \frac{\|\mathbf{A}\mathbf{u}\|}{\|\mathbf{u}\|}.\end{aligned}$$

Il faut aussi se rappeler que la norme d'une matrice vaut la plus grande de ses valeurs propres pour écrire finalement une expression permettant de sélectionner une valeur de α n'entraînant pas la divergence de la méthode itérative.

$$\begin{aligned}|1 - \alpha\lambda_j| &\leq 1 \\ &\downarrow \\ \alpha\lambda_{max} &\leq 2 \\ &\downarrow \\ \alpha &\leq \frac{2}{\lambda_{max}}\end{aligned}\tag{3.10}$$

En prenant comme valeur de $\alpha = 1/\lambda_{max}$, on effectue un choix qui semble optimal et on obtient que la norme :

$$\begin{aligned}\|\boldsymbol{\delta} - \alpha\mathbf{A}\| &= 1 - \frac{\lambda_{min}}{\lambda_{max}} \\ &= 1 - \frac{1}{\kappa(\mathbf{A})}\end{aligned}\tag{3.11}$$

où $\kappa(\mathbf{A})$ est le nombre de condition ou conditionnement de la matrice \mathbf{A} . On peut maintenant estimer le nombre d'itérations n requis pour réduire la norme de l'erreur initiale $\|\mathbf{e}_0\|$ d'un facteur $\epsilon > 0$ donné. En d'autres mots, il s'agira de trouver n tel que

$$\begin{aligned}\left(1 - \frac{1}{\kappa(\mathbf{A})}\right)^n &\leq \epsilon \\ \log\left(\frac{1}{\epsilon}\right) &\leq \underbrace{-n \log\left(1 - \frac{1}{\kappa(\mathbf{A})}\right)}_{\leq \frac{1}{\kappa(\mathbf{A})}} \\ \log\left(\frac{1}{\epsilon}\right)\kappa(\mathbf{A}) &\leq n\end{aligned}$$

Nous pouvons en conclure que le nombre requis d'itérations dans la méthode de la plus grande pente avec un pas α choisi de manière adéquate est proportionnel au conditionnement de la matrice et au nombre de décimales dans le facteur de réduction ϵ . Or, la matrice générée par la discrétisation d'un problème elliptique du second ordre par une méthode d'éléments finis classiques a un conditionnement qui est de l'ordre de h^{-2} . En d'autres mots, plus le maillage sera fin pour un problème donné, plus grand sera le nombre d'itérations requis et cela de manière quadratique. C'est un très mauvais résultat !

3.2.2 Méthode des gradients conjugués

Nous allons maintenant esquisser une présentation d'une méthode itérative nettement plus efficace. Il s'agit de la méthode des gradients conjugués. Dans cette méthode, les pas sont choisis de manière optimale et les directions de recherche sont conjuguées au sens suivant

$$\mathbf{d}^i \cdot \mathbf{A} \mathbf{d}^j = 0 \quad i \neq j$$

Comme la matrice \mathbf{A} est définie positive, on peut observer que ceci définit un produit scalaire pour les vecteurs de \mathbb{R}^n . On peut également observer qu'il est possible d'en déduire une norme énergétique. On voit donc qu'on utilise ici fondamentalement une approche semblable à celle qui a été sous-jacente à la théorie des éléments finis...

La méthode des gradients conjugués est définie par le schéma suivant

$$\begin{aligned} \alpha^k &= -\frac{\langle \mathbf{r}^k, \mathbf{r}^k \rangle}{\langle \mathbf{A} \mathbf{d}^k, \mathbf{r}^k \rangle} \\ \mathbf{r}^{k+1} &= \mathbf{r}^k + \alpha^k \mathbf{A} \mathbf{d}^k \\ \beta^k &= \frac{\langle \mathbf{r}^{k+1}, \mathbf{r}^{k+1} \rangle}{\langle \mathbf{r}^k, \mathbf{r}^k \rangle} \\ \mathbf{x}^{k+1} &= \mathbf{x}^k + \alpha^k \mathbf{d}^k \\ \mathbf{d}^{k+1} &= \mathbf{r}^{k+1} + \beta^k \mathbf{d}^k \end{aligned}$$

On utilise comme première direction $\mathbf{d}^0 = \mathbf{r}^0$ et donc l'opposé de la direction de la plus grande pente. En pratique, comme la direction de descente est multipliée par un facteur, on aurait aussi pu sélectionner le résidu dans l'algorithme de la plus grande pente et calculer un coefficient optimal qui aurait alors été négatif : c'est donc juste une question de convention.

Il est important d'insister sur le fait qu'il est parfaitement possible d'évaluer en calculant directement tous les vecteurs à un niveau local et de n'assembler que des expressions

vectérielles ! En d'autres mots, il n'est pas requis de construire la matrice globale \mathbf{A} .

Si nous comparons ces expressions avec celles de la méthode de la plus grande pente, on observe que l'on choisit un pas optimal et que la nouvelle direction de recherche est maintenant une combinaison linéaire de la précédente et du nouveau résidu. On peut également observer que cette combinaison linéaire est réalisée de manière à ce que \mathbf{d}^k soit conjugué à \mathbf{d}^{k-1} . Il serait ensuite possible d'effectuer une étape essentielle dans la compréhension de cette méthode en remarquant que \mathbf{d}^k est également conjuguée avec toutes les directions précédentes. Mais cela sort du cadre de cette introduction.

Par contre, il est essentiel de savoir qu'il serait à nouveau possible d'évaluer le nombre d'itérations requises pour réduire la norme de l'erreur initiale $\|\mathbf{e}_0\|$ d'un facteur $\epsilon > 0$ donné. Une analyse permettrait d'obtenir le résultat suivant

$$\frac{1}{2} \log\left(\frac{2}{\epsilon}\right) \sqrt{\kappa(\mathbf{A})} \leq n$$

On y observe que le nombre requis d'itérations est maintenant proportionnel à $\sqrt{\kappa(\mathbf{A})}$. Résultat qui doit être comparé avec la proportionnalité à $\kappa(\mathbf{A})$ dans le cas de la méthode de la plus grande pente. Ainsi, la méthode des gradients conjugués est de loin plus rapide et efficace que la méthode de la plus grande pente. A nouveau, pour une application de type éléments finis, le nombre d'itérations sera de l'ordre de h^{-1} pour les gradients conjugués et de l'ordre de h^{-2} pour la méthode de la plus grande pente.

Le principe de l'algorithme est le suivant :

$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha^k \mathbf{d}^k$	avec α^k tel que $\langle \mathbf{r}^{k+1} \mathbf{r}^k \rangle = 0$
$\mathbf{d}^{k+1} = \mathbf{r}^{k+1} + \beta^k \mathbf{d}^k$	avec β^k tel que $\langle \mathbf{d}^{k+1} \mathbf{A} \mathbf{d}^k \rangle = 0$

On peut montrer qu'un tel algorithme converge en (au plus) n étapes pour une matrice définie positive.

Calcul de α_k

$$\begin{aligned}
 \mathbf{x}^{k+1} &= \mathbf{x}^k + \alpha^k \mathbf{d}^k \\
 &\downarrow \text{En multipliant par } \mathbf{A} \text{ et en soustrayant } \mathbf{b} \\
 \mathbf{r}^{k+1} &= \mathbf{r}^k + \alpha^k \mathbf{A} \mathbf{d}^k \\
 &\downarrow \text{En effectuant le produit scalaire avec } \mathbf{r}^k \\
 \underbrace{\langle \mathbf{r}^{k+1} \mathbf{r}^k \rangle}_{=0} &= \langle \mathbf{r}^k \mathbf{r}^k \rangle + \alpha^k \langle \mathbf{r}^k \mathbf{A} \mathbf{d}^k \rangle \\
 &\downarrow \\
 \alpha^k &= - \frac{\langle \mathbf{r}^k \mathbf{r}^k \rangle}{\langle \mathbf{A} \mathbf{d}^k \mathbf{r}^k \rangle}
 \end{aligned}$$

Calcul de β_k

On observe tout d'abord que

$$\begin{aligned}
 \mathbf{d}^k &= \mathbf{r}^k + \beta^{k-1} \mathbf{d}^{k-1} \\
 &\downarrow \text{En effectuant le produit scalaire avec } \mathbf{A} \mathbf{d}^k \\
 \langle \mathbf{d}^k \mathbf{A} \mathbf{d}^k \rangle &= \langle \mathbf{r}^k \mathbf{A} \mathbf{d}^k \rangle + \beta^{k-1} \underbrace{\langle \mathbf{d}^k \mathbf{A} \mathbf{d}^{k-1} \rangle}_{=0} \\
 &\downarrow \\
 \langle \mathbf{d}^k \mathbf{A} \mathbf{d}^k \rangle &= \langle \mathbf{r}^k \mathbf{A} \mathbf{d}^k \rangle
 \end{aligned}$$

Ensuite, on refait la même opération pour l'itération suivante :

$$\begin{aligned}
\mathbf{d}^{k+1} &= \mathbf{r}^{k+1} + \beta^k \mathbf{d}^k \\
&\downarrow \text{En effectuant le produit scalaire avec } \mathbf{Ad}^k \\
\underbrace{\langle \mathbf{d}^{k+1} \mathbf{Ad}^k \rangle}_{=0} &= \langle \mathbf{r}^{k+1} \mathbf{Ad}^k \rangle + \beta^k \langle \mathbf{d}^k \mathbf{Ad}^k \rangle \\
&\downarrow \\
\beta^k &= -\frac{\langle \mathbf{r}^{k+1} \mathbf{Ad}^k \rangle}{\langle \mathbf{d}^k \mathbf{Ad}^k \rangle} \\
&\downarrow \text{Car } \mathbf{r}^{k+1} - \mathbf{r}^k = \alpha^k \mathbf{Ad}^k \\
\beta^k &= -\frac{1}{\alpha^k} \frac{\langle \mathbf{r}^{k+1} \mathbf{r}^{k+1} \rangle - \overbrace{\langle \mathbf{r}^{k+1} \mathbf{r}^k \rangle}^{=0}}{\langle \mathbf{d}^k \mathbf{Ad}^k \rangle} \\
&\downarrow \text{Car } \langle \mathbf{d}^k \mathbf{Ad}^k \rangle = \langle \mathbf{r}^k \mathbf{Ad}^k \rangle \\
\beta^k &= -\frac{1}{\alpha^k} \frac{\langle \mathbf{r}^{k+1} \mathbf{r}^{k+1} \rangle}{\langle \mathbf{r}^k \mathbf{Ad}^k \rangle} \\
&\downarrow \text{Car } \alpha^k = -\frac{\langle \mathbf{r}^k \mathbf{r}^k \rangle}{\langle \mathbf{Ad}^k \mathbf{r}^k \rangle} \\
\beta^k &= \frac{\langle \mathbf{r}^{k+1} \mathbf{r}^{k+1} \rangle}{\langle \mathbf{r}^k \mathbf{r}^k \rangle}
\end{aligned}$$

Implémentation

Finalement, l'implémentation à réaliser est la suivante :

$$\begin{aligned}
\alpha^k &= -\frac{\langle \mathbf{r}^k, \mathbf{r}^k \rangle}{\langle \mathbf{A} \mathbf{d}^k, \mathbf{r}^k \rangle} \\
\mathbf{r}^{k+1} &= \mathbf{r}^k + \alpha^k \mathbf{A} \mathbf{d}^k \\
\beta^k &= \frac{\langle \mathbf{r}^{k+1}, \mathbf{r}^{k+1} \rangle}{\langle \mathbf{r}^k, \mathbf{r}^k \rangle} \\
\mathbf{x}^{k+1} &= \mathbf{x}^k + \alpha^k \mathbf{d}^k \\
\mathbf{d}^{k+1} &= \mathbf{r}^{k+1} + \beta^k \mathbf{d}^k
\end{aligned}$$

Attention, on n'effectue pas le produit matriciel $\mathbf{A} \mathbf{d}^k$, mais on assemble un vecteur $\mathbf{s}^k = \mathbf{A} \mathbf{d}^k$, tout comme on assemble un vecteur $\mathbf{r}^k = \mathbf{A} \mathbf{x}^k - \mathbf{b}$! Tout l'intérêt d'un algorithme itératif est de ne pas nécessiter le stockage d'une matrice de grande taille : c'est une implémentation *matrix-free* qu'on veut réaliser :-)

3.2.3 Préconditionnement

Reprenons notre problème de minimisation (3.4).

$$\begin{aligned}
&\text{Trouver } \mathbf{x} \in \mathbb{R}^n \text{ tel que} \\
J(\mathbf{x}) &= \min_{\mathbf{v} \in \mathbb{R}^n} \underbrace{\left(\frac{1}{2} \mathbf{v} \cdot \mathbf{A} \mathbf{v} - \mathbf{b} \cdot \mathbf{v} \right)}_{J(\mathbf{v})}
\end{aligned}$$

Introduisons maintenant une matrice carrée \mathbf{E} non singulière et introduisons également le changement de variable

$$\begin{aligned}
\mathbf{y} &= \mathbf{E} \mathbf{x} \\
\mathbf{x} &= \mathbf{E}^{-1} \mathbf{y}
\end{aligned} \tag{3.12}$$

Il est alors possible de réécrire notre fonction à minimiser en terme de \mathbf{y}

$$\begin{aligned}
J(\mathbf{x}) &= \frac{1}{2} \mathbf{x} \cdot \mathbf{A} \mathbf{x} - \mathbf{b} \cdot \mathbf{x} \\
&= \frac{1}{2} (\mathbf{E}^{-1} \mathbf{y}) \cdot \mathbf{A} (\mathbf{E}^{-1} \mathbf{y}) - \mathbf{b} \cdot (\mathbf{E}^{-1} \mathbf{y}) \\
&= \frac{1}{2} \mathbf{y} \cdot \underbrace{(\mathbf{E}^{-T} \mathbf{A} \mathbf{E}^{-1})}_{\tilde{\mathbf{A}}} \mathbf{y} - \underbrace{(\mathbf{E}^{-T} \mathbf{b})}_{\tilde{\mathbf{b}}} \cdot \mathbf{y} \\
&= \tilde{J}(\mathbf{y})
\end{aligned}$$

et d'écrire un nouveau problème de minimum

Trouver $\mathbf{y} \in \mathbb{R}^n$ tel que

$$\tilde{J}(\mathbf{y}) = \min_{\mathbf{v} \in \mathbb{R}^n} \underbrace{\left(\frac{1}{2} \mathbf{v} \cdot \tilde{\mathbf{A}} \mathbf{v} - \tilde{\mathbf{b}} \cdot \mathbf{v} \right)}_{\tilde{J}(\mathbf{v})} \quad (3.13)$$

Il est évidemment possible d'utiliser notre méthode des gradients conjugués sur ce nouveau problème en lieu et place du problème original. Jusqu'à présent, il semble qu'on ait simplement remplacé un problème par un problème strictement équivalent... Toutefois, la démarche peut devenir très intéressante si $\kappa(\tilde{\mathbf{A}}) \ll \kappa(\mathbf{A})$. En effet, la méthode itérative convergera beaucoup plus vite sur le nouveau problème que sur le problème original. Evidemment, une nouvelle question arrive : comment choisir cette fameuse matrice \mathbf{E} ?

Observons tout d'abord que l'application de la méthode de la plus grande pente pour le problème modifié implique que

$$\begin{aligned}
\mathbf{y}^{k+1} &= \mathbf{y}^k - \alpha (\tilde{\mathbf{A}} \mathbf{y}^k - \tilde{\mathbf{b}}) \\
&\downarrow \\
\mathbf{x}^{k+1} &= \mathbf{x}^k - \alpha \underbrace{\mathbf{E}^{-1} \mathbf{E}^{-T}}_{\mathbf{C}^{-1}} (\mathbf{A} \mathbf{x}^k - \mathbf{b})
\end{aligned}$$

où \mathbf{C} est définie comme étant $\mathbf{E}^T \mathbf{E}$. En d'autres mots, appliquer la méthode de la plus grande pente sur le problème modifié est équivalent à appliquer le schéma suivant sur le problème original

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha \mathbf{C}^{-1} (\mathbf{A} \mathbf{x}^k - \mathbf{b}) \quad (3.14)$$

On dira que le schéma (3.14) est la version préconditionnée de la méthode de la plus grande pente, avec la matrice \mathbf{C} comme préconditionneur. Pour obtenir \mathbf{x}^{k+1} à partir de \mathbf{x}^k , il faut résoudre un système linéaire de la forme

$$\mathbf{C}(\mathbf{x}^{k+1} - \mathbf{x}^k) = -\alpha(\mathbf{A}\mathbf{x}^k - \mathbf{b}) \quad (3.15)$$

Il est maintenant possible d'énoncer les propriétés souhaitées pour le préconditionneur $\mathbf{C} = \mathbf{E}^T \mathbf{E}$:

- Tout d'abord, le conditionnement de $\mathbf{E}^{-T} \mathbf{A} \mathbf{E}^{-1}$ doit être le plus proche de l'unité. En tous cas, il doit au moins être nettement plus modeste que le nombre de condition de \mathbf{A} .
- Résoudre un système $\mathbf{C}\mathbf{x} = \mathbf{b}$ doit être économique, par rapport au système original.
- Obtenir le préconditionneur doit être peu coûteux.

Ces demandes sont contradictoires et difficiles à satisfaire et c'est souvent l'art du numéricien et l'expérience accumulée qui permettent de trouver de bons préconditionneurs. Il s'agit parfois de secrets de fabrication que les concepteurs de logiciels ne dévoilent pas aux utilisateurs.

A titre d'exemple, supposons que l'on choisisse comme préconditionneur \mathbf{A} . On observe qu'un tel choix permet d'obtenir le meilleur conditionnement possible. En effet, on peut alors montrer que \mathbf{E} correspond à la matrice \mathbf{B}^T de la décomposition de Cholesky et que la matrice modifiée $\tilde{\mathbf{A}}$ n'est rien d'autre que la matrice identité ! Par contre, résoudre $\mathbf{C}\mathbf{x} = \mathbf{b}$ est strictement équivalent à résoudre le système original. En d'autres mots, on converge en une itération, mais pour effectuer cette itération, il faut résoudre le problème original.

Un compromis relativement efficace est d'avoir un préconditionneur $\mathbf{C} = \mathbf{E}^T \mathbf{E}$ avec une matrice \mathbf{E} la plus creuse possible afin que la résolution du système $\mathbf{C}\mathbf{x} = \mathbf{b}$ soit très économique. Par contre, on souhaite qu'elle ne soit pas vraiment totalement creuse ou diagonale, car il faut que $\mathbf{E}^T \mathbf{E}$ soit une relative bonne approximation de \mathbf{A} . Une façon de réaliser ce compromis est d'exiger que \mathbf{E} ait exactement la même structure creuse que celle de \mathbf{A} . En d'autres mots, nous acceptons qu'un élément e_{ij} puisse être non nul uniquement si l'élément correspondant a_{ij} est non nul : on interdit au phénomène de *fill-in* de se réaliser. Avec un tel préconditionneur, on peut alors réaliser une élimination gaussienne modifiée où les éléments non nuls, apparaissant dans le processus d'élimination à des destinations interdites, sont froidement remplacés par des zéros. Une telle procédure appelée *factorisation incomplète* requiert seulement un nombre d'opérations de l'ordre de $\mathcal{O}(n)$ et permet d'obtenir une factorisation approchée de la matrice correspondant souvent à une réduction exceptionnelle du nombre de condition.

Chapitre 4

Théorie de la meilleure approximation

Dans ce chapitre, nous donnons une petite introduction à la théorie mathématique sous-jacente à la méthode des éléments finis appliquée à des problèmes elliptiques. Typiquement, il s'agit de la conduction thermique, de l'élasticité linéaire, la théorie des cordes et des membranes ainsi que la théorie des poutres et des coques. Tous ces problèmes peuvent être écrits sous la formulation abstraite suivante :

$$\begin{array}{l} \text{Trouver } u \in \mathcal{U} \text{ tel que} \\ J(u) = \min_{v \in \mathcal{U}} \underbrace{\frac{1}{2} a(v, v) - b(v)}_{J(v)}, \end{array} \quad (4.1)$$

où nous n'avons pas rigoureusement défini l'espace \mathcal{U} jusqu'à présent.

Nous allons d'abord introduire le concept d'espace d'Hilbert pour lequel il existe un théorème d'existence et d'unicité de la solution d'un problème elliptique. Ensuite, nous introduirons les espaces de Sobolev et quelques résultats théoriques liant ces espaces aux espaces usuels de fonctions $C^n(\Omega)$. Les espaces \mathcal{U} et $\hat{\mathcal{U}}$ apparaîtront comme des exemples d'espaces de Sobolev et d'Hilbert. Ces espaces et les résultats théoriques qui y sont associés permettront alors de construire des bornes ou des estimations de l'erreur de discrétisation.

Dans la recherche d'une solution approchée, il est toujours souhaitable de pouvoir estimer l'erreur afin de donner le degré de fiabilité de la solution discrète obtenue. On distingue deux types d'estimation pour l'erreur :

- On peut estimer *a priori* l'erreur avant de calculer la solution discrète. Une telle estimation sera relativement imprécise et ne fournira que des informations sur le comportement asymptotique ou le taux de convergence de l'erreur de la méthode numérique.

- On peut aussi estimer *a posteriori* l'erreur après avoir calculé la solution discrète dont on pourra maintenant tenir compte. Une telle estimation sera clairement plus fiable.

4.1 Espaces d'Hilbert

Avant d'introduire le concept d'espaces d'Hilbert, effectuons quelques rappels élémentaires d'algèbre linéaire. Soit \mathcal{U} un espace vectoriel muni d'une norme $\|\cdot\|$. Considérons une forme bilinéaire et symétrique a et une forme linéaire b .

b est une forme linéaire si	$b(\alpha u + \beta v) = \alpha b(u) + \beta b(v)$ $\forall \alpha, \beta \in \mathcal{R}, \forall u, v \in \mathcal{U}$
-------------------------------	---

a est une forme bilinéaire si	$a(\alpha u + \beta v, w) = \alpha a(u, w) + \beta a(v, w)$ $a(w, \alpha u + \beta v) = \alpha a(w, u) + \beta a(w, v)$ $\forall \alpha, \beta \in \mathcal{R}, \forall u, v, w \in \mathcal{U}$
---------------------------------	--

Pour de telles formes linéaires ou bilinéaires, on dit habituellement :

a est une forme symétrique si	$a(u, v) = a(v, u) \quad \forall u, v \in \mathcal{U}$
---------------------------------	--

b est une forme continue si	$\exists c > 0$ tel que $ b(u) \leq c \ u\ $ $\forall u \in \mathcal{U}$
-------------------------------	--

a est une forme continue si	$\exists c > 0$ tel que $ a(u, v) \leq c \ u\ \ v\ $ $\forall u, v \in \mathcal{U}$
-------------------------------	--

a est une forme coercive ou \mathcal{U} -elliptique si	$\exists \alpha > 0$ tel que $a(u, u) \geq \alpha \ u\ ^2$ $\forall u \in \mathcal{U}$
--	---

a est une forme définie positive ou est un produit scalaire pour \mathcal{U} si	a symétrique, $a(u, u) \geq 0 \quad \forall u \neq 0 \in \mathcal{U}$ $a(u, u) = 0 \Rightarrow u = 0$
--	---

On peut montrer que la discrétisation d'équations aux dérivées partielles de type elliptique par la méthode de Galerkin engendre une forme a continue et coercive pour la norme $\|\cdot\|$ de l'espace fonctionnel dans lequel on recherche la solution. La continuité et

la coercivité impliquent que la forme a est définie positive et définit un produit scalaire et une norme particulière définie par

$$\begin{aligned} \langle u, v \rangle_* &= a(u, v), \\ \|v\|_*^2 &= a(v, v) \end{aligned} \tag{4.2}$$

Par contre, le fait que a soit définie positive implique seulement le fait que a est continue et coercive par rapport à la norme énergétique $\|\cdot\|_*$. Un espace vectoriel muni d'un produit scalaire et de la norme qui y est associée est qualifié d'espace d'Hilbert si cet espace est complet pour la norme considérée¹.

Inégalité de Cauchy

Une des propriétés utiles (et qui sera donc souvent mise à contribution) du produit scalaire $\langle \cdot, \cdot \rangle$ associé à une norme $\|\cdot\|$ est l'inégalité de Cauchy :

$$\langle u, v \rangle \leq \|u\| \|v\| \tag{4.3}$$

L'inégalité de Cauchy est obtenue en considérant simplement une combinaison linéaire de deux éléments $u + \lambda v$ (λ est un réel quelconque):

$$\begin{array}{ccc} 0 & \leq & \langle u + \lambda v, u + \lambda v \rangle \\ \downarrow & & \\ 0 & \leq & \langle u, u \rangle + 2\lambda \langle u, v \rangle + \lambda^2 \langle v, v \rangle \end{array}$$

L'expression de droite est un polynôme du second degré en λ qui sera toujours positif seulement si son discriminant ($\langle u, v \rangle^2 - \langle u, u \rangle \langle v, v \rangle$) est strictement négatif. Cette condition correspond simplement à l'inégalité de Cauchy.

Théorème d'existence et d'unicité de Lax-Milgram

L'intérêt majeur de travailler dans un espace d'Hilbert est de disposer alors d'un résultat théorique d'existence et d'unicité sous la forme du théorème de Lax-Milgram.

¹Un espace \mathcal{U} est complet pour une norme $\|\cdot\|$ si toute suite de Cauchy par rapport à $\|\cdot\|$ converge. Oui, mais c'est quoi une suite de Cauchy ? Une suite de Cauchy $\{u_1, u_2, \dots\}$ est une suite telle que $\forall \varepsilon \exists n \forall i, j > n \ \|u_i - u_j\| < \varepsilon$. Oui, mais c'est quoi une suite qui converge ? Une suite $\{u_1, u_2, \dots\}$ converge vers u si $\|u - u_i\| \rightarrow 0$ lorsque $i \rightarrow \infty$. Notez que si le concept d'espace complet ne vous intéresse absolument pas, vous pouvez l'oublier et vous contenter de considérer bêtement un espace d'Hilbert comme un espace vectoriel muni d'un produit scalaire. Mais, évidemment, c'est moins joli d'un point de vue intellectuel...

Si	\mathcal{U} est un espace d'Hilbert a est une forme bilinéaire continue et coercive b est une forme linéaire continue,	(4.4)
alors,	<div style="border: 1px solid black; padding: 10px; margin: 10px auto; width: 80%;"> Trouver $u(\mathbf{x}) \in \mathcal{U}$ tel que $a(\hat{u}, u) = b(\hat{u}), \quad \forall \hat{u} \in \mathcal{U},$ </div>	
le problème abstrait	a une solution unique qui dépend continûment du terme source b ($\ u\ \leq \frac{1}{\alpha} \ b\ $).	

où la norme d'une forme b est définie par l'expression

$$\|b\| = \sup_{v \in \mathcal{U}} \frac{|b(v)|}{\|v\|}.$$

Pour le lecteur vraiment intéressé, la démonstration du théorème de Lax-Milgram est fournie dans le livre de Ciarlet (pages 8-9). Toutefois, pour pouvoir utiliser ce théorème en toute rigueur, il faudrait, pour chacune de nos applications, définir précisément l'espace utilisé (ce qui est relativement intuitif) et démontrer que a est bien continue (en général, c'est trivial) et coercive (ce qui est souvent plus laborieux).

4.2 Espaces de Sobolev

Nous allons maintenant introduire les espaces d'Hilbert qu'il est naturel d'employer pour la formulation variationnelle de problèmes aux limites. Tout d'abord, commençons par introduire l'espace des fonctions carré-intégrables sur Ω .

$$L_2(\Omega) = \{v \text{ tel que } \langle v^2 \rangle < \infty\} \quad (4.5)$$

où notre notation $\langle \cdot \rangle$ représente toujours l'intégration sur le domaine Ω . Cet espace L_2 est un espace d'Hilbert² avec le produit scalaire $\langle \cdot \rangle$. Une fonction typique de L_2 apparaît comme une fonction continue par morceaux, éventuellement non bornée, mais dont l'intégrale du carré est bornée.

Il est ensuite possible de construire une famille d'espaces de la manière suivante.

$$H^m(\Omega) = \{v \text{ tel que } D^{\alpha}v \in L_2(\Omega), |\alpha| \leq m\} \quad (4.6)$$

²Pour vraiment apprécier la définition de L_2 et réaliser que cet espace est complet, il faudrait introduire la notion d'intégrale de Lebesgue.

où la notation multi-indicielle $D^{\alpha}v$ pour $\Omega \subset \mathcal{R}^n$ représente :

$$\begin{aligned} D^{\alpha}v &= \frac{\partial^{\alpha_1+\alpha_2+\dots+\alpha_n}v}{\partial x_1^{\alpha_1}\partial x_2^{\alpha_2}\dots\partial x_n^{\alpha_n}} & \forall \alpha &= (\alpha_1, \alpha_2, \dots, \alpha_n) \\ |\alpha| &= \sum_{i=1}^n \alpha_i \end{aligned} \quad (4.7)$$

Ces espaces H^m sont les exemples les plus connus d'*espaces de Sobolev*. En d'autres mots, ces espaces contiennent les fonctions dont les dérivées jusqu'à l'ordre m sont carré-intégrables. Il est clair que $H^{m+1}(\Omega) \subset H^m(\Omega)$ et que $H^0(\Omega) = L_2(\Omega)$. Ces espaces permettent d'introduire une caractérisation rigoureuse de l'idée intuitive de fonction "lisse". Ainsi, certaines singularités sont exclues par le critère d'intégration au carré. Typiquement, sur un intervalle ouvert $\Omega =]0, 1[$, on notera que $x^{-1/4}$ appartient à L_2 , mais que $x^{-1/2}$ n'y appartient pas ³.

Il est aussi important de noter que les dérivées introduites dans la définition des espaces de Sobolev sont des généralisations du concept habituel de dérivées. Typiquement, on utilise des dérivées au sens des distributions : par exemple, la dérivée faible d'une fonction échelon est la distribution de Dirac, mais ceci sort largement du cadre de cette introduction. On peut observer que la distribution de Dirac appartient en fait à H^{-1} , espace qui contient les distributions dont les primitives au sens des distributions sont carré-intégrables.

Finalement, on peut démontrer que les espaces de Sobolev H^m sont des espaces de Hilbert pour le produit scalaire et la norme associés donnés par :

$$\begin{aligned} \langle v, w \rangle_m &= \sum_{|\alpha| \leq m} \int_{\Omega} D^{\alpha}v D^{\alpha}w d\Omega. \\ \|w\|_m^2 &= \langle w, w \rangle_m. \end{aligned} \quad (4.8)$$

La notation $\langle \cdot, \cdot \rangle$ introduite dans les chapitres précédents correspond maintenant clairement au produit scalaire associé à $H^0(\Omega)$. En général, lorsque nous omettons de le préciser, on parlera des normes et du produit scalaire associés à $H^0(\Omega) = L_2(\Omega)$. Typiquement, on parlera de norme et de produit scalaire de L_2 .

Théorèmes d'immersion de Sobolev

Une dernière question qu'il convient de considérer est la relation entre les espaces de Sobolev et les espaces classiques $C^k(\Omega)$ contenant les fonctions dont les dérivées jusqu'à l'ordre k sont continues. La réponse est fournie par un des théorèmes d'immersion de Sobolev :

³Sans rire, ceci vous est laissé à titre d'exercice...

$$\begin{array}{ll}
\text{Si} & \Omega \subset \mathcal{R}^n, \\
& 2k > n, \\
\text{alors} & H^k(\Omega) \subset \{w \mid w \in C^0(\overline{\Omega}), \\
& \exists c \mid w(x) \mid < c, \quad \forall x \in \Omega\}
\end{array} \tag{4.9}$$

Ce théorème déconseille (à juste titre, d'ailleurs) d'imposer une force ponctuelle sur une membrane dans le cas bidimensionnel, si on souhaite conserver une solution continue. En effet, la solution d'un tel problème appartient seulement à l'espace $H^1(\Omega)$ et comme l'inégalité $2k > n$ du théorème n'est pas satisfaite, la continuité de la solution n'est pas garantie. En pratique, la solution analytique d'un problème de membrane sous charge ponctuelle tend vers une valeur infinie en ce point et n'y est donc pas continue. Ceci contredit l'hypothèse de petits déplacements que contient implicitement le modèle à la base des équations. Par contre, imposer une charge ponctuelle sur la corde est acceptable, car dans le cas unidimensionnel, le lien entre les espaces de fonctions continues d'ordre m et les espaces de Sobolev est nettement plus fort :

$$\begin{array}{ll}
\text{Si} & \Omega \subset \mathcal{R}, \\
\text{alors} & H^{k+1}(\Omega) \subset \{w \mid w \in C^k(\overline{\Omega}), \\
& \exists c \mid w(x) \mid < c, \quad \forall x \in \Omega\}
\end{array} \tag{4.10}$$

Problème elliptique modèle : vérification détaillée des hypothèses

A titre d'exemple, considérons simplement le problème elliptique modèle défini à partir de l'équation de Poisson.

$$\begin{array}{l}
\text{Trouver } u(\mathbf{x}) \in \mathcal{U} \text{ tel que} \\
J(u) = \min_{v \in \mathcal{U}} \underbrace{\left(\frac{1}{2} a(v, v) - b(v) \right)}_{J(v)},
\end{array} \tag{4.11}$$

où nous définissons maintenant les espaces et les formes comme suit :

$$\begin{aligned}\mathcal{U} &= \{w \in H^1(\Omega) \text{ et } w = 0 \text{ sur } \partial\Omega\} \\ a(u, v) &= \langle \nabla u \cdot k \nabla v \rangle \\ b(u) &= \langle fu \rangle\end{aligned}$$

où k est une constante positive et f est une fonction quelconque de L_2 . Pour montrer l'existence et l'unicité du problème abstrait défini par (4.1), il faut montrer que $a(\cdot, \cdot)$ est un produit scalaire pour l'espace \mathcal{U} .

Il est évident que a est bilinéaire et symétrique. On obtient immédiatement la continuité de a en observant que

$$\begin{aligned}|a(u, v)| &\leq \|\nabla u\|_0 \|\nabla v\|_0 \\ &\downarrow \\ &\leq \|u\|_1 \|v\|_1\end{aligned}$$

Pour obtenir la coercivité de a , il suffit de montrer qu'il existe une constante C telle que $\|v\|_0 \leq C\|\nabla v\|_0 \forall v \in \mathcal{U}$. Pour éviter les aspects techniques, nous allons nous limiter

ici au cas unidimensionnel ^{4 5} et écrire que :

$$\begin{aligned}
v(x) - \underbrace{v(0)}_{=0} &= \int_0^x \frac{dv}{dx}(t) dt & \forall x \in \Omega \\
&\downarrow \text{Par l'inégalité de Cauchy } \langle v, x, 1 \rangle \leq \|v, x\| \|1\| \\
(v(x))^2 &\leq \underbrace{\int_0^x dt}_{\leq C} \underbrace{\int_0^x \left(\frac{dv}{dx}(t) \right)^2 dt}_{\leq \left\| \frac{dv}{dx} \right\|_0^2} & \forall x \in \Omega \\
&\downarrow \text{En intégrant sur } \Omega \\
\|v\|_0^2 &\leq C_1 \left\| \frac{dv}{dx} \right\|_0^2 \\
&\downarrow \\
kC_2 \underbrace{\left(\|v\|_0^2 + \left\| \frac{dv}{dx} \right\|_0^2 \right)}_{\|v\|_1^2} &\leq \underbrace{k \left\| \frac{dv}{dx} \right\|_0^2}_{a(v, v)}
\end{aligned}$$

Il est utile de noter que nous avons besoin de la condition à l'extrémité $v(0) = 0$, pour pouvoir contrôler la norme de la fonction à partir de la norme de la dérivée. En d'autres mots, on a besoin d'un *point fixe*. On obtient ainsi la raison mathématique sous-jacente à l'intuition physique que l'espace Γ_D ne peut être vide.

4.3 Estimations d'erreur a priori

L'obtention de l'estimation d'erreur a priori est effectuée en deux étapes :

- D'abord, nous allons obtenir une borne d'erreur sur l'erreur commise $\tilde{e} = u - \tilde{u}^h$ par l'interpolation \tilde{u}^h de la solution exacte u dans un sous-espace $\mathcal{U}^h \subset \mathcal{U}$

⁴La preuve pour le cas bidimensionnel est obtenue de manière totalement analogue.

⁵Dans les majorations, toutes les constantes seront simplement notées C , même si elles sont différentes d'une ligne à l'autre. En général, lors de chaque majoration, de telles constantes grandissent ou diminuent progressivement rendant le résultat final de moins en moins utile, pour le numéricien. Bref, se méfier parfois de l'élégance trompeuse de certaines démonstrations...

- Ensuite, nous montrerons que u^h est la meilleure approximation de la solution discrète et que son erreur $e = u - u^h$ est donc inférieure à l'erreur d'interpolation \tilde{e} et donc aux bornes de cette erreur d'interpolation.

4.3.1 Erreur de l'interpolation : lemme de Bramble-Hilbert

Afin de tenter de rester simple, limitons-nous au cas unidimensionnel et considérons une interpolation polynomiale par morceaux de degré p d'une fonction u appartenant à $C^{p+1}(\Omega)$.

$$\tilde{u}^h(x) = \sum_{j=1}^n u(X_j), \tau_j(x) \quad (4.12)$$

Sur l'élément parent $\hat{\Omega}$, on écrit la relation suivante pour une dérivée $i \leq p$ de l'erreur d'interpolation évaluée en tout point ξ :

$$\begin{aligned}
\frac{d^i \tilde{e}}{d\xi^i} &= \frac{d^i u}{d\xi^i} - \frac{d^i \tilde{u}^h}{d\xi^i} \\
&\downarrow \text{En ajoutant et soustrayant } u^t \text{ qui est le développement en série de Taylor d'ordre } p \text{ de la fonction } u \text{ autour de } \xi \\
&= \frac{d^i u}{d\xi^i} - \frac{d^i u^t}{d\xi^i} + \frac{d^i u^t}{d\xi^i} - \frac{d^i \tilde{u}^h}{d\xi^i} \\
&\downarrow \text{Car l'interpolation d'ordre } p \text{ de } u^t \text{ est lui-même} \\
&= \frac{d^i u}{d\xi^i} - \frac{d^i u^t}{d\xi^i} + \frac{d^i u^t}{d\xi^i} - \frac{d^i \tilde{u}^h}{d\xi^i} \\
&= \frac{d^i u}{d\xi^i} - \frac{d^i u^t}{d\xi^i} + \sum_{j=0}^p (u^t(\Xi_j) - u(\Xi_j)) \underbrace{\frac{d^i \phi_j}{d\xi^i}}_{\leq C} \\
&\downarrow \text{En vertu du théorème du reste du développement de Taylor} \\
&\leq C \max_{\xi \in \hat{\Omega}} \left| \frac{d^{p+1} u}{d\xi^{p+1}} \right|
\end{aligned}$$

On peut alors en déduire le résultat classique pour l'estimation de l'erreur d'interpolation

comme suit

$$\begin{aligned}
\|\tilde{e}\|_m^2 &= \sum_{e=1}^N \sum_{i=0}^m \int_{\Omega_e} \frac{d^i \tilde{e}}{dx^i} \frac{d^i \tilde{e}}{dx^i} dx, \\
&= \sum_{e=1}^N \sum_{i=0}^m \left(\frac{h}{2}\right)^{1-2i} \int_{-1}^1 \frac{d^i \tilde{e}}{d\xi^i} \frac{d^i \tilde{e}}{d\xi^i} d\xi, \\
&\leq C^2 \sum_{e=1}^N \sum_{i=0}^m \left(\frac{h}{2}\right)^{1-2i} \int_{-1}^1 \max_{\xi \in \hat{\Omega}} \left| \frac{d^{p+1} u}{d\xi^{p+1}} \right|^2 d\xi, \\
&\leq C^2 \sum_{e=1}^N \sum_{i=0}^m \left(\frac{h}{2}\right)^{1-2i} \left(\frac{2}{h}\right)^{1-2(p+1)} \int_{\Omega_e} \max_{x \in \Omega_e} \left| \frac{d^{p+1} u}{dx^{p+1}} \right|^2 dx, \\
&\leq C^2 \sum_{e=1}^N \sum_{i=0}^m \left(\frac{h}{2}\right)^{2(p+1-i)} \|u\|_{p+1}^2,
\end{aligned}$$

Observons à nouveau que la constante C (qui est le même symbole pour alléger les notations) grandit lors de chaque majoration et peut donc rendre totalement inutilisable une telle estimation. En faisant tendre la taille de l'élément vers zéro, on obtient l'estimation asymptotique suivante

$$\|\tilde{e}\|_m \leq Ch^{p+1-m} \|u\|_{p+1},$$

lorsque h tend vers 0.

(4.13)

En général, nous ne connaissons ni la constante, ni la norme de la solution exacte dans ce résultat. Il faut également mentionner que ce résultat suppose que la solution exacte appartient à l'espace de Sobolev d'ordre $p+1$. Toutefois, ce résultat nous fournit le taux de convergence asymptotique de la méthode.

Pour une interpolation linéaire par morceaux, on peut écrire :

$$\begin{aligned}
\|\tilde{e}\|_0 &\leq C_0 h^2 \|u\|_2, \\
\|\tilde{e}\|_1 &\leq C_1 h^1 \|u\|_2, \\
\|\tilde{e}\|_* &\leq C_1 h^1 \|u\|_2.
\end{aligned}$$

Ce qui signifie que l'erreur d'interpolation sur la valeur de u diminuera asymptotiquement de manière quadratique, tandis que cette même erreur sur la pente ou sur l'énergie diminuera de manière linéaire.

4.3.2 Théorème de la meilleure approximation

Dans cette section, nous allons établir pourquoi la méthode des éléments finis fonctionne parfaitement lorsque le problème aux équations partielles peut être identifié comme un problème elliptique. Dans ce cas, on peut montrer que la méthode fournira la meilleure approximation possible de la solution exacte.

Rappelons que le problème exact et le problème discret peuvent être écrits sous une forme générique de recherche d'un minimum

<p>Trouver $u \in \mathcal{U}$ tel que</p> $J(u) = \min_{v \in \mathcal{U}} \underbrace{\frac{1}{2} a(v, v) - b(v)}_{J(v)},$	(4.14)
<p>Trouver $u^h \in \mathcal{U}^h \subset \mathcal{U}$ tel que</p> $J(u^h) = \min_{v^h \in \mathcal{U}^h} \underbrace{\frac{1}{2} a(v^h, v^h) - b(v^h)}_{J(v^h)},$	

Ecrivons les conditions de stationnarité des deux problèmes en choisissant une même fonction arbitraire \hat{u}^h pour les deux problèmes.

- Condition de stationnarité de la solution exacte u

$$a(u, \hat{u}) = b(\hat{u}), \quad \forall \hat{u} \in \hat{\mathcal{U}},$$

$$\downarrow \quad \text{Car } \hat{\mathcal{U}}^h \text{ est inclus dans } \hat{\mathcal{U}},$$

$$a(u, \hat{u}^h) = b(\hat{u}^h), \quad \forall \hat{u}^h \in \hat{\mathcal{U}}^h,$$

On voit qu'il est essentiel que les espaces discrets soient inclus dans les espaces originaux. C'est le critère de conformité qui justifie l'introduction des éléments hermitiens pour l'exemple de la poutre, ou d'éléments continus pour l'exemple de la corde. Il est toutefois assui possible d'utiliser des éléments non conformes dans des formulations mixtes.

- Condition de stationnarité de la solution discrète u^h

$$a(u^h, \widehat{u}^h) = b(\widehat{u}^h), \quad \forall \widehat{u}^h \in \widehat{\mathcal{U}}^h,$$

En notant *l'erreur de discrétisation* par $e = u - u^h$, nous pouvons écrire que l'erreur est orthogonale à tout élément de $\widehat{\mathcal{U}}^h$ au sens du produit scalaire énergétique $a(\cdot, \cdot)$. En effet, en soustrayant les deux conditions de stationnarité entre elles, on obtient immédiatement :

$$\boxed{a(\underbrace{u - u^h}_e, \widehat{u}^h) = 0, \quad \forall \widehat{u}^h \in \widehat{\mathcal{U}}^h,} \quad (4.15)$$

De cette propriété d'orthogonalité de l'erreur, on peut déduire le théorème de la meilleure approximation en énergie en écrivant simplement

$$\begin{aligned} a(u - v^h, u - v^h) &= a(u - u^h + u^h - v^h, u - u^h + u^h - v^h), \\ &\downarrow \text{En vertu de la définition de } e, \\ &= a(e + u^h - v^h, e + u^h - v^h), \\ &\downarrow \text{En vertu de la bilinéarité de } a, \\ &= a(e, e) + a(u^h - v^h, u^h - v^h) + \underbrace{2a(e, u^h - v^h)}_{= 0} \\ &\downarrow \text{Car } u^h - v^h \text{ est un élément de } \widehat{\mathcal{U}}^h, \\ &= a(e, e) + \underbrace{a(u^h - v^h, u^h - v^h)}_{\geq 0} \\ &\geq a(e, e) \end{aligned}$$

Il est alors possible d'énoncer le théorème de la meilleure approximation :

$$\boxed{\|u - v^h\|_* \geq \underbrace{\|u - u^h\|_*}_e, \quad \forall v^h \in \mathcal{U}^h,} \quad (4.16)$$

On vient donc de montrer que l'approximation u^h est la meilleure approximation au sens de la norme énergétique qu'il est possible de trouver dans l'espace discret. Deux corollaires sont souvent utiles :

$$\boxed{\|u - v^h\|_*^2 = \underbrace{\|u - u^h\|_*^2}_e + \|u^h - v^h\|_*^2, \quad \forall v^h \in \mathcal{U}^h,} \quad (4.17)$$

$$\boxed{\|u\|_*^2 = \underbrace{\|u - u^h\|_*^2}_e + \|u^h\|_*^2, \quad \text{si } \widehat{\mathcal{U}}^h = \mathcal{U}^h,} \quad (4.18)$$

Ces deux résultats permettent de réaliser une interprétation de la méthode des éléments finis en termes géométriques dans l'espace \mathcal{U} . Il suffit de représenter symboliquement l'espace \mathcal{U} comme le plan dont l'origine est la fonction identiquement nulle, et les espaces $\widehat{\mathcal{U}}^h$ et \mathcal{U}^h comme des droites parallèles dont la première passe par l'origine. En accord avec (4.16), u^h est la meilleure approximation de u , puisqu'il s'agit de sa projection. La méthode des éléments finis apparaît comme une méthode de projection puisqu'elle permet d'obtenir la projection d'une fonction totalement inconnue u , en connaissant uniquement le problème dont elle est la solution.

4.3.3 Lemme de Cea

Pour vraiment pouvoir tirer profit de (4.16), il faut encore établir que cette meilleure approximation au sens de l'énergie correspond également à une meilleure approximation au sens de la norme usuelle de \mathcal{U} , dans le cas qui nous concerne une norme de Sobolev. Ceci est uniquement rendu possible par le caractère elliptique du problème considéré qui implique que la forme a est continue et coercive. En d'autres mots, il convient maintenant d'observer que $a(\cdot)$ a bien le statut d'un produit scalaire et qu'un résultat obtenu pour une telle norme est valable pour toutes les autres normes de l'espace \mathcal{U} .

On observe ainsi

$$\|e\|^2 \leq \frac{1}{\alpha} \|e\|_*^2 \leq \frac{1}{\alpha} \|\tilde{e}\|_*^2 \leq \frac{c}{\alpha} \|\tilde{e}\|^2,$$

\uparrow En vertu de la continuité de a ,
 \uparrow Car u^h est la meilleure approximation énergétique,
 \uparrow En vertu de la coercivité de a ,

On obtient ainsi ce qui est connu comme le lemme de Cea

$\|e\|^2 \leq \frac{c}{\alpha} \|\tilde{e}\|^2$

(4.19)

C'est la dernière brique de l'édifice mathématique permettant d'obtenir une estimation d'erreur a priori. Il suffit en effet alors d'observer

$$\|e\|_m^2 \leq \frac{c}{\alpha} \|\tilde{e}\|_m^2 \leq \frac{c}{\alpha} C^2 h^{2(p+1-m)} \|u\|_{p+1}^2,$$

\uparrow Estimation de l'erreur d'interpolation,
 \uparrow En vertu du lemme de Cea,

Ce qui permet d'obtenir l'estimation classique de l'erreur d'une méthode d'éléments finis pour un problème elliptique

$\|e\|_m^2 \leq \frac{c}{\alpha} C^2 h^{2(p+1-m)} \|u\|_{p+1}^2,$

(4.20)

A nouveau, lorsque la solution exacte n'est pas suffisamment régulière pour appartenir à H^{p+1} , il n'est alors possible que d'écrire

$$\|e\|_m^2 \leq \frac{c}{\alpha} C^2 h^{2(\min(p+1,r)-m)} \|u\|_r^2,$$

où r correspond à l'indice de l'espace de Sobolev d'ordre le plus élevé contenant la solution exacte. Ce nombre r caractérise la régularité de notre solution. On voit ainsi que si la solution n'est pas suffisamment régulière, on perdra tout le bénéfice du taux de convergence exponentiel des méthodes d'ordre élevé ⁶.

4.4 Estimation d'erreur a posteriori

Nous souhaitons à présent calculer une estimation de l'erreur a posteriori afin d'avoir une idée plus précise de la fiabilité de notre solution discrète après l'avoir obtenue. Supposons donc que nous ayons un maillage composé de N éléments Ω_e couvrant le domaine Ω et que nous disposions de la solution discrète u^h pour ce maillage et le problème elliptique modèle décrit par :

<p>Trouver $u(\mathbf{x}) \in \hat{\mathcal{U}}$ tel que</p> $\underbrace{\langle (\nabla \hat{u}) \cdot (k \nabla u) \rangle}_{a(\hat{u}, u)} = \underbrace{\langle \hat{u} f \rangle}_{b(\hat{u})}, \quad \forall \hat{u} \in \hat{\mathcal{U}},$	(4.21)
<p>Trouver $u^h(\mathbf{x}) \in \hat{\mathcal{U}}^h$ tel que</p> $\underbrace{\langle (\nabla \hat{u}^h) \cdot (k \nabla u^h) \rangle}_{a(\hat{u}^h, u^h)} = \underbrace{\langle \hat{u}^h f \rangle}_{b(\hat{u}^h)}, \quad \forall \hat{u}^h \in \hat{\mathcal{U}}^h,$	

On déduit immédiatement que l'erreur de discrétisation $e(\mathbf{x})$ satisfait à la relation :

$$a(\hat{u}, \underbrace{u - u^h}_e) = -a(\hat{u}, u^h) + b(\hat{u}), \quad \forall \hat{u} \in \hat{\mathcal{U}},$$

Nous souhaitons maintenant évaluer de manière approchée les normes énergétiques

⁶Au grand dam de nombreux chercheurs qui restent encore aujourd'hui attirés par le taux de convergence théoriquement exponentiel des méthodes d'ordre élevé.

globale et locale de l'erreur de cette discrétisation données par

$$\begin{aligned}
\| \underbrace{u - u^h}_e \|_*^2 &= a(e, e) \\
&= \sum_{e=1}^N \underbrace{\int_{\Omega_e} (\nabla e) \cdot (k \nabla e) \, d\Omega}_{a_e(e, e)}, \\
&= \sum_{e=1}^N \|e\|_{*, \Omega_e}^2
\end{aligned}$$

et pour lesquelles, nous allons tenter d'obtenir des estimations a posteriori

$$\begin{aligned}
\theta^h &\approx \theta = \|e\|_* \\
\theta_e^h &\approx \theta_e = \|e\|_{*, \Omega_e}
\end{aligned} \tag{4.22}$$

Une manière élémentaire de procéder est de construire une approximation e^{h+} de l'erreur e (ou une approximation u^{h+} de la solution exacte u) d'un ordre supérieur à celui de l'approximation u^h , et de résoudre le problème discret qui y est associé. Cela consiste par exemple à augmenter le degré de l'approximation et à comparer la solution u^{h+} à la solution u^h . En appliquant (4.17) en considérant u^{h+} comme la solution et u^h un élément quelconque de \mathcal{U}^{h+} , on obtient immédiatement

$$\begin{aligned}
\underbrace{\|u^{h+} - u^h\|_*^2}_{(\theta^h)^2} + \|u - u^{h+}\|_*^2 &= \underbrace{\|u - u^h\|_*^2}_{\theta^2}, & \forall u^h \in \mathcal{U}^h \subset \mathcal{U}^{h+} \\
\downarrow & \\
\theta^h &\leq \theta
\end{aligned}$$

Ce qui est un résultat assez décevant, puisque l'estimation d'erreur a posteriori peut théoriquement fournir une valeur nulle, alors que l'erreur réelle pourrait être très importante... Toutefois, en pratique, une telle approche fonctionne en général parfaitement bien : l'exemple typique est la technique de raffinement successif des maillages et la comparaison des solutions correspondantes. Reprenons (4.17) et considérons des espaces \mathcal{U}^{h+} de plus en plus grand, ou en d'autres mots, considérons la limite asymptotique de $h+ \rightarrow 0$ si nous divisons chaque élément en sous-éléments (ou $p+ \rightarrow 0$, si nous augmentons le degré

des fonctions de forme) tout en gardant inchangé u^h . Nous pouvons alors écrire

$$\underbrace{\|u^{h+} - u^h\|_*^2}_{(\theta^h)^2} + \underbrace{\|u - u^{h+}\|_*^2}_{\rightarrow 0} = \underbrace{\|u - u^h\|_*^2}_{\theta^2}, \quad \forall u^h \in \mathcal{U}^h \subset \mathcal{U}^{h+}$$

\downarrow

En considérant la limite asymptotique $h+ \rightarrow 0$

$$\theta^h \rightarrow \theta$$

C'est déjà plus sympathique : notre estimation convergera vers l'erreur exacte si nous ajoutons suffisamment de degrés de liberté !

Le vrai défi est d'obtenir une estimation d'erreur qui soit une vraie norme de l'erreur e , c'est-à-dire telle que

$$C_1 \|e\|_* \leq \theta \leq C_2 \|e\|_*$$

où les valeurs des constantes doivent être aussi proches de l'unité que possible pour que la valeur numérique fournie par l'estimation d'erreur a posteriori soit pratiquement utilisable. On parlera typiquement de *sharp error estimations*. Le second défi est d'estimer ensuite l'erreur sur les données finales du calcul : erreur en un point spécifique, sur un flux global de chaleur. Mais ceci sort du cadre de cet exposé.

L'inconvénient d'une telle approche est que l'estimation de l'erreur nécessite davantage d'efforts que le calcul de la première solution. Toutefois, on peut obtenir plus efficacement une approximation a posteriori de l'erreur. Il suffit de choisir un espace \mathcal{U}^{h+} des fonctions de forme avec un support limité à un seul élément. Des telles fonctions de forme sont appelées *fonctions bulles*. Dans ce cas, le calcul de l'estimateur d'erreur peut être fait de manière locale sur chaque élément car toutes les équations associées aux degrés de liberté présents sur un élément sont totalement découplées des autres équations. Il n'y a pas de système global à résoudre. Toutefois, il faut désormais observer que \mathcal{U}^h n'est alors plus inclus dans \mathcal{U}^{h+} : en d'autres mots, il n'est plus certain que l'estimation d'erreur ainsi calculée convergera vers la vraie erreur.

4.4.1 Stratégies adaptatives

Une stratégie adaptative peut être comparée à un paradigme de contrôle optimal dans lequel l'erreur de discrétisation est contrôlée par un schéma adaptatif qui orchestre la répartition de la taille des éléments afin d'obtenir un niveau de précision fixé à un coût minimal. Un tel objectif est obtenu avec un maillage qui présente une distribution uniforme de l'erreur.

Il convient donc tout d'abord de se fixer un objectif de précision θ . Ensuite la stratégie adaptative peut être vue comme la succession des étapes suivantes :

- 1 On calcule une première solution discrète u^h .
- 2 On estime a posteriori l'erreur θ^h .
- 3 Si $\theta^h > \theta$, nous raffinons le maillage afin d'atteindre l'objectif de précision avec un nombre minimum de valeurs nodales.
- 4 On recalcule une solution discrète u^{h+1} sur le nouveau maillage et on recommence le processus jusqu'à satisfaction de l'objectif.

L'utilisation conjointe des estimations d'erreur a priori et a posteriori permet de proposer une stratégie simple et efficace pour l'obtention totalement automatique de maillages quasi-optimaux. Nous décidons de modifier le maillage en divisant chaque élément Ω_e en un nombre n_e de sous-éléments, nombre qu'il s'agit dès lors de prédire.

- D'une part, on souhaite atteindre l'objectif de précision avec une distribution uniforme de l'erreur. Dans le cas unidimensionnel, cela revient à demander sur chaque élément Ω_e

$$\theta_e^2(n_e) = \frac{\theta^2}{N}, \quad (4.23)$$

où la fonction $\theta_e(n_e)$ fournit la norme énergétique sur Ω_e de l'erreur lorsque cet élément est divisé en n_e sous-éléments.

- D'autre part, l'estimation a posteriori sur le premier maillage permet d'écrire

$$\theta_e^2(1) \approx (\theta^h)^2. \quad (4.24)$$

- Finalement, on estime le taux de convergence asymptotique α de la norme énergétique par l'estimation a priori et on peut extrapoler l'estimation *a posteriori* lorsque l'on introduira un nombre fixé de sous-éléments,

$$\begin{aligned} \theta_e^2(1) &\approx h_e^{2\alpha} C^2, \\ \theta_e^2(n_e) &\approx \left(\frac{h_e}{n_e}\right)^{2\alpha} C^2 \approx n_e^{-2\alpha} \theta_e^2(1) \end{aligned} \quad (4.25)$$

En éliminant $\theta_e(n_e)$ de (4.23), (4.24) et (4.25), nous déduisons que le nombre de sous-éléments à introduire au sein de chaque élément Ω_e est l'entier le plus proche de

$$\left((N) \frac{(\theta^h)^2}{\theta^2} \right)^{\frac{1}{2\alpha}}. \quad (4.26)$$

Exemple élémentaire

Considérons le problème de la corde résolu avec des fonctions de forme linéaires. Nous introduisons une unique fonction de forme supplémentaire pour u^{h+} sur l'élément parent :

$$\phi_e^{h+}(\xi) = (1 - \xi)(1 + \xi),$$

et la norme énergétique de l'erreur a posteriori est directement obtenue sur chaque élément en calculant :

$$\theta_e^h = \frac{(-a(u^h, \phi_e^{h+}) + \langle f, \phi_e^{h+} \rangle)}{\sqrt{a(\phi_e^{h+}, \phi_e^{h+})}}.$$

Le taux de convergence $\alpha = 1$. On obtient donc le nombre de divisions à effectuer dans chaque élément pour obtenir une erreur globale inférieure à θ comme l'entier le plus proche de

$$\sqrt{N} \frac{\theta_e^h}{\theta}$$

Chapitre 5

Méthodes d'éléments finis pour des problèmes d'advection-diffusion

Jusqu'à présent, nous avons appliqué la méthodologie des éléments finis à des problèmes linéaires elliptiques. Cela permettait d'obtenir des résultats discrets avec des propriétés tout à fait satisfaisantes. Nous souhaitons maintenant considérer des problèmes qui ont un caractère hyperbolique prononcé, c'est-à-dire des problèmes d'advection-diffusion avec peu ou pas de diffusion. De tels problèmes sont courants en mécanique des fluides, en dynamique des gaz ou en propagation des ondes.

L'utilisation des éléments finis dans une formulation usuelle (c'est-à-dire l'application de la technique de Galerkin) pour de tels problèmes hyperboliques fournit souvent des résultats inacceptables, contrairement à ce que nous avons observé dans le cas de problèmes elliptiques. Plus précisément, on observe que les méthodes classiques d'éléments finis pour des problèmes hyperboliques ne sont pas adéquates lorsque la solution exacte n'est pas *lisse*. Si la solution exacte présente un saut ou une discontinuité, alors la solution discrète sera polluée par des oscillations parasites même à une grande distance du saut. C'est évidemment une difficulté majeure puisque dans la plupart des applications avec des équations hyperboliques, la solution présente des sauts ou des chocs.

Toutefois, la mise au point de nouvelles formulations pour les méthodes d'éléments finis a permis de contourner partiellement cette difficulté et d'obtenir des solutions discrètes acceptables. Dans ce chapitre, nous présentons des techniques spécialisées pour utiliser les éléments finis pour des équations avec un caractère hyperbolique prononcé.

Nous allons successivement considérer le cas d'une équation scalaire de transport (ou équation purement hyperbolique du premier ordre), puis d'une équation scalaire d'advection-diffusion (ou formellement, une équation elliptique du second ordre qui contient un terme hyperbolique du premier ordre très important).

5.1 Equation scalaire de transport

A titre d'exemple de problème purement hyperbolique, considérons le problème suivant :

Trouver $u(\mathbf{x})$ tel que

$$\begin{aligned}\boldsymbol{\beta} \cdot \nabla u &= f, & \forall x \in \Omega, \\ u &= 0, & \forall x \in \Gamma_-\end{aligned}$$

(5.1)

où l'inconnue u est un scalaire représentant, par exemple, une concentration, le vecteur $\boldsymbol{\beta} = (\beta_x, \beta_y, \beta_z)$ est une vitesse donnée qui peut éventuellement être une fonction de la position.

Les courbes caractéristiques d'une telle équation sont données par

$$\frac{d\mathbf{x}}{ds}(s) = \boldsymbol{\beta}(\mathbf{x}, s), \quad (5.2)$$

où s est le paramètre de la courbe. Si nous supposons que $\boldsymbol{\beta}$ est Lipschitz-continu, on peut montrer qu'en chaque point de Ω , il passe une et une seule courbe caractéristique. Il est donc possible d'obtenir la solution de l'équation scalaire de transport en intégrant le long des caractéristiques, si une valeur de la solution est fournie sur la courbe de la solution. C'est pourquoi on choisit classiquement d'imposer une condition aux limites sur une partie de la frontière (en général, le morceau par lequel les caractéristiques entrent dans le domaine) qu'on notera Γ_- et qu'on définira par

$$\Gamma_- = \{\mathbf{x} \in \partial\Omega \text{ tel que } \mathbf{n}(\mathbf{x}) \cdot \boldsymbol{\beta}(\mathbf{x}) \leq 0\} \quad (5.3)$$

où $\mathbf{n}(\mathbf{x})$ représente la normale sortant en un point \mathbf{x} de la frontière.

Exemple vraiment élémentaire

Il est important d'observer que la solution du problème (5.2) peut être discontinue. Ainsi, si la concentration prescrite sur Γ_- présente un saut ou une discontinuité en un point \mathbf{x}_c , alors la solution sera discontinue tout le long de la courbe caractéristique passant par \mathbf{x}_c . Afin d'illustrer ce point, considérons simplement l'exemple de \mathbb{R}^2

$$\begin{aligned}\frac{\partial u}{\partial x}(x, y) &= 0, & 0 < x < 2, \ 0 < y < 2, \\ u(0, y) &= 1, & 0 < y < 1, \\ u(0, y) &= 0, & 1 < y < 2.\end{aligned} \quad (5.4)$$

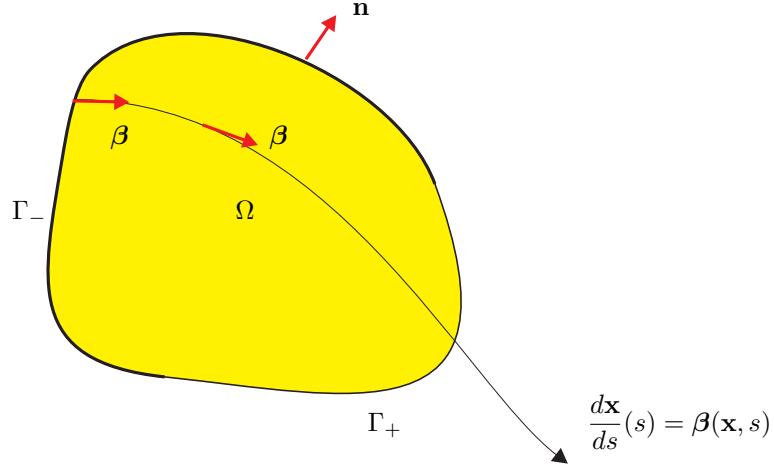


Figure 5.1: Conditions aux limites pour une équation scalaire de transport.

Cela correspond à $\beta = (1, 0)$ et $f = 0$. Les courbes caractéristiques sont ici toutes les droites horizontales. La solution analytique du problème est clairement donnée par

$$\begin{aligned} u(x, y) &= 1, & 0 < x < 2, 0 < y < 1, \\ u(x, y) &= 0, & 0 < x < 2, 1 < y < 2. \end{aligned} \quad (5.5)$$

5.1.1 A propos de la méthode des caractéristiques

La résolution d'équations scalaires de transport dans un espace à plusieurs dimensions ou d'équations vectorielles¹ de transport dans un espace à une dimension est très souvent réalisée par la méthode des caractéristiques. En pratique, cela revient à trouver les courbes caractéristiques et ensuite à intégrer le long de ces courbes. Dans les deux cas, on fait appel aux méthodes numériques d'intégration d'équations différentielles ordinaires : citons à titre d'exemple, les méthodes de Runge-Kutta, Euler...

En théorie, c'est la meilleure approche pour des équations purement hyperboliques. Toutefois, elle n'est pas facile à mettre en oeuvre pour la résolution d'un système d'équations dans un espace à plusieurs dimensions et elle n'est pas utilisable dans le cas d'une équation d'advection-diffusion. En outre, cette méthode des caractéristiques ne peut pas être construite sur un maillage d'éléments finis qui n'est pas aligné a priori avec les courbes caractéristiques.

Nous allons donc maintenant tenter de construire une méthode numérique basée sur une grille fixée. Typiquement, il s'agit de méthodes d'éléments finis ou de différences finies. L'utilisation d'un maillage fixe permet une programmation plus facile du schéma numérique, mais sera la source de difficulté numérique si la solution exacte n'est pas lisse

¹c'est-à-dire un système d'équations linéaires.

et présente des discontinuités. L'usage des différences finies centrées conventionnelles ou de formulation de Galerkin pour les éléments finis produiront des solutions numériques oscillantes.

5.1.2 Méthode de Galerkin

Introduisons donc une approximation polynomiale par morceaux u^h continue sous la forme habituelle :

$$u^h(x) = \sum_{j=1}^n U_j \tau_j(x) \quad (5.6)$$

où les $\tau_j(x)$ sont les fonctions de forme globales. Par exemple, supposons que ces fonctions sont linéaires par morceaux et continues.

Pour obtenir une formulation discrète, nous allons nous rappeler que les éléments finis sont une méthode de résidus pondérés. Comme a priori, les éléments de \mathcal{U}^h (y compris, u^h le meilleur d'entre eux, en général) ne satisfont pas l'équation différentielle de (5.1), ils donnent lieu à un résidu non nul

$$r^h = \beta \cdot \nabla u^h - f. \quad (5.7)$$

La méthode des résidus pondérés consiste alors à sélectionner n fonctions poids w_i et évaluer les valeurs nodales afin de minimiser les n intégrales du produit du résidu et de chaque fonction poids.

$$\langle w_i r^h \rangle = 0, \quad i = 1, \dots, n, \quad (5.8)$$

Parmi les manières de minimiser le résidu, la technique classique dite de Galerkin consiste à annuler en moyenne le produit des résidus avec les fonctions de forme. En d'autres mots, on sélectionne $w_i = \tau_i$. Ce choix influence fortement la précision (et donc la fiabilité) de l'approximation produite. Pour des problèmes elliptiques, la technique de Galerkin est optimale.

Ce n'est malheureusement plus le cas pour des problèmes hyperboliques : la technique de Galerkin n'est plus le meilleur choix pour de tels problèmes. Afin d'illustrer cela,

effectuons un peu d'algèbre

$$\begin{aligned}
& \langle \tau_i r^h \rangle = 0, \quad i = 1, \dots, n, \\
& \quad \downarrow \\
& \text{Par définition de } r^h, \\
& \langle \tau_i (\boldsymbol{\beta} \cdot \nabla u^h) \rangle - \langle \tau_i f \rangle = 0, \quad i = 1, \dots, n, \\
& \quad \downarrow \\
& \text{Par définition de } u^h, \\
& \sum_{j=1}^n \underbrace{\langle \tau_i \boldsymbol{\beta} \cdot \nabla \tau_j \rangle}_{A_{ij}} U_j - \underbrace{\langle \tau_i f \rangle}_{B_i} = 0, \quad i = 1, \dots, n,
\end{aligned}$$

afin d'écrire un problème discret correspondant à la technique de Galerkin ²

Trouver $U_j \in \mathbb{R}^n$ tel que

$$\sum_{j=1}^n \underbrace{\langle \tau_i \boldsymbol{\beta} \cdot \nabla \tau_j \rangle}_{A_{ij}} U_j = \underbrace{\langle \tau_i f \rangle}_{B_i}, \quad i = 1, \dots, n,$$

(5.9)

On peut facilement ³ observer que la matrice A_{ij} n'est plus une matrice symétrique définie positive. En d'autres mots, la formulation discrète obtenue par l'application de la technique de Galerkin ne fournit plus la solution d'un problème de minimisation. Les éléments finis et l'application de la méthode de Galerkin sortent du cadre de la théorie de la meilleure approximation : plus rien ne nous dit qu'une telle formulation discrète forme un problème bien posé.

Et il est facile de montrer, sur un exemple élémentaire, que le problème discret construit avec la technique de Galerkin est un problème mal posé pour certains termes sources f . C'est l'objet de la section suivante qui illustrera ce point dans le cadre simplifié d'un problème unidimensionnel.

²Il faut évidemment retirer les noeuds sur lesquels une condition aux limites doit être appliquée.

³Ceci est laissé à titre d'exercice *facile* avec une attention spéciale pour les étudiants en mathématiques appliquées.

5.1.3 Cas unidimensionnel

Afin d'illustrer les performances parfois modestes de la technique de Galerkin pour des équations hyperboliques, nous allons nous restreindre au cas unidimensionnel. Supposons donc que nous ayons à résoudre le problème suivant :

$$\begin{aligned}\frac{du}{dx} &= f, & \forall x \in \Omega =]0, 1[, \\ u(0) &= 0,\end{aligned}\tag{5.10}$$

Ecrivons maintenant une formulation faible en utilisant la technique de Galerkin et en s'inspirant de (5.9).

Trouver $u(x) \in \mathcal{U}$ tel que

$$\langle \hat{u} \frac{du}{dx} \rangle = \langle \hat{u} f \rangle, \quad \forall \hat{u} \in \mathcal{U},$$

(5.11)

On observe que \mathcal{U} peut être défini comme étant un sous-espace de $L_2(\Omega)$ contenant des fonctions s'annulant en $x = 0$. On voit donc qu'on peut avoir des solutions discontinues à un tel problème, si le terme source f est une distribution de Dirac. Dans un tel cas, la solution sera la fonction échelon : ce qui est bien une fonction discontinue.

Une approximation polynomiale par morceaux continue ou discontinue ?

Il n'est pas évident que le choix d'une approximation continue soit optimal pour un tel problème. Nous allons donc considérer deux approximations linéaires par morceaux possibles de u

- Introduisons d'abord une approximation polynomiale par morceaux u^h continue de la forme habituelle

$$u^h(x) = \sum_{j=1}^n U_j \tau_j(x) \tag{5.12}$$

où les U_j sont les *valeurs nodales globales*, tandis que les $\tau_j(x)$ sont les *fonctions de forme globales* que nous supposons linéaires et continues.

- Introduisons ensuite une approximation polynomiale par morceaux discontinue dont l'expression u_e^h sur chaque élément Ω_e est donnée par

$$u_e^h(x) = \sum_{j=1}^2 U_j^e \phi_j(x) \quad (5.13)$$

où les U_j^e sont les *valeurs nodales locales*, tandis que les $\phi_j(x)$ sont les *fonctions de forme locales* que nous supposerons linéaires. On n'impose plus que les valeurs nodales en un même noeud soient identiques. A un noeud global j situé entre l'élément Ω_e et l'élément Ω_{e+1} , nous distinguerons $U_j^+ = U_1^{e+1}$ et $U_j^- = U_2^e$. Ces valeurs restent distinctes, bien qu'elles soient supposées être l'approximation à droite et à gauche de la valeur exacte $u(X_j)$ et il faudra donc bien exprimer d'une certaine manière qu'elles ne sont pas censées être totalement différentes... La forme globale de l'approximation linéaire par morceaux et discontinue peut s'écrire sous la forme

$$u_{disc}^h(x) = \sum_{j=1}^{n-1} U_j^+ \tau_j^+(x) + \sum_{j=2}^n U_j^- \tau_j^-(x) \quad (5.14)$$

On observe immédiatement que la fonction échelon pourra être représentée de manière parfaite par une approximation discontinue, tandis qu'elle ne pourra qu'être imparfaitement représentée par une approximation continue.

Méthode de Galerkin et une approximation continue

Afin de satisfaire la condition frontière, on utilise un espace discret \mathcal{U}^h ne contenant que des fonctions s'annulant en $x = 0$ et on écrit

Trouver $U_j \in \mathbb{R}^{n-1}$ tel que

$$\sum_{j=1}^n \underbrace{\langle \tau_i \tau_{j,x} \rangle}_{A_{ij}} U_j = \underbrace{\langle \tau_i f \rangle}_{B_i}, \quad i = 1, \dots, n, \quad (5.15)$$

Nous avons simplement imposé que $U_1 = 0$, et le problème discret consiste maintenant à trouver la solution d'un système discret de $n - 1$ équations à $n - 1$ inconnues.

Si le terme source est une fonction relativement lisse, nous obtiendrons une solution discrète satisfaisante. Mais, utiliser la même procédure avec un terme source qui est une distribution de Dirac provoquera l'apparition d'oscillations. Mathématiquement, on peut montrer que la solution discrète ainsi obtenue convergera vers la solution exacte, uniquement si le terme source est suffisamment lisse.

Méthodes de Petrov-Galerkin et une approximation continue

L'idée de base des méthodes de Petrov-Galerkin est d'obtenir un système discret dont la matrice serait définie positive ou tendrait à s'en rapprocher. Dans un tel cas de figure, on espère intuitivement que le problème discret correspondrait, à nouveau, à un problème de minimisation et serait à nouveau un problème bien posé.

Une manière de réaliser un tel objectif serait d'écrire les équations discrètes (5.9) en ne remplaçant pas w_i par τ_i

Trouver $U_j \in \mathbb{R}^{n-1}$ tel que

$$\sum_{j=2}^n \underbrace{\langle w_i \tau_{j,x} \rangle}_{A_{ij}} U_j = \underbrace{\langle w_i f \rangle}_{B_i}, \quad i = 2, \dots, n,$$

(5.16)

On définit les *formulations dites de Petrov-Galerkin* comme des formulations de résidus pondérés où les poids ne sont pas les fonctions de forme. A l'opposé, les *formulations dites de Galerkin* ou *formulations de Bubnov-Galerkin*, ou encore *formulations classiques de Galerkin* sont caractérisées comme des formulations de résidus pondérés où les poids sont les fonctions de forme.

Une première manière de réaliser notre objectif serait de sélectionner $w_i = \tau_{i,x}$. Dans une telle formulation, on observe immédiatement que la matrice discrète sera symétrique définie positive et donc qu'un tel problème discret correspond bien à un problème de minimisation. On peut alors constater que cette formulation correspond à l'application de la méthode de Galerkin du problème elliptique suivant :

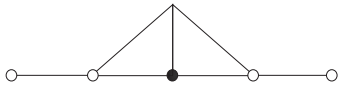
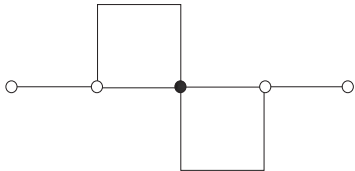

$$\begin{aligned} \frac{d^2 u}{dx^2} + \frac{df}{dx} &= 0, \quad \forall x \in \Omega, \\ u(0) &= 0, \\ u_{,x}(1) &= 0, \end{aligned} \tag{5.17}$$

qui est presque équivalent au problème original, à l'exception d'une condition limite de Neumann additionnelle à la sortie qui est totalement arbitraire et va donc générer beaucoup d'effets indésirables. Tout est parfait, à l'exception du fait que l'on va converger vers une mauvaise solution...

Il existe d'autres possibilités. Physiquement, on peut observer que l'information se propage le long du domaine de gauche à droite : il semble donc relativement logique d'évaluer U_j en ne tenant compte que de la contribution gauche de ce résidu. Cela peut être réalisé en utilisant, comme fonction de pondération discrète, une fonction de forme constante sur l'élément situé à gauche du noeud j . Une telle manière de procéder consiste

à créer des équations discrètes correspondant à des différences finies amont ou à un schéma d'Euler pour l'équation différentielle ordinaire (5.11).

Pour ces trois fonctions de pondération, l'équation discrète d'un résidu associé au noeud j est donnée afin d'observer l'analogie avec des schémas de différences finies.

<p><i>Galerkin</i> $w_i = \tau_i$</p> 	<p><i>Différences finies centrées</i></p> <p>Simple et donc tentant... Oscillations numériques si f n'est pas lisse !</p> $\frac{U_{i+1} - U_{i-1}}{2h} = \frac{F_{i+1} + 4F_i + F_{i-1}}{6},$
<p><i>Petrov-Galerkin</i> $w_i = \tau_{i,x}$</p> 	<p><i>Différences finies centrées d'ordre deux</i></p> <p>Mathématiquement, tentant Condition frontière parasite !</p> $\frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} = \frac{F_{i+1} - F_{i-1}}{2h},$
<p><i>Petrov-Galerkin</i> $w_i = \tau_{i-1}^{cst}$</p> 	<p><i>Différences finies amont</i></p> <p>Quasiment optimal... Correspond à une intégration le long de la caractéristique, Pas d'oscillation numérique</p> $\frac{U_i - U_{i-1}}{h} = \frac{F_i + F_{i-1}}{2},$

Méthode de Galerkin avec des conditions frontières faiblement respectées

Considérons d'abord une approximation continue. Dire que la condition frontière n'est pas strictement respectée consiste à utiliser un espace discret \mathcal{U}^h ne satisfaisant pas a priori les conditions aux limites. En d'autres mots, il peut contenir des fonctions ne s'annulant

pas en $x = 0$. Le respect des conditions aux limites sera faiblement imposé en modifiant la formulation discrète comme suit :

Trouver $U_j \in \mathbb{R}^n$ tel que

$$\sum_{j=1}^n \langle \tau_1 \tau_{j,x} \rangle U_j + U_1 = \langle \tau_1 f \rangle , \quad (5.18)$$

$$\sum_{j=1}^n \langle \tau_i \tau_{j,x} \rangle U_j = \langle \tau_i f \rangle , \quad i = 2, \dots, n,$$

où le terme additionnel correspond à un terme $\ll u^h \tau_i \gg_-$ où $\ll . \gg_-$ correspond à l'intégration sur Γ_- . On ajoute donc aux résidus liés à l'équation différentielle un résidu calculé sur la frontière correspondant à la condition frontière à l'entrée. Ce terme additionnel s'annulera si la solution discrète satisfait les conditions aux limites. A nouveau, on peut montrer que la solution discrète ainsi obtenue converge vers la solution exacte, si le terme source est suffisamment lisse.

Cette manière d'imposer les conditions aux limites est plus coûteuse et plus complexe que notre manière habituelle de procéder. En pratique, son intérêt n'est pas directement évident pour une approximation continue. Toutefois, elle permet de construire des équations discrètes pour l'obtention des valeurs nodales d'une approximation discontinue.

Considérons d'abord une approximation discontinue avec un et un seul élément (en fait, il n'y a alors pas de différence entre approximations continue et discontinue !). Le problème discret consiste à trouver la solution du système discret de 2 équations à 2 inconnues. Les deux équations s'écrivent :

$$\begin{aligned} \sum_{j=1}^2 A_{1j}^1 U_j^1 + U_1^1 &= B_1^1 \\ \sum_{j=1}^2 A_{2j}^1 U_j^1 &= B_2^1 \end{aligned} \quad (5.19)$$

où la matrice A_{ij}^1 et le vecteur B_i^1 sont donnés par les expressions de (5.9) où l'intégration est effectuée seulement sur Ω_1 .

Pour une approximation discontinue avec plusieurs éléments, l'utilisation de (5.18) sur chaque élément en prenant comme condition à l'entrée d'un élément, la valeur fournie à la sortie de l'élément précédent, fournit le système d'équation suivant pour l'élément Ω_{e+1} :

$$\begin{aligned}
\sum_{j=1}^2 A_{1j}^{e+1} U_j^{e+1} + (U_1^{e+1} - U_2^e) &= B_1^{e+1} \\
\sum_{j=1}^2 A_{2j}^{e+1} U_j^{e+1} &= B_2^{e+1}
\end{aligned} \tag{5.20}$$

où on peut observer que le couplage entre les divers éléments n'est réalisé que par le biais d'une condition de raccord faiblement imposée.

Cette manière de procéder est appelée classiquement *formulation de Galerkin discontinue*. Cette appellation est en partie trompeuse. En effet, il est exact qu'au sein d'un élément, on utilise les fonctions de forme comme poids pour les résidus. Par contre, les résidus liés au raccord entre deux éléments sont attribués systématiquement à l'élément de droite. En d'autres mots, on observe qu'on effectue une sorte d'intégration décentrée de la condition de raccord : ce qui ne correspond pas à l'idée d'une méthode de Galerkin qui consiste à réaliser une pondération centrée des résidus !

Afin d'établir un parallèle entre une formulation basée sur une approximation discontinue et le monde des différences finies, considérons une approximation constante par morceau où U^e est la valeur nodale associée à un élément Ω_e . Le système (5.20) se réduit à une simple équation

$$\frac{U^{e+1} - U^e}{h} = F^e$$

qui correspond à un schéma d'Euler pour l'intégration d'une équation différentielle ordinaire ou à des différences finies amont.

Finalement, il nous faut mentionner qu'on peut utiliser des techniques de Petrov-Galerkin pour la pondération du résidu lié à l'équation différentielle au sein de chaque élément et que cela améliore encore la précision et la stabilité du schéma. Une telle approche est appelée une *formulation discontinue de Petrov-Galerkin*.

5.2 Equation scalaire d'advection-diffusion

La *formulation forte* d'un problème scalaire d'advection-diffusion peut s'écrire comme suit :

Trouver $u(\mathbf{x}) \in \mathcal{U}_s$ tel que

$$\begin{aligned} \boldsymbol{\beta} \cdot \nabla u - \nabla \cdot (\epsilon \nabla u) &= f, & \forall \mathbf{x} \in \Omega, \\ \mathbf{n} \cdot (\epsilon \nabla u) &= g, & \forall \mathbf{x} \in \Gamma_N, \\ u &= t, & \forall \mathbf{x} \in \Gamma_D, \end{aligned}$$

(5.21)

où $\mathbf{n} = (n_x, n_y)$ représente la normale sortante de la courbe Γ et ϵ est un paramètre constant. Sur Γ_D , la valeur de u est imposée à t , tandis que sur Γ_N , la valeur du flux $\mathbf{n} \cdot (\epsilon \nabla u)$ est imposée à la valeur de g . Formellement, un tel problème est elliptique et il faut donc prescrire des conditions aux limites sur l'ensemble de la frontière. On distingue toujours les conditions de Dirichlet sur une partie de la frontière Γ_D et les conditions de Neumann le long de la partie Γ_N de la frontière ($\partial\Omega = \Gamma_D \cup \Gamma_N$ et $\Gamma_D \cap \Gamma_N = \emptyset$). Le vecteur $\boldsymbol{\beta} = (\beta_x, \beta_y, \beta_z)$ est une vitesse donnée qui peut éventuellement être une fonction de la position.

Les dimensions typiques d'un tel problème sont une vitesse caractéristique β , le coefficient de diffusion ϵ , ainsi qu'une longueur caractéristique L . On peut alors définir un nombre de Péclet pour un tel problème

$$Pe = \frac{\beta L}{\epsilon}. \quad (5.22)$$

Ce nombre mesure l'importance relative du terme de diffusion par rapport au terme de transport. Il est donc infini, si on considère le cas $\epsilon = 0$ qui correspond à l'équation scalaire de transport et est nul pour le cas $\boldsymbol{\beta} = 0$ qui correspond à l'équation de Poisson.

Tout d'abord, rappelons quelques propriétés de base sur la régularité de la solution exacte du problème (5.21). Comme on vient de le remarquer, la solution du cas limite ($\epsilon = 0$) peut être discontinue, en particulier, le long des caractéristiques si la donnée initiale le long de la frontière Γ_- est discontinue. Dans le problème complet (5.21), la solution sera toujours continue à l'intérieur du domaine Ω , même si les conditions aux limites sont discontinues et même si la valeur non nulle de ϵ est petite. On observera que le saut présent dans le cas limite le long d'une caractéristique sera repris de manière continue dans une région dont l'épaisseur adimensionnelle est donnée par

$$\mathcal{O}\left(\sqrt{\frac{\epsilon}{\beta L}}\right).$$

Si la valeur de ϵ est très petite, cette région sera très étroite et sera le théâtre de changements très rapides de la valeur de u . On y observe donc de très importants gradients, susceptibles de générer des problèmes numériques et nécessitant beaucoup de points de discrétisation pour pouvoir approcher la solution exacte avec précision. Cette région est

appelé *couche limite*. Cette analyse de régularité est typiquement effectuée lors de l'étude des couches limites laminares en mécanique des fluides où on obtient un résultat très semblable avec une taille caractéristique de la couche limite proportionnelle à $\mathcal{O}(Re^{-1/2})$.

Il faut aussi mentionner qu'à la sortie du domaine Γ_+ , une condition aux limites arbitraire peut maintenant être imposée et ne pas du tout coïncider avec la solution que l'on obtiendrait pour $\epsilon = 0$. On observe la présence d'une fine région d'épaisseur adimensionnelle

$$\mathcal{O}\left(\frac{\epsilon}{\beta L}\right)$$

où le raccord continu doit s'effectuer. On parle ici de *couche limite de sortie*.

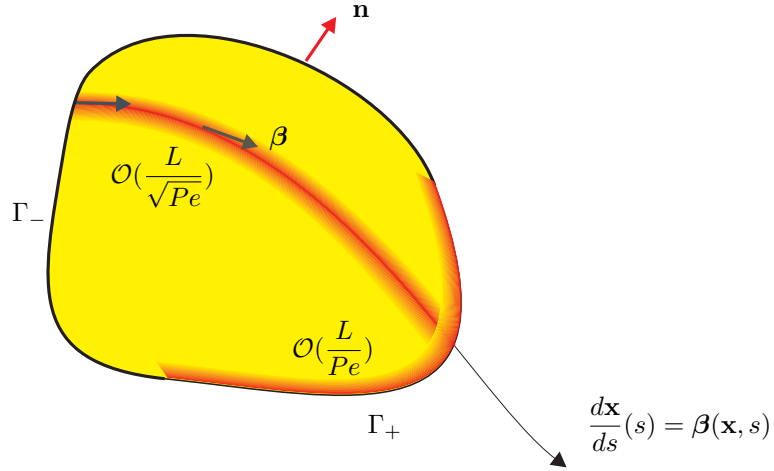


Figure 5.2: Couches limites dans un problème d'advection-diffusion.

5.2.1 Méthodes de Petrov-Galerkin

Nous allons utiliser une approximation continue de la solution : puisque cette dernière est continue, c'est un choix qui peut paraître naturel. Et pour obtenir une formulation discrète stable, on recourt à une formulation discrète de Petrov-Galerkin de l'équation d'advection-diffusion

Trouver $U_j \in \mathbb{R}^n$ tel que

$$\sum_{j=1}^n \underbrace{\langle w_i \beta \cdot \nabla \tau_j + \epsilon \nabla w_i \cdot \nabla \tau_j \rangle}_{A_{ij}} U_j = \underbrace{\langle w_i f \rangle + \ll w_i g \gg_N}_{B_i}, \quad i = 1, \dots, n,$$

(5.23)

avec $w_i = \tau_i$ pour la méthode classique de Galerkin. Mais il est plus efficace d'utiliser comme fonction de poids

$$w_i = \tau_i + \zeta \boldsymbol{\beta} \cdot \nabla \tau_i$$

où le paramètre ζ est une constante judicieusement choisie. Le choix de cette constante sera illustré dans un cas unidimensionnel modèle. Il faut mentionner ici que la valeur extraite de l'analyse unidimensionnelle est souvent utilisée dans les dimensions supérieures où une telle analyse n'est plus possible. L'utilisation d'une telle formulation permet l'obtention de résultats nettement meilleurs qu'avec la méthode classique de Galerkin pour un maillage identique. Par contre, il faut mentionner que pour un maillage suffisamment fin, la méthode de Galerkin usuelle fonctionnera correctement.

Pour illustrer le choix d'une bonne valeur pour le paramètre ζ , nous allons maintenant utiliser l'équivalence qui existe entre les éléments finis et les différences finies dans le cas unidimensionnel et considérer successivement le cas des différences finies centrées, le cas des différences finies décentrées et celui qui correspond à une formulation hybride strictement équivalente à notre formulation de Petrov-Galerkin pour une approximation linéaire par morceaux appliquée à un problème modèle unidimensionnel.

5.2.2 Cas unidimensionnel

Supposons donc que nous ayons à résoudre un problème unidimensionnel modèle :

$$\begin{aligned} \beta \frac{du}{dx} - \epsilon \frac{d^2u}{dx^2} &= 0, & \forall x \in \Omega =]0, L[, \\ u(0) &= u_0, \\ u(L) &= u_L, \end{aligned} \tag{5.24}$$

avec le nombre de Péclet habituel.

$$Pe = \frac{\beta L}{\epsilon}.$$

La solution analytique de ce problème (5.24) est :

$$\frac{u - u_0}{u_L - u_0} = \frac{\exp(Pe x/L) - 1}{\exp(Pe) - 1} \tag{5.25}$$

A bas nombre de Péclet, la solution tend vers une droite puisque la diffusion est l'effet dominant. Par contre, à haut nombre de Péclet, une couche limite apparaît près de $x = L$, puisque maintenant ce sont les effets de transport ou d'advection qui dominent.

Différences finies centrées ou méthode de Galerkin

Pour obtenir la solution numérique approchée pour notre problème unidimensionnel, utilisons d'abord des différences finies centrées ou appliquons la technique de Galerkin pour des éléments finis uniformes et de taille h .

$$\beta \frac{U_{i+1} - U_{i-1}}{2h} = \epsilon \frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} \quad i = 1, \dots, n-1 \quad (5.26)$$

avec $U_0 = u_0$ et $U_n = u_L$.

Il est alors relativement facile d'obtenir une expression analytique de la solution discrète obtenue par les équations (5.26). Il suffit d'exprimer les valeurs nodales U_i comme une expression polynomiale du type $Ar^i + B$ et d'effectuer un peu d'algèbre à partir de (5.26) :

$$\beta \frac{U_{i+1} - U_{i-1}}{2h} = \epsilon \frac{U_{i+1} - 2U_i + U_{i-1}}{h^2}$$

En remplaçant U_i par $Ar^i + B$, ↓

$$Ar^{i-1} \frac{\beta h}{2\epsilon} (r^2 - 1) = Ar^{i-1} (r^2 - 2r + 1)$$

En remplaçant $\frac{\beta h}{\epsilon}$ par Pe^h , ↓

$$0 = \left(1 - \frac{Pe^h}{2}\right) r^2 - 2r + \left(1 + \frac{Pe^h}{2}\right)$$

où le nombre Pe^h est appelé le nombre de Péclet de maille. Il est alors facile de voir que r est solution d'une équation du second degré et vaut (en excluant la valeur triviale $r = 1$) :

$$r = \frac{1 + \sqrt{1 - (1 + \frac{Pe^h}{2})(1 - \frac{Pe^h}{2})}}{\left(1 - \frac{Pe^h}{2}\right)} = \frac{\left(1 + \frac{Pe^h}{2}\right)}{\left(1 - \frac{Pe^h}{2}\right)}$$

L'utilisation des conditions aux limites (sur U_0 et U_n) permet de déterminer les valeurs de A et de B et d'obtenir l'expression finale :

$$\frac{U_i - u_0}{u_L - u_0} = \frac{\left(\frac{1 + Pe^h/2}{1 - Pe^h/2}\right)^i - 1}{\left(\frac{1 + Pe^h/2}{1 - Pe^h/2}\right)^n - 1} \quad (5.27)$$

Comme on observe que :

$$\lim_{h \rightarrow 0} \left(\frac{1 + Pe^h/2}{1 - Pe^h/2}\right) = \exp(Pe^h),$$

on en déduit que l'expression analytique de la solution discrète converge vers la solution exacte, lorsqu'on raffine le maillage. Pour éviter un comportement oscillant de la solution discrète, il faut que r soit toujours du même signe. Cela fixe un critère de taille maximale pour le pas du maillage en fonction du problème considéré :

$$\begin{array}{c} \frac{Pe^h}{2} < 1 \\ \downarrow \\ h < \frac{2\epsilon}{\beta} \end{array}$$

Pour un problème d'advection-diffusion, la méthode de Galerkin ne produira pas d'oscillations si le maillage est suffisamment raffiné. Toutefois, pour un problème à très haut nombre de Péclet, le raffinement requis peut devenir totalement impraticable. Le même problème se posera pour les simulations d'écoulements à très haut nombre de Reynolds.

Différences décentrées pour le terme de transport

Afin d'obtenir un schéma plus efficace, on pourrait être tenté par l'utilisation d'une différence amont pour le terme d'advection. Dans le monde des différences finies, c'est une stratégie courante pour obtenir une solution discrète sans oscillations.

Cela revient à remplacer (5.26) par l'expression suivante

$$\beta \frac{U_i - U_{i-1}}{h} = \epsilon \frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} \quad i = 1, \dots, n-1 \quad (5.28)$$

On peut à nouveau effectuer un peu d'algèbre pour trouver une forme analytique pour la solution discrète correspondant au système (5.28).

$$\beta \frac{U_i - U_{i-1}}{h} = \epsilon \frac{U_{i+1} - 2U_i + U_{i-1}}{h^2}$$

En remplaçant U_i par $Ar^i + B$, ↓

$$Ar^{i-1} \frac{\beta h}{\epsilon} (r - 1) = Ar^{i-1} (r^2 - 2r + 1)$$

En remplaçant $\frac{\beta h}{\epsilon}$ par Pe^h , ↓

$$0 = r^2 - 2 \left(1 + \frac{Pe^h}{2}\right) r + (1 + Pe^h)$$

En résolvant l'équation du second degré, ↓

$$r = \underbrace{\left(1 + \frac{Pe^h}{2}\right) + \sqrt{\left(1 + \frac{Pe^h}{2}\right)^2 - (1 + Pe^h)}}_{1 + Pe^h}$$

En utilisant les conditions sur U_0 et U_n , ↓

$$\frac{U_i - u_0}{u_L - u_0} = \frac{(1 + Pe^h)^i - 1}{(1 + Pe^h)^n - 1}$$

Comme r est maintenant toujours positif, on n'observe jamais d'oscillations parasites, même si le nombre de Péclet de maille est supérieur à un. Il n'y a donc aucune condition sur le maillage pour éviter l'apparition de telles oscillations numériques : on dit qu'un tel schéma est numériquement stable.

Par contre, l'inconvénient d'une telle approche est que la solution numérique est trop diffusive. En d'autres mots, les gradients prononcés de la solution exacte seront lissés ou gommés dans la solution numérique. Pour illustrer cette diffusion numérique, il suffit d'observer que l'équation (5.28) peut être interprétée comme suit

$$\begin{aligned}
\beta \frac{U_i - U_{i-1}}{h} &= \epsilon \frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} \\
&\downarrow \\
\beta \frac{U_{i+1} - U_{i-1}}{2h} - \left(\frac{\beta h}{2} \right) \frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} &= \epsilon \frac{U_{i+1} - 2U_i + U_{i-1}}{h^2}
\end{aligned}$$

En comparant les équations (5.28) et (5.26), on peut tirer une importante conclusion : l'utilisation d'une différence finie amont pour le terme de transport est équivalent à ajouter un terme de diffusion numérique $\epsilon^h = \frac{\beta h}{2}$ à un problème d'advection-diffusion discrétisé de manière centrée. En conséquence, la solution numérique associée à (5.28) correspond à un nombre de Péclet plus modeste $Pe = \frac{\beta L}{\epsilon + \epsilon^h}$ que celui de la solution exacte $Pe = \frac{\beta L}{\epsilon}$. Les différences décentrées pour le terme de transport conduisent au lissage des gradients prononcés, en particulier sur des grilles peu raffinées, puisque le rapport de la diffusion numérique et de la diffusion physique est donné par le nombre de Péclet de maille.

En termes d'éléments finis, cela revient à utiliser des fonctions de poids distinctes pour les termes de diffusion et le terme de transport. Une telle stratégie permet d'obtenir des formulations dites inconsistantes qui sont numériquement stables, mais qui sont relativement imprécises. Comme le terme additionnel est proportionnel à h , ces formulations inconsistantes ont une précision qui est au mieux linéaire.

Schéma hybride ou méthode de Petrov-Galerkin

Comme l'usage d'une différence amont permet d'obtenir la stabilité numérique au détriment de la précision, un bon compromis semble le recours à un schéma hybride :

$$(1 - \zeta) \beta \frac{U_{i+1} - U_{i-1}}{2h} + \zeta \beta \frac{U_i - U_{i-1}}{h} = \epsilon \frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} \quad i = 1, \dots, n-1 \quad (5.29)$$

où le paramètre ζ est choisi judicieusement.

On peut à nouveau effectuer un peu d'algèbre pour trouver une forme analytique de la solution discrète correspondant au système (5.29).

$$(1 - \zeta)\beta \frac{U_{i+1} - U_{i-1}}{2h} + \zeta\beta \frac{U_i - U_{i-1}}{h} - \epsilon \frac{U_{i+1} + 2U_i - U_{i+1}}{h^2} = 0$$

En remplaçant U_i par $Ar^i + B$,
et en remplaçant $\frac{\beta h}{\epsilon}$ par Pe^h ,

$$\left(1 - (1 - \zeta)\frac{Pe^h}{2}\right)r^2 - \left(1 + (1 - \zeta)\frac{Pe^h}{2} + \zeta\frac{Pe^h}{2}\right)r + (1 + \zeta Pe^h) = 0$$

En résolvant l'équation du second degré,

$$\frac{1 + (1 + \zeta)Pe^h/2}{1 - (1 - \zeta)Pe^h/2} = r$$

Comment sélectionner une valeur adéquate de ζ ? On va imposer que la valeur analytique discrète au noeud soit égale à la solution exacte. Cela revient à écrire que :

$$U_i = u(ih)$$

↓

$$\left(\frac{1 + (1 + \zeta)\frac{Pe^h}{2}}{1 - (1 - \zeta)\frac{Pe^h}{2}}\right)^i = (\exp(Pe^h))^i$$

$$(1 + (1 + \zeta)\frac{Pe^h}{2}) = \exp(Pe^h)(1 - (1 - \zeta)\frac{Pe^h}{2})$$

$$\zeta = \frac{\exp(Pe^h)(1 - \frac{Pe^h}{2}) - (1 + \frac{Pe^h}{2})}{\frac{Pe^h}{2}(1 - \exp(Pe^h))}$$

$$\zeta = -\frac{1 + \exp(Pe^h)}{1 - \exp(Pe^h)} - \frac{2}{Pe^h}$$

On obtient ainsi la valeur optimale du paramètre ζ par la relation

$$\boxed{\zeta = \coth\left(\frac{Pe^h}{2}\right) - \frac{2}{Pe^h}} \quad (5.30)$$

Pour cette valeur du paramètre ζ , le schéma hybride fournit la solution exacte aux valeurs nodales. Finalement, il est possible de montrer que ce schéma hybride fournit exactement les mêmes équations discrètes qu'un schéma de Petrov-Galerkin pour des fonctions de forme linéaires par morceaux et des poids :

$$w_i = \tau_i + \zeta \frac{h}{2} \tau_{i,x}.$$