Université Aix Marseille 1

Master de mathématiques

Analyse numérique des équations aux dérivées partielles

Raphaèle Herbin

3 décembre 2008

Table des matières

Introduction

L'analyse numérique des équations aux dérivées partielles

Pour aborder le calcul numérique (à l'aide d'un outil informatique) des solutions d'un problème "réel", on passe par les étapes suivantes :

1. Description qualitative des phénomènes physiques.
   Cette étape est effectuée par des spécialistes des phénomènes que l'on veut quantifier (ingénieurs, chimistes, biologistes etc.....)
2. Modélisation
   Il s'agit, à partir de la description qualitative précédente, d'écrire un modèle mathématique. On supposera ici que ce modèle amène à un système d'EDP (équations aux dérivées partielles). Dans la plupart des cas, on ne saura pas calculer une solution analytique, explicite, du modèle ; on devra faire appel à des techniques de résolution approchée.
3. Analyse mathématique
   Même si l'on ne sait pas trouver une solution explicite du modèle, il est important d'en étudier les propriétés mathématiques, dans la mesure du possible. Il est bon de se poser les questions suivantes :
   - Le problème est-il bien posé ? c'est-à-dire y–a–t'il existence et unicité de la solution ?
   - Les propriétés physiques auxquelles on s'attend sont elles satisfaites par les solutions du modèle mathématique ? Si u est une concentration, par exemple, peut-on prouver qu'elle est toujours positive ?
   - Y a-t-il continuité de la solution par rapport aux données ?
4. Discrétisation et résolution numérique
   Un problème posé sur un domaine continu (espace - temps) n'est pas résoluble tel quel par un ordinateur, qui ne peut traiter qu'un nombre fini d'inconnues. Pour se ramener à un problème en dimension finie, on discrétise l'espace et/ou le temps. Si le problème original est linéaire on obtient un système linéaire. Si le problème original est non linéaire (par exemple s'il s'agit de la minimisation d'une fonction) on aura un système non linéaire à résoudre par une méthode *ad hoc* (méthode de Newton...)
5. Analyse numérique
   Une fois le problème discret obtenu, il est raisonnable de se demander si la solution de ce problème est proche, et en quel sens, du problème continu. De même, si on doit mettre en oeuvre une méthode itérative pour le traitement des non-linéarités, il faut étudier la convergence de la méthode itérative proposée.
6. Mise en oeuvre, programmation et analyse des résultats
   La partie mise en oeuvre est une grosse consommatrice de temps. Actuellement, de nombreux codes commerciaux existent, qui permettent en théorie de résoudre "tous" les problèmes. Il faut

cependant procéder à une analyse critique des résultats obtenus par ces codes, qui ne sont pas toujours compatibles avec les propriétés physiques attendues...

Principales méthodes de discrétisation

Méthodes de différences finies et volumes finis

On considère un domaine physique $\Omega \subset \mathbb{R}^d$, où d est la dimension de l'espace. Le principe des méthodes de différences finies consiste à se donner un certain nombre de points du domaine, qu'on notera $(x_1 \dots x_N) \subset (\mathbb{R}^d)^N$. On approche alors l'opérateur différentiel en espace en chacun des x_i par des quotients différentiels. Il faut alors discrétiser la dérivéen en temps : on pourra par exemple considérer un schéma d'Euler explicite ou implicite pour la discrétisation en temps.
Les méthodes de volumes finis sont adaptées aux équations de conservation et utilisées en mécanique des fluides depuis plusieurs décennies. Le principe consiste à découper le domaine Ω en des "volumes de contrôle" ; on intègre ensuite l'équation de conservation sur les volumes de contrôle ; on approche alors les flux sur les bords du volume de contrôle par une technique de différences finies.

Méthodes variationnelles, méthodes d'éléments finis

On met le problème d'équations aux dérivées partielles sous forme variationnelle :

$$\begin{cases} a(u,v) = (f,v)_H, \quad \forall v \in H, \\ u \in H, \end{cases}$$

où H est un espace de Hilbert bien choisi (par exemple parce qu'il y a existence et unicité de la solution dans cet espace), $(\cdot,\cdot)_H$ le produit scalaire sur H et a une forme bilinéaire sur H.. La discrétisation consiste à remplacer H par un sous espace de dimension finie H_k, construit par exemple à l'aide de fonctions de base éléments finis qu'on introduira plus loin :

$$\begin{cases} a(u_k,v_k) = (f,v_k)_H, \quad \forall v \in H_k, \\ u_k \in H_k. \end{cases}$$

Méthodes spectrales

L'idée de ces méthodes est de chercher un solution approchée sous forme d'un développement sur une certaine famille de fonctions. On peut par exemple écrire la solution approchée sous la forme : $u = \sum_{i=1}^{n} \alpha(u) p_i$ p_i fonction polynomiales, on choisit la base p_i de manière à ce que $\alpha_{i'}$ et p_i' soient faciles à calculer. Ces dernières méthodes sont réputées coûteuses, mais précises. Elles sont le plus souvent utilisées comme aide à la compréhension des phénomènes physiques.

Types d'équations aux dérivées partielles

Il existe une classification des équations aux dérivées partielles linéaires du second ordre. Considérons par exemple une équation aux dérivées partielles écrite sous la forme :

$$Au_{xx} + Bu_{yy} + Cu_{xy} + Du_x + Eu_y + F = 0 \tag{0.0.1}$$

L'appellation "elliptique", "parabolique" ou "hyperbolique" d'une équation aux dérivées partielles (0.0.1) correspond à la nature de la conique décrite par l'équation caractéristique correspondante, c'est-à-dire :

$$Ax^2 + By^2 + Cxy + Dx + Ey + F = 0.$$

Donnons maintenant des exemples d'équations elliptiques, paraboliques et hyperboliques.

Problèmes elliptiques

L'équation elliptique modèle est

$$-\Delta u = f, \tag{0.0.2}$$

où $\Delta u = \partial_1^2 u + \partial_2^2$, ∂_i désignant la dérivée partielle par rapport à la i-ème variable (et donc ∂_i^2 la dérivée partielle d'ordre 2 par rapport à la i-ème variable). Cette équation modélise par exemple le phénomène de conduction de la chaleur stationnaire (c.à.d. en régime permanent). En élasticité, on rencontre également l'équation du bi-laplacien, c.à.d. :

$$-\Delta^2 u = f \tag{0.0.3}$$

L'équation (0.0.2) peut être discrétisée par différences finies, volumes finis où éléments finis. On verra par la suite que les méthodes des différences finies sont limitées à des domaines géométriques 'simples". L'équation (0.0.3) est le plus souvent discrétisée par éléments finis, pour des raisons de précision.

Problèmes paraboliques

L'équation parabolique modèle est

$$u_t - \Delta u = f, \tag{0.0.4}$$

où u_t désigne la dérivée partielle de u par rapport au temps (u est donc une fonction de x, variable d'espace, et de t, variable de temps). Cette équation modélise par exemple la conduction de la chaleur en régime instationnaire. Cette équation parabolique comporte deux opérateurs : la dérivée d'ordre 1 en temps est, de manière usuelle, discrétisée par différences finies, tandis que le traitement de l'opérateur différentiel d'ordre 2 en espace est effectué comme pour l'équation (0.0.2).

Problèmes hyperboliques

Les équations de type hyperbolique interviennent principalement en mécanique des fluides (aéronautique, écoulements diphasiques, modélisation de rupture de barrage et d'avalanches). Elles sont souvent obtenues en négligeant les phénomènes de diffusion (parce qu'ils sont faibles) dans les équations de conservation de la mécanique. L'exemple le plus classique d'équation hyperbolique linéaire est l'équation de transport (ou d'advection).

$$u_t - u_x = 0 t \in \mathbb{R}_+, x \in \mathbb{R}, \tag{0.0.5}$$

avec condition initiale :

$$u(x,0) = u_0(x). \tag{0.0.6}$$

Dans le cas où la condition initiale u_0 est suffisamment régulière, il est facile de voir que la fonction :

$$u(x,t) = u_0(x + t), \tag{0.0.7}$$

est solution de (0.0.5)-(0.0.6). Si u_0 est non régulière (par exemple discontinue, nous verrons qu'il y a encore moyen de montrer que la fonction définie par (0.0.7) est solution en un sens que nous qualifierons de "faible".

Si l'équation est non linéaire, i.e.

$$u_t + (f(u))_x = 0, t \in \mathbb{R} + , x \in \mathbb{R}, \tag{0.0.8}$$

avec par exemple $f(u) = u^2$, et condition initiale (0.0.6), on peut encore définir des solutions faibles, mais leur calcul est plus difficile. Les équations hyperboliques sont discrétisées de manière usuelle par la méthode des volumes finis. Les discrétisations par éléments finis mènent à des schémas instables (c'est–à–dire que les solutions discrètes ne vérifient pas les propriétés physiques souhaitées).

Chapitre 1

Méthodes de différences finies et volumes finis pour les problèmes elliptiques

1.1 Principe des deux méthodes

1.1.1 Cas de la dimension 1

On considère le problème unidimensionnel

$$-u''(x) = f(x), \quad \forall x \in]0,1[, \tag{1.1.1}$$

$$u(0) = u(1) = 0, \tag{1.1.2}$$

où $f \in C([0,1])$. Les conditions aux limites (1.1.2) considérées ici sont dites de type Dirichlet homogène (le terme homogène désigne les conditions nulles). Cette équation modélise par exemple la diffusion de la chaleur dans un barreau conducteur chauffé (terme source f) dont les deux extrémités sont plongées dans de la glace.

Méthode de différences finies.

Soit $(x_k)_{k=0,\ldots,N+1}$ une subdivision de [0,1], avec :

$$x_0 = 0 < x_1 < x_2 < \ldots < x_N < x_{N+1} = 1.$$

Pour $i = 0, \ldots, N$, on note $h_{i+1/2} = x_{i+1} - x_i$ et on définit le "pas" du maillage par :

$$h = \max_{i=0,\ldots,N} h_{i+1/2}. \tag{1.1.3}$$

Pour simplifier l'exposé, on se limitera dans un premier temps à un pas constant :

$$h_{i+1/2} = h \qquad \forall i \in [0,N].$$

On écrit l'équation aux dérivées partielles (1.1.1) aux points x_i

$$-u''(x_i) = f(x_i), \qquad \forall i = 1, \ldots, N,$$

Effectuons un développement de Taylor en x_i :

$$u(x_{i+1}) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\zeta_i),$$

$$u(x_{i-1}) = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\eta_i),$$

avec $\zeta_i \in [x_i, x_{i+1}]$, $\eta_i \in [x_{i-1}, x_i]$. En additionnant, on obtient :

$$u(x_{i+1}) + u(x_{i-1}) = 2u(x_i) + h^2u''(x_i) + O(h^2)$$

Il semble donc raisonnable d'approcher la dérivée seconde $-u''(x_i)$ par le "quotient différentiel"

$$\frac{2u(x_i) - u(x_{i-1}) - u(x_{i+1})}{h^2}.$$

Sous des hypothèses de régularité sur u, on peut montrer (voir lemme 1.12 page 16) que cette approximation est d'ordre 2 au sens

$$R_i = u''(x_i) + \frac{2u(x_i) - u(x_{i-1}) - u(x_{i+1})}{h^2} = O(h^2)$$

On appelle erreur de consistance au point x_i la quantité R_i.

Méthode des volumes finis.

On ne se donne plus des points mais des volumes de contrôle K_i, $\quad i = 1, \ldots, N$, avec $K_i =]x_{i-1/2}, x_{i+1/2}[$, et on note $h_i = x_{i+1/2} - x_{i-1/2}$. Pour chaque volume de contrôle K_i, on se donne un point $x_i \in K_i =]x_{i-1/2}, x_{i+1/2}[$. On pourra considérer par exemple (mais ce n'est pas le seul point possible) : $x_i = 1/2\left(x_{i+1/2} + x_{i-1/2}\right)$. On intègre l'équation $-u'' = f$ sur K_i :

$$\int_{x_{i-1/2}}^{x_{i+1/2}} -u''(x)dx = \int_{x_{i-1/2}}^{x_{i+1/2}} f(x)dx$$

et $f_i = \frac{1}{h_i}\int_{x_{i-1/2}}^{x_{i+1/2}} f(x)dx$. On obtient :

$$-u'(x_{i+1/2}) + u'(x_{i-1/2}) = h_i f_i, \quad i = 1, \ldots, N.$$

On cherche donc a approcher les flux $-u'(x_{i+1/2})$ aux interfaces $x_{i+1/2}$ des mailles. Notons que l'opérateur à approcher est ici d'ordre 1, alors qu'il était d'ordre 2 en différences finies pour la même équation.
On se donne une inconnue par maille (ou volume de contrôle i), qu'on note u_i, et on espère approcher ainsi la valeur $u(x_i)$ (ou $\frac{1}{h_i}\int_{K_i} u$). On approche $u'(x_{i+1/2})$ par le quotient différentiel

$$\frac{u(x_{i+1}) - u(x_i)}{h_{i+1/2}}.$$

Le schéma numérique s'écrit donc :

$$-\frac{u_{i+1}-u_i}{h_{i+1/2}}+\frac{u_i-u_{i-1}}{h_{i-1/2}}=h_i f_i \qquad i=2,\ldots,N-1. \tag{1.1.4}$$

Pour la première et N-ième equations, on tient compte des conditions aux limites (1.1.2), et on $u'(0)$ (resp. $u'(1)$) par $\frac{u(x_1)}{h_{1/2}}$ (resp. $\frac{u(x_N)}{h_{N+1/2}}$, ce qui donne comme première et dernière équations du schéma numérique :

$$-\frac{u_2-u_1}{h_{3/2}}+\frac{u_1}{h_{1/2}}=h_1 f_1, \tag{1.1.5}$$

$$-\frac{u_N-u_{N-1}}{h_{N-1/2}}-\frac{u_N}{h_{N+1/2}}=h_N f_N, \tag{1.1.6}$$

Remarque 1.1 *Si le pas du maillage est constant : $h_i = h$, $\forall i = 1,\ldots,N$ (on dit aussi que le maillage est uniforme), on peut montrer (exercice 1 page 32) que les équations des schémas volumes finis et différences finies aux conditions de bord et au second membre près. Si le maillage n'est pas régulier, ceci n'est plus verifié.*

1.1.2 Cas de la dimension 2 ou 3

On considère maintenant le problème (0.0.2) en dimension 2 ou 3, sur un ouvert borné Ω de $\mathbb{R}^d$, $d = 2$ ou 3, avec conditions aux limites de Dirichlet homogènes qui s'écrivent maintenant :

$$u(x)=0, \forall\ x\in\partial\Omega, \tag{1.1.7}$$

où $\partial\Omega$ désigne la frontière de Ω.

Méthode de différences finies.

Supposons (pour simplifier) que le domaine Ω soit un carré (c.à.d. $d = 2$, le cas rectangulaire se traite tout aussi facilement). On se donne un pas de maillage constant h et des points $x_{i,j} = (ih,jh), i = 1,\ldots,N$, $i = 1,\ldots,N$. En effectuant les développements limités de Taylor (comme au paragraphe 1.1.1 page 8) dans les deux directions (voir exercice 14), on approche $-\partial_i^2 u(x_{i,j})$ (resp. $-\partial_j^2 u(x_{i,j})$) par

$$\frac{2u(x_{i,j})-u(x_{i+1,j})-u(x_{i-1,j})}{h^2} \text{ (resp. par } \frac{2u(x_{i,j})-u(x_{i,j+1})-u(x_{i,j-1})}{h^2}).$$

Ce type d'approche est limité à des géométries simples. Pour mailler des géométries compliqués, il est en général plus facile d'utiliser des triangles (tétraèdres en dimension 3), auquel cas la méthode des différences finies est plus difficile à généraliser.

Méthode de volumes finis.

On suppose maintenant que Ω est un ouvert polygonal de $\mathbb{R}^2$, et on se donne un maillage $\mathcal{T}$ de Ω, c.à.d., en gros, un découpage de Ω en volumes de contrôle polygônaux K. En intégrant l'équation (0.0.2) sur K, on obtient :

$$\int_K -\Delta u dx = \int_K f dx.$$

Par la formule de Stokes, on peut réécrire cette équation :

$$-\int_{\partial K} \nabla u(x).\mathbf{n}_K(x) d\gamma(x) = \int_K f(x)dx,$$

où $d\gamma(x)$ désigne l'intégrale par rapport à la mesure uni-dimensionnelle sur le bord de l'ouvert Ω, et où $\mathbf{n}_K$ désigne le vecteur normal unitaire à ∂K extérieur à K. Comme K est polygonal, on peut décomposer ∂K en arêtes σ qui sont des segments de droite, et en appelant $\mathcal{E}_K$ l'ensemble des arêtes de ∂K, on a donc :

$$-\sum_{\sigma\in\mathcal{E}_K} \int_\sigma \nabla u.\mathbf{n}_{K,\sigma} d\gamma(x) = \int_K f(x)dx,$$

où $\mathbf{n}_{K,\sigma}$ désigne le vecteur normal unitaire à σ extérieur à K (noter que ce vecteur est constant sur σ). On cherche donc maintenant à approcher la dérivée normale $\nabla u.\mathbf{n}_{K,\sigma}$ de manière consistante sur chaque arête σ. On se donne donc des inconnues discrètes notées $(u_K)_{K\in\mathcal{T}}$, qui, on l'espère vont s'avérer être des approximations de $u(x_K)$. Pour une arête $\sigma = K|L$ séparant les volumes de contrôle K et L, il est tentant d'approcher la dérivée normale $\nabla u.\mathbf{n}_{K,\sigma}$ par le quotient différentiel

$$\frac{u(x_L) - u(x_K)}{d_{K,L}},$$

où $d_{K,L}$ est la distance entre les points x_K et x_L. Cependant, cette approximation ne pourra être justifiée que si la direction du vecteur défini par les deux points x_K et x_L est la même que celle de la normale $\mathbf{n}_{K,\sigma}$, c.à.d. si le segment de droite $x_K x_L$ est orthogonal à l'arête $K|L$. Pour un maillage triangulaire à angles strictement inférieurs à $\pi/2$, ceci est facile à obtenir en choisissant les points x_K comme intersection des médiatrices du triangle K, voir Figure 1.1.

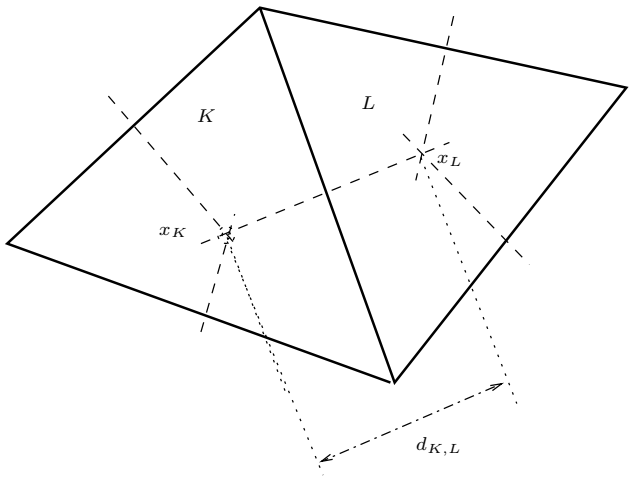

FIG. 1.1 – *Exemple de volumes de contrôle pour la méthode des volumes finis en deux dimensions d'espace*

On se placera ici dans ce cas, et on verra plus loin d'autres possibilités. on approche donc $\nabla u.n_K|_\sigma$ par $\dfrac{u(x_L) - u(x_K)}{d_{K,L}}$ et en notant $|\sigma|$ la longueur de l'arête σ, on approche :

$$\int_\sigma \nabla u.n_K d\gamma \text{ par } F_{K,\sigma} = |\sigma|\frac{u_L - u_K}{d_{K,L}}, \text{ pour tout } \sigma \in \mathcal{E}_K \text{ et pour tout } K \in \mathcal{T}.$$

Le schéma volumes finis s'écrit donc

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} = |K| f_K, \tag{1.1.8}$$

où $|K|$ est la mesure de K, et $f_K = \frac{1}{|K|}\int_K f(x)dx$, et où les flux numériques $F_{K,\sigma}$ sont définis (en tenant compte des conditions limites pour les arêtes du bord) par :

$$F_{K,\sigma} = \begin{cases} -|\sigma|\dfrac{u_L - u_K}{d_{K,L}} \text{ si } \sigma = K|L, \\ -|\sigma|\dfrac{u_K}{d_{K,\sigma}} \text{ si } \sigma \subset \partial\Omega \text{ et } \sigma \in \mathcal{E}_K, \end{cases} \tag{1.1.9}$$

où $d_{K,\sigma}$ = distance entre x_K et σ

Comparaison des méthodes

Cette introduction aux différences finies et volumes finis nous permet de remarquer que les différences finies sont particulièrement bien adaptées dans le cas de domaines rectangulaires ou parallèlepipédiques, pour lesquels on peut facilement définir des maillages structurés (cartésiens dans le cas présent) c.à.d. dont on peut indexer les mailles par un ordre (i,j) naturel.
Dans le cas de domaines plus complexes, on maille souvent à l'aide de triangles (ou tétraèdres) et dans ce cas la méthode des différences finies ne se généralise pas facilement. On a alors recours soit aux volumes finis, dont on vient de donner le principe, soit aux éléments finis, que nous aborderons ultérieurement.

1.1.3 Questions d'analyse numérique

Voici un certain nombre de questions, qui sont typiquement du domaine de l'analyse numérique, auxquelles nous tenterons de répondre dans la suite :

1. Le problème qu'on a obtenu en dimension finie, (avec des inconnues localisées aux noeuds du maillage dans le cas de la méthode des différences finies et dans les mailles dans le cas de la méthode des volumes finis) admet-il une (unique) solution ? On montrera que oui.
2. La solution du problème discret converge-t-elle vers la solution du problème continu lorsque le pas du maillage h tend vers 0 ? Dans le cas des différences finies en une dimension d'espace, le pas du maillage est défini par

   $$h = \sup_{i=1\ldots N} |x_{i+1} - x_i|. \tag{1.1.10}$$

   Dans le cas des volumes finis en une dimension d'espace, il est défini par :

   $$h = \sup_{i=1\ldots N} |x_{i+1/2} - x_{i-1/2}|. \tag{1.1.11}$$

   en deux dimensions d'espace, le pas h est défini par

   $$h = \sup_{K\in\mathcal{T}} \text{diam}(K), \text{ avec diam}(K) = \sup_{x,y\in K} d(x,y),$$

   où $\mathcal{T}$, le maillage, est l'ensemble des volumes de contrôle K. Notons que la réponse à cette question n'est pas évidente *a priori*. La solution discrète peut converger vers la solution continue, elle peut aussi converger mais vers autre chose que la solution du problème continu, et enfin elle peut ne pas converger du tout.

1.2 Etude de la méthode différences finies pour un problème elliptique unidimensionnel

On cherche à discrétiser le problème aux limites, suivant :

$$\begin{cases} -u''(x) + c(x)u(x) = f(x), & 0 < x < 1, \\ u(0) = u(1) = 0, \end{cases} \tag{1.2.12}$$

où $c \in C([0,1],\mathbb{R}_+)$, et $c \in C([0,1],\mathbb{R})$, qui peut modéliser par exemple un phénomène de diffusion - réaction d'une espèce chimique. On se donne un pas du maillage constant $h = \frac{1}{N+1}$, et une subdivision de $]0,1[$, notée $(x_k)_{k=0,\ldots,N+1}$, avec : $x_0 = 0 < x_1 < x_2 < \ldots < x_N < x_{N+1} = 1$. Soit u_i l'inconnue discrète associée au noeud i $(i = 1,\ldots,N)$. On pose $u_0 = u_{N+1} = 0$. On obtient les équations discrètes en approchant $u''(x_i)$ par quotient différentiel par développement de Taylor, comme on l'a vu au paragraphe 1.1.1 page 8.

$$\begin{cases} \frac{1}{h^2}(2u_i - u_{i-1} - u_{i+1}) + c_i u_i = f_i, \quad i = 1,\ldots,N, \\ u_0 = u_{N+1} = 0. \end{cases} \tag{1.2.13}$$

avec $c_i = c(x_i)$ et $f_i = f(x_i)$. On peut écrire ces équations sous forme matricielle :

$$A_h U_h = b_h, \text{ avec } U_h = \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix} \text{ et } b_h = \begin{pmatrix} f_1 \\ \vdots \\ f_N \end{pmatrix} \tag{1.2.14}$$

$$\text{et } A_h = \begin{pmatrix} 2+c_1h^2 & -1 & 0 & \ldots & 0 \\ -1 & 2c_2h^2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2+c_{N-1}h^2 & -1 \\ 0 & \ldots & 0 & -1 & 2+c_Nh^2 \end{pmatrix}. \tag{1.2.15}$$

Les questions suivantes surgissent alors naturellement :

1. Le système (1.2.14) admet-il un unique solution ?
2. A-t-on convergence de U_h vers u et en quel sens ?

Nous allons répondre par l'affirmative à ces deux questions. Commençons par la première.

Proposition 1.2 *Soit $c = (c_1,\ldots,c_N)^t \in \mathbb{R}^N$ tel que $c_i \geq 0$ pour $i = 1,\ldots,N$; alors la matrice A_h définie par (1.2.15) est symétrique définie positive, et donc inversible.*

Démonstration : La matrice A_h est évidemment symétrique. Montrons qu'elle est définie positive. Soit $v = (v_1 \ldots v_N)^t$, on pose $v_0 = v_{N+1} = 0$. Calculons le produit scalaire $A_h v \cdot v = v^t A_h v$. On a :

$$A_h v \cdot v = \frac{1}{h^2}\ (v_1 \ldots v_N) \begin{pmatrix} 2+c_1h^2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & & -1 \\ 0 & & -1 & 2+c_Nh^2 \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ \vdots \\ v_N \end{pmatrix},$$

c'est-à -dire :

$$A_h v \cdot v = \frac{1}{h^2} \sum_{i=1}^{N} v_i(-v_{i-1} + (2 + c_i h^2)v_i - v_{i+1}).$$

On a donc, par changement d'indice :

$$A_h v \cdot v = \frac{1}{h^2} \left[\sum_{i=1}^{N} (-v_{i-1}v_i) + \sum_{i=1}^{N} (2 + c_i h^2)v_i^2 - \sum_{j=2}^{N+1} v_{j-1}v_j \right].$$

Et comme on a posé $v_0 = 0$ et $v_{N+1} = 0$, on peut écrire :

$$A_h v \cdot v = \frac{1}{h^2} \sum_{i=1}^{N} (2 + c_i h^2)v_i^2 + \frac{1}{h^2} \sum_{i=1}^{N} (-2v_i v_{i-1}),$$

soit encore :

$$A_h v \cdot v = \sum_{i=1}^{N} c_i v_i^2 + \frac{1}{h^2} \sum_{i=1}^{N} (-2v_i v_{i-1} + v_i^2 + v_{i-1}^2) + v_N^2.$$

On a donc finalement :

$$A_h v \cdot v = \sum_{i=1}^{N} c_i v_i^2 + \frac{1}{h^2} \sum_{i=1}^{N} (v_i - v_{i-1})^2 + v_N^2 \geq 0, \, \forall v = (v_1, \ldots, v_N) \in \mathbb{R}^N.$$

Si on suppose $A_h v \cdot v = 0$, on a alors

$$\sum_{i=1}^{N} c_i h^2 v_i^2 = 0 \text{ et } v_i - v_{i-1} = 0, \qquad \forall i = 1 \ldots N.$$

On a donc $v_1 = v_2 = \ldots = v_N = v_0 = v_{N+1} = 0$. Remarquons que ces égalités sont vérifiées même si les c_i sont nuls. Ceci démontre que la matrice A_h est bien définie.

Remarque 1.3 (Existence et unicité de la solution) *On a montré ci-dessus que A_h est symétrique définie positive, donc inversible, ce qui entraîne l'existence et l'unicité de la solution de (1.2.14). On aurait pu aussi démontrer l'existence et l'unicité de la solution de (1.2.14) directement, en montrant que $Ker(A_h) = 0$ (voir exercice 2 page 32). On rappelle qu'en dimension finie, toute application linéaire injective ou surjective est bijective. On en déduit ainsi l'existence de la solution du système (1.2.14).*

Remarque 1.4 (Caractère défini et conditions limites) *Dans la démonstration de la proposition 1.2, si $c_i > 0$ pour tout $i = 1, \ldots, N$ le terme $\sum_{i=1}^{N} c_i h^2 v_i^2 = 0$ permet de conclure que $v_i = 0$ pour tout $i = 1, \ldots, N$. Par contre, si $c_i \geq 0$ (ou même $c_i = 0$ pour tout $i = 1, \ldots, N$, c'est grâce aux conditions au limites de Dirichlet homogènes (représentées par le fait qu'on pose $v_0 = 0$ et $v_{N+1} = 0$ ce qui permet d'écrire alors les équations 1 et N sous la même forme que l'équation i) qu'on peut montrer que que $v_i = 0$, pour tout $i = 1, \ldots, N$, car $v_i = v_{i-1}$, pour tout $i = 1, \ldots, N$, et $v_0 = 0$. En particulier, la matrice de discrétisation de $-u''$ par différences finies avec conditions aux limites de Neumann homogènes :*

$$\begin{cases} -u'' = f, \\ u'(0) = u'(1) = 0. \end{cases} \tag{1.2.16}$$

donne une matrice A_h qui est symétrique et positive, mais non définie (voir exercice 11 page 36). De fait la solution du problème continu (1.2.16) n'est pas unique, puisque les fonctions constantes sur [0,1] *sont solutions de (1.2.16).*

Nous allons maintenant nous préoccuper de la question de la convergence.

Définition 1.5 (Matrices monotones) *Soit $A \in \mathcal{M}_N(\mathbb{R})$, de coefficients $a_{i,j}$, $i = 1,\ldots,N$ et $j = 1,\ldots,N$. On dit que A est positive (ou $A \geq 0$) si $a_{i,j} \geq 0$, $\forall\ i,j = 1,\ldots,N$. On dit que A est monotone si A est inversible et $A^{-1} \geq 0$.*

L'avantage des schémas à matrices monotones est de satisfaire la propriété de conservation de la positivité, qui peut être cruciale dans les applications physiques :

Définition 1.6 (Conservation de la positivité) *Soit $A \in \mathcal{M}_N(\mathbb{R})$, de coefficients $a_{i,j}$, $i = 1,\ldots,N$ et $j = 1,\ldots,N$; on dit que A conserve la positivité si $Av \geq 0$ entraîne $v \geq 0$ (les inégalités s'entendent composante par composante).*

On a en effet la proposition suivante :

Proposition 1.7 (Monotonie et positivité) *Soit $A \in \mathcal{M}_N(\mathbb{R})$. Alors A conserve la positivité si et seulement si A est monotone.*

Démonstration : Supposons d'abord que A conserve la positivité, et montrons que A inversible et que A^{-1} a des coefficients ≥ 0. Si x est tel que $Ax = 0$, alors $Ax \geq 0$ et donc, par hypothèse, $x \geq 0$. Mais on a aussi $Ax \leq 0$, soit $A(-x) \geq 0$ et donc par hypothèse, $x \leq 0$. On en déduit $x = 0$, ce qui prouve que A est inversible. La conservation de la positivité donne alors que $y \geq 0 \Rightarrow A^{-1}y \geq 0$. En prenant $y = e_1$ on obtient que la première colonne de A^{-1} est positive, puis en prenant $y = e_i$ on obtient que la i-ème colonne de A^{-1} est positive, pour $i = 2,\ldots,N$. Donc A^{-1} a tous ses coefficients positifs.

Réciproquement, supposons maintenant que A est inversible et que A^{-1} a des coefficients positifs. Soit $x \in \mathbb{R}^N$ tel que $Ax = y \geq 0$, alors $x = A^{-1}y \geq 0$. Donc A conserve la positivité. ∎

Remarque 1.8 (Principe du maximum) *On appelle principe du maximum continu le fait que si $f \geq 0$ alors le minimum de la fonction u solution du problème (1.2.12) page 13 est atteint sur les bords. Cette propriété mathématique correspond à l'intuition physique qu'on peut avoir du phénomène : si on chauffe un barreau tout en maintenant ses deux extrémités à une température fixe, la température aux points intérieurs du barreau sera supérieure à celle des extrémités. Il est donc souhaitable que la solution approchée satisfasse la même propriété (voir exercice 5 page 34 à ce sujet).*

Lemme 1.9 *Soit $c = (c_1,\ldots,c_N)^t \in \mathbb{R}^N$, et $A_h \in \mathcal{M}_N(\mathbb{R})$ définie par (1.2.15). Si $c_i \geq 0$ pour tout $i = 1,\ldots,N$, alors A_h est monotone.*

Démonstration : On va montrer que si $v \in \mathbb{R}^N$, $A_h v \geq 0$ alors $v \geq 0$. On peut alors utiliser la proposition 1.7 pour conclure. Soit $v = (v_1,\ldots,v_N)^t \in \mathbb{R}^N$. Posons $v_0 = v_{N+1} = 0$.. Supposons que $A_h v \geq 0$. On a donc

$$-\frac{1}{h^2}v_{i-1} + \left(\frac{2}{h^2} + c_i\right)v_i - \frac{1}{h^2}v_{i+1} \geq 0, \quad i = 1,\ldots,N \tag{1.2.17}$$

Soit

$$p = \min\left\{i \in \{1,\ldots,N\}; v_p = \min_{j=1,\ldots,N} v_j\right\}.$$

Supposons que $\min_{j=1,\ldots,N} v_j < 0$. On a alors $p \geq 1$ et :

$$\frac{1}{h^2}(v_p - v_{p-1}) + c_p v_p + \frac{1}{h^2}(v_p - v_{p-1}) \geq 0.$$

On en déduit que

$$\frac{2}{h^2}c_p v_p \geq \frac{1}{h^2}(v_{p-1} - v_p) + \frac{1}{h^2}(v_{p+1} - v_p) \geq 0.$$

Si $c_p > 0$, on a donc $v_p \geq 0$, et donc $v_i \geq 0, \quad \forall i = 1, \ldots, N$. Si $c_p = 0$, on doit alors avoir $v_{p-1} = v_p = v_{p+1}$ ce qui est impossible car p est le plus petit indice j tel que $v_j = \min_{i=1,\ldots,N} v_i$. Donc dans ce cas le minimum ne peut pas être atteint pour $j = p > 1$. On a ainsi finalement montré que $\min_{i\in\{1,\ldots,N\}} v_i \geq 0$, on a donc $v \geq 0$. ■

Définition 1.10 (Erreur de consistance) *On appelle erreur de consistance la quantité obtenue en remplaçant l'inconnue par la solution exacte dans le schéma numérique. Dans le cas du schéma (1.2.13), l'erreur de consistance au point x_i est donc définie par :*

$$R_i = \frac{1}{h^2}(2u(x_i) - u(x_{i-1}) - u(x_{i+1})) + c(x_i)u(x_i) - f(x_i). \tag{1.2.18}$$

L'erreur de consistance R_i est donc l'erreur qu'on commet en remplaçant l'opérateur $-u''$ par le quotient différentiel

$$\frac{1}{h^2}(2u(x_i) - u(x_{i-1}) - u(x_{i+1})).$$

Cette erreur peut être évaluée si u est suffisamment régulière, en effectuant des développements de Taylor.

Définition 1.11 (Ordre du schéma) *On dit qu'un schéma de discrétisation à N points de discrétisation est d'ordre p s'il existe $C \in \mathbb{R}$, ne dépendant que de la solution exacte, tel que l'erreur de consistance satisfasse:*

$$\max_{i=1,\ldots,N}(R_i) < ch^p,$$

où h est le le pas du maillage défini par (1.1.3) (c.à.d. le maximum des écarts $x_{i+1} - x_i$). On dit qu'un schéma de discrétisation est consistant si

$$\max_{i=1,\ldots N}(R_i) \to 0 \text{ lorsque } h \to 0,$$

où N est le nombre de points de discrétisation.

Lemme 1.12 *Si la solution de (1.2.12) vérifie $u \in C^4([0,1])$, alors le schéma (1.2.13) est consistant d'ordre 2, et on a plus precisément :*

$$|R_i| \leq \frac{h^2}{12} \sup_{[0,1]} |u^{(4)}|, \ \forall i = 1, \ldots, N. \tag{1.2.19}$$

Démonstration : Par développement de Taylor, on a :

$$u(x_{i+1}) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\xi_i)$$

$$u(x_{i-1}) = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\eta_i)$$

En additionnant ces deux égalités, on obtient que :

$$\frac{1}{h^2}(u(x_{i+1}) + u(x_i) - 2u(x_i)) = u''(x_i) + \frac{h^2}{24}(u^{(4)}(\xi_i) + u^{(4)}(\eta_i)),$$

ce qui entraîne que :

$$|R_i| \leq \frac{h^2}{12} \sup_{[0,1]} |u^{(4)}|. \tag{1.2.20}$$

■

Remarque 1.13 (Sur l'erreur de consistance)

1. *Si on note* $\bar{U}_h : (u(x_i))_{i=1\dots N}$ *le vecteur dont les composantes sont les valeurs exactes de la solution de (1.2.12), et* $U_h = (u_1 \dots u_N)^t$ *la solution de (1.2.13), on a :*

$$R = A_h(U_h - \bar{U}_h). \tag{1.2.21}$$

2. *On peut remarquer que si* $u^{(4)} = 0$*, les développements de Taylor effectués ci-dessus se résument à :*

$$-u''(x_i) = \frac{2u(x_i) - u(x_{i-1}) - u(x_{i+1})}{h^2},$$

et on a donc $R_i = 0$*, pour tout* $i = 1, \dots, N$*, et donc* $u_i = u(x_i)$*, pour tout* $i = 1 \dots N$*. Dans ce cas (rare!), le schéma de discrétisation donne la valeur exacte de la solution en* x_i*, pour tout* $i = 1, \dots, N$*. Cette remarque est bien utile lors de la phse de validation de méthodes et numériques et/ou programmes informatiques pour la résolution de l'équation (1.2.12). En effet, si on choisit* f *telle que la solution soir un polynôme de degré inférieur ou égal à 3, alors on doit avoir une erreur entre solution exacte et approchée inférieure à l'erreur machine.*

La preuve de convergence du schéma utilise la notion de consistance, ainsi qu'une notion de stabilité, que nous introduisons maintenant :

Proposition 1.14 *On dit que le schéma (1.2.13) est stable, au sens où la matrice de discrétisation* A_h *satisfait :*

$$\|A_h^{-1}\|_\infty \leq \frac{1}{8}. \tag{1.2.22}$$

On peut réécrire cette inégalité comme une estimation sur les solutions du système (1.2.14) :

$$\|U_h\| \leq \frac{1}{8}\|f\|_\infty. \tag{1.2.23}$$

Démonstration : On rappelle que par définition, si $M \in \mathcal{M}_N(\mathbb{R})$,

$$\|M\|_\infty = \sup_{\substack{v \in \mathbb{R}^N \\ v \neq 0}} \frac{\|M\|_\infty}{\|v\|_\infty}, \text{ avec } \|v\|_\infty = \sup_{i=1,\dots,N} |v_i|.$$

Pour montrer que $\|A_h^{-1}\|_\infty \leq \frac{1}{8}$, on décompose la matrice A_h sous la forme $A_h = A_{0h} + diag(c_i)$ où A_{0h} est la matrice de discrétisation de l'opérateur $-u''$ avec conditions aux limites de Dirichlet homogènes, et

$$A_{0h} = \begin{bmatrix} \frac{2}{h^2} & -\frac{1}{h^2} & & 0 & \\ -\frac{1}{h^2} & & \ddots & & \\ & & & \ddots & -\frac{1}{h^2} \\ 0 & & & -\frac{1}{h^2} & \frac{2}{h^2} \end{bmatrix} \tag{1.2.24}$$

et $diag(c_i)$ désigne la matrice diagonale de coefficients diagonaux c_i. Les matrices A_{0h} et A_h sont inversibles, et on a :

$$A_{0h}^{-1} - A_h^{-1} = A_{0h}^{-1} A_h A_h^{-1} - A_{0h}^{-1} A_{0h} A_h^{-1} = A_{0h}^{-1}(A_h - A_{0h})A_h^{-1}.$$

Comme $diag(c_i) \geq 0$, on a $A_h \geq A_{0h}$, et comme A_{0h} et A_h sont monotones, on en déduit que :

$$0 \leq A_h^{-1} \leq A_{0h}^{-1}, \text{ (composante par composante).}$$

On peut maintenant remarquer que si $B \in \mathcal{M}_N(\mathbb{R})$, et si $B \geq 0$ (c.à.d. $B_{ij} \geq 0$ pour tout i et j), on a

$$\|B\|_\infty = \sup_{\substack{v \in \mathbb{R}^N \\ \|v\|=1}} \sup_{i=1,\ldots,N} |(Bv)_i| = \sup_{\substack{v \in \mathbb{R}^N \\ \|v\|=1}} \sup_{i=1,\ldots,N} \left| \sum_{j=1}^N B_{ij} v_j \right| \|B\|_\infty = \sup_{i=1,\ldots,N} \sum_{j=1}^N B_{ij}.$$

On a donc $\|A_h^{-1}\| = \sup_{i=1,\ldots,N} \sum_{j=1}^N (A_h^{-1})_{ij} \leq \sup_{i=1,\ldots,N} \sum_{j=1}^N (A_{0h}^{-1})_{ij}$ car $A_h^{-1} \leq A_{0h}^{-1}$; d'où on déduit que $\|A_h^{-1}\|_\infty \leq \|A_{0h}^{-1}\|_\infty$. Il ne reste plus qu'à estimer $\|A_{0h}^{-1}\|_\infty$. Comme $A_{0h}^{-1} \geq 0$, on a

$$\|A_{0h}^{-1}\|_\infty = \|A_{0h}^{-1} e\|_\infty \text{ avec } e = (1, \ldots, 1)^t.$$

Soit $d = A_{0h}^{-1} e \in \mathbb{R}^N$. On veut calculer $\|d\|_\infty$, où d vérifie $A_{0h} d = e$. Or le système linéaire $A_{0h} d = e$ n'est autre que la discrétisation par différences finies du problème

$$\begin{cases} -u'' = 1 \\ u(0) = u(1) = 0 \end{cases}, \tag{1.2.25}$$

dont la solution exacte est :

$$u_0(x) = \frac{x(1-x)}{2},$$

qui vérifie $u_0^{(4)}(x) = 0$. On en conclut, par la remarque 1.13, que

$$u_0(x_i) = d_i, \quad \forall i = 1 \ldots N.$$

Donc $\|d\|_\infty = \sup_{i=1,N} \dfrac{ih(ih-1)}{2}$ où $h = \dfrac{1}{N+1}$ est le pas de discrétisation Ceci entraîne que

$$\|d\|_\infty \leq \sup_{[0,1]} \left| \frac{x(x-1)}{2} \right| = \frac{1}{8}, \text{ et donc que } \|A_h^{-1}\|_\infty \leq \frac{1}{8}.$$

■

Remarque 1.15 (Sur la stabilité) *Noter que l'inégalité (1.2.23) donne une estimation sur les solutions approchées indépendantes du pas de maillage. C'est ce type d'estimation qu'on recherchera par la suite pour la discrétisation d'autres problèmes comme garant de la stabilité d'un schéma numérique.*

Définition 1.16 (Erreur de discrétisation) *On appelle erreur de discrétisation en x_i, la différence entre la solution exacte en x_i et la i-ème composante de la solution donnée par le schéma numérique*

$$e_i = u(x_i) - u_i, \quad \forall i = 1, \ldots, N. \tag{1.2.26}$$

Théorème 1.17 *Soit u la solution exacte de*

$$\begin{cases} -u'' + cu = f, \\ u(0) = u(1) = 0. \end{cases}$$

On suppose $u \in C^4([0,1])$. Soit u_h la solution de (1.2.13). Alors l'erreur de discrétisation définie par (1.2.26) satisfait

$$\max_{i=1,\dots,N} |e_i| \leq \frac{1}{96} \|u^{(4)}\|_\infty h^2.$$

Le schéma est donc convergent d'ordre 2.

Démonstration : Soit $U_h = (U_1, \dots, U_n)^t$ et $\bar{U}_h = (u(x_1), \dots, u(x_N))^t$, on cherche à majorer $\|\bar{U}_h - U_h\|_\infty$. On a $A(\bar{U}_h - U_h) = R$ où R est l'erreur de consistance (voir remarque 1.13). On a donc

$$\|\bar{U}_h - U_h\|_\infty \leq \|A_h^{-1}\|_\infty \|R\|_\infty \leq \frac{1}{8} \times \frac{1}{12} \|u^{(4)}\|_\infty = \frac{1}{96} \|u^{(4)}\|_\infty$$

■

Remarque 1.18 (Sur la convergence) *On peut remarquer que la preuve de la convergence s'appuie sur la stabilité (elle-même déduite de la conservation de la positivité) et sur la consistance. Dans certains livres d'analyse numérique, vous trouverez la "formule" : stabilité + consistance $\Longrightarrow$ convergence. Il faut toutefois prendre garde au fait que ces notions de stabilité et convergence peuvent être variables d'un type de méthode à un autre (comme nous le verrons en étudiant la méthode des volumes finis, par exemple).*

Remarque 1.19 (Contrôle des erreurs d'arrondi) *On cherche à calculer la solution approchée de $-u'' = f$. Le second membre f est donc une donnée du problème. Supposons que des erreurs soient commises sur cette donnée (par exemple des erreurs d'arrondi, ou des erreurs de mesure). On obtient alors un nouveau système, qui s'écrit $A_h \tilde{U}_h = b_h + \varepsilon_h$, où ε_h représente la discrétisation des erreurs commises sur le second membre. Si on résout $A_h \tilde{U}_h = b_h + \varepsilon_h$ au lieu de $A_h U_h = b_h$, l'erreur commise sur la solution du système s'écrit*

$$E_h = \tilde{U}_h - U_h = A_h^{-1} \varepsilon_h.$$

On en déduit que

$$\|E_h\|_\infty \leq \frac{1}{8} \|\varepsilon_h\|_\infty.$$

On a donc une borne d'erreur sur l'erreur qu'on obtient sur la solution du système par rapport à l'erreur commise sur le second membre.

1.3 Schéma volumes finis pour un problème elliptique en une dimension d'espace

1.3.1 Origine du Schéma

On va étudier la discrétisation par volumes finis du problème (1.1.1)–(1.1.2), qu'on rappelle ici :

$$\begin{cases} -u_{xx} = f, \qquad x \in]0,1[, \\ u(0) = u(1) = 0. \end{cases} \tag{1.3.27}$$

Définition 1.20 (Maillage volumes finis) *On appelle maillage volumes finis de l'intervalle* [0,1]*, un ensemble de* N *mailles* $(K_i)_{i=1,\dots,N}$*, telles que* $K_i =]x_{i-1/2}, x_{i+1/2}[$*, avec* $x_{1/2} = 0 < x_{\frac{3}{2}} < x_{i-1/2} < x_{i+1/2} < \ldots < x_{N+1/2} = 1$*, et on note* $K_i = x_{i+1/2} - x_{i-1/2}$*. On se donne également* N *points* $(x_i)_{i=1,\dots,N}$ *situés dans les mailles* K_i*. On a donc :*

$$0 = x_{1/2} < x_1 < x_{\frac{3}{2}} < \ldots < x_{i-1/2} < x_i < x_{i+1/2} < \ldots < x_{N+1/2} = 1.$$

On notera $h_{i+1/2} = x_{i+1} - x_i$*, et* $h = \max_{i=1,\dots,N}$*, et pour des questions de notations, on posera également* $x_0 = 0$ *et* $x_{N+1} = 1$.
On intègre (1.1.1) sur $K_i = x_{i+1/2} - x_{i-1/2}$, et on obtient :

$$-u_x(x_i + 1/2) + u_x(x_i - 1/2) = \int_{K_i} f(x)dx. \tag{1.3.28}$$

On pose : $f_i = \frac{1}{h_i} \int_{k_i} f(x)dx$, et on introduit les inconnues discrètes $(u_i)_{i=1\dots N}$ (une par maille) et les équations discrètes du schéma numérique :

$$F_{i+1/2} - F_{i-1/2} = h_i f_i, \qquad i = 1, \dots, N, \tag{1.3.29}$$

où $F_{i+1/2}$ est le flux numérique en $x_{i+1/2}$ qui devrait être une approximation raisonnable de $-u_x(x_{i+1/2})$.
On pose alors :

$$F_{i+1/2} = -\frac{u_{i+1} - u_i}{h_{i+1/2}}, \qquad i = 1, \dots, N,$$

$$F_{1/2} = -\frac{u_1}{h_{1/2}}, \; F_{N+1/2} = \frac{u_N}{h_{N+1/2}},$$

pour tenir compte des conditions aux limites de Dirichlet homogènes $u(0) = u(1) = 0$. On peut aussi écrire :

$$F_{i+1/2} = -\frac{u_{i+1} - u_i}{h_{i+1/2}}, \qquad i = 0, \dots, N, \tag{1.3.30}$$

$$\text{en posant } u_0 = u_{N+1} = 0. \tag{1.3.31}$$

On peut écrire le système linéaire obtenu sur $(u_1, \dots, u_N)^t$ sous la forme

$$A_h U_h = b_h, \tag{1.3.32}$$

avec

$$(A_h)_i = \frac{1}{h_i} \left[\frac{-1}{h_{i+1/2}}(u_{i+1} - u_i) + \frac{1}{h_{i-1/2}}(u_i - u_{i-1}) \right] \text{ et } (b_h)_i = f_i.$$

Remarque 1.21 (Non consistance au sens des différences finies)
L'approximation de $-u''(x_i)$ *par*

$$\frac{1}{h_i} \left[\frac{-1}{h_{i+1/2}}(u(x_{i+1}) - u(x_i)) + \frac{1}{h_{i-1/2}}(u(x_i) - u(x_{i-1}) \right]$$

n'est pas consistante dans le cas général : voir exercice 8.

On peut montrer que les deux schémas différences finies et volumes sont identiques "au bord près" dans le cas d'un maillage uniforme avec x_i : centre de la maille voir exercice 1 page 32.

1.3.2 Analyse mathématique du schéma.

On va démontrer ici qu'il existe une unique solution $(u_1, \dots u_N)^t$ au schéma (1.3.29)–(1.3.31), et que cette solution, et que cette solution converge, en un certain sens, vers la solution de problème continu (1.3.27) lorsque le pas du maillage tend vers 0.

Proposition 1.22 (Existence de la solution du schéma volumes finis) *Soit $f \in C([0,1])$ et $u \in C^2([0,1])$ solution de (1.3.27). Soit $(K_i)_{i=1,\dots N}$ le maillage par la définition 1.20 page 20. Alors il existe une unique solution $u_h = (u_1, \dots, u_N)^t$ de (1.3.29)–(1.3.31).*

Démonstration : Le schéma s'écrit

$$-\frac{u_{i+1}-u_i}{h_{i+1/2}} + \frac{u_i - u_{i-1}}{h_{i-1/2}} = h_i f_i, \quad i = 1, \dots, N.$$

(où on a posé $u_0 = 0$ et $u_{N+1} = 0$) En multipliant par u_i et en sommant de $i = 1$ à N, on obtient donc :

$$\sum_{i=1}^{N} -\frac{u_{i+1}-u_i}{h_{i+1/2}} u_i + \sum_{i=1}^{N} \frac{u_i - u_{i-1}}{h_{i-1/2}} u_i = \sum_{i=1}^{N} h_i f_i u_i.$$

En effectuant un changement d'indice sur la deuxième somme, on obtient :

$$\sum_{i=1}^{N} -\frac{u_{i+1}-u_i}{h_{i+1/2}} u_i + \sum_{i=0}^{N-1} \frac{u_{i+1} - u_i}{h_{i+1/2}} u_{i+1} = \sum_{i=1}^{N} h_i f_i u_i;$$

en regroupant les sommes, on a donc :

$$\sum_{i=1}^{N} \frac{(u_{i+1}-u_i)^2}{h_{i+1/2}} + \frac{u_1^2}{h_{1/2}} + \frac{u_N^2}{h_{N+1/2}} = \sum_{i=1}^{N} h_i f_i u_i.$$

Si $f_i = 0$ pour tout $i = 1, \dots, N$, on a bien alors $u_i = 0$ pour tout $i = 1 \dots N$. Ceci démontre l'unicité de $(u_i)_{i=1\dots N}$ solution de (1.3.29)–(1.3.31), et donc son existence, puisque le système (1.3.29)–(1.3.31) est un système linéaire carré d'ordre N. (On rappelle qu'une matrice carrée d'ordre N est inversible si et seulement si son noyau est réduit à $\{0\}$. ∎

Lemme 1.23 (Consistance des flux) *Soit $u \in C^2([0,1])$ solution de (1.3.27). On se donne une subdivision de $[0,1]$. On appelle $\bar{F}_{i+1/2} = -u_x(x_{i+1/2})$ le flux exact en $x_{i+1/2}$, et $F^*_{i+1/2} = -\frac{u(x_{i+1})-u(x_i)}{h_{i+1/2}}$ le quotient différentiel qui approche la dérivée première $-u_x(x_{i+1/2})$. On dit que le flux numérique $F_{i+1/2} = -\frac{u_{i+1}-u_i}{h_{i+1/2}}$ est consistant s'il existe $C \in \mathbb{R}_+$ ne dépendant que de u telle que l'erreur de consistance sur le flux, définie par :*

$$R_{i+1/2} = \bar{F}_{i+1/2} - F^*_{i+1/2},$$

vérifie

$$|R_{i+1/2}| \leq Ch. \tag{1.3.33}$$

La démonstration de ce résultat s'effectue facilement à l'aide de développements de Taylor. On peut aussi montrer (voir exercice 9 page 36) que si $x_{i+1/2}$ est au centre de l'intervalle $[x_i x_{i+1}]$, l'erreur de consistance sur les flux est d'ordre 2, *i.e.* il existe $C \in \mathbb{R}_+$ ne dépendant que de u telle que $R_{i+1/2} \leq Ch^2$. Notez que cette propriété de consistance est vraie sur les flux, et non pas sur l'opérateur $-u''$ (voir remarque 1.21).

Définition 1.24 (Conservativité) *On dit que le schéma volumes finis (1.3.29)–(1.3.31) est conservatif, au sens où, lorsqu'on considère une interface $x_{i+1/2}$ entre deux mailles K_i et K_{i+1}, le flux numérique entrant dans une maille est égal à celui sortant de l'autre.*

C'est grâce à la conservativité et à la consistance des flux qu'on va montrer la convergence du schéma volumes finis.

Théorème 1.25 (Convergence du schéma volumes finis) *On suppose que la solution u de (1.3.27) vérifie $u \in C^2([0,1])$. On pose pour $e_i = u(x_i) - u_i$ pour $i = 1,\ldots,N$, et $e_0 = e_{N+1} = 0$. Il existe $C \geq 0$ ne dépendant que de u tel que :*

$$\sum_{i=0}^{N} \frac{(e_{i+1} - e_i)^2}{h} \leq Ch^2, \tag{1.3.34}$$

$$\sum_{i=1}^{N} h e_i^2 \leq Ch^2 \tag{1.3.35}$$

$$\max_{i=1\ldots N} |e_i| \leq Ch. \tag{1.3.36}$$

(On rappelle que $h = \sup_{i=1\ldots N} h_i$.)

Démonstration : Ecrivons le schéma volumes finis (1.3.29) :

$$F_{i+1/2} - F_{i-1/2} = h_i f_i,$$

l'équation exacte intégrée sur la maille K_i (1.3.28) :

$$\bar{F}_{i+1/2} - \bar{F}_{i-1/2} = h_i f_i,$$

où $\bar{F}_{i+1/2}$ est défini dans le lemme 1.23, et soustrayons :

$$\bar{F}_{i+1/2} - F_{i+1/2} - \bar{F}_{i-1/2} + F_{i-1/2} = 0.$$

En introduisant $R_{i+1/2} = \overline{F}_{i+1/2} - F^*_{i+1/2}$, on obtient :

$$F^*_{i+1/2} - F_{i+1/2} - F^*_{i-1/2} + F_{i-1/2} = -R_{i+1/2} + R_{i-1/2}$$

ce qui s'écrit encore, au vu de la définition de e_i,

$$-\frac{1}{h_{i+1/2}}(e_{i+1} - e_i) + \frac{1}{h_{i-1/2}}(e_i - e_{i-1}) = -R_{i+1/2} + R_{i-1/2}.$$

On multiplie cette dernière égalité par e_i et on somme de 1 à N :

$$\sum_{i=1}^{N} -\frac{1}{h_{i+1/2}}(e_{i+1} - e_i)e_i + \sum_{i=1}^{N} \frac{1}{h_{i-1/2}}(e_i - e_{i-1})e_i \sum_{i=1}^{N} -R_{i+1/2}e_i + \sum_{i=1}^{N} R_{i-1/2}e_i,$$

ce qui s'écrit encore :

$$\sum_{i=1}^{N} -\frac{1}{h_{i+1/2}}(e_{i+1} - e_i)e_i + \sum_{i=0}^{N-1} \frac{1}{h_{i+1/2}}(e_{i+1} - e_i)e_{i+1} \sum_{i=1}^{N} -R_{i+1/2}e_i + \sum_{i=0}^{N-1} R_{i+1/2}e_{i+1}$$

En réordonnant les termes, on obtient, en remarquant que $e_0 = 0$ et $e_{N+1} = 0$:

$$\sum_{i=0}^{N} \frac{(e_{i+1} - e_i)^2}{h_{i+1/2}} = \sum_{i=0}^{N} R_{i+1/2}(e_{i+1} - e_i).$$

Or, $R_{i+1/2} \leq C\ \ h$ (par le lemme 1.23). On a donc

$$\sum_{i=0}^{N} \frac{(e_{i+1}-e_i)^2}{h_{i+1/2}} \leq C\ h \sum_{i=0}^{N} \frac{|e_{i+1}-e_i|}{\sqrt{h_{i+1/2}}} \sqrt{h_{i+1/2}},$$

et, par l'inégalité de Cauchy-Schwarz :

$$\sum_{i=0}^{N} \frac{(e_{i+1}-e_i)^2}{h_{i+1/2}} \leq C\ h \left(\sum_{i=0}^{N} \frac{|e_{i+1}-e_i|^2}{h_{i+1/2}} \right)^{1/2} \times \left(\sum_{i=0}^{N} h_{i+1/2} \right)^{1/2}.$$

En remarquant que $\sum_{i=0}^{N} h_{i+1/2} = 1$, on déduit que :

$$\sum_{i=0}^{N} \frac{(e_{i+1}-e_i)^2}{h_{i+1/2}} \leq C\ h \left(\sum \frac{|e_{i+1}-e_i|^2}{h_{i+1/2}} \right)^{1/2},$$

et donc

$$\left(\sum_{i=0}^{N} \frac{(e_{i+1}-e_i)^2}{h_{i+1}} \right)^{1/2} \leq C\ h.$$

On a ainsi démontré (1.3.34). Démontrons maintenant (1.3.36). Pour obtenir une majoration de $|e_i|$ par $C\ h$, on remarque que :

$$|e_i| = \left| \sum_{j=1}^{i} e_j - e_{j-1} \right| \leq \sum_{j=1}^{i} |e_j - e_{j-1}| \leq \sum_{j=1}^{N} |e_j - e_{j-1}|.$$

On en déduit, par l'inégalité de Cauchy Schwarz, que :

$$|e_i| \leq \left(\sum \frac{|e_j - e_{j-1}|^2}{h_{i+1/2}} \right)^{1/2} \left(\sum h_{i+1/2} \right)^{1/2},$$

ce qui entraîne $\max_{i=1\ldots N} |e_i| \leq C\ h$. Notons que de cette estimation, on déduit immédiatement l'estimation (1.3.35).

■

Remarque 1.26 (Espaces fonctionnels et normes discrètes) *On rappelle qu'une fonction u de $L^2(]0,1[)$ admet une dérivée faible dans $L^2(]0,1[)$ s'il existe $v \in L^2(]0,1[)$ telle que*

$$\int_{]0,1[} u(x)\varphi'(x)dx = -\int_{]0,1[} v(x)\varphi(x)dx, \tag{1.3.37}$$

pour toute fonction $\varphi \in C_c^1(]0,1[)$, où $C_c^1(]0,1[)$ désigne l'espace des fonctions de classe C^1 à support compact dans $]0,1[$. On peut montrer que v est unique, voir par exemple [1]. On notera $v = Du$. On peut remarquer que si $u \in C^1(]0,1[)$, alors $Du = u'$, dérivée classique. On note $H^1(]0,1[)$ l'ensemble des fonctions de $L^2(]0,1[)$ qui admettent une dérivée faible dans $L^2(]0,1[)$: $H^1(]0,1[) = \{u \in L^2(]0,1[)\ ;\ Du \in L^2(]0,1[)\}$. On a $H^1(]0,1[) \subset C(]0,1[)$ et on définit

$$H_0^1(]0,1[) = \{u \in H^1(]0,1[)\ ; u(0) = u(1) = 0\}.$$

Pour $u \in H^1(]0,1[)$, on note :

$$\|u\|_{H_0^1} = \left(\int_0^1 (Du(x))^2 dx \right)^{1/2}.$$

C'est une norme sur H_0^1 qui est équivalente à la norme $\|.\|_{H^1}$ définie par $\|u\|_{H^1} = \left(\int u^2(x)dx + \int (Du)^2(x)dx \right)^{1/2}$, ce qui se démontre grâce à l'inégalité de Poincaré :

$$\|u\|_{L^2(]0,1[)} \leq \|Du\|_{L^2(]0,1[)} \text{ pour tout } u \in H_0^1(]0,1[). \tag{1.3.38}$$

Soit maintenant $\mathcal{T}$ un maillage volumes finis de [0,1] (voir définition 1.20), on note $X(\mathcal{T})$ l'ensemble des fonctions de [0,1] dans $\mathbb{R}$, constantes par maille de ce maillage. Pour $v \in X(\mathcal{T})$, on note v_i la valeur de v sur la maille i ; on peut écrire les normes L^2 et L^∞ de v :

$$\|v\|^2_{L^2(]0,1[)} = \sum_{i=1}^N h_i v_i^2,$$

et

$$\|v\|_{L^\infty(]0,1[)} = \max_{i=1}^N |v_i|.$$

Par contre, la fonction v étant constante par maille, elle n'est pas dérivable au sens classique, ni même au sens faible On peut toutefois définir une norme H^1 discrète de v de la manière suivante :

$$|v|_{1,\mathcal{T}} = \left(\sum_{i=0}^N h_{i+1/2} (\frac{v_{i+1} - v_i}{h_{i+1/2}})^2 \right)^{1/2}$$

On peut définir une sorte de "dérivée discrète" de v par les pentes

$$p_{i+1/2} = \frac{v_{i+1} - v_i}{h_{i+1/2}}.$$

On peut alors définir une $D_\mathcal{T} v$, fonction constante par intervalle et égale à $p_{i+1/2}$ sur l'intervalle x_i,x_{i+1}. La norme L^2 de $D_\mathcal{T} v$ est donc définie par :

$$\|D_\mathcal{T} v\|^2_{L^2(]0,1[)} = \sum_{i=0}^N h_{i+1/2} p^2_{i+1/2} = \sum_{i=0}^N \sum_{i=0}^N h_{i+1/2} \frac{(v_{i+1} - v_i)^2}{h_{i+1/2}}.$$

On peut montrer (Exercice 13) que si $u_\mathcal{T} :]0,1[\longrightarrow \mathbb{R}$ est définie par $u_\mathcal{T}(x) = u_i \quad \forall x \in K_i$ où $(u_i)_{i=1,\ldots,N}$ solution de (1.3.29)–(1.3.31), alors $|u_\mathcal{T}|_{1,\mathcal{T}}$ converge dans $L^2(]0,1[)$ lorsque h tend vers 0, vers $\|Du\|_{L^2(]0,1[)}$, où u est la solution de (1.3.27).

Remarque 1.27 (Dimensions supérieures) *En une dimension d'espace, on a obtenu une estimation d'erreur en norme "H_0^1 discrète" et en norme L^∞. En dimension supérieure ou égale à 2, on aura une estimation en h, en norme H_0^1 discrète, en norme L^2, mais pas en norme L^∞. Ceci tient au fait que l'injection de Sobolev $H^1(]0,1[) \subset C(]0,1[)$ n'est vraie qu'en dimension 1. La démonstration de l'estimation d'erreur en norme L^2 (1.3.35) se prouve alors directement à partir de l'estimation en norme H_0^1 discrète, grâce à une "inégalité de Poincaré discrète", équivalent discret de la célèbre inégalité de Poincaré continue*[1] *(voir (1.3.38) pour la dimension 1.*

1. Soit Ω un ouvert borné de $\mathbb{R}^N$, et $u \in H_0^1(\Omega$, alors $\|u\|_{L^2(\Omega)} \leq \text{diam}(\Omega)\|Du\|_{L^2(]\Omega[)}\}$.

Prise en compte de discontinuités

On considère ici un barreau conducteur constitué de deux matériaux de conductivités λ_1 et λ_2 différentes, et dont les extrémités sont plongées dans de la glace. On suppose que le barreau est de longueur 1, que le matériau de conductivité λ_1 (resp. λ_2) occupe le domaine $\Omega_1 =]0,1/2[$ (resp. $\Omega_2 =]1/2,1[$). Le problème de conduction de la chaleur s'écrit alors :

$$\left\{ \begin{array}{ll} (-\lambda_1(x)u_x)_x = f(x) & x \in]0,1/2[\\ (-\lambda_2(x)u_x)_x = f(x) & x \in]1/2,1[\\ u(0) = u(1) = 0, & \\ -(\lambda_1 u_x)(1/2) = -(\lambda_2 u_x)(1/2) & \end{array} \right. \tag{1.3.39}$$

Remarque 1.28 *La dernière égalité traduit la conservation du flux de chaleur à l'interface $x = .5$. On peut noter que comme λ est discontinu en ce point, la dérivée u_x le sera forcément elle aussi.*

On choisit de discrétiser le problème par volumes finis. On se donne un maillage volumes finis comme défini par la définition 1.20 page 20, en choisissant les mailles telles que la discontinuité de λ soit située sur un interface de deux mailles qu'on note K_k et K_{k+1}. On a donc, avec les notations du paragraphe (1.1.1) $x_{k+1/2} = 0.5$. La discrétisation par volumes finis s'écrit alors

$$F_{i+1/2} - F_{i-1/2} = h_i f_i, \quad i = 1, \ldots, N,$$

où les flux numériques $F_{i+1/2}$ sont donnés par

$$F_{i+1/2} = \lambda_* \frac{u_{i+1} - u_i}{h_{i+1/2}}, \text{ avec } \lambda_* = \left\{ \begin{array}{l} \lambda_1 \text{ si } x_{i+1/2} > 0.5, \\ \lambda_2 \text{ si } x_{i+1/2} < 0.5. \end{array} \right.$$

Il ne reste donc plus qu'à calculer le flux $F_{k+1/2}$, approximation de $(\lambda u_x)(x_{k+1/2})$ (avec $x_{k+1/2} = 0.5$). On introduit pour cela une inconnue auxiliaire $u_{k+1/2}$ que l'on pourra éliminer plus tard, et on écrit une discrétisation du flux de part et d'autre de l'interface.

$$F_{k+1/2} = -\lambda_1 \frac{u_{k+1/2} - u_k}{h_k^+}, \text{ avec } h_k^+ = x_{k+1/2} - x_k,$$

$$F_{k+1/2} = -\lambda_2 \frac{u_{k+1} - u_{k+1/2}}{h_{k+1}^-} \text{ avec } h_{k+1}^- = x_{k+1} - x_{k+1/2}.$$

L'élimination (et le calcul) de l'inconnue se fait en écrivant la conservation du flux numérique :

$$-\lambda_1 \frac{u_{k+1/2} - u_k}{h_k^+} = -\lambda_2 \frac{u_{k+1} - u_{k+1/2}}{h_{k+1}^-}$$

On en déduit la valeur de $u_{k+1/2}$

$$u_{k+1/2} = \frac{\frac{\lambda_1}{h_k^+} u_k + \frac{\lambda_2}{h_{k+1}^-} u_{k+1}}{\frac{\lambda_1}{h_k^+} + \frac{\lambda_2}{h_{k+1}^-}}$$

On remplace $u_{k+1/2}$ par cette valeur dans l'expression du flux $F_{k+1/2}$, et on obtient :

$$F_{k+1/2} = \frac{\lambda_1 \ \lambda_2}{h_k^+ \lambda_2 + h_{k+1}^- \lambda_1} (u_{k+1} - u_k).$$

Si le maillage est uniforme, on obtient

$$F_{k+1/2} = \frac{2\lambda_1\lambda_2}{\lambda_1+\lambda_2}\left(\frac{u_{i+1}-u_i}{h}\right).$$

Le flux est donc calculé en faisant intervenir la moyenne harmonique des conductivités λ_1 et λ_2. Notons que lorsque $\lambda_1 = \lambda_2$, on retrouve la formule habituelle du flux.

1.4 Exemples de discrétisation par différences finies ou volumes finis des problèmes elliptiques en dimension 2.

1.4.1 Différences finies

On considère maintenant le problème de diffusion dans un ouvert Ω de $\mathbb{R}^2$:

$$\left\{\begin{array}{ll} -\Delta u = f \text{ dans } \Omega, \\ u = 0 \qquad \text{sur } \partial\Omega. \end{array}\right. \tag{1.4.40}$$

Le problème est bien posé au sens où : Si $f \in C^1(\Omega)$, alors il existe une unique solution $u \in C(\bar{\Omega}) \cap C^2(\Omega)$, solution de (1.4.40). Si $f \in L^2(\Omega)$ alors il existe une unique fonction $u \in H^2(\Omega)$ au sens faible [2] de (1.4.40), c.à.d. qui vérifie :

$$\left\{\begin{array}{l} u \in H_0^1(\Omega), \\ \displaystyle\int_\Omega \nabla u(x)\nabla v(x)dx = \int_\Omega f(x)v(x)dx, \forall v \in H_0^1(\Omega). \end{array}\right. \tag{1.4.41}$$

On peut montrer (voir cours Equations aux dérivées partielles) que si $u \in C^2(\Omega)$, alors u est solution de (1.4.40) si et seulement si u est solution faible de (1.4.40). Pour discrétiser le problème, on se donne un certain nombre de points, alignés dans les directions x et y, comme représentés sur la figure 1.2 (on prend un pas de maillage uniforme et égal à h). Certains de ces points sont à l'intérieur du domaine Ω, d'autres sont situés sur la frontière $\partial\Omega$.
Comme en une dimension d'espace, les inconnues discrètes sont associées aux noeuds du maillage. On note $\{P_i, i \in I\}$ les points de discrétisation, et on écrit l'équation aux dérivées partielles en ces points :

$$-\Delta u(P_i) - \frac{\partial^2 u}{\partial x^2}(P_i) - \frac{\partial^2 u}{\partial y^2}(P_i) = f(P_i).$$

1er cas :
Dans le cas de points points "vraiment intérieurs", tel que le point P_1 sur la figure 1.2, *i.e.* dont tous les points voisins sont situés à l'intérieur de Ω, les quotients différentiels

$$\frac{2u(P_1) - u(P_2) - u(P_3)}{h^2} \text{ et } \frac{2u(P_1) - u(P_5) - u(P_4)}{h^2}$$

sont des approximations consistantes à l'ordre 2 de $-\partial_1^2 u(P_1)$ et $-\partial_2^2 u(P_1)$.
Par contre, pour un point "proche" du bord tel que le point $\tilde{P}_1$, les mêmes approximations (avec les points $\tilde{P}_2$, $\tilde{P}_3$, $\tilde{P}_4$ et $\tilde{P}_5$) ne seront que d'ordre 1 en raison des différences de distance entre les points (faire les développements de Taylor pour s'en convaincre.
Une telle discrétisation amène à un système linéaire $A_h U_h = b_h$, où la structure de A_h (en particulier sa "largeur de bande", c.à.d. le nombre de diagonales non nulles) dépend de la numérotation des noeuds. On peut montrer que la matrice A_h est monotone et le schéma est stable. De la consistance et la stabilité, on déduit, comme en une dimension d'espace, la convergence du schéma.

2. Par définition, $H^2(\Omega)$ est l'ensemble des fonctions de $L^2(\Omega)$ qui admet des dérivées faibles jusqu'à l'ordre 2 dans $L^2(\Omega)$.

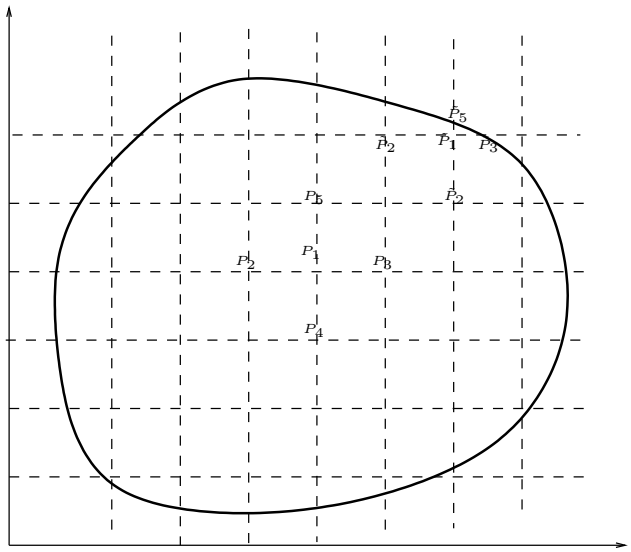

FIG. 1.2 – *Discrétisation différences finies bi-dimensionnelle*

1.4.2 Volumes finis

Le problème modèle

On considère le problème modèle suivant (par exemple de conduction de la chaleur) :

$$-\mathrm{div}(\lambda_i \nabla u(x)) = f(x) \qquad x \in \Omega_i,\, i = 1{,}2 \tag{1.4.42}$$

où $\lambda_1 > 0$, $\lambda_2 > 0$ sont les conductivités thermiques dans les domaines Ω_1 et avec Ω_2, avec $\Omega_1 =]0{,}1[\times]0{,}1[$ et $\Omega_2 =]0{,}1[\times]1{,}2[$. On appelle $\Gamma_1 =]0{,}1[\times\{0\}$, $\Gamma_2 = \{1\}\times]0{,}2[$, $\Gamma_3 =]0{,}1[\times\{2\}$, et $\Gamma_4 = \{0\}\times]0{,}2[$ les frontières extérieures de Ω, et on note $I =]0{,}1[\times\{1\}$ l'interface entre Ω_1 et Ω_2 (voir Figure 1.3). Dans la suite, on notera λ la conductivité thermique sur Ω, avec $\lambda|_{\Omega_i} = \lambda_i$, $i = 1{,}2$.
On va considérer plusieurs types de conditions aux limites, en essayant d'expliquer leur sens physique. On rappelle que le flux de chaleur par diffusion est égal $\mathbf{q}$ est donné par la loi de Fourier: $\mathbf{q} = -\lambda \nabla u \cdot \mathbf{n}$, où $\mathbf{n}$ est le vecteur normal unitaire à la surface à travers laquelle on calcule le flux.

1. Conditions aux limites de type Fourier (Robin) sur $\Gamma_1 \cup \Gamma_3$: On suppose qu'il existe un transfert thermique entre les parois Γ_1 et Γ_3 et l'extérieur. Ce transfert est décrit par la condition de Fourier (Robin dans la littérature anglo-saxonne), qui exprime que le flux transféré est proportionnel à la différence de température entre l'extérieur et l'intérieur :

$$-\lambda \nabla u \cdot \mathbf{n}(x) = \alpha(u(x) - u_{ext}),, \forall x \in \Gamma_1 \cup \Gamma_3. \tag{1.4.43}$$

   où $\alpha > 0$ est le coefficient de transfert thermique, $\mathbf{n}$ le vecteur unitaire normal à $\partial\Omega$ extérieur à Ω, et u_{ext} est la température extérieure (donnée).
2. Conditions aux limites de type Neumann sur Γ_2 On suppose que la paroi Γ_2 est parfaitement isolée, et que le flux de chaleur à travers cette paroi est donc nul. Ceci se traduit par une condition dite "de Neumann homogène" :

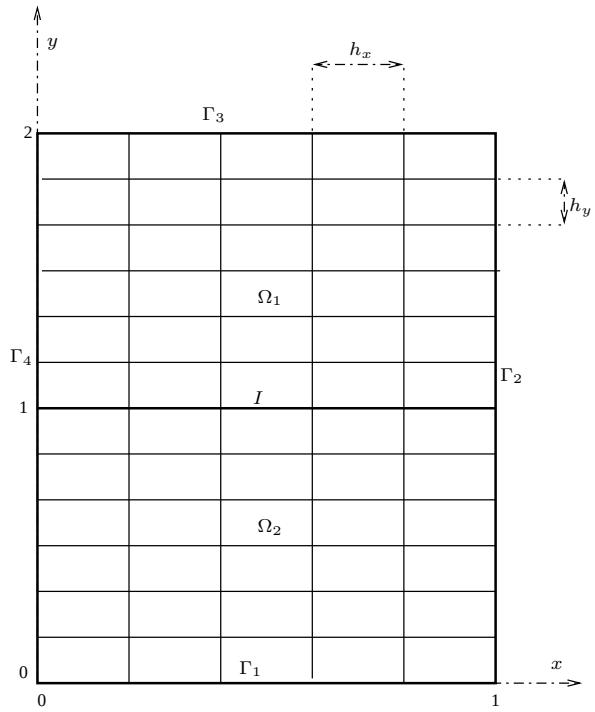

FIG. 1.3 – *Domaine d'étude*

$$-\lambda \nabla u \cdot \mathbf{n} = 0 \qquad \forall x \in \Gamma_2. \tag{1.4.44}$$

3. Conditions aux limites de type Dirichlet sur Γ_4 Sur la paroi Γ_4, on suppose que la température est fixée. Ceci est une condition assez difficile à obtenir expérimentalement pour un problème de type chaleur, mais qu'on peut rencontrer dans d'autres problèmes pratiques.

$$u(x) = g(x), \quad \forall\, x \in \Gamma_4. \tag{1.4.45}$$

4. Conditions sur l'interface I : On suppose que l'interface I est par exemple le siège d'une réaction chimique surfacique θ qui provoque un dégagement de chaleur surfacique. On a donc un saut du flux de chaleur au travers de l'interface I. Ceci se traduit par la condition de saut suivante :

$$-\lambda_1 \nabla u_1(x) \cdot \mathbf{n}_1 - \lambda_2 \nabla u_2(x) \cdot \mathbf{n}_2 = \theta(x), \quad x \in I. \tag{1.4.46}$$

où $\mathbf{n}_i$ désigne le vecteur unitaire normal à I et extérieur à Ω_i, et θ est une fonction donnée.

Discrétisation par volumes finis

On se donne un maillage "admissible" $\mathcal{T}$ de Ω

$$\bar{\Omega} = \bigcup_{K \in \mathcal{T}} \bar{K}.$$

Par "admissible", on entend un maillage tel qu'il existe des points $(x_K)_{K\in\mathcal{T}}$ situés dans les mailles, tels que chaque segment $x_K x_L$ soit orthogonal à l'arête $K|L$ séparant la maille K de la maille L, comme visible sur la figure 1.4.

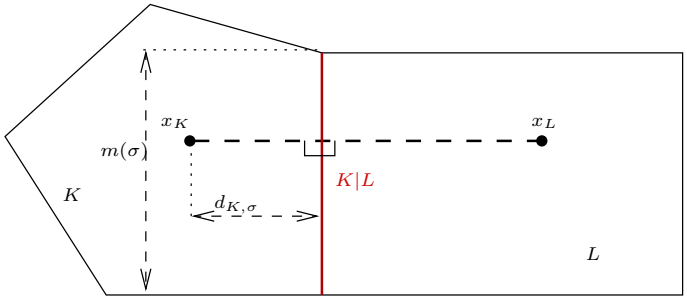

FIG. 1.4 – *Condition d'orthogonalité pour un maillage volumes finis*

Cette condition est nécessaire pour obtenir une approximation consistante du flux de diffusion (c'est-à-dire de la dérivée normale sur l'arête $K|L$), voir remarque 1.29. Dans le cas présent, le domaine representé sur la figure 1.3 étant rectangulaire, cette condition est particulièrement facile à vérifier en prenant un maillage rectangulaire. Par souci de simplicité, on prendra ce maillage uniforme, et on notera $h_x = 1/n$ le pas de discrétisation dans la direction x et $h_y = 1/p$ le pas de discrétisation dans la direction y. Le maillage est donc choisi de telle sorte que l'interface I coïncide avec un ensemble d'arêtes du maillage qu'on notera $\mathcal{E}_I$. On a donc

$$\bar{I} = \bigcup_{\sigma\in\mathcal{E}_I} \bar{\sigma},$$

où le signe $\bar{}$ désigne l'adhérence de l'ensemble. On se donne ensuite des inconnues discrètes $(u_K)_{K\in\mathcal{T}}$ associées aux mailles et $(u_\sigma)_{\sigma\in\mathcal{E}}$ associées aux arêtes.
Pour obtenir le schéma volumes finis, on commence par établir les bilans par maille en intégrant l'équation sur chaque maille K (notons que ceci est faisable en raison du fait que l'équation est sous forme conservative, c'est-à-dire sous la forme : $-\text{div}(\text{flux}) = f$). On obtient donc :

$$\int_K -\text{div}(\lambda_i \nabla u(x))dx = \int_K f(x)dx,$$

soit encore, par la formule de Stokes,

$$\int_{\partial K} -\lambda_i \nabla u(x).\mathbf{n}(x)d\gamma(x) = m(K)f_K,$$

où $\mathbf{n}$ est le vecteur unitaire normal à $\partial\Omega$, extérieur à Ω, et γ désigne le symbole d'intégration sur la frontière. On décompose ensuite le bord de chaque maille K en arêtes du maillage : $\partial K = \bigcup_{\sigma\in\mathcal{E}_K} \bar{\sigma}$ où $\mathcal{E}_K$ représente l'ensemble des arêtes de K. On obtient alors :

$$\sum_{\sigma\in\mathcal{E}_K} \int_\sigma -\lambda_i \nabla u.\mathbf{n}_{K,\sigma} d\gamma(x) = m(K)f_K$$

où $\mathbf{n}_{K,\sigma}$ est le vecteur unitaire normal à σ extérieur à K. On écrit alors une "équation approchée" :

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} = m(K) f_K,$$

où $F_{K,\sigma}$ est le flux numérique à travers σ, qui approche le flux exact $F^*_{K,\sigma} = \int_\sigma -\lambda_i \nabla u.\mathbf{n}_{K,\sigma} d\gamma(x)$. Pour obtenir le schéma numérique, il nous reste à exprimer le flux numérique $F_{K,\sigma}$ en fonction des inconnues discrètes $(u_K)_{K\in\mathcal{T}}$ associées aux mailles et $(u_\sigma)_{\sigma\in\mathcal{E}}$ associées aux arêtes (ces dernières seront ensuite éliminées) :

$$F_{K,\sigma} = -\lambda_i \frac{u_\sigma - u_K}{d_{K,\sigma}} m(\sigma), \tag{1.4.47}$$

où $d_{K,\sigma}$ est la distance du point x_K à l'arête σ et $m(\sigma)$ est la longueur de l'arête σ (voir Figure 1.4). L'équation associée à l'inconnue u_K est donc :

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} = m(K) f_K.$$

On a ainsi obtenu autant d'équations que de mailles. Il nous reste maintenant à écrire une équation pour chaque arête, afin d'obtenir autant d'équations que d'inconnues.
En ce qui concerne les arêtes intérieures, on écrit la conservativité du flux, ce qui nous permettra d'éliminer les inconnues associées aux arêtes internes. Soit $\sigma = K|L \subset \Omega_i$, On a alors :

$$F_{K,\sigma} = -F_{L,\sigma}. \tag{1.4.48}$$

On vérifiera par le calcul (cf. exercice 16 page 41) que, après élimination de u_σ, ceci donne

$$F_{K,\sigma} = -F_{L,\sigma} = \lambda_i \frac{m(\sigma)}{d_\sigma}(u_K - u_L), \tag{1.4.49}$$

où $d_\sigma = d(x_K,x_L)$.
Remarque 1.29 (Consistance du flux) *On appelle erreur de consistance associée au flux (1.4.47) l'expression:*

$$R_{K,\sigma} = -\frac{1}{m(\sigma)} \int_\sigma \nabla u(x)\cdot \mathbf{n}_{K,\sigma} d\gamma(x) - F^*_{K,\sigma}, \text{ où } F^*_{K,\sigma} = -\lambda_i \frac{u(x_\sigma) - u(x_K)}{d_{K,\sigma}} m(\sigma),$$

où x_σ est l'intersection de σ avec l'arête $K|L$, u la solution exacte.
On dit que le flux numérique donné par l'expression (1.4.47) est consistant si

$$\lim_{h(\mathcal{T})\to 0} \max_{K\in\mathcal{T},\sigma\in K} |R_{K,\sigma}| = 0,$$

où $h(\mathcal{T})$ est le pas du maillage, i.e. $h(\mathcal{T}) = \max_{K\in\mathcal{T}} diam(K)$, avec $diam(K) = \sup_{(x,y)\in K^2} d(x,y)$. On vérifie facilement que si u est suffisamment régulière et si le segment $x_K x_L$ est colinéaire au vecteur normal$\mathbf{n}$, alors le flux numérique est consistant. Cette propriété, alliée a la propirété de conservativité des flux, permet de démontrer la convergence du schéma, comme on l'a fait dans le cas unidimensionnel.

Remarque 1.30 (Cas du maillage cartésien de la figure 1.3) *Dans le cas du maillage carésien considéré pour notre problème, il est naturel de choisir les points x_K comme les centres de gravité des mailles. Comme le maillage est uniforme, on a donc $d_{K,\sigma} = \frac{h_x}{2}$ (resp. $\frac{h_y}{2}$) et $|\sigma| = h_y$ (resp. $|\sigma| = h_x$) pour une arête σ verticale (resp. horizontale).*

Ecrivons maintenant la discrétisation des conditions aux limites et interface :

1. Condition de Neumann sur Γ_2 Sur Γ_2, on a la condition de Neumann (1.4.44) : $\lambda_i \nabla u \cdot \mathbf{n} = 0$, qu'on discrétise par : $\sigma \in \mathcal{E}_K$ et $\sigma \subset \Gamma_2$,$F_{K,\sigma} = 0$.
2. Condition de Dirichlet sur Γ_4 La discrétisation de la condition de Dirichlet (1.4.45) peut s'effectuer de la manière suivante :
$$u_\sigma = \frac{1}{m(\sigma)} \int_\sigma g(y) d\gamma(y).$$
L'expression du flux numérique est alors :
$$F_{K,\sigma} = -\lambda_i \frac{u_\sigma - u_K}{d_{K,\sigma}} m(\sigma).$$
3. Condition de Fourier sur $\Gamma_1 \cup \Gamma_3$ Sur $\Gamma_1 \cup \Gamma_3$ on a la condition de Fourier (1.4.43) :
$$-\lambda_i \nabla u \cdot \mathbf{n} = \alpha(u(x) - u_{ext}) \qquad \forall x \in \Gamma_1 \cup \Gamma_3$$
qu'on discrétise par
$$F_{K,\sigma} = -m(\sigma)\lambda_i \frac{u_\sigma - u_K}{d_{K,\sigma}} = m(\sigma)\alpha(u_\sigma - u_{ext}) \text{ pour } \sigma \subset \Gamma_1 \cup \Gamma_3.$$
Après élimination de u_σ (cf. exercice 16 page 41), on obtient :
$$F_{K,\sigma} = \frac{\alpha\lambda_i m(\sigma)}{\lambda_i + \alpha d_{K,\sigma}}(u_K - u_{ext}). \tag{1.4.50}$$
4. Condition de saut pour le flux sur I Si $\sigma = K|L \in \mathcal{E}_I$, la discrétisation de la condition de saut (1.4.46) se discrétise facilement en écrivant :
$$F_{K,\sigma} + F_{L,\sigma} = \theta_\sigma, \text{ avec } \theta_\sigma = \frac{1}{|\sigma|} \int_\sigma \theta(x) d\gamma(x). \tag{1.4.51}$$
3.
Après élimination de l'inconnue u_σ (voir exercice 16 page 41), on obtient
$$F_{K,\sigma} = \frac{\lambda_1 m(\sigma)}{\lambda_1 d_{L,\sigma} + \lambda_2 d_{K,\sigma}} \left[\lambda_2(u_K - u_L) + d_{L,\sigma}\theta_\sigma\right]. \tag{1.4.52}$$

On a ainsi éliminé toutes les inconnues u_σ, ce qui permet d'obtenir un système linéaire dont les inconnues sont les valeurs $(u_K)_{K\in\mathcal{T}}$.

Remarque 1.31 (Implantation informatique de la méthode) *Lors de l'implantation informatique, la matrice du système linéaire est construite "par arête" (contrairement à une matrice éléments finis, dont nous verrons plus tard la construction "par élément"), c.à.d. que pour chaque arête, on additionne la contribution du flux au coefficient de la matrice correspondant à l'équation et à l'inconnue concernées.*

1.5 Exercices

Exercice 1 (Comparaison différences finies- volumes finis) *Suggestions en page 42, corrigé en page 44.*

On considère le problème :

$$\begin{array}{l} -u''(x) = f(x),\ x \in]0,1[, \\ u(0) = a, u(1) = b, \end{array} \tag{1.5.53}$$

Ecrire les schémas de différences finies et volumes finis avec pas constant pour le problème (1.5.53), et comparer les schémas ainsi obtenus.

Exercice 2 (Conditionnement "efficace".) *Suggestions en page 42, corrigé en page 44.*

Soit $f \in C([0,1])$. Soit $N \in \mathbb{N}^\star$, N impair. On pose $h = 1/(N+1)$. Soit A la matrice définie par (1.2.24) page 17, issue d'une discrétisation par différences finies (vue en cours) du problème (1.3.27) page 19.

Pour $u \in \mathbb{R}^N$, on note $u_1, \ldots, u_N$ les composantes de u. Pour $u \in \mathbb{R}^N$, on dit que $u \geq 0$ si $u_i \geq 0$ pour tout $i \in \{1, \ldots, N\}$. Pour $u, v \in \mathbb{R}^N$, on note $u \cdot v = \sum_{i=1}^N u_i v_i$.

On munit $\mathbb{R}^N$ de la norme suivante : pour $u \in \mathbb{R}^N$, $\|u\| = \max\{|u_i|,\ i \in \{1, \ldots, N\}\}$. On munit alors $\mathcal{M}_N(\mathbb{R})$ de la norme induite, également notée $\|\cdot\|$, c'est-à-dire $\|B\| = \max\{\|Bu\|, u \in \mathbb{R}^N \text{ t.q. } \|u\| = 1\}$, pour tout $B \in \mathcal{M}_N(\mathbb{R})$.

Partie I Conditionnement de la matrice et borne sur l'erreur relative

1. (Existence et positivité de A^{-1}) Soient $b \in \mathbb{R}^N$ et $u \in \mathbb{R}^N$ t.q. $Au = b$. Remarquer que $Au = b$ peut s'écrire :

$$\left\{ \begin{array}{l} \frac{1}{h^2}(u_i - u_{i-1}) + \frac{1}{h^2}(u_i - u_{i+1}) = b_i, \forall i \in \{1, \ldots, N\}, \\ u_0 = u_{N+1} = 0. \end{array} \right. \tag{1.5.54}$$

   Montrer que $b \geq 0 \Rightarrow u \geq 0$. [On pourra considérer $p \in \{0, \ldots, N+1\}$ t.q. $u_p = \min\{u_j,\ j \in \{0, \ldots, N+1\}$.]

   En déduire que A est inversible.
2. (Préliminaire...) On considère la fonction $\varphi \in C([0,1], \mathbb{R})$ définie par $\varphi(x) = (1/2)x(1-x)$ pour tout $x \in [0,1]$. On définit alors $\phi \in \mathbb{R}^N$ par $\phi_i = \phi(ih)$ pour tout $i \in \{1, \ldots, N\}$. Montrer que $(A\phi)_i = 1$ pour tout $i \in \{1, \ldots, N\}$.
3. (calcul de $\|A^{-1}\|$) Soient $b \in \mathbb{R}^N$ et $u \in \mathbb{R}^N$ t.q. $Au = b$. Montrer que $\|u\| \leq (1/8)\|b\|$ [Calculer $A(u \pm \|b\|\phi)$ avec ϕ défini à la question 2 et utiliser la question 1]. En déduire que $\|A^{-1}\| \leq 1/8$ puis montrer que $\|A^{-1}\| = 1/8$.
4. (calcul de $\|A\|$) Montrer que $\|A\| = \frac{4}{h^2}$.
5. (Conditionnement pour la norme $\|\cdot\|$). Calculer $\|A^{-1}\|\|A\|$. Soient $b, \delta_b \in \mathbb{R}^N$. Soient $u, \delta_u \in \mathbb{R}^N$ t.q. $Au = b$ et $A(u + \delta_u) = b + \delta_b$. Montrer que $\dfrac{\|\delta_u\|}{\|u\|} \leq \|A^{-1}\|\|A\| \dfrac{\|\delta_b\|}{\|b\|}$.

   Montrer qu'un choix convenable de b et δ_b donne l'égalité dans l'inégalité précédente.

Partie II Borne réaliste sur l'erreur relative: Conditionnement "efficace"

On se donne maintenant $f \in C([0,1], \mathbb{R})$ et on suppose (pour simplifier...) que $f(x) > 0$ pour tout $x \in]0,1[$. On prend alors, dans cette partie, $b_i = f(ih)$ pour tout $i \in \{1, \ldots, N\}$. On considère aussi le vecteur φ défini à la question 2 de la partie I.

1. Montrer que $h \sum_{i=1}^N b_i \varphi_i \to \int_0^1 f(x)\phi(x)dx$ quand $N \to \infty$ et que $\sum_{i=1}^N b_i \varphi_i > 0$ pour tout N. En déduire qu'il existe $\alpha > 0$, ne dépendant que de f, t.q. $h \sum_{i=1}^N b_i \varphi_i \geq \alpha$ pour tout $N \in \mathbb{N}^\star$.

2. Soit $u \in \mathbb{R}^N$ t.q. $Au = b$. Montrer que $N\|u\| \geq \sum_{i=1}^N u_i = u \cdot A\varphi \geq \frac{\alpha}{h}$ (avec α donné à la question 1).
   Soit $\delta_b \in \mathbb{R}^N$ et $\delta_u \in \mathbb{R}^N$ t.q. $A(u+\delta_u) = b + \delta_b$. Montrer que $\frac{\|\delta_u\|}{\|u\|} \leq \frac{\|f\|_{L^\infty(]0,1[)}}{8\alpha} \frac{\|\delta_b\|}{\|b\|}$.
3. Comparer $\|A^{-1}\|\|A\|$ (question I.5) et $\frac{\|f\|_{L^\infty(]0,1[)}}{8\alpha}$ (question II.2) quand N est "grand" (ou quand $N \to \infty$).

Exercice 3 (Conditionnement, réaction diffusion 1d.) *Corrigé en page 47.*

On s'intéresse au conditionnement pour la norme euclidienne de la matrice issue d'une discrétisation par Différences Finies du problème aux limites suivant :

$$\begin{array}{l} -u''(x) + u(x) = f(x),\ x \in]0,1[, \\ u(0) = u(1) = 0. \end{array} \tag{1.5.55}$$

Soit $N \in \mathbb{N}^\star$. On note $U = (u_j)_{j=1\ldots N}$ une "valeur approchée" de la solution u du problème (1.5.55) aux points $\left(\frac{j}{N+1}\right)_{j=1\ldots N}$. On rappelle que la discrétisation par différences finies de ce problème consiste à chercher U comme solution du système linéaire $AU = \left(f(\frac{j}{N+1})\right)_{j=1\ldots N}$ où la matrice $A \in M_N(\mathbb{R})$ est définie par $A = (N+1)^2 B + Id$, Id désigne la matrice identité et

$$B = \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \ldots & 0 & -1 & 2 \end{pmatrix}$$

1. (Valeurs propres de la matrice B.)
   On rappelle que le problème aux valeurs propres

   $$\begin{array}{l} -u''(x) = \lambda u(x),\ x \in]0,1[, \\ u(0) = u(1) = 0. \end{array} \tag{1.5.56}$$

   admet la famille $(\lambda_k, u_k)_{k \in \mathbb{N}^*}$, $\lambda_k = (k\pi)^2$ et $u_k(x) = \sin(k\pi x)$ comme solution. Montrer que les vecteurs $U_k = \left(u_k(\frac{j}{N+1})\right)_{j=1\ldots N}$ sont des vecteurs propres de la matrice B. En déduire toutes les valeurs propres de la matrice B.
2. En déduire les valeurs propres de la matrice A.
3. En déduire le conditionnement pour la norme euclidienne de la matrice A.

Exercice 4 (Erreur de consistance) *Suggestions en page 42, corrigé en page 47.*

On considère la discrétisation à pas constant par le schéma aux différences finies symétrique à trois points (vu en cours) du problème (1.3.27) page 19, avec $f \in C([0,1])$. Soit $N \in \mathbb{N}^\star$, N impair. On pose $h = 1/(N+1)$. On note u la solution exacte, $x_i = ih$, pour $i = 1, \ldots, N$ les points de discrétisation, et $(u_i)_{i=1,\ldots,N}$ la solution du système discrétisé.

1. Montrer que si f est constante, alors

   $$\max_{1 \leq i \leq N} |u_i - u(x_i)| = 0.$$

2. Soit N fixé, et $\max_{1\leq i\leq N} |u_i - u(x_i)| = 0$. A-t-on forcément que f est constante sur [0,1] ? (justifier la réponse.)

Exercice 5 (Principe du maximum) *Suggestions en page 42, corrigé en page 47*

On considère le problème :

$$\begin{cases} -u''(x) + c(x)u(x) = f(x), & 0 < x < 1, \\ u(0) = a, u(1) = b, \end{cases} \tag{1.5.57}$$

où $c \in C([0,1],\mathbb{R}_+)$, et $c \in C([0,1],\mathbb{R})$, et $(a,b) \in \mathbb{R}^2$.
1. Donner la discrétisation par différences finies de ce problème. On appelle U_h la solution approchée (c.à.d. $U_h = (u_1,\ldots,u_N)^t$, où u_i est l'inconnue discrète en x_i.
2. On suppose ici que $c = 0$. Montrer que $u_i \geq \min(a,b)$, pour tout $i = 1,\ldots,N$.

Exercice 6 (Positivité et principe du Maximum) *Corrigé en page 48.*

Soient $v \in C([0,1],\mathbb{R}_+)$ et $a_0,a_1 \in \mathbb{R}$.

1. On considère le problème suivant :

$$\begin{cases} -u_{xx}(x) + v(x)u_x(x) = 0, x \in]0,1[, \\ u(0) = a_0, u(1) = a_1. \end{cases} \tag{1.5.58}$$

On admettra qu'il existe une unique solution $u \in C([0,1],\mathbb{R}) \cap C^2(]0,1[,\mathbb{R})$ à ce problème. On cherche à approcher cette solution par une méthode de différences finies. On se donne un pas de maillage $h = \frac{1}{N+1}$ uniforme, des inconnues discrètes $u_1,\ldots,u_N$ censées approcher les valeurs $u(x_1),\ldots,u(x_N)$. On considère le schéma aux différences finies suivant:

$$\begin{cases} \dfrac{1}{h^2}(2u_i - u_{i+1} - u_{i-1}) + \dfrac{1}{h}v_i(u_i - u_{i-1}) = 0,\ i = 1,\ldots,N \\ u(0) = a_0, u(1) = a_1, \end{cases} \tag{1.5.59}$$

où $v_i = v(x_i)$, pour $i = 1,\ldots,N$.

1.1 Expliquez en quoi ce schéma est "décentré amont".

1.2 Montrer que le système (1.5.59) s'écrit sous la forme $MU = b$ avec $U = (u_1,\ldots,u_N)^t$, $b \in \mathbb{R}^N$, et M est une matrice telle que :

(a) $MU \geq 0 \Rightarrow U \geq 0$ (les inégalités s'entendent composante par composante),
(b) M est inversible,
(c) Si U est solution de $MU = b$ alors $\min(a_0,a_1) \leq u_i \leq \max(a_0,a_1)$.

1.3 Montrer que M est une M–matrice, c. à.d. que M vérifie :

(a) $m_{i,i} > 0$ pour $i = 1,\ldots,n$;
(b) $m_{i,j} \leq 0$ pour $i,j = 1,\ldots,n$, $i \neq j$;
(c) M est inversible ;
(d) $M^{-1} \geq 0$;

2. On suppose maintenant que $v \in C([0,1],\mathbb{R}_+) \cap C^1(]0,1[,\mathbb{R})$, et on considère le problème :

$$\begin{cases} -u_{xx}(x) + (vu)_x(x) = 0, x \in]0,1[, \\ u(0) = a_0, u(1) = a_1. \end{cases} \tag{1.5.60}$$

On admettra qu'il existe une unique solution $u \in C([0,1],\mathbb{R}) \cap C^2(]0,1[,\mathbb{R})$ à ce problème. On cherche ici encore à approcher cette solution par une méthode de différences finies. On se donne un pas de maillage $h = \frac{1}{N+1}$ uniforme, des inconnues discrètes $u_1,\ldots,u_N$ censées approcher les valeurs $u(x_1),\ldots,u(x_N)$. On considère le schéma aux différences finies suivant:

$$\begin{cases} \dfrac{2u_i - u_{i+1} - u_{i-1}}{h^2} + \dfrac{1}{h}(v_{i+\frac{1}{2}}u_i - v_{i-\frac{1}{2}}u_{i-1}) = 0,\ i = 1,\ldots,N \\ u(0) = a_0, u(1) = a_1, \end{cases} \tag{1.5.61}$$

où $v_{i+\frac{1}{2}} = v(\frac{x_i+x_{i+1}}{2})$, pour $i = 0,\ldots,N$.

2.1 Expliquez en quoi ce schéma est "décentré amont".

2.2 Montrer que le système (1.5.61) s'écrit sous la forme $MU = b$ avec $U = (u_1,\ldots,u_N)^t$, , $b \in \mathbb{R}^N$,

2.3 Pour $U = (u_1,\ldots,u_N)^t$ et $W = (w_1,\ldots,w_N)^t \in \mathbb{R}^N$, calculer $MU \cdot W$, et en déduire l'expression de $(M^tW)_i$, pour $i = 1,\ldots,N$ (on distinguera les cas $i = 2,\ldots,N-1$, $i = 1$ et $i = N$.

2.4 Soit $W \in \mathbb{R}^N$;

2.4. (a) montrer que si $M^tW \geq 0$ alors $W \geq 0$; en déduire que si $U \in \mathbb{R}^N$ est tel que $MU \geq 0$ alors $U \geq 0$.

2.4. (b) en déduire que si $U \in \mathbb{R}^N$ est tel que $MU \geq 0$ alors $U \geq 0$.

2.5 Montrer que M est une M-matrice.

2.6 Montrer que U solution de (1.5.61) peut ne pas vérifier $\min(a_0,a_1) \leq u_i \leq \max(a_0,a_1)$.

Exercice 7 (Problème elliptique 1d, discrétisation par différences finies) [3] *Suggestions en page 42, corrigé en page 48.*

Soit $f \in C^2([0,1])$. On s'intéresse au problème suivant:

$$\begin{array}{l} -u_{xx}(x) + \frac{1}{1+x}u_x(x) = f(x),\ x \in]0,1[, \\ u(0) = a\ u(1) = b. \end{array} \tag{1.5.62}$$

On admet que ce problème admet une et une seule solution u et on suppose que $u \in C^4(]0,1[)$. On cherche une solution approchée de (3.4.41) par la méthode des différences finies. Soit $n \in \mathbb{N}^*$, et $h = \frac{1}{N+1}$. On note u_i la valeur approchée recherchée de u au point ih, pour $i = 0,\cdots,N+1$.
On utilise les approximations centrées les plus simples de u_x et u_{xx} aux points ih,$i = 1,\cdots,n$ On pose $u_h = (u_1,\cdots,u_n)^t$.

1. Montrer que u_h est solution d'un système linéaire de la forme $A_hu_h = b_h$; donner A_h et b_h.

2. Montrer que le schéma numérique obtenu est consistant et donner une majoration de l'erreur de consistance (on rappelle que l'on a supposé $u \in C^4$).

3. Soit $v \in \mathbb{R}^n$, montrer que $A_hv \geq 0 \Rightarrow v \geq 0$ (ceci s'entend composante par composante). Cette propriété s'appelle conservation de la positivité. En déduire que A_h est monotone.

4. On définit θ par

$$\theta(x) = -\frac{1}{2}(1+x)^2 ln(1+x) + \frac{2}{3}(x^2+2x)ln2,\ x \in [0,1].$$

4.a. Montrer qu'il existe $C \geq 0$, indépendante de h, t.q.

3. Cet exercice est tiré du livre *Exercices d'analyse numérique matricielle et d'optimisation*, de P.G. Ciarlet et J.M. Thomas, Collection Mathématiques pour la maîtrise, Masson, 1982

$$\max_{1\leq i\leq n} |\frac{1}{h^2}(-\theta_{i-1}+2\theta_i-\theta_{i+1})+\frac{1}{2h(1+ih)}(\theta_{i+1}-\theta_{i-1})-1| \leq Ch^2,$$

avec $\theta_i = \theta(x_i)$, $i = 0,\ldots,n+1$.

4.b On pose $\theta_h = (\theta_1,\cdots,)^t$. Montrer que $(A_h\theta_h)_i \geq 1 - Ch^2$, pour $i = 1,\ldots,N$.

4.c Montrer qu'il existe $M \geq 0$ ne dépendant pas de h t.q. $\|A_h^{-1}\|_\infty \leq M$.

5. Montrer la convergence, en un sens à définir, de u_h vers u.

6. Que peut on dire si $u \notin C^4$, mais seulement $u \in C^2$ ou C^3 ?

7. On remplace dans (3.4.41) $\frac{1}{1+x}$ par $\alpha u_x(x)$, avec α donné (par exemple $\alpha = 100$). On utilise pour approcher (3.4.41) le même principe que précédemment (approximations centrées de u_x et u_{xx}. Que peut on dire sur la consistance, la stabilité, la convergence du schéma numérique?

Exercice 8 (Non consistance des volumes finis) *Suggestions en page 43, corrigé en page 52*

Montrer que la discrétisation de l'opérateur $-u''$ par le schéma volumes finis n'est pas toujours consistante au sens des différences finies, *i.e.* que l'erreur de consistance définie par (voir remarque 1.21 page 20)

$$R_i = \frac{1}{h_i}\left[\frac{-1}{h_{i+1/2}}(u(x_{i+1})-u(x_i))+\frac{1}{h_{i-1/2}}(u(x_i)-u(x_{i-1}))\right] - u''(x_i)$$

ne tend pas toujours vers 0 lorsque h tend vers 0.

Exercice 9 (Consistance des flux) *Corrigé en page 53 Corrigé en page 53*

Montrer que le flux défini par (1.3.30) est consistant d'ordre 1 dans le cas général, et qu'il est d'ordre 2 si $x_{i+1/2} = (x_{i+1} + x_i)/2$.

Exercice 10 (Conditions aux limites de Neumann) *Suggestions en page 43, corrigé en page 53*

On considère ici l'équation le problème de diffusion réaction avec conditions aux limites de Neumann homogènes (correspondant à une condition physique de flux nul sur le bord) :

$$\left\{\begin{array}{l} -u''(x)+cu(x) = f(x),\ x \in]0,1[, \\ u'(0) = u'(1) = 0, \end{array}\right. \tag{1.5.63}$$

avec $c \in \mathbb{R}_+^*$, et $f \in C([0,1])$. Donner la discrétisation de ce problème par

1. différences finies,
2. volumes finis

Montrer que les matrices obtenues ne sont pas inversibles. Proposer une maniére de faire en sorte que le problème soit bien posé, compatible avec ce qu'on connaît du problème continu.

Exercice 11 (Conditions aux limites de Fourier (ou Robin)) *Suggestions en page 43, corrigé en page 54*

On considère le problème :

$$\left\{\begin{array}{l} -u''(x)+cu(x) = f(x),\ x \in]0,1[, \\ u'(0)-\alpha(u-\tilde{u}) = 0, \\ u'(1)+\alpha(u-\tilde{u}) = 0, \end{array}\right. \tag{1.5.64}$$

avec $c \in \mathbb{R}_+$, $f \in C([0,1])$, $\alpha \in \mathbb{R}_+^*$, et $\tilde{u} \in \mathbb{R}$.

Donner la discrétisation de ce problème par

1. différences finies,

2. volumes finis

Dans les deux cas, écrire le schéma sous la forme d'un système linéaire de N équations à N inconnues, en explicitant matrice et second membre (N est le nombre de noeuds internes en différences finies, de mailles en volumes finis).

Exercice 12 (Problème elliptique 1d, discrétisation par volumes finis) *Suggestions en page 43, corrigé en page 55*

Soient $a,b \geq 0$, $c,d \in \mathbb{R}$ et $f \in C([0,1],\mathbb{R})$; on cherche à approcher la solution u du problème suivant :

$$-u_{xx}(x) + au_x(x) + b(u(x) - f(x)) = 0,\ x \in [0,1], \tag{1.5.65}$$

$$u(0) = c,\ u(1) = d. \tag{1.5.66}$$

On suppose (mais il n'est pas interdit d'expliquer pourquoi...) que (1.5.65)-(1.5.66) admet une solution unique $u \in C^2([0,1],\mathbb{R})$.

Soient $N \in \mathbb{N}^\star$ et $h_1,\ldots,h_N > 0$ t.q. $\sum_{i=1}^N h_i = 1$. On pose $x_{\frac{1}{2}} = 0$, $x_{i+\frac{1}{2}} = x_{i-\frac{1}{2}} + h_i$, pour $i = 1,\ldots,N$ (de sorte que $x_{N+\frac{1}{2}} = 1$), $h_{i+\frac{1}{2}} = \frac{h_{i+1}+h_i}{2}$, pour $i = 1,\ldots,N-1$, et $f_i = \frac{1}{h_i}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)dx$, pour $i = 1,\ldots,N$.

Pour approcher la solution u de (1.5.65)-(1.5.66), on propose le schéma numérique suivant :

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} + bh_iu_i = bh_if_i,\ i \in \{1,\ldots,N\}, \tag{1.5.67}$$

avec $(F_{i+\frac{1}{2}})_{i\in\{0,\ldots,N\}}$ donné par les expressions suivantes :

$$F_{i+\frac{1}{2}} = -\frac{u_{i+1}-u_i}{h_{i+\frac{1}{2}}} + au_i,\ = i \in \{1,\ldots,N-1\}, \tag{1.5.68}$$

$$F_{\frac{1}{2}} = -\frac{u_1 - c}{\frac{h_1}{2}} + ac,\ F_{N+\frac{1}{2}} = -\frac{d-u_N}{\frac{h_N}{2}} + au_N. \tag{1.5.69}$$

En tenant compte des expressions (1.5.68) et (1.5.69), le schéma numérique (1.5.67) donne donc un système de N équations à N inconnues (les inconnues sont $u_1,\ldots,u_N$).

1. Expliquer comment, à partir de (1.5.65) et (1.5.66), on obtient ce schéma numérique.
2. (Existence de la solution approchée.)
   (a) On suppose ici que $c = d = 0$ et $f_i = 0$ pour tout $i \in \{1,\ldots,N\}$. Montrer qu'il existe un unique vecteur $U = (u_1,\ldots,u_N)^t \in \mathbb{R}^N$ solution de (1.5.67). Ce vecteur est obtenu en prenant $u_i = 0$, pour tout $i \in \{1,\ldots,N\}$. (On rappelle que dans (1.5.67) les termes $F_{i+\frac{1}{2}}$ et $F_{i-\frac{1}{2}}$ sont donnés par (1.5.68) et (1.5.69).)
   (b) On revient maintenant au cas général (c'est à dire $c,d \in \mathbb{R}$ et $f \in C([0,1],\mathbb{R})$. Montrer qu'il existe un unique vecteur $U = (u_1,\ldots,u_N)^t \in \mathbb{R}^N$ solution de (1.5.67). (On rappelle, encore une fois, que dans (1.5.67) les termes $F_{i+\frac{1}{2}}$ et $F_{i-\frac{1}{2}}$ sont donnés par (1.5.68) et (1.5.69).)

Soient $\alpha,\beta > 0$. On suppose, dans tout la suite de l'exercice, qu'il existe $h > 0$ tel que $\alpha h \leq h_i \leq \beta h$, pour tout $i \in \{1,\ldots,N\}$. On note $\overline{u}_i = \frac{1}{h_i}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x)dx$, pour $i = 1,\ldots,N$. (On rappelle que u est la solution exacte de (1.5.65)-(1.5.66).)

3. (Non consistance du schéma au sens des différences finies)
   (a) Montrer que le système peut se mettre sous la forme $AU = B$, où B est définie par

$$B_1 = bf_1 + \frac{2c}{h_1^2} + \frac{ac}{h_1},$$

$$B_i = bf_i,\ i = 2,...,N-1,$$

$$B_N = bf_n + \frac{2d}{h_N^2}.$$

   (b) On pose $\bar{R} = A\bar{U} - B$ avec $\bar{U} = (\bar{u}_1, \ldots, \bar{u}_N)^t$. Vérifier que pour tout $i \in \{1,...,N\}$, $\bar{R}_i$ peut se mettre sous la forme :

$$\bar{R}_i = \bar{R}_i^1 + \bar{R}_i^2$$

   où $\sup_{i=1,..,N} \mid \bar{R}_i^1 \mid \leq C_1$ et $\sup_{i=1,..,N} \mid \bar{R}_i^2 \mid \leq C_2 h$.

   (c) On se restreint dans cette question au cas où $a = 0$, $b > 0$, $f = 0$, $c = 1$, $d = e^{\sqrt{b}}$, $N = 2q$, $h_i = h$ si i est pair et $h_i = \frac{h}{2}$ si i est impair, avec $h = \frac{2}{3N}$.
   Montrer que $\|\bar{R}\|_\infty$ ne tend pas vers 0 avec h.

4. (Consistance des flux.) En choisissant convenablement $(\bar{F}_{i+\frac{1}{2}})_{i\in\{0,...,N\}}$, montrer que :

$$\bar{F}_{i+\frac{1}{2}} - \bar{F}_{i-\frac{1}{2}} + bh_i\overline{u}_i = bh_i f_i,\ i \in \{1, \ldots, N\}, \tag{1.5.70}$$

   et que $(\bar{F}_{i+\frac{1}{2}})_{i\in\{0,...,N\}}$ vérifie les égalités suivantes :

$$\bar{F}_{i+\frac{1}{2}} = -\frac{\overline{u}_{i+1} - \overline{u}_i}{h_{i+\frac{1}{2}}} + a\overline{u}_i + R_{i+\frac{1}{2}},\ i \in \{1, \ldots, N-1\}, \tag{1.5.71}$$

$$\bar{F}_{\frac{1}{2}} = -\frac{\overline{u}_1 - c}{\frac{h_1}{2}} + ac + R_{\frac{1}{2}},\ \bar{F}_{N+\frac{1}{2}} = -\frac{d - \overline{u}_N}{\frac{h_N}{2}} + au_N + R_{N+\frac{1}{2}}, \tag{1.5.72}$$

   avec,

$$|R_{i+\frac{1}{2}}| \leq C_1 h,\ i \in \{0, \ldots, N\}, \tag{1.5.73}$$

   où $C_1 \in \mathbb{R}$, et C_1 ne dépend que de α,β, et u.

5. (Estimation d'erreur.) On pose $e_i = \overline{u}_i - u_i$, pour $i \in \{1, \ldots, N\}$ et $E = (e_1, \ldots, e_N)^t$.
   (a) Montrer que E est solution du système (de N équations) suivant :

$$G_{i+\frac{1}{2}} - G_{i-\frac{1}{2}} + bh_i e_i = 0,\ i \in \{1, \ldots, N\}, \tag{1.5.74}$$

   avec $(G_{i+\frac{1}{2}})_{i\in\{0,...,N\}}$ donné par les expressions suivantes :

$$G_{i+\frac{1}{2}} = -\frac{e_{i+1} - e_i}{h_{i+\frac{1}{2}}} + ae_i + R_{i+\frac{1}{2}},\ i \in \{1, \ldots, N-1\}, \tag{1.5.75}$$

$$G_{\frac{1}{2}} = -\frac{e_1}{\frac{h_1}{2}} + R_{\frac{1}{2}},\ G_{N+\frac{1}{2}} = -\frac{-e_N}{\frac{h_N}{2}} + ae_N + R_{N+\frac{1}{2}}, \tag{1.5.76}$$

(b) En multipliant (1.5.74) par e_i et en sommant sur $i = 1,\ldots,N$, montrer qu'il existe $C_2 \in \mathbb{R}$, ne dépendant que de α,β, et u tel que :

$$\sum_{i=0}^{N}(e_{i+1} - e_i)^2 \leq C_2 h^3, \tag{1.5.77}$$

avec $e_0 = e_{N+1} = 0$.

(c) Montrer qu'il existe $C_3 \in \mathbb{R}$, ne dépendant que de α,β, et u tel que :

$$|e_i| \leq C_3 h, \text{ pour tout } i \in \{1,\ldots,N\}. \tag{1.5.78}$$

6. (Principe du maximum.) On suppose, dans cette question, que $f(x) \leq d \leq c$, pour tout $x \in [0,1]$. Montrer que $u_i \leq c$, pour tout $i \in \{1,\ldots,N\}$. (On peut aussi montrer que $u(x) \leq c$, pour tout $x \in [0,1]$.)
7. On remplace, dans cette question, (1.5.68) et (1.5.69) par :

$$F_{i+\frac{1}{2}} = -\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + au_{i+1},\ i \in \{1,\ldots,N-1\}, \tag{1.5.79}$$

$$F_{\frac{1}{2}} = -\frac{u_1 - c}{\frac{h_1}{2}} + au_1,\ F_{N+\frac{1}{2}} = -\frac{d - u_N}{\frac{h_N}{2}} + ad. \tag{1.5.80}$$

Analyser brièvement le nouveau schéma obtenu (existence de la solution approchée, consistance des flux, estimation d'erreur, principe du maximum).

Exercice 13 (Convergence de la norme H^1 discrète)

Montrer que si $u_{\mathcal{T}}$:$]0,1[\longrightarrow \mathbb{R}$ est définie par $u_{\mathcal{T}}(x) = u_i \quad \forall x \in K_i$ où $(u_i)_{i=1,\ldots,N}$ solution de (1.3.29)–(1.3.31), alors $|u_{\mathcal{T}}|_{1,\mathcal{T}}$ converge dans $L^2(]0,1[)$ lorsque h tend vers 0, vers $\|Du\|_{L^2(]0,1[)}$, où u est la solution de (1.3.27).

Exercice 14 (Discrétisation 2D par différences finies)

Ecrire le système linéaire obtenu lorsqu'on discrétise le problème

$$\begin{cases} -\Delta u = f \text{ dans } \Omega =]0,1[\times]0,1[, \\ u = 0 \text{ sur } \partial\Omega. \end{cases} \tag{1.5.81}$$

par différences finis avec un pas uniforme $h = 1/N$ dans les deux directions d'espace. Montrer l'existence et l'unicité de la solution du système linéaire obtenu.

Exercice 15 (Problème elliptique 2d, discrétisation par DF) *Corrigé en page 1.7 page 57*

Soit $\Omega =]0,1[^2 \subset \mathbb{R}^2$. On se propose d'étudier deux schémas numériques pour le problème suivant :

$$\left\{ \begin{array}{ll} -\Delta u(x,y) + k\dfrac{\partial u}{\partial x}(x,y) = f(x,y), & (x,y) \in \Omega, \\ u = 0, & \text{sur } \partial\Omega, \end{array} \right. \tag{1.5.82}$$

où $k > 0$ est un réel donné et $f \in C(\bar{\Omega})$ est donnée. On note u la solution exacte de (1.5.82) et on suppose que $u \in C^4(\bar{\Omega})$.

1. (Principe du maximum)
   Montrer que pour tout $\varphi \in C^1(\bar{\Omega})$ t.q. $\varphi = 0$ sur $\partial\Omega$, on a :

   $$\int_\Omega \nabla u(x) \cdot \nabla \varphi(x)\ dx + \int_\Omega k\frac{\partial u}{\partial x}(x)\varphi(x)\ dx = \int_\Omega f(x)\varphi(x)\ dx.$$

   En déduire que si $f \leq 0$ sur $\bar{\Omega}$, on a alors $u \leq 0$ sur $\bar{\Omega}$.
   Soit $N \in \mathbb{N}$, on pose $h = \frac{1}{N+1}$, et $u_{i,j}$ est la valeur approchée recherchée de $u(ih,jh)$, $(i,j) \in \{0,...,N+1\}^2$. On pose $f_{i,j} = f(ih,jh)$, pour tout $(i,j) \in \{1,...,N\}^2$. On s'intéresse à deux schémas de la forme :

   $$\left\{ \begin{array}{l} a_0 u_{i,j} - a_1 u_{i-1,j} - a_2 u_{i+1,j} - a_3 u_{i,j-1} - a_4 u_{i,j+1} = f_{i,j},\ \forall (i,j) \in \{1,...,N\}^2, \\ u_{i,j} = 0,\ (i,j) \in \gamma, \end{array} \right. \tag{1.5.83}$$

   où a_0,a_1,a_2,a_3,a_4 sont données (ce sont des fonctions données de h) et $\gamma = \{(i,j),\ (ih,jh) \in \partial\Omega\}$ (γ dépend aussi de h). Le premier schéma, schéma [I], correspond au choix suivant des a_i :

   $$a_0 = \frac{4}{h^2},\ a_1 = \frac{1}{h^2} + \frac{k}{2h},\ a_2 = \frac{1}{h^2} - \frac{k}{2h},\ a_3 = a_4 = \frac{1}{h^2}.$$

   Le deuxième schéma, schéma [II], correspond au choix suivant des a_i :

   $$a_0 = \frac{4}{h^2} + \frac{k}{h},\ a_1 = \frac{1}{h^2} + \frac{k}{h},\ a_2 = a_3 = a_4 = \frac{1}{h^2}.$$

2. (Consistance)
   Donner une majoration de l'erreur de consistance en fonction de k, h et des dérivées de u, pour les schémas [I] et [II]. Donner l'ordre des schémas [I] et [II].
3. (Principe du maximum discret)
   Dans le cas du schéma [II] montrer que si $(w_{i,j})$ vérifie :

   $$a_0 w_{i,j} - a_1 w_{i-1,j} - a_2 w_{i+1,j} - a_3 w_{i,j-1} - a_4 w_{i,j+1} \leq 0,\ \forall (i,j) \in \{1,...,N\}^2,$$

   on a alors

   $$w_{i,j} \leq \max_{(n,m)\in\gamma} (w_{n,m}),\ \forall (i,j) \in \{1,...,N\}^2.$$

   Montrer que ceci est aussi vrai dans le cas du schéma [I] si h vérifie une condition à déterminer.

4. (Stabilité)
   Montrer que le schéma [II] et le schéma [I] sous la condition trouvée en 3. sont stables (au sens $||U||_\infty \leq C||f||_\infty$, avec une constante C à déterminer explicitement, où $U = \{u_{i,j}\}_{(i,j)\in\{0,\ldots,N+1\}^2}$ est solution de (1.5.83). [On pourra utiliser la fonction $\phi(x,y) = \frac{1}{2}y^2$].
   En déduire que dans le cas du schéma [II] et du schéma [I] sous la condition trouvée en 3. le problème (1.5.83) admet, pour tout f, une et une seule solution.
5. (Convergence)
   Les schémas [I] et [II] sont-ils convergents ? (au sens $\max_{(i,j)\in\{0,\ldots,N+1\}^2}(|u_{i,j} - u(ih,jh)|) \to 0$ quand $h \to 0$). Quel est l'ordre de convergence de chacun des schémas?
6. (Commentaires)
   Quels sont, à votre avis, les avantages respectifs des schémas [I] et [II] ?

Exercice 16 (Elimination des inconnues d'arêtes.) *Suggestions en page 43, corrigé en page 62*

On se place ici dans le cadre des hypothèses et notations du paragraphe 1.4.2 page 27

1. Pour chaque arête interne $\sigma = K|L$, calculer la valeur u_σ en fonction de u_K et u_L et en déduire que les flux numériques $F_{K,\sigma}$ et $F_{L,\sigma}$ vérifient bien (1.4.49)

2. Pour chaque arête $\sigma \subset \Gamma_1 \cup \Gamma_3$, telle que $\sigma \in \mathcal{E}_K$, calculer u_σ en fonction de u_K et montrer que $F_{K,\sigma}$ vérifie bien (1.4.50)

3. Pour chaque arête $\sigma \in \mathcal{E}_I$, avec $\sigma = K|L \qquad K \in \Omega_1$, calculer la valeur u_σ en fonction de u_K et u_L et en déduire que les flux numériques $F_{K,\sigma}$ et $F_{L,\sigma}$ vérifient bien (1.4.52)

4. Ecrire le système linéaire que satisfont les inconnues $(u_K)_{K\in\mathcal{T}}$.

Exercice 17 (Implantation de la méthode des volumes finis.)

On considère le problème de conduction du courant électrique

$$-div(\mu_i \nabla\phi(x)) = 0 \qquad x \in \Omega_i,\ i = 1,2 \tag{1.5.84}$$

où ϕ représente le potentiel électrique, $j = -\mu\nabla\phi(x)$ est donc le courant électrique, $\mu_1 > 0$, $\mu_2 > 0$ sont les conductivités thermiques dans les domaines Ω_1 et avec Ω_2, avec $\Omega_1 =]0,1[\times]0,1[$ et $\Omega_2 =]0,1[\times]1,2[$. On appelle $\Gamma_1 =]0,1[\times\{0\}$, $\Gamma_2 = \{1\}\times]0,2[$, $\Gamma_3 =]0,1[\times\{2\}$, et $\Gamma_4 = \{0\}\times]0,2[$ les frontières extérieures de Ω, et on note $I =]0,1[\times\{0\}$ l'interface entre Ω_1 et Ω_2 (voir Figure 1.3). Dans la suite, on notera μ la conductivité électrique sur Ω, avec $\mu|_{\Omega_i} = \mu_i$, $i = 1,2$.
On suppose que les frontières Γ_2 et Γ_4 sont parfaitement isolées. Le potentiel électrique étant défini à une constante près, on impose que sa moyenne soit nulle sur le domaine, pour que le problème soit bien posé. La conservation du courant électrique impose que

$$\int_{\Gamma_1} j\cdot\mathbf{n} + \int_{\Gamma_3} j\cdot\mathbf{n} = 0,$$

où $\mathbf{n}$ désigne le vecteur unitaire normal à la frontière $\partial\Omega$ et extérieure à Ω.
Enfin, on suppose que l'interface I est le siège d'une réaction électrochimique qui induit un saut de potentiel. On a donc pour tout point de l'interface I:

$$\phi_2(x) - \phi_1(x) = \psi(x),\ \forall x \in I,$$

où ϕ_i désigne la restriction de ϕ au sous domaine i. La fonction ϕ est donc discontinue sur l'interface I. Notons que, par contre, le courant électrique est conservé et on a donc

$$(-\mu\nabla\phi\cdot\mathbf{n})|_2(x)+(-\mu\nabla\phi\cdot\mathbf{n})|_1(x)=0,\ \forall x\in I.$$

1. Ecrire le problème complet, avec conditions aux limites.
2. Discrétiser le problème par la méthode des volumes finis, avec un maillage rectangulaire uniforme, (considérer deux inconnues discrètes pour chaque arête de l'interface) et écrire le système linéaire obtenu sur les inconnues discrètes.

1.6 Suggestions pour les exercices

Exercice 1 page 32 (Comparaison différences finies- volumes finis)

On rappelle que le schéma différences finies s'obtient en écrivant l'équation en chaque point de discrétisation, et en approchant les dérivées par des quotients différentiels, alors que le schéma volumes finis s'obtient en intégrant l'équation sur chaque maille et en approchant les flux par des quotients différentiels.

Exercice 2 page 32 (Conditionnement efficace)

Partie 1
1. Pour montrer que A est inversible, utiliser le théorème du rang.

2. Utiliser le fait que φ est un polynôme de degré 2.

3. Pour montrer que $\|A^{-1}\|=\dfrac{1}{8}$, remarquer que le maximum de φ est atteint en $x=.5$, qui correspond à un point de discrétisation car N est impair.

Partie 2 Conditionnement efficace
1. Utiliser la convergence uniforme. 2. Utiliser le fait que $A\phi=(1\dots1)^t$.

Exercice 4 page 33 (Erreur de consistance)

1. Utiliser l'erreur de consistance.
2. Trouver un contre-exemple.

Exercice 5 page 34 (Principe du maximum)

2. Poser $u_0=a$, $u_{N+1}=b$. Considérer $p=\min\{i=0,\cdots,N+1\ ;\ u_p=\min_{j=0,\cdots,N+1}u_j\}$. Montrer que $p=0$ ou $N+1$.

Exercice 7 page 35 (Différences finies pour un problème elliptique)

Questions 1 à 3 : application directe des méthodes de démonstration vues en cours (paragraphe 1.2 page 13).
Question 4 : La fonction θ est introduite pour montrer une majoration de $\|A^{-1}\|$, puisqu'on a plus $A\Phi=1$, où $\Phi_i=\varphi(x_i)$ est la et φ est la fonction "miracle" dans le cas $-u''=f$. Une fois qu'on a montré les

bonnes propriétés de la fonction θ (questions 4.a et 4.b), on raisonne comme dans le cours pour la question 4.c (voir démonstration de la proposition 1.14 page 17.

Exercice 8 (Non consistance des volumes finis)

Prendre f constante et égale à 1 et prendre $h_i = h/2$ pour i pair et $h_i = h$ pour i impair.

Exercice 10 page 36 (Conditions aux limites de Neumann)

1. En différences finies, écrire les équations internes de manière habituelle, et éliminez les inconnues qui apparaissent au bord u_0 et u_{N+1} en discrétisant convenablement les conditions aux limites. En volumes finis, c'est encore plus simples (flux nul au bord...)

2. Remarquer les constantes sont solutions du problème continu, et chercher alors par exemple une solution à moyenne nulle.

Exercice 11 page 36 (Conditions aux limites de Fourier (ou Robin) et Neumann)

Ecrire les équations internes de manière habituelle, et éliminez les inconnues qui apparaissent au bord u_0 et u_{N+1} en discrétisant convenablement les conditions aux limites.

Exercice 12 page 37 (Volumes finis 1D)

1. Pour justifier le schéma: écrire les bilans par maille, et approchez les flux par des quotients différentiels de manière consistante.

2 (a) On pourra, par exemple, multiplier (1.5.67) par u_i et sommer pour $i = 1, \ldots, N$, puis conclure en remarquant, en particulier, que $\sum_{i=1}^{N}(u_i - u_{i-1})u_i = \frac{1}{2}\sum_{i=1}^{N+1}(u_i - u_{i-1})^2$, avec $u_0 = u_{N+1} = 0$.

2 (b) Pensez au miracle de la dimension finie...

4. Effectuer les développements de Taylor

5. (b) (c) Se débarasser des termes de convection en remarquant qu'ils ont "le bon signe", et s'inspirer de la démonstration du théorème 1.25 page 22.

Exercice 14 (Discrétisation 2D par différences finies)

Adapter le cas unidimensionnel, en faisant attention aux conditions limites. Pour montrer l'existence et unicité, calculer le noyau de la matrice.

Exercice 16 (Elimination des inconnues d'arêtes)

1. Ecrire la conservativité du flux : $F_{K,\sigma} = -F_{L,\sigma}$ et en déduire la valeur de u_σ.

2. Trouver la valeur de u_σ qui vérifie

$$-m(\sigma)\lambda_i \frac{u_\sigma - u_K}{d_{K,\sigma}} = m(\sigma)\alpha(u_\sigma - u_{ext}).$$

3. Remplacer $F_{K,\sigma}$ et $F_{L,\sigma}$ par leurs expressions dans (1.4.51) et en déduire la valeur de u_σ.

4. Adopter l'ordre lexicographique pour la numérotation des mailles, puis établir l'équation de chaque maille, en commenąnt par les mailles interne.

1.7 Corrigés des exercices

Corrigé de l'exercice 1 page 32

Le schéma différences finies pour l'équation (1.5.53) s'écrit :

$$\begin{cases} \frac{1}{h^2}\left(2u_i - u_{i-1} - u_{i+1}\right) = f_i, \quad i = 1, \ldots, N, \\ \\ u_0 = a, \quad u_{N+1} = b. \end{cases}$$

Le schéma volumes finis pour la même équation s'écrit :

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} = hf_i, \quad i = 1, \ldots, N \tag{1.7.85}$$

avec $F_{i+\frac{1}{2}} = -\dfrac{u_{i+1} - u_i}{h}, \quad i = 1, \ldots, N-1$ et $F_{\frac{1}{2}} = -\dfrac{u_1 - a}{\frac{h}{2}}$.

$$F_{N+\frac{1}{2}} = -\frac{b - u_N}{\frac{h}{2}}$$

En remplaçant les expressions des flux dans l'équation (1.7.85). On obtient :

$$\frac{1}{h^2}\left(2u_i - u_{i+1} - u_{i-1}\right) = f_i, \quad i = 2, \ldots, N-1$$

$$\frac{1}{h^2}\left(3u_1 - 2u_2 - a\right) = 2f_1$$

$$\frac{1}{h^2}\left(3u_N - 2u_{N-1} - b\right) = 2f_N,$$

La différence entre les deux schémas réside dans la première et dernière équations.

Corrigé de l'exercice 2 page 32 (Conditionnement "efficace")

Partie I

1. Soit $u = (u_1 \ldots u_N)^t$. On a

$$Au = b \Leftrightarrow \begin{cases} \dfrac{1}{h^2}(u_i - u_{i-1}) + \dfrac{1}{h^2}(u_i - u_{i+1}) = b_i, \ \forall i = 1, \ldots N, \\ u_0 = u_{N+1} = 0. \end{cases}$$

Supposons $b_i \geq 0$, $\forall i = 1, \ldots, N$, et soit $p \in \{0, \ldots, N+1\}$ tel que $u_p = \min(u_i, \ i = 0, \ldots, N+1)$.
Si $p = 0$ ou $N+1$, alors $u_i \geq 0 \ \forall i = 0, N+1$ et donc $u \geq 0$.

Si $p \in \{1,\ldots,N\}$, alors

$$\frac{1}{h^2}(u_p - u_{p-1}) + \frac{1}{h^2}(u_p - u_{p+1}) \geq 0$$

et comme $u_p - u_{p-1} < 0$ et $u_p - u_{p+1} \leq 0$, on aboutit à une contradiction.

Montrons maintenant que A est inversible. On vient de montrer que si $Au \geq 0$ alors $u \geq 0$. On en déduit par linéarité que si $Au \leq 0$ alors $u \leq 0$, et donc que si $Au = 0$ alors $u = 0$. Ceci démontre que l'application linéaire représentée par la matrice A est injective donc bijective (car on est en dimension finie).

2. Soit $\varphi \in C([0,1],\mathbb{R})$ tel que $\varphi(x) = 1/2x(1-x)$ et $\phi_i = \varphi(x_i), i = 1,N$, où $x_i = ih$.
$(A\phi)_i$ est le développement de Taylor à l'ordre 2 de $\varphi''(x_i)$, et comme φ est un polynôme de degré 2, ce développement est exact. Donc $(A\phi)_i = \varphi''(x_i) = 1$.

3Soient $b \in \mathbb{R}^N$ et $u \in \mathbb{R}^N$ tels que $Au = b$. On a :

$$(A(u \pm \|b\|\varphi))_i = (Au)_i \pm \|b\|(A\phi)_i = b_i \pm \|b\|.$$

Prenons d'abord $\tilde{b}_i = b_i + \|b\| \geq 0$, alors par la question (1),

$$u_i + \|b\|\phi_i \geq 0 \quad \forall i = 1,\ldots,N.$$

Si maintenant on prend $\bar{b}_i = b_i - \|b\| \leq 0$, alors

$$u_i - \|b\|\phi_i \leq 0 \quad \forall i = 1,\ldots,N.$$

On a donc $-\|b\|\phi_i \leq \|b\|\phi_i$.
On en déduit que $\|u\|_\infty \leq \|b\| \ \|\phi\|_\infty$; or $\|\phi\|_\infty = \frac{1}{8}$. D'où $\|u\|_\infty \leq \frac{1}{8}\|b\|$.
On peut alors écrire que pour tout $b \in \mathbb{R}^N$,

$$\|A^{-1}b\|_\infty \leq \frac{1}{8}\|b\|, \text{ donc } \frac{\|A^{-1}b\|_\infty}{\|b\|_\infty} \leq \frac{1}{8}, \text{ d'où } \|A^{-1}\| \leq \frac{1}{8}.$$

On montre que $\|A^{-1}\| = \frac{1}{8}$ en prenant le vecteur b défini par $b(x_i) = 1$, $\forall i = 1,\ldots,N$. On a en effet $A^{-1}b = \phi$, et comme N est impair, $\exists i \in \{1,\ldots,N\}$ tel que $x_i = \frac{1}{2}$;or $\|\varphi\|_\infty = \varphi(\frac{1}{2}) = \frac{1}{8}$.

4. Par définition, on a $\|A\| = \sup_{\|x\|_\infty = 1} \|Ax\|$, et donc $\|A\| = \max_{i=1,N} \sum_{j=1,N} |a_{i,j}|$, d'où le résultat.

5. Grâce aux questions 3 et 4, on a, par définition du conditionnement pour la norme $\|\cdot\|$, $\text{cond}(A) = \|A\|\|A^{-1}\| = \frac{1}{2h^2}$.

Comme $A\delta_u = \delta_b$, on a:

$$\|\delta_u\| \leq \|A^{-1}\|\delta_b\|\frac{\|b\|}{\|b\|} \leq \|A^{-1}\|\delta_b\|\frac{\|A\|\|u\|}{\|b\|},$$

d'où le résultat.

Pour obtenir l'égalité, il suffit de prendre $b = Au$ où u est tel que $\|u\| = 1$ et $\|Au\| = \|A\|$, et δ_b tel que $\|\delta_b\| = 1$ et $\|A^{-1}\delta_b\| = \|A^{-1}\|$. On obtient alors

$$\frac{\|\delta_b\|}{\|b\|} = \frac{1}{\|A\|} \text{ et } \frac{\|\delta u\|}{\|u\|} = \|A^{-1}\|.$$

D'où l'égalité.

Partie 2 Conditionnement efficace

1. Soient $\varphi^{(h)}$ et $f^{(h)}$ les fonctions constantes par morceaux définies par

$$\begin{array}{ll} \varphi^h(x) & = \left\{ \begin{array}{l} \varphi(ih) = \phi_i \text{ si } x \in]x_i - \frac{h}{2}, x_i + \frac{h}{2}[, i = 1, \ldots, N, \\ 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in]1 - \frac{h}{2}, 1]. \end{array} \right. \quad \text{et} \\ f^{(h)}(x) & = \left\{ \begin{array}{l} f(ih) = b_i \text{ si } x \in]x_i - \frac{h}{2},\ x_i + \frac{h}{2}[, \\ f(ih) = 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in]1 - \frac{h}{2}, 1]. \end{array} \right. \end{array}$$

Comme $f \in C([0,1], \mathbb{R})$ et $\varphi \in C^2([0,1], \mathbb{R})$, la fonction f_h (resp. φ_h) converge uniformément vers f (resp. φ) lorsque $h \to 0$. On a donc

$$h \sum_{i=1}^{N} b_i \varphi_i = \int_0^1 f^{(h)}(x) \varphi^{(h)}(x) dx \to \int_0^1 f(x) \varphi(x) dx \text{ lorsque } h \to 0.$$

Comme $b_i > 0$ et $f_i > 0$ $\forall i = 1, \ldots, N$, on a évidemment

$$S_N = \sum_{i=1}^{N} b_i \varphi_i > 0 \text{ et } S_N \to \int_0^1 f(x) \varphi(x) dx = \beta > 0 \text{ lorsque } h \to 0.$$

Donc il existe $N_0 \in \mathbb{N}$ tel que si $N \geq N_0$, $S_N \geq \frac{\beta}{2}$, et donc $S_N \geq \alpha = \min(S_0, S_1 \ldots S_{N_0}, \frac{\beta}{2}) > 0$.

2. On a $N\|u\| = N \sup_{i=1,N} |u_i| \geq \sum_{i=1}^{N} u_i$. D'autre part, $A\varphi = (1 \ldots 1)^t$ donc $u \cdot A\varphi = \sum_{i=1}^{N} u_i$; or $u \cdot A\varphi = A^t u \cdot \varphi = Au \cdot \varphi$ car A est symétrique. Donc $u \cdot A\varphi = \sum_{i=1}^{N} b_i \varphi_i \geq \frac{\alpha}{h}$ d'après la question 1. Comme $\delta_u = A^{-1}\delta_b$, on a donc $\|\delta_u\| \leq \|A^{-1}\| \ \|\delta_b\|$; et comme $N\|u\| \geq \frac{\alpha}{h}$, on obtient : $\frac{\|\delta_u\|}{\|u\|} \leq \frac{1}{8} \frac{hN}{\alpha} \|\delta_b\| \frac{\|f\|_\infty}{\|b\|}$. Or $hN = 1$ et on a donc bien :

$$\frac{\|\delta_u\|}{\|u\|} \leq \frac{\|f\|_\infty}{8\alpha} \frac{\|\delta_b\|}{\|b\|}.$$

3. Le conditionnement $\text{cond}(A)$ calculé dans la partie 1 est d'ordre $1/h^2$, et donc tend vers l'infini lorsque le pas du maillage tend vers 0, alors qu'on vient de montrer dans la partie 2 que la variation relative $\frac{\|\delta_u\|}{\|u\|}$ est inférieure à une constante multipliée par la variation relative de $\frac{\|\delta_b\|}{\|b\|}$. Cette dernière information est nettement plus utile et réjouissante pour la résolution effective du système linéaire.

Corrigé de l'exercice 3 page 33 (Conditionnement, réaction diffusion 1d)

1. Pour $k = 1$ à N, calculons BU_k :

   $$(BU_k)_j = -sink\pi(j-1)h + 2\sin k\pi(jh) - \sin k\pi(j+1)h, \text{ où } h = \frac{1}{N+1}.$$

   En utilisant le fait que $\sin(a+b) = \sin a\cos b + \cos a \sin b$ pour développer $\sin k\pi(1-j)h$ et $\sin k\pi(j+1)h$, on obtient (après calculs) :

   $$(BU_k)_j = \lambda_k (U_k)_j, \quad j = 1, \ldots, N,$$

   où $\lambda_k = 2(1-\cos k\pi h) = 2(1-\cos\frac{k\pi}{N+1})$. On peut remarquer que pour $k = 1, \ldots, N$, les valeurs λ_k sont distinctes.
   On a donc trouvé les N valeurs propres $\lambda_1 \ldots \lambda_N$ de B associées aux vecteurs propres $U_1, \ldots, U_N$ de $\mathbb{R}^N$ tels que $(U_k)_j = \sin\frac{k\pi j}{N+1}$, $j = 1, \ldots, N$.
2. Comme $A = Id + \frac{1}{h^2}B$, les valeurs propres de la matrice A sont les valeurs $\mu_i = 1 + \frac{1}{h^2}\lambda_i$.
3. Comme A est symétrique, le conditionnement de A est donné par

   $$cond_2(A) = \frac{\mu_N}{\mu_1} = \frac{1 + \frac{2}{h^2}(1 - \cos\frac{N\pi}{N+1})}{1 + \frac{2}{h^2}(1 - \cos\frac{\pi}{N+1})}.$$

Corrigé de l'exercice 4 page 33 (Erreur de consistance)

1. Si f est constante, alors $-u''$ est constante, et donc les dérivées d'ordre supérieur de u sont nulles. Donc par l'estimation (1.2.19) page 16 sur l'erreur de consistance, on a $R_i = 0$ pour tout $i = 1, \ldots, N$. Si on appelle U le vecteur de composantes u_i et $\bar{U}$ le vecteur de $\mathbb{R}^N$ de composantes $u(x_i)$, on peut remarquer facilement que $U - \bar{U} = A^{-1}R$, où R est le vecteur de composantes R_i. On a donc $U - \bar{U} = 0$, c.q.f.d.
2. Il est facile de voir que f n'est pas forcément constante, en prenant $f(x) = \sin 2\pi x$, et $h = 1/2$, on n'a donc qu'une seule inconnue u_1 qui vérifie $u_1 = 0$, et on a également $u(1/2) = \sin\pi = 0$.

Corrigé de l'exercice 5 page 34

1. On se donne une discrétisation $(x_i)_{i=1,\ldots N}$ de [0,1], avec un pas constant h. On approche $u''(x_i)$ par le quotient différentiel $\frac{1}{h^2}(2u(x_i) - u(x_{i-1}) - u(x_{i+1}))$.
Le schéma résultant s'écrit donc :

$$\begin{cases} \frac{1}{h^2}(2u_i - u_{i-1} - u_{i+1}) + c_i u_i = f_i, \quad i = 1, \ldots, N. \\ u_0 = a, u_{N+1} = b. \end{cases}$$

2. Soit $p = \min\{i; u_i = \min_{j=0,N+1} u_j\}$.
Supposons que $1 < p < N$,, alors on a

$$\frac{1}{h^2}(u_p - u_{p-1}) + \frac{1}{h^2}(u_p - u_{p+1}) = f_p.$$

Or $u_p = \min u_j$, on a donc $u_p - u_{p-1} < 0$, puisque $p = \min\{i; u_i = \min_{j=0,\dots N+1} u_j\}$, et $u_p - u_{p+1} \leq 0$.Comme $f_p \geq 0$, on aboutit à une contradiction. On en déduit que $p = 0$ pu $N+1$, ce qui prouve que $u_i \geq \min(a,b)$.

Corrigé de l'exercice 6 page 34

Corrigé en cours de rédaction.

Corrigé de l'exercice 7 page 35

1. On se donne un pas constant $h = x_{i+1} - x_i = \frac{1}{N+1}$. Comme $u \in C^4(]0,1[)$, un développement de Taylor à l'ordre 4 aux points x_{i+1} et x_i donne :

$$\left\{ \begin{array}{l} u(x_{i+1}) = u(x_i) + hu'(x_i) + \dfrac{h^2}{2}\, u''(x_i) + \dfrac{h^3}{6}\, u^{(3)}(x_i) + \dfrac{h^4}{24}\, u^{(4)}(\eta_i) \\ u(x_{i-1}) = u(x_i) - hu'(x_i) + \dfrac{h^2}{2}\, u''(x_i) - \dfrac{h^3}{6}\, u^{(3)}(x_i) + \dfrac{h^4}{24}\, u^{(4)}(\zeta_i), \end{array} \right.$$

où η_i (resp. ζ_i) appartient à l'intervalle $[x_i,x_{i+1}]$ (resp. $[x_{i-1},x_i]$). En effectuant la somme et la différence de ces deux lignes, et on trouve :

$$\left\{ \begin{array}{l} -u''(x_i) = \dfrac{2u(x_i) - u(x_{i-1}) - u(x_{i+1})}{h^2} + \dfrac{h^2}{24}(u^{(4)}(\eta_i) + u^{(4)}(\zeta_i)) \\ u'(x_i) = \dfrac{u(x_{i+1}) - u(x_{i-1})}{2h} + \dfrac{h^3}{3}u^{(3)}(x_i) + \dfrac{h^2}{24}(u^{(4)}(\eta_i) - u^{(4)}(\zeta_i)). \end{array} \right. \tag{1.7.86}$$

On prend alors en compte les conditions limites du système de type Dirichlet. On introduit les valeurs $u_0 = a$ et $u_{N+1} = b$, et on obtient alors le schéma suivant :

$$\left\{ \begin{array}{l} \dfrac{2u_i - u_{i-1} - u_{i+1}}{h^2} + \left(\dfrac{1}{1+ih}\right)\, \dfrac{u_{i+1} - u_{i-1}}{2h} = f_i,\ i = 1,\dots,N, \\ u_0 = a, u_{N+1} = b. \end{array} \right.$$

On en déduit la forme matricielle du système $A_h u_h = b_h$

$$A_h = \begin{pmatrix} \frac{2}{h^2} & -\frac{1}{h^2} + \frac{1}{2h(1+h)} & 0 & \dots & 0 \\ -\frac{1}{h^2} + \frac{1}{2h(1+h)} & \frac{2}{h^2} & & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & \\ \vdots & & \ddots & \ddots & -\frac{1}{h^2} + \frac{1}{2h(1+(N-1)h)} \\ 0 & \dots & 0 & -\frac{1}{h^2} + \frac{1}{2h(1+Nh)} & \frac{2}{h^2} \end{pmatrix},$$

soit encore

$$A_h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \dots & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & & \\ \vdots & & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & & 0 & -1 & 2 \end{pmatrix} + \frac{1}{2h}\begin{pmatrix} 0 & \frac{1}{1+h} & 0 & \dots & \dots & 0 \\ -\frac{1}{1+2h} & \ddots & \frac{1}{1+2h} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & & \\ \vdots & & \ddots & \ddots & \ddots & \frac{1}{1+(N-1)h} \\ 0 & \dots & & 0 & \frac{1}{1+Nh} & 0 \end{pmatrix}$$

avec

$$b_h = \begin{pmatrix} f_1 + a\left(\frac{1}{h^2} + \frac{1}{2h(1+h)}\right) \\ f_2 \\ \vdots \\ \vdots \\ f_{N-1} \\ f_N + b\left(\frac{1}{h^2} - \frac{1}{2h(1+Nh)}\right) \end{pmatrix} \text{ et } u_h = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix}.$$

2. Majorons l'erreur de consistance :

$$R_1 = \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} - u''(x_i)$$
$$R_2 = \frac{u(x_{i+1}) - u(x_{i-1})}{2h} - u'(x_i)$$

La première égalité de (1.7.86) entraîne que :

$$\|R^1\|_\infty \leq \frac{h^2}{12} \sup_{[0;1]} |u^{(4)}|.$$

En effectuant alors les développements limités jusqu'à l'ordre 3, on obtient de la même manière que :

$$\|R^2\|_\infty \leq \frac{h^2}{6} \sup_{[0;1]} \|u^{(3)}\|$$

D'où on déduit finalement que :

$$\|R\|_\infty \leq \frac{h^2}{12}\left(\sup_{[0,1]} |u^{(4)}| + 2\sup_{[0,1]} |u^{(3)}|\right). \tag{1.7.87}$$

3. Par hypothèse on suppose $A_h v \geq 0$, et on va montrer que $v \geq 0$ pour $v \in \mathbb{R}^N$. L'hypothèse sur la matrice A_h nous permet d'écrire pour la i-ème ligne l'inéquation suivante:

$$\left(\frac{-1}{h^2} - \frac{1}{2h(1+ih)}\right) u_{i-1} + \frac{2}{h^2} u_i + \left(\frac{-1}{h^2} + \frac{1}{2h(1+ih)}\right) u_{i+1} \geq 0 \tag{1.7.88}$$

Soit

$$p = \min\left\{i \in \{1, \dots, N\}; v_p = \min_{j=1,\dots,N} v_j\right\}.$$

Supposons d'abord que $p = 1$. On a alors :

$$v_1 \leq v_j, \forall\ j = 1, \ldots, N.$$

Mais l'inéquation (1.7.88) pour $p = 1$ s'écrit encore :

$$\frac{1}{h^2}(v_1 - v_2) + \left(\frac{1}{h^2} + c_1\right) v_1 \geq 0,$$

On a donc

$$v_1 \geq \frac{1}{1 + c_1 h^2}(v_2 - v_1) \geq 0,$$

ce qui montre que $\min_{j=1,\ldots,N} v_j = v_1 \geq 0$ si $p = 1$. Un raisonnement similaire permet de montrer que si $p = N$, on a $v_n \geq 0$ et donc $\min\limits_{j=1,\ldots,N} v_j \geq 0$.
Supposons enfin que $p \in \{2, \ldots, N-1\}$ on a alors, par la deuxième inéquation de (1.2.17) :

$$\frac{1}{h^2}(v_p - v_{p-1}) + \frac{1}{h^2}(v_p - v_{p+1}) + \frac{1}{2h(1+ph)}(v_{p+1} - v_p + v_p - v_{p-1}) \geq 0.$$

On en déduit que

$$\left(\frac{1}{h^2} - \frac{1}{2h(1+ph)}\right)(v_p - v_{p-1}) + \left(\frac{1}{h^2} + \frac{1}{2h(1+ph)}\right)(v_p - v_{p+1}) \geq 0.$$

Or $h < 1$, donc $\frac{1}{h^2} - \frac{1}{2h(1+ph)} > 0$. Les deux termes du membre de gauche de l'équation ci–dessus sont donc négatifs ou nuls. On doit donc avoir : $v_{p-1} = v_p = v_{p+1}$ ce qui est impossible car p est le plus petit indice j tel que $v_j = \min_{i=1,\ldots,N} v_i$. Donc dans ce cas le minimum ne peut pas être atteint pour $j = p > 1$. On a ainsi finalement montré que $\min\limits_{i\in\{1,\ldots,N\}} v_i \geq 0$, on a donc $v \geq 0$.
La matrice A_h est monotone si et seulement si elle est inversible et que $A_h^{-1} \geq 0$. Pour vérifier que A_h est inversible il suffit de remarquer que si $A_h v = 0$, alors $A_h v \geq 0$ et $A_h(-v) \geq 0$, ce qui entraîne que $v = 0$, par conservation de la positivité. On en déduit que $Ker(A_h) = \{0\}$, et donc A_h est inversible. Pour montrer que A_h est monotone, il nous rester à prouver que $A_h^{-1} \geq 0$.
On se donne pour cela $b \in \mathbb{R}^n$ tel que $A_h b = e_i$, où e_i est le vecteur de la base canonique de $\mathbb{R}^n$, ainsi $e_i \geq 0$. D'apres la conservation de la positivité, on a $b \geq 0$. Cependant b est le ième colonne de A_h^{-1}, donc b est solution de $A_h^{-1} e_i = b$ tel que $b \geq 0$ donc $(A_h^{-1}) \geq 0$.

4.a.On calcule les dérivées successives de θ:

$$\begin{cases} \theta'(x) = -\frac{1}{2}(1+x) - (1+x)ln(1+x) + \frac{4}{3}(1+x)\ln 2, \\ \theta''(x) = -\frac{1}{2} - 1 - \ln(1+x) + \frac{4}{3}\ \ln 2, \\ \theta^{(3)}(x) = \frac{-1}{1+x}, \\ \theta^{(4)}(x) = \frac{1}{(1+x)^2}. \end{cases} \tag{1.7.89}$$

On vérifie alors que $\theta(x)$ est solution de l'équation :

$$-\theta_{xx}(x) + \frac{1}{1+x}\theta_x(x) = f(x),$$

avec $f(x) = \frac{1}{2} + 1 + \ln(1+x) - \frac{4}{3}\ln 2 - \frac{1}{2} - \ln(1+x) + \frac{4}{3}\ln 2 = 1$. De plus $\theta(0) = 0$ et $\theta(1) = 0$, donc $A_h \theta_h = b_h$. On a donc :

$$\max_{1\leq i\leq n}|\frac{1}{h^2}(-\theta_{i-1}+2\theta_i-\theta_{i+1})+\frac{1}{2h(1+ih)}(\theta_{i+1}-\theta_{i-1})-1|\leq\frac{h^2}{12}\left(\sup_{[0;1]}|\theta^{(4)}|+2\sup_{[0;1]}|\theta^{(3)}|\right)$$

et comme $\sup_{[0;1]}|\theta^{(4)}|=1$ et $\sup_{[0;1]}|\theta^{(3)}|=1$, on a

$$\max_{1\leq i\leq n}|\frac{1}{h^2}(-\theta_{i-1}+2\theta_i-\theta_{i+1})+\frac{1}{2h(1+ih)}(\theta_{i+1}-\theta_{i-1})-1|\leq\frac{h^2}{4}.$$

4.b On a d'aprés la question précedente que: $\max_{1\leq i\leq n}|(A_h\theta_h)-f_i|\leq\frac{h^2}{4}$, avec $f_i=1$, et donc :

$$\frac{h^2}{4}\leq(A_h\theta_h)_i-1\leq\frac{h^2}{4}$$

ce qui entraîne

$$\Rightarrow(A_h\theta_h)_i\geq 1-\frac{h^2}{4}.$$

4.c Par définition de la norme on a :

$$\|B\|_\infty=\sup_{v\neq 0}\frac{\|Bv\|}{\|v\|}=\sup_{i\in(1;.;N)}\sum_{j=1}^{N}|B_{i,j}|,$$

et comme A_h^{-1} est une matrice positive, on a :

$$\|A_h^{-1}\|_\infty=\sup_{i\in(1;.;N)}\sum_{j=1}^{N}(A_h^{-1})_{i,j}.$$

On se donne

$$v=(1-\frac{1}{4}h^2;...;1-\frac{1}{4}h^2),$$

et on note $d=A_h^{-1}v$ on a donc $A_hd=v$. D'aprés la question 4b on a :

$$(A_h\theta_h)_i\geq v_i,\ \forall i\in(1,\ldots,N)$$

ce qui peut encore s'écrire :

$$A_h(\theta_h-A_h^{-1}v)\geq 0.$$

Par conservation de la positivité (question 3), on en déduit que

$$\theta_h-A_h^{-1}v\geq 0,$$

soit encore

$$(\theta_h)_i\geq(1-\frac{1}{4}h^2)(A_h^{-1}e)_i\text{ avec }e=(1,\ldots,1)^t$$

Or $e\geq 0$, $A_h^{-1}e\geq 0$, et $1-\frac{1}{4}h^2>0$. On en déduit que $(A_h^{-1}e)_i\leq\frac{1}{1-\frac{1}{4}h^2}(\theta_h)_i$, soit encore :

$$\|A_h^{-1}e\|_\infty=\|A_h^{-1}\|_\infty\leq\frac{1}{1-\frac{1}{4}h^2}\|\theta_h\|_\infty.$$

La fonction θ est continue et bornée sur $[0;1]$; il existe donc K tel que $|\theta(x)| \leq K$. De plus $\frac{1}{1-\frac{1}{4}h^2} \leq \frac{4}{3}$. On en déduit que

$$\|A_h^{-1}\|_\infty \leq M, \tag{1.7.90}$$

avec $M = \frac{4}{3}K$.

5. Dans la question 2, on a montré que le schéma est consistant d'ordre 2 avec la majoration suivante :

$$\|R^h\|_\infty \leq \frac{h^2}{12}\left(\sup_{[0;1]}\|u^{(4)}\| + 2.\sup_{[0;1]}\|u^{(3)}\|\right)$$

Soit $\bar{u} = (u(x_1),\ldots,u(x_N))^t$ le vecteur dont les composantes sont les valeurs de la solution exacte aux points de discrétisation. Par définition, on a :

$$A_h u_h - A_h \bar{u} = b_h - (b_h + R^h),$$

et donc l'erreur de discrétisation vérifie :

$$e_h = u - u_h = A_h^{-1} R^h. \tag{1.7.91}$$

Pour montrer la convergence du schéma il suffit de montrer que e_h tend vers 0 avec h. Ceci se déduit de la stabilité du schéma ainsi que la consistance. De (1.7.91), on déduit :

$$\|e_h\|_\infty = \|A_h^{-1}R^h\|_\infty \leq \|A_h^{-1}\|_\infty \, \|R^h\|_\infty,$$

et donc, grâce à la stabilité (1.7.90) et la consistance (1.7.87), on obtient : $\|e_h\|_\infty \leq K\ h^2$, ce qui prouve la convergence.

Corrigé de l'exercice 8 page 36 (Non consistance des volumes finis)

Par développement de Taylor, pour $i = 1,\ldots,N$, il existe $\xi_i \in [x_i,x_{i+1}]$ tel que :

$$u(x_{i+1}) = u(x_i) + h_{i+\frac{1}{2}}u_x(x_i) + \frac{1}{2}h_{i+\frac{1}{2}}^2 u_{xx}(x_i) + \frac{1}{6}h_{i+\frac{1}{2}}^3 u_{xxx}(\xi_i),$$

et donc

$$R_i = -\frac{1}{h_i}\frac{h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}}}{2}u_{xx}(x_i) + u_{xx}(x_i) + \rho_i,\ \ i = 1,\ldots,N, \tag{1.7.92}$$

où $|\rho_i| \leq Ch$, C ne dépendant que de la dérivée troisième de u. Il est facile de voir que, en général, R_i, ne tend pas vers 0 lorsque h tend vers 0 (sauf dans des cas particuliers). En effet, prenons par exemple $f \equiv 1$, $h_i = h$ pour i pair, $h_i = h/2$ pour i impair, and $x_i = (x_{i+1/2} + x_{i-1/2})/2$, pour $i = 1,\ldots,N$. On a dans ce cas $u'' \equiv -1$, $u''' \equiv 0$, et donc:

$$R_i = -\frac{1}{4} \text{ si } i \text{ est pair, et } R_i = +\frac{1}{2} \text{ si } i \text{ est impair.}$$

On en conclut que $\sup\{|R_i|,i = 1,\ldots,N\} \not\to 0$ as $h \to 0$.

Corrigé de l'exercice 9 page 36

Par développement de Taylor, pour $i = 1, \ldots, N$, il existe $\xi_i \in x_{i+\frac{1}{2}}, x_{i+1}$ $\eta_i \in [x_i, x_{i+\frac{1}{2}}]$ tels que :

$$u(x_{i+1}) = u(x_{i+\frac{1}{2}}) + (x_{i+1} - x_{i+\frac{1}{2}})u'(x_{i+\frac{1}{2}}) + \frac{1}{2}(x_{i+1} - x_{i+\frac{1}{2}})^2 u''(x_{i+\frac{1}{2}}) + \frac{1}{6}(x_{i+1} - x_{i+\frac{1}{2}})^3 u'''(\eta_i),$$

$$u(x_i) = u(x_{i+\frac{1}{2}}) + (x_i - x_{i+\frac{1}{2}})u'(x_{i+\frac{1}{2}}) + \frac{1}{2}(x_i - x_{i+\frac{1}{2}})^2 u''(x_{i+\frac{1}{2}}) + \frac{1}{6}(x_i - x_{i+\frac{1}{2}})^3 u'''(\eta_i),$$

Par soustraction, on en déduit que :

$$u(x_{i+1}) - u(x_i) = (x_{i+1} - x_i)u'(x_{i+\frac{1}{2}}) + \frac{1}{2}(x_{i+1} - x_i)(x_{i+1} + x_i - 2x_{i+\frac{1}{2}})u''(x_{i+\frac{1}{2}}) + \tilde{\rho}_{i+\frac{1}{2}},$$

où

$$\tilde{\rho}_{i+\frac{1}{2}} = \frac{1}{6}((x_{i+1} - x_{i+\frac{1}{2}})^3 u'''(\xi_i) - (x_i - x_{i+\frac{1}{2}})^3 u'''(\eta_i))$$

et donc, en posant $F^*_{i+\frac{1}{2}} = -\dfrac{u(x_{i+1}) - u(x_i)}{h_{i+\frac{1}{2}}}$, on obtient après simplifications :

$$F^*_{i+\frac{1}{2}} = -u'(x_{i+\frac{1}{2}}) - \frac{1}{2}\left[x_{i+1} + x_i - 2x_{i+\frac{1}{2}}\right] u''(x_{i+\frac{1}{2}}) + \rho_{i+\frac{1}{2}}$$

avec $\left|\rho_{i+\frac{1}{2}}\right| \leq Ch^2$, où $h = \max\limits_{i=1,\ldots,N} h_i$ et C ne dépend que de u'''.

Dans le cas où $x_{i+\frac{1}{2}} = \frac{x_{i+1}+x_i}{2}$, on a donc $\left|F^*_{i+\frac{1}{2}} + u'(x_{i+\frac{1}{2}}\right| \leq \rho_{i+\frac{1}{2}}$, et donc le flux est consistant d'ordre 2.

Dans le cas général, on peut seulement majorer $\dfrac{1}{2}\left(x_{i+1} + x_i - 2x_{i+\frac{1}{2}}\right)$ par h, on a donc un flux consistant d'ordre 1.

Corrigé de l'exercice 10 page 36

1. On se donne une discrétisation $(x_i)_{i=1,\ldots,N}$ de l'intervalle [0,1], de pas constant et égal à h. On écrit l'équation en chaque point x_i, et on remplace $-u''(x_i)$ par le quotient différentiel habituel. En appelant $u_1, \ldots, u_N$ les inconnues localisées aux points $x_1, \ldots, x_N$, et u_0, u_{N+1} les inconnues auxiliaires localisées en $x = 0$ et $x = 1$, on obtient les équations discrètes associées aux inconnues $i = 1, \ldots, N$.

$$\frac{1}{h^2}(2u_i - u_{i-1} - u_{i+1}) + c_i u_i = f_i,$$

avec $c_i = c(x_i)$ et $f_i = f(x_i)$. Il reste à déterminer u_0 et u_{N+1}. Ceci se fait en approchant la dérivée $u'(0)$(resp.$u'(1)$) par $\dfrac{1}{h}\left(u(x_1) - u(0)\right)$ $\left(\text{ resp.}\dfrac{1}{h}(u(1) - u(x_N)\right)$.

Comme $u'(0) = 0$ et $u'(1) = 0$, on obtient donc que $u_0 = u_1$ et $u_{N+1} = u_N$. Le schéma différences finies s'écrit donc :

$$\begin{cases} \dfrac{1}{h^2}(u_1 - u_2) + c_1 u_1 = f_1 \\ \dfrac{1}{h^2}(2u_i - u_{i-1} - u_{i+1}) + c_i u_i = f_i, \quad i = 2, \ldots, N-1, \\ \dfrac{1}{h^2}(u_N - u_{N-1}) + c_N u_N = f_N. \end{cases}$$

2. On se donne un maillage volumes finis, et on intègre l'équation sur chaque maille, ce qui donne le schéma

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} = h_i f_i, i = 1, \ldots, N,$$

où $F_{i+\frac{1}{2}}$ est le flux numérique à l'interface $x_{i+\frac{1}{2}}$. Pour $i = 1, \ldots, N-1$, ce flux numérique est donné par

$$F_{i+\frac{1}{2}} = \frac{u_{i+1} - u_i}{h_{i\frac{1}{2}}}.$$

Pour $i = 0$ et $i = N+1$, on se sert des conditions de Neumann, qui imposent un flux nul. On écrit donc :

$$F_{\frac{1}{2}} = F_{N+\frac{1}{2}} = 0.$$

Corrigé de l'exercice 11 page 36

1. La discrétisation par différences finies donne comme i-ème équation (voir par exemple exercice 10 page 36 :

$$\frac{1}{h^2}(2u_i - u_{i-1} - u_{i+1}) + c_i u_i = f_i, i = 1, \ldots, N.$$

Il reste donc à déterminer les inconnues u_0 et u_{N+1} à l'aide de la discrétisation des conditions aux limites, qu'on approche par :

$$\frac{u_1 - u_0}{h} + \alpha(u_0 - \widetilde{u}) = 0,$$

$$\frac{u_{N+1} - u_N}{h} + \alpha(u_{N+1} - \widetilde{u}) = 0$$

où u_0 et u_{N+1} sont les valeurs approchées en x_0 et x_{N+1}, on a donc par élimination :

$$u_0 = \frac{1}{\alpha - \frac{1}{h}} \left(\alpha\widetilde{u} - \frac{u_1}{h} \right) \text{ et } u_{N+1} = \frac{1}{\alpha + \frac{1}{h}} \left(\alpha\widetilde{u} + \frac{u_N}{h} \right).$$

Ce qui termine la définition du schéma.
2. Par volumes finis, la discrétisation de l'équation s'écrit

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} = h_i f_i, i = 1, \ldots, N,$$

et les seuls flux "nouveaux" sont encore $F_{1/2}$ et $F_{N+\frac{1}{2}}$, qu'on obtient à partir de la discrétisation des conditions aux limites. Ceci peut se faire de plusieurs manières.
On peut, par exemple, discrétiser la condition aux limites en 0 par :

$$F_{1/2} + \alpha(u_0 - \widetilde{u}) = 0, \text{ avec } F_{1/2} = \frac{u_1 - u_0}{\frac{h_1}{2}}.$$

On a ddans ce cas : $-\alpha(u_0 - \widetilde{u}) \times \dfrac{h_1}{2} = -u_1 + u_0$, d'où on déduit que $u_0 = \dfrac{\alpha\widetilde{u}h_1 + 2u_1}{\alpha h_1 + 2}$, et qui conduit à l'expression suivante pour $F_{1/2}$:

$$F_{1/2} = \frac{\alpha}{\alpha h_1 + 2} \left(2(u_1 - \widetilde{u}) - \alpha h_1 \widetilde{u} \right).$$

Le calcul est semblable pour $F_{N+1/2}$

Corrigé de l'exercice 12 page 37

1. On intègre (1.5.65) sur une maille $[x_{i-1/2},x_{i+1/2}]$ et on obtient :

$$-u'(x_{i+1/2}) + u'(x_{i-1/2}) + a[u(x_{i+1/2}) - u(x_{i-1/2})] + b\int_{x_{i-1/2}}^{x_{i+1/2}} u(x)dx = bh_i f_i. \tag{1.7.93}$$

Pour justifier le schéma numérique proposé on remarque que :

$$u(x_{i+1}) = u(x_{i+1/2}) + (x_{i+1} - x_{i+1/2})u'(x_{i+1/2}) + \frac{1}{2}(x_{i+1} - x_{i+1/2})^2 u''(\xi_i), \text{ avec } \xi_i \in [x_{i+1/2},x_{i+1}],$$

et de même

$$u(x_i) = u(x_{i+1/2}) + (x_i - x_{i+1/2})u'(x_{i+1/2}) + \frac{1}{2}(x_i - x_{i+1/2})^2 u''(\gamma_i), \text{ avec } \gamma_i \in [x_{i-1/2},x_i],$$

dont on déduit :

$$u(x_{i+1}) - u(x_i) = h_{i+1/2}u'(x_{i+1/2}) + \frac{1}{8}(h_{i+1}^2 u''(\xi_i) - h_i^2 u''(\gamma_i)).$$

De plus en utilisant le fait que x_i est le milieu de $[x_{i-1/2},x_{i+1/2}]$ on a (voir démonstration plus loin)

$$\int_{x_{i-1/2}}^{x_{i+1/2}} u dx = u(x_i)h_i + \frac{1}{24}u''(\alpha_i)h_i^3 \tag{1.7.94}$$

D'où le schéma numérique.
Démontrons la formule (1.7.94). Pour cela il suffit (par changement de variable) de démontrer que si $u \in C^2(\mathbb{R})$, alors pour tout $\alpha \geq 0$, on a:

$$\int_{-\alpha}^{\alpha} u dx = 2\alpha u(0) + \frac{1}{3}u''(\alpha_i)\alpha^3. \tag{1.7.95}$$

Pour cela, on utilise une formule de type Taylor avec reste intégral, qu'on obtient en remarquant que si on pose $\varphi(t) = u(tx)$, alors $\varphi'(t) = xu(tx)$, et $\varphi''(t) = x^2u''(tx)$. Or $\varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)dt$, et par inégration par parties, on obtient donc :

$$\varphi(1) = \varphi(0) + \varphi'(0) + \int_0^1 \varphi''(t)(1-t)ds.$$

On en déduit alors que

$$u(x) = u(0) + xu'(0) + \int_0^1 x^2u''(tx)(1-t)dt.$$

En intégrant entre $-\alpha$ et α, on obtient alors:

$$\int_{-\alpha}^{\alpha} u(x)dx = 2\alpha u(0) + A, \text{ avec } A = \int_0^1 x^2u''(tx)(1-t)dt\ dx.$$

Comme la fonction u'' est continue elle est minorée et majorée sur $[-\alpha,\alpha]$. Soient donc $m = \min_{[-\alpha,\alpha]} u''$ et $M = \min_{[-\alpha,\alpha]} u''$. Ces deux valeurs sont atteintes par u'' puisqu'elle est continue. On a donc $u''([-\alpha,\alpha]) =$

$[m,M]$. De plus, la fonction $(x,t) \mapsto x^2(1-t)$ est positive ou nulle sur $[-\alpha,\alpha]\times[0,1]$. On peut donc minorer et majorer A de la manière suivante

$$m \int_0^1 x^2(1-t)dt\ dx \leq A \leq M \int_0^1 x^2(1-t)dt\ dx.$$

Or $\int_0^1 x^2(1-t)dt\ dx = \frac{1}{3}\alpha^3$. On en déduit que $\frac{1}{3}\alpha^3 m \leq A \leq \frac{1}{3}\alpha^3 M$, et donc que $A = \frac{1}{3}\alpha^3\gamma$, avec $\gamma \in [m,M]$; mais comme u'' est continue, elle prend toutes les valeurs entre m et M, il existe donc $\beta \in [-\alpha,\alpha]$ tel que $\gamma = u''(\beta)$, ce qui termine la preuve de (1.7.95).

2 (a). On multiplie (1.5.67) par u_i et on somme pour $i = 1,\ldots,N$. On obtient après changement d'indice que

$$\sum_{i=0}^{i=N} \frac{(u_{i+1}-u_i)^2}{h_{i+1/2}} + \frac{a}{2}\sum_{i=0}^{i=N}(u_{i+1}-u_i)^2 + b\sum_{i=0}^{i=N} u_i^2 h_i = 0.$$

Ce qui donne $u_i = 0$ pour tout $i = 1\ldots N$, d'où en mettant le schéma sous la forme matricielle $AU = B$ on déduit que l'application linéaire représentée par la matrice A est injective donc bijective (grâce au fait qu'on est en dimension finie) et donc que (1.5.67) admet une unique solution.

3 (a). Evident.
(b). On pose $\bar{R} = A\bar{U} - B$. On a donc $R_i = R_i^{(1)} + R_i^{(2)}$, avec

$$R_i^{(1)} = -\frac{1}{h_i}\left[\left(\frac{\bar{u}_{i+1}-\bar{u}_i}{h_{i+1/2}} - u'(x_{i+1/2})\right) - \left(\frac{\bar{u}_i-\bar{u}_{i-1}}{h_{i-1/2}} - u'(x_{i-1/2})\right)\right],$$

$$R_i^{(2)} = \frac{a}{h_i}[(\bar{u}_i - u(x_{i+1/2})) - (\bar{u}_{i-1} - u(x_{i+1/2}))].$$

De plus on remarque que

$$\bar{u}_i = \frac{1}{h_i}\int_{x_{i-1/2}}^{x_{i+1/2}} u dx = u(x_{i+1/2}) - \frac{1}{2}u'(x_{i+1/2})h_i + \frac{1}{6}u''(x_{i+1/2})h_i^2 - \frac{1}{24}u^{(3)}(d_i)h_i^3 \text{ avec } d_i \in [x_{i-1/2},x_{i+1/2}],$$

$$\bar{u}_{i+1} = \frac{1}{h_{i+1}}\int_{x_{i+1/2}}^{x_{i+3/2}} u dx = u(x_{i+1/2}) + \frac{1}{2}u'(x_{i+1/2})h_{i+1} + \frac{1}{6}u''(x_{i+1/2})h_{i+1}^2 - \frac{1}{24}u^{(3)}(\delta_i)h_{i+1}^3 \text{ avec } \delta_i \in [x_{i+1/2},x_{i+3/2}].$$

Ce qui implique que :

$$\frac{\bar{u}_{i+1}-\bar{u}_i}{h_{i+1/2}} = u'(x_{i+1/2}) + \frac{1}{3}u''(x_{i+1/2})(h_{i+1}-h_i) + \frac{1}{24}\ \frac{1}{h_{i+1/2}}\left[u^{(3)}(\delta_i)h_{i+1}^3 + u^{(3)}(d_i)h_i^3\right]$$

et donc

$$\frac{1}{h_i}\left[\frac{\bar{u}_{i+1}-\bar{u}_i}{h_{i+1/2}} - u'(x_{i+1/2})\right] = S_i + K_i,$$

avec

$$|S_i| = |\frac{1}{3}u''(x_{i+1/2})\left(\frac{h_{i+1}}{h_i} - 1\right) + \frac{1}{24}| \leq Ch \text{ et } |K_i| = |\frac{1}{h_i h_{i+1/2}}\left[u^{(3)}(\delta_i)h_{i+1}^3 + u^{(3)}(d_i)h_i^3\right]| \leq Ch,$$

où C ne dépend que de u. De plus si on pose : $L_i = \frac{1}{h_i}(\bar{u}_i - u(x_{i+1/2})$, par développement de Taylor, il existe $\tilde{C}$ ne dépendant que de u telle que $|L_i| \leq Ch$. Finalement on conclut que $|R_i^{(1)}| = |-S_i + S_{i+1}| \leq C_1$ et $|R_i^{(2)}| = | - K_i + K_{i-1} + a(L_i - L_{i-1})| \leq C_2 h$.
4. Reprendre les résultats précédents... Pour $|R_{i+1/2}| \leq Ch$ reprendre calcul du 3 $|R_{i+1/2}| = |h_i(-S_i - K_i + L_i)|$.
5 (a). On pose $e_i = \bar{u}_i - u_i$. Cette définition implique que e_i est solution du système (1.5.74)-(1.5.76).
(b). Un calcul similaire à celui de la question 2. donne que

$$b\sum_1^N h_i e_i + \sum_{i=0}^{i=N} \frac{(e_{i+1} - e_i)^2}{h_{i+1/2}} + \frac{a}{2}\sum_{i=0}^{i=N}(e_{i+1} - e_i)^2 = \sum_{i=0}^{i=N} R_{i+1/2}(e_{i+1} - e_i)$$

D'où en utilisant le fait que
$\alpha h \leq h_i \leq \beta h$ et l'inégalité de Cauchy-Schwarz on déduit que

$$\frac{1}{\beta h}\sum_{i=0}^{i=N}(e_{i+1} - e_i)^2 \leq \left(\sum_{i=0}^{i=N}(e_{i+1} - e_i)^2\right)^{1/2}\left(\sum_{i=0}^{i=N} R_{i+1/2}^2\right)^{1/2}$$

et en utilisant (1.5.73), et le fait que $\sum_{i=0}^{i=N} h_i = 1$ entraine $N \leq \frac{1}{\alpha h}$, on déduit :

$$\sum_{i=0}^{i=N}(e_{i+1} - e_i)^2 \leq C_1 \frac{\beta}{\alpha} h^3.$$

En remarquant que $e_i = \sum_{j=0}^{j=i-1}(e_{j+1} - e_j)$ on a pour tout $0 < i \leq N$ que

$$|e_i| \leq \left(\sum_{j=0}^{j=i}(e_{j+1} - e_j)^2\right)^{1/2} i^{1/2} \leq \left(C_1 \frac{\beta}{\alpha} h^3\right)^{1/2} N^{1/2}$$

et donc $|e_i| \leq \frac{\sqrt{C_1 \beta}}{\alpha} h$, pour tout $0 < i \leq N$.

Corrigé de l'exercice 15 page 40

On note (x,y) les coordonnées d'un point de $\mathbb{R}^2$.

1. En multipliant la première équation de (1.5.82) par φ et en intégrant par parties, on trouve, pour tout $\varphi \in C^1(\bar{\Omega})$ t.q. $\varphi = 0$ sur $\partial\Omega$:

$$\int_\Omega \nabla u(x,y) \cdot \nabla \varphi(x,y)\ dxdy + \int_\Omega k \frac{\partial u(x,y)}{\partial x}\varphi(x,y)\ dx = \int_\Omega f(x,y)\varphi(x,y)\ dxdy. \tag{1.7.96}$$

On suppose maintenant que $f \leq 0$ sur Ω. On se donne une fonction $\psi \in C^1(\mathbb{R},\mathbb{R})$ t.q. :

$$\psi(s) = 0, \quad \text{si } s \leq 0,$$
$$\psi(s) > 0, \quad \text{si } s > 0.$$

(On peut choisir, par exemple, $\psi(s) = s^2$ pour $s > 0$ et $\psi(s) = 0$ pour $s \leq 0$) et on prend dans (1.7.96) $\varphi = \psi \circ u$. On obtient ainsi :

$$\int_\Omega \psi'(u(x,y)) \left|\nabla u(x,y)\right|^2 \, dxdy + \int_\Omega k \frac{\partial u}{\partial x}(x,y)\psi(u(x,y))dx = \int_\Omega f(x,y)\psi(u(x,y)) \, dxdy \leq 0. \quad (1.7.97)$$

En notant G la primitive de ψ s'annulant en 0, on a : $\dfrac{\partial}{\partial x} G(u(x,y)) = \psi(u(x,y)) \dfrac{\partial u}{\partial x}(x,y)$. Comme $u = 0$ sur $\partial\Omega$, on obtient donc :

$$\int_\Omega k \frac{\partial u}{\partial x}(x,y)\psi(u(x,y)) \, dxdy = \int_\Omega k \frac{\partial}{\partial x} G(u(x,y)) \, dxdy = \int_{\partial\Omega} kG(u(x,y)) n_x \, d\gamma(x,y) = 0,$$

où n_x désigne la première composante du vecteur normal n à $\partial\Omega$ extérieur à Ω, et $d\gamma(x,y)$ le symbole d'intégration par rapport à la mesure de Lebesgue unidimensionnelle sur $\partial\Omega$. De (1.7.97) on déduit alors :

$$\int_\Omega \psi'(u(x,y)) \left|\nabla u(x,y)\right|^2 \, dxdy \leq 0,$$

et donc, comme $\psi' \geq 0$ et que la fonction $(x,y) \mapsto \psi'(u(x,y))|\nabla u(x,y)|^2$ est continue :

$$\psi'(u(x,y))|\nabla u(x,y)|^2 = 0, \forall (x,y) \in \bar{\Omega}$$

Ceci donne aussi

$$\nabla \psi(u(x,y)) = 0, \forall (x,y) \in \bar{\Omega}.$$

La fonction $\psi \circ u$ est donc constante sur $\bar{\Omega}$, comme elle est nulle sur $\partial\Omega$, elle est nulle sur $\bar{\Omega}$, ce qui donne

$$u \leq 0 \text{ sur } \bar{\Omega}$$

2. On s'intéresse ici à la consistance au sens des différences finies. On pose donc

$$\bar{u}_{i,j} = u(ih, jh) \text{ pour } i,j \in \{0, \ldots, N+1\}^2.$$

On a bien $\bar{u}_{i,j} = 0$ pour $(i,j) \in \gamma$, et pour $(i,j) \in \{1, \ldots, N\}^2$, on pose :

$$R_{ij} = a_0 \bar{u}_{i,j} - a_1 \bar{u}_{i-1,j} - a_2 \bar{u}_{i+1,j} - a_3 \bar{u}_{i,j-1} - a_4 \bar{u}_{i,j+1} - f_{i,j}.$$

On rappelle que u est solution de (2.5.84), R_j est donc l'erreur de consistance. Dans le cas du schéma [I] on a :

$$R_{ij} = \frac{2\bar{u}_{ij} - \bar{u}_{i+1,j} - \bar{u}_{i-1,j}}{h^2} + \frac{2\bar{u}_{ij} - \bar{u}_{i,j+1} - \bar{u}_{i,j-1}}{h^2} + k \frac{\bar{u}_{i+1,j} - \bar{u}_{i-1,j}}{2h} - f_{ij}.$$

Comme $u \in C^4(\bar{\Omega})$, il existe $\xi_{ij} \in]0,1[$ et $\eta_{ij} \in]0,1[$ t.q.

$$\bar{u}_{i+1,j} = \bar{u}_{i,j} + h \frac{\partial u}{\partial x}(ih,jh) + \frac{h^2}{2} \frac{\partial^2 u}{\partial^2 x}(ih,jh) + \frac{h^3}{6} \frac{\partial^3 u}{\partial^3 x}(ih,jh) + \frac{h^4}{24} \frac{\partial^4 u}{\partial^4 x}(ih + \xi_{ij} h, jh)$$

$$\bar{u}_{i-1,j} = \bar{u}_{i,j} - h \frac{\partial u}{\partial x}(ih,jh) + \frac{h^2}{2} \frac{\partial^2 u}{\partial^2 x}(ih,jh) - \frac{h^3}{6} \frac{\partial^3 u}{\partial^3 x}(ih,jh) + \frac{h^4}{24} \frac{\partial^4 u}{\partial^4 x}(ih + \eta_{ij} h, jh).$$

On obtient des formules analogues pour $\bar{u}_{i,j+1}$ et $\bar{u}_{i,j-1}$, et on en déduit

$$|R_{i,j}| \leq \frac{h^2}{12} \|u_{xxxx}\|_\infty + \frac{h^2}{12} \|u_{yyyy}\|_\infty + k \frac{h^2}{6} \|u_{xxx}\|_\infty$$

où $\|u_{xxxx}\|_\infty$ désigne la norme uniforme sur $\bar{\Omega}$ de la dérivée h^2 de u par rapport à x (notations analogues pour $\|u_{yyyy}\|_\infty$ et $\|u_{xxx}\|_\infty$). On obtient finalement

$$|R_{ij}| \leq C_1 h^2,$$

où C_1 ne dépend que de u et k. Comme pour h petit, on a $h^2 \leq h$, on en déduit que schéma [I] est donc d'ordre 2.
Pour le schéma [II], on a :

$$R_{ij} = \frac{2\bar{u}_{ij} - \bar{u}_{i+1,j} - \bar{u}_{i-1,j}}{h^2} + \frac{2\bar{u}_{ij} - \bar{u}_{i,j+1} - \bar{u}_{i,j-1}}{h^2} + k\frac{\bar{u}_{ij} - \bar{u}_{i-1,j}}{h} - f_{ij}.$$

D'où l'on déduit

$$|R_{ij}| \leq \frac{h^2}{12}\|u_{xxxx}\|_\infty + \frac{h^2}{12}\|u_{yyyy}\|_\infty + \frac{kh}{2}\|u_{xx}\|_\infty,$$

et donc

$$|R_{ij}| \leq C_2 h$$

où C_2 ne dépend que de u et k. Le schéma [II] est donc d'ordre 1.

3. Dans le cas du schéma [II], la famille des w_{ij},i,$j \in \{0,\ldots,N+1\}^2$ vérifie :

$$\frac{1}{h^2}(w_{ij} - w_{i+1,j}) + \left(\frac{1}{h^2} + \frac{k}{h}\right)(w_{ij} - w_{i-1,j}) + \frac{1}{h^2}(w_{ij} - w_{i,j+1}) + \frac{1}{h^2}(w_{ij} - w_{i,j-1}) \leq 0, \forall i,j \in \{1,\ldots,N\}$$

On pose $M = \max\{w_{i,j},(i,j) \in \{0,\ldots,N+1\}^2\}$ et $m = \max\{w_{i,j},(i,j) \in \gamma\}$. Noter que $\gamma = \{0,\ldots,N+1\}^2 \setminus \{1,\ldots,N\}^2$. On a bien sûr $m \leq M$ et il reste donc à montrer que $M \leq m$. Soit $A = \{(i,j) \in \{0,\ldots,N+1\}^2, w_{i,j} = M\}$ et soit $(\bar{i},\bar{j}) \in A$ tel que $\bar{i} = \max\{i,(i,j) \in A\}$ et $\bar{j} = \max\{i,(i,j) \in A\}$. On distingue deux cas :

1. Si $\bar{i} \in \{0,N+1\}$ ou $\bar{j} \in \{0,\ldots,N+1\}$, on a alors $(\bar{i},\bar{j}) \in \gamma$ et donc $M = w_{\bar{i},\bar{j}} \leq m$.
2. Sinon, on a $\bar{i} \notin \{0,N+1\}$ et $\bar{j} \notin \{0,N+1\}$, et donc $(\bar{i},\bar{j}) \in \{1,\ldots,N\}^2$. On en déduit que :

$$\frac{1}{h^2}\left(w_{\bar{i},\bar{j}} - w_{\bar{i}+1,\bar{j}}\right) + \left(\frac{1}{h^2} + \frac{k}{h}\right)\left(w_{\bar{i},\bar{j}} - w_{\bar{i}-1,\bar{j}}\right) + \frac{1}{h^2}\left(w_{\bar{i},\bar{j}} - w_{\bar{i},\bar{j}+1}\right) + \frac{1}{h^2}\left(w_{\bar{i},\bar{j}} - w_{\bar{i},\bar{j}-1}\right) \leq 0,$$

ce qui est impossible car $w_{\bar{i},\bar{j}} = M$ et donc

$$\begin{aligned} w_{\bar{i},\bar{j}} - w_{\bar{i},\bar{j}} - 1 &\geq 0, \\ w_{\bar{i},\bar{j}} - w_{\bar{i},\bar{j}} + 1 &\geq 0, \\ w_{\bar{i},\bar{j}} - w_{\bar{i}} - 1,\bar{j} &\geq 0, \\ w_{\bar{i},\bar{j}} - w_{\bar{i}} + 1,\bar{j} &> 0, \end{aligned}$$

noter que la dernière inégalité est bien stricte car $(\bar{i}+1,\bar{j}) \notin A$ (c'est l'intérêt du choix de $\bar{i}$). On a donc bien montré que $M \leq m$.

Dans le cas du schéma [II], si on a $\bar{i} \notin \{0,N+1\}$ et $\bar{j} \notin \{0,N+1\}$, et donc $(\bar{i},\bar{j}) \in \{1,\ldots,N\}^2$ le même raisonnement que celui du schéma 1 donne :

$$\begin{aligned} &\left(\tfrac{1}{h^2} - \tfrac{k}{2h}\right)\left(u_{\bar{i}\bar{j}} - u_{\bar{i}+1,\bar{j}}\right) + \left(\tfrac{1}{h^2} + \tfrac{h}{2h}\right)\left(u_{\bar{i},\bar{j}} - u_{\bar{i}-1,\bar{j}}\right), \\ &+\tfrac{1}{h^2}\left(u_{\bar{i}\bar{j}} - u_{\bar{i},\bar{j}+1}\right) + \frac{1}{h^2}\left(u_{\bar{i}\bar{j}} - u_{\bar{i},\bar{j}-1}\right) \leq 0. \end{aligned}$$

On ne peut conclure à une contradiction que si $\frac{1}{h^2} - \frac{k}{2h} \geq 0$. Le schéma [II] vérifie

$$w_{i,j} \leq \max_{(k,\ell)\in\gamma}(w_{k,\ell}) \qquad \forall i,j \in \{1,\ldots,N\}^2$$

lorsque h vérifie la condition (dite Condition de stabilité) :

$$h \leq \frac{2}{k} \tag{1.7.98}$$

4. La fonction ϕ vérifie

$$\phi_x = 0$$
$$\phi_y = y, \qquad \phi_{yy} = 1,$$

et donc $-\Delta\phi + k\frac{\partial\phi}{\partial x} = -1$. On pose maintenant $\phi_{i,j} = \phi(ih,jh)$ pour $i,j \in \{a,\ldots,N+1\}^2$ (Noter que ϕ ne vérifie pas la condition $\phi_{i,j} = 0$ si $(i,j) \in \gamma$) Comme $\phi_{xx} = \phi_{xxx} = \phi_{xxxx} = \phi_{yyyy} = 0$, les calculs de la question 2 montrent que pour les schémas [I] et [II],

$$a_0\phi_{i,j} - a_1\phi_{i-1,j} - a_2\phi_{i+1,j} - a_3\phi_{i,j-1} - a_4\phi_{i,j+1} = -1$$

pour $i,j \in \{1,\ldots,N\}^2$.
En posant $w_{i,j} = u_{i,j} + C\phi_{i,j}$ pour $(1,j) \in \{0,\ldots,N+1\}^2$ (et U solution de (1.5.83)) on a donc

$$a_0w_{i,j} - a_1w_{i-1,j} - a_2w_{i+1,j} - a_3w_{i,j-1} - a_4w_{i,j+1} = f_{ij} - C \quad \forall i,j \in \{1,\ldots,N\}$$

On prend $C = \|f\|_\infty$, de sorte que $f_{i,j} - C \leq 0$ pour tout (i,j) pour le schéma [II] et pour le schéma [I] avec $h \leq 2/k$, la question 3 donne alors pour $(i,j) \in \{1,\ldots,N\}^2$,

$$w_{ij} \leq \max\{w_{k\ell},(k\ell) \in \gamma\} \leq \frac{C}{2},$$

car $u_{i,j} = 0$ si $(i,j) \in \gamma$ et $-\max_\Omega \phi = \frac{1}{2}$. On en déduit pour $(i,j) \in \{1,\ldots,N\}^2$,

$$w_{ij} \leq \frac{C}{2} = \frac{1}{2}\|f\|_\infty.$$

Pour montrer que $-w_{ij} \leq \frac{1}{2}\|f\|_\infty$, on prend maintenant $w_{ij} = C\phi_{ij} - u_{i,j}$ pour $(i,j) \in \{0,\ldots,N+1\}^2$, avec $C = \|f\|_\infty$. On a donc

$$a_0w_{i,j} - a_1w_{i-1,j} - a_2w_{i+1,j} - a_3w_{i,j-1} - a_4w_{i,j+1} = -C - f_{i,j} \leq 0, \forall i,j \in \{1,\ldots,N\}.$$

Ici encore, pour le schéma [II] ou le schéma [I] avec la condition $h \leq \frac{2}{k}$, la question 3 donne

$$w_{ij} \leq \max\{w_{k\ell},(k\ell) \in \gamma\} = \frac{C}{2}$$

donc $u_{ij} \geq -\frac{C}{2} = -\frac{\|f\|_\infty}{2}$ pour tout$(i,j) \in \{1,\ldots,N\}^2$. Pour le schéma [II] ou le schéma [I] avec la condition $h \leq \frac{2}{k}$, on a donc :

$$\|U\|_\infty \leq \frac{1}{2}\|f\|_\infty. \tag{1.7.99}$$

Le système (1.5.83) peut être vu comme un système linéaire de N^2 équation, à N^2 inconnues (qui sont les $u_{i,j}$ pour $(i,j) \in \{1,\ldots,N\}^2$). Si le second membre de ce système linéaire est nul, l'inégalité (1.7.99)(I) prouve que la solution est nulle. Le système (1.5.83) admet donc, pour tout f, au plus une solution. Ceci est suffisant pour affirmer qu'il admet, pour tout f, une et une seule solution.

5. pour $(i,j) \in \{0,\ldots,N+1\}^2$ on pose

$$e_{ij} = u(ih,jh) - u_{i,j}.$$

On a donc, pour les schémas [I] et [II], avec les notations de la question 2 :

$$a_0 e_{ij} - a_1 e_{i-1,j} - a_2 e_{i+1,j} - a_3 e_{i,j-1} - a_4 e_{i,j+1} = R_{ij}, \quad \forall i,j \in \{1,\ldots,N\}^2.$$

avec les questions 2 et 4, on a donc, pour le schéma [I], si $h \leq \dfrac{2}{k}$:

$$\max\{|e_{ij}|,(i,j) \in \{1,\ldots,N\}^2\} \leq \frac{1}{2} C_1 h^2,$$

où C_1 et C_2 ne dépendent que de u et k (et sont données à la question 2). Les 2 schémas sont convergents. Le schéma [I] converge en "h^2" et le schéma [II] en "h".

6. Le schéma [1] converge plus vite mais a une condition de stabilité $k \leq \frac{2}{h}$. Le schéma [II] est inconditionnellement stable.

Corrigé de l'exercice 16 page 41

1. On a vu au paragraphe 1.4.2 page 27 que si σ est une arête du volume de contrôle K, alors le flux numérique $F_{K,\sigma}$ s'écrit :

$$F_{K,\sigma} = \lambda_i \frac{u_\sigma - u_K}{d_{K,\sigma}} m(\sigma).$$

On cherche à éliminer les inconnues auxiliaires u_σ. Pour cela, si σ est une arête interne, $\sigma = K|L$, on écrit la conservativité du flux numérique :

$$F_{K,\sigma} = -F_{L,\sigma},$$

Ce qui entraîne, si σ n'est pas une arête de l'interface I, que :

$$-\lambda_i \frac{u_\sigma - u_K}{d_{K,\sigma}} m(\sigma) = \lambda_i \frac{u_\sigma - u_L}{d_{L,\sigma}} m(\sigma)$$

On en déduit que

$$u_\sigma \left(\frac{1}{d_{K,\sigma}} + \frac{1}{d_{L,\sigma}} \right) = \frac{u_K}{d_{K,\sigma}} + \frac{u_L}{d_{L,\sigma}},$$

soit encore que

$$u_\sigma = \frac{d_{K,\sigma} d_{L,\sigma}}{d_\sigma} \left(\frac{u_K}{d_{K,\sigma}} + \frac{u_L}{d_{L,\sigma}} \right).$$

Remplaçons alors dans (1.7). On obtient :

$$\begin{aligned} F_{K,\sigma} &= \lambda_i \left(\frac{d_{L,\sigma}}{d_\sigma} \left(\frac{u_K}{d_{K,\sigma}} + \frac{u_L}{d_{L,\sigma}} \right) - \frac{u_K}{d_{K,\sigma}} \right) \\ &= -\frac{\lambda_i}{d_\sigma} \left(\frac{d_{L,\sigma}}{d_{K,\sigma}} u_K + u_L - u_K - \frac{d_{L,\sigma}}{d_{K,\sigma}} u_K \right) \end{aligned}$$

On obtient donc finalement bien la formule (1.4.49).

2. Considérons maintenant le cas d'une arête $\sigma \subset \Gamma_1 \cup \Gamma_3$, où l'on a une condition de Fourier, qu'on a discrétisée par :

$$F_{K,\sigma} = -m(\sigma)\lambda_i \frac{u_\sigma - u_K}{d_{K,\sigma}} = m(\sigma)\alpha(u_\sigma - u_{ext}).$$

On a donc

$$u_\sigma = \frac{1}{\alpha + \frac{\lambda_i}{d_{K,\sigma}}} \left(\frac{\lambda_i u_K}{d_{K,\sigma}} + \alpha u_{ext} \right)$$

On remplace cette expression dans l'égalité précédente. Il vient :

$$F_{K,\sigma} = \frac{m(\sigma)\alpha}{\alpha + \frac{\lambda_i}{d_{K,\sigma}}} \left(\frac{\lambda_i}{d_{K,\sigma}} u_K + \alpha u_{ext} - \alpha u_{ext} - \frac{\lambda_i}{d_{K,\sigma}} u_{ext} \right),$$

Ce qui, après simplifications, donne exactement (1.4.50).

3. Considérons maintenant une arête $\sigma = K|L$ appartenant à l'interface I. La discrétisation de la condition de saut de flux sur I. S'écrit :

$$F_{K,\sigma} + F_{L,\sigma} = \int_\sigma \theta(x) d\gamma(x) = m(\sigma)\theta_\sigma$$

Supposons que K (resp. L) soit situé dans le milieu de conductivité (resp. λ_2). En remplaçant $F_{K,\sigma}$ et $F_{L,\sigma}$ par leurs expressions, on obtient :

$$-\lambda_i m(\sigma)\frac{u_\sigma - u_K}{d_{K,\sigma}} - \lambda_2 m(\sigma)\frac{u_\sigma - u_L}{d_{L,\sigma}} = m(\sigma)\theta_\sigma.$$

On en déduit que

$$u_\sigma\left(\frac{\lambda_1}{d_{K,\sigma}} + \frac{\lambda_2}{d_{L,\sigma}}\right) = \left(\frac{\lambda_1 u_K}{d_{K,\sigma}} + \frac{\lambda_2 u_L}{d_{L,\sigma}} - \theta_\sigma\right).$$

En remplaçant u_σ dans l'expression de $F_{K,\sigma}$, on obtient :

$$F_{K,\sigma} = -\frac{m(\sigma)}{d_{K,\sigma}}\lambda_1 \frac{1}{\frac{\lambda_1}{d_{K,\sigma}} + \frac{\lambda_2}{d_{L,\sigma}}}\left(\frac{\lambda_1 u_K}{d_{K,\sigma}} + \frac{\lambda_2 u_L}{d_{L,\sigma}} - \theta_\sigma - \frac{\lambda_1 u_K}{d_{K,\sigma}} - \frac{\lambda_2 u_K}{d_{L,\sigma}}\right).$$

En simplifiant, on obtient :

$$F_{K\sigma} = -\frac{m(\sigma)\lambda_1}{\lambda_1 d_{L,\sigma} + \lambda_2 d_{K,\sigma}}\left(\lambda_2 u_L - \lambda_2 u_K - d_{L,\sigma}\theta_\sigma\right),$$

ce qui est exactement (1.4.52). On obtient alors l'expression de $F_{L,\sigma}$:

$$F_{L,\sigma} = m(\sigma)\theta_\sigma - F_{K,\sigma},$$

ce qui donne, après simplifications :

$$F_{L,\sigma} = \frac{\lambda_2 m(\sigma)}{\lambda_1 d_{L,\sigma} + \lambda_2 d_{K,\sigma}}\left[\lambda_1(u_L - u_K) + d_{K,\sigma}\theta_\sigma\right].$$

On vérifie bien que $F_{K\sigma} + F_{L,\sigma} = m(\sigma)\theta_\sigma$.

4. Le système linéaire que satisfont les inconnues $(u_K)_{K\in\mathcal{M}}$ s'écrit

$$AU = b$$

avc $U = (u_K)_{K\in\mathcal{T}}$. Pour construire les matrices A et b, il faut se donner une numérotation des mailles. On suppose qu'on a $n \times 2p$ mailles ; on considère un maillage uniforme du type de celui décrit sur la figure 1.3 page 28 et on note $h_x = \frac{1}{n}$ (resp. $h_y = \frac{1}{p}$) la longueur de la maille dans la direction x (resp. y). Comme le maillage est cartésien, il est facile de numéroter les mailles dans l'ordre "lexicographique" ; c'est-à-dire que la k-ième maille a comme centre le point $x_{i,j} = ((i-\frac{1}{2})h_x,(j-\frac{1}{2})h_y)$, avec $k = n(j-1)+i$. On peut donc déterminer le numéro de la maille (et de l'inconnue associée) k à partir de la numérotation cartésienne (i,j) de la maille.

$$k = n(j-1) + i$$

Remarquons que, comme on a choisi un maillage uniforme, on a pour tout $K \in \mathcal{T}$: $m(K) = h_x h_y$, pour toute arête intérieure verticale σ : $d_\sigma = h_x m(\sigma) = h_y$ et pour toute arête intérieure horizontale,

$d_\sigma = h_y$ et $m(\sigma) = h_x$. Pour chaque numéro de maille, nous allons maintenant construire l'équation correspondante.
Mailles internes $i = 2,\ldots,n-1$; $j = 2,\ldots,p-1,p+1,\ldots,2p-1$.
L'équation associée à une maille interne K s'écrit

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} = m(K) f_K.$$

Avec l'expression de $F_{K,\sigma}$ donnée par (1.4.49), ceci amène à :

$$2\lambda_m\left(\frac{h_x}{h_y}+\frac{h_y}{h_x}\right)u_k - \lambda_m\frac{h_x}{h_y}(u_{k-n}+u_{k+n}) - \lambda_m\frac{h_y}{h_x}(u_{k+1}+u_{k-1}) = h_x h_y f_k,$$

avec $m = 1$ si $j \le p-1$ et $m = 2$ si $j \ge p+1$.
Mailles du bord Γ_2 Les mailles du bord Γ_2 sont repérées par les indices (n,j),$j = 2$ à $p-1$, $j = p+1$ à $2p-1$, (on exclut pour l'instant les coins).
L'équation des flux est la même que pour les mailles internes, mais le flux sur la frontière Γ_2 est nul. Ceci donne :

$$\lambda_m\left(2\frac{h_x}{h_y}+\frac{h_y}{h_x}\right)u_k - \lambda_m\frac{h_x}{h_y}(u_{k-n}+u_{k+n}) - \lambda_m\frac{h_y}{h_x}u_{k-1} = h_x h_y f_k,$$

avec $k = n(j-1)+n$,$j = 2$ à $p-1$,$j = p+1$ à $2p-1$ et $m = 1$ si $j \le p-1$,$m = 2$ si $j \ge p+1$.
Mailles de bord Γ_4 Les mailles du bord Γ_4 sont repérées par les indices $(1,j)$,$j = 2$ à $p-1$,$j = p+1$ à $2p-1$. Pour ces mailles, il faut tenir compte du fait que sur une arête de Γ_4, le flux $F_{K,\sigma}$ est donné par :

$$F_{K,\sigma} = -\lambda_m\frac{g_\sigma - u_K}{d_{K,\sigma}}m(\sigma)$$

avec $g_\sigma = \dfrac{1}{m(\sigma)}\displaystyle\int g(y)d\gamma(y)$.
D'où on tire l'équation relative à la maille $k = n(j-1)+1$,$j = 2,\ldots,p-1,p+1,\ldots,2p-1$:

$$\lambda_m\left(2\frac{h_x}{h_y}+3\frac{h_y}{h_x}\right)u_k - \lambda_m\frac{h_x}{h_y}(u_{k-n}+u_{k+n}) - \lambda_m\frac{h_y}{h_x}u_{k+1} = h_x h_y f_k + 2\frac{h_y}{h_x}\lambda_m g_j,$$

avec $g_j = g_{\sigma_j}$ et $m = 1$ si $j \le p-1$,$m = 2$ si $j \ge p+1$.

Mailles du bord $\Gamma_1 \cup \Gamma_3$ Pour $j = 1$, où $j = 2p$,$i = 2\ldots n-1$. On tient compte ici de la condition de Fourier sur la maille qui appartient au bord, pour laquelle l'expression du flux est :

$$F_{K,\sigma} = \frac{\alpha\lambda_m m(\sigma)}{\lambda_m + \alpha d_{K,\sigma}}(u_K - u_{ext}).$$

Pour une arête σ horizontale, on note : $C_{F,\sigma} = \dfrac{\alpha m(\sigma)}{\lambda_m + \alpha d_{K,\sigma}}$. Notons que $C_{F,\sigma}$ est égal à

$$C_F = \frac{2\alpha h_x}{2\lambda_m + \alpha h_y}.$$

Notons que ce coefficient ne dépend pas de σ.

Les équations s'écrivent donc :

$$\lambda_1\left(2\tfrac{h_y}{h_x}+\tfrac{h_x}{h_y}+C_F\right)u_k-\lambda_1\tfrac{h_x}{h_y}u_{k+n}-\lambda_1\tfrac{h_y}{h_x}(u_{k+1}+u_{k-1})=h_xh_yf_k+\lambda_1C_Fu_{ext},\quad k=2,\ldots,n-1,$$

$$\lambda_2\left(\tfrac{2h_y}{h_x}+\tfrac{h_x}{h_y}+C_F\right)u_k-\lambda_2\tfrac{h_x}{h_y}u_{k-n}-\lambda_2\tfrac{h_y}{h_x}(u_{k+1}+u_{k-1})=h_xh_yf_k+\lambda_2C_Fu_{ext},\quad k=2n(p-1)+2,\ldots,2np-1,$$

Mailles des coins extérieurs : Il suffit de synthétiser les calculs déjà faits :
• coin sud-est : $i=1$,$j=1$,$k=1$; un bord Dirichlet, un bord Fourier :

$$\lambda_1\left(\frac{3h_y}{h_x}+\frac{h_x}{h_y}+C_F\right)u_1-\lambda_1\frac{h_y}{h_x}u_2-\lambda_1\frac{h_x}{h_y}u_{n+1}=h_xh_yf_1+\lambda_1C_Fu_{ext}+\frac{2h_y}{h_x}\lambda_1g_1$$

• coin sud-ouest : $i=1n$,$j=1$,$k=n$; un bord Fourier, un bord Neumann :

$$\lambda_1\left(\frac{h_y}{h_x}+\frac{h_x}{h_y}+C_F\right)u_1-\lambda_1\frac{h_y}{h_x}u_{n-1}-\lambda_1\frac{h_x}{h_y}u_{2n}=h_xh_yf_n+\lambda_1C_Fu_{ext}$$

• coin nord-ouest : $i=2n$,$j=2p$,$k=2np$.
On a encore un bord Fourier, un bord Neumann, et l'équation s'écrit :

$$\lambda_2\left(\frac{h_y}{h_x}+\frac{h_x}{h_y}+C_F\right)u_{2np}-\lambda_2\frac{h_y}{h_x}u_{2np-1}-\lambda_2\frac{h_x}{h_y}u_{2n(p-1)}=h_xh_yf_{2np}+\lambda_2C_Fu_{ext}$$

• coin nord-est : $i=1$,$j=2p\quad k=n(2p-1)+1$ un bord Dirichlet, un bord Fourier :

$$\lambda_2\left(\frac{3h_y}{h_x}+\frac{h_x}{h_y}+C_F\right)u_k-\lambda_2\frac{h_y}{h_x}u_{k+1}-\lambda_2\frac{h_x}{h_y}(u_{k-n}=h_xh_yf_k+\lambda_2C_Fu_{ext},+\frac{2h_y}{h_x}\lambda_2g_k.$$

Interface L'expression du flux sur une arête de l'interface est donnée par (1.4.52). On pose, pour chaque arête σ de l'interface,

$$s_{I,\sigma}=\frac{m(\sigma)}{\lambda_1d_{L,\sigma}+\lambda_2d_{K,\sigma}}.$$

Notons que dans le cas du maillage uniforme considéré, ce coefficient est égal à :

$$s_I=\frac{2h_x}{(\lambda_1+\lambda_2)h_y},$$

et qu'il est indépendant de l'arête σ. Tenant compte de ce flux, on obtient, pour $k=n(p-1)+i$,$i=2,\ldots,N-1$

$$\lambda_1\left(\frac{2h_y}{h_x}+\frac{h_x}{h_y}+\lambda_2S_I\right)u_k-\lambda_1\frac{h_y}{h_x}u_{k+1}-\lambda_1\frac{h_y}{h_x}u_{k-1}-\lambda_1\frac{h_x}{h_y}u_{k-n}-\lambda_1S_Iu_{k+n}=h_xh_yf_k+\lambda_1S_I\frac{h_y}{2}\theta_i,$$

avec

$$\theta_i=\int_{\sigma_i}\theta(x)d\gamma(x).$$

Et de même, pour $k=np+i$,$i=2,\ldots,N-1$,

$$\lambda_1\left(\frac{2h_y}{h_x}+\frac{h_x}{h_y}+\lambda_1 S_I\right)u_k-\lambda_2\frac{h_y}{h_x}u_{k+1}-\lambda_2\frac{h_y}{h_x}u_{k-1}-\lambda_2\frac{h_x}{h_y}u_{k+n})-\lambda_2 S_I u_{k-n}=h_x h_y f_k+\lambda_2 S_I\frac{h_y}{2}\theta_i.$$

Il ne reste plus qu'à traiter les coins des interfaces.

• $i=1$,$j=p \quad k=n(p-1)+1$. Dirichlet sous l'interface

$$\lambda_1\left(\frac{3h_y}{h_x}+\frac{h_x}{h_y}+\lambda_2 s_I\right)u_k-\lambda_1\frac{h_y}{h_x}u_{k+1}-\lambda_1\frac{h_x}{h_y}u_{k+n}-\lambda_1 s_I u_{k+n}=h_x h_y f_k+\lambda_1 S_I\frac{h_y}{2}\theta_i+\frac{2h_y}{h_x}\lambda_1 g_j$$

• $i=1$,$j=p+1 \quad k=np+1$, Dirichlet, dessus de l'interface

$$\lambda_2\left(\frac{3h_y}{h_x}+\frac{h_x}{h_y}+\lambda_1 s_I\right)u_k-\lambda_2\frac{h_y}{h_x}u_{k+1}-\lambda_2\frac{h_x}{h_y}u_{k+n}-\lambda_2 s_I u_{k-n}=h_x h_y f_k+\lambda_2 S_I\frac{h_y}{2}\theta_i+\frac{2h_y}{h_x}\lambda_2 g_j$$

• $i=n$,$j=p$,$k=n(p-1)+n$. Neumann, sous l'interface.

$$\lambda_1\left(\frac{h_y}{h_x}+\frac{h_x}{h_y}+\lambda_2 s_I\right)u_k-\lambda_1\frac{h_y}{h_x}u_{k-1}-\lambda_1\frac{h_x}{h_y}u_{k-n}-\lambda_1 s_I u_{k+n}=h_x h_y f_k+\lambda_1 S_I\frac{h_y}{2}\theta_i$$

• $i=n$,$j=p+1$,$k=np+n$, Neuman, dessus de l'interface

$$\lambda_2\left(\frac{h_y}{h_x}+\frac{h_x}{h_y}+\lambda_1 s_I\right)u_k-\lambda_2\frac{h_y}{h_x}u_{k-1}-\lambda_2\frac{h_x}{h_y}u_{k+n}-\lambda_2 s_I u_{k-n}=h_x h_y f_k+\lambda_2 S_I\frac{h_y}{2}\theta_i.$$

On a ainsi obtenu $2np$ équations à $2np$ inconnues. Notons que chaque équation fait intervenir au plus 5 inconnues.

Corrigé de l'exercice 17 page 41

1. Le problème complet s'écrit :

$$\begin{cases} -div(\mu_i\nabla\phi)(x)=0, & x\in\Omega_i, \quad i=1,2\\ \nabla\phi(x)\cdot\mathbf{n}(x)=0, & x\in\Gamma_2\cup\Gamma_4,\\ \displaystyle\int_{\Gamma_1}\mu_1\nabla\phi(x)\cdot\mathbf{n}(x)\,d\gamma(x)+\int_{\Gamma_3}\mu_2\nabla\phi(x)\cdot\mathbf{n}(x)\,d\gamma(x)=0,\\ \phi_2(x)-\phi_1(x)=0, & \forall x\in I,\\ -(\mu\nabla\phi\cdot\mathbf{n})|_2(x)-(\mu\nabla\phi\cdot\mathbf{n})|_1(x)=0, & \forall x\in I. \end{cases}$$

2. On se donne le même maillage rectangulaire uniforme qu'à l'exercice précédent. On note ϕ_K l'inconnue associée à la maille K (ou ϕ_k si on la référence la maille K par son numéro $k=n(j-1)+i$, où $i\in\{1,\ldots,n\}$ et $j\in\{1,\ldots,2p\}$. Pour une maille intérieure, l'équation obtenue est la même que (1.7) en remplaçant λ_m par μ_m.

Etudions maintenant le cas d'une maille proche de l'interface. Comme indiqué, on va considérer deux inconnues discrètes par arête de l'interface. Soient K et L ayant en commun l'arête $\widetilde{\sigma}\subset I$, K est située au dessous de L. Les équations associées à K et L s'écrivent alors

$$\sum_{\sigma\in\xi_K}F_{K,\sigma}=0 \text{ et } \sum_{\sigma\in\xi_L}F_{L,\sigma}=0.$$

Pour les arêtes $\sigma \in \xi_K$ autres que $\widetilde{\sigma}$, le flux s'écrit de manière habituelle

$$F_{K,\sigma} = \mu_1 \frac{\phi_K - \phi_M}{d_\sigma}, \text{ avec } \sigma = K|M.$$

Pour l'arête $\sigma = \widetilde{\sigma}$, on a $F_{K,\sigma} = \mu_1 \frac{\phi_K - \phi_\sigma}{d_\sigma} m(\sigma)$ et $F_{L,\sigma} = \mu_2 \frac{\phi_L - \phi_\sigma^+}{d_\sigma} m(\sigma)$, où les deux inconnues discrètes ϕ_σ^+ et ϕ_σ^- sont reliées par les relations :

$$\phi_\sigma^+ - \phi_\sigma^- = \psi_\sigma \left(= \frac{1}{m(\sigma)} \int_\sigma \psi(x) d\gamma(x) \right)$$

$$F_{K,\sigma} + F_{L,\sigma} = 0.$$

On peut alors éléminer ϕ_σ^+ et ϕ_σ^- ; en utilisant par exemple $\phi_\sigma^+ = \psi_\sigma + \phi_\sigma^-$ et en remplaçant dans la deuxième équation, on obtient :

$$-\mu_1 \frac{\phi_\sigma^- - \phi_K}{d_{K,\sigma}} + \mu_2 \frac{\phi_\sigma^- + \psi_\sigma - \phi_L}{d_{L,\sigma}} = 0,$$

ce qui donne :

$$\phi_{\sigma^-} = \frac{1}{\frac{\mu_1}{d_{K,\sigma}} + \frac{\mu_2}{d_{L,\sigma}}} \left(\frac{\mu_1}{d_{K,\sigma}} \phi_K + \frac{\mu_2}{d_{L,\sigma}} \phi_L - \frac{\mu_2}{d_{L,\sigma}} \psi_\sigma . \right).$$

En remplaçant cette expression dans les flux, on obtient :

$$F_{K,\sigma} = -F_{L,\sigma} = m(\sigma) \frac{\mu_1 \mu_2}{\mu_1 d_{L,\sigma} + \mu_2 d_{K,\sigma}} (\phi_K - \phi_L + \psi_\sigma)$$

On peut alors écrire l'équation discrète associée à une maille de numéro k située sous l'interface (avec $k = n(p-1) + i, \quad i = 2, \ldots, n-1$). On pose :

$$\frac{\mu_1 \mu_2}{\mu_1 d_{L,\sigma} + \mu_2 d_{K,\sigma}} = \frac{\mu_I}{d_\sigma}$$

(μ_I est donc la moyenne harmonique pondérée entre μ et μ_2). Notons que pour une arête de I, $d_\sigma = h_y$, et $m(\sigma) = h_x$. L'équation associée à la maille k s'écrit donc :

$$\left(2\mu_1 \frac{h_y}{h_x} + \mu_1 \frac{h_x}{h_y} + \frac{\mu_I h_x}{h_y} \right) u_k - \mu_1 \frac{h_y}{h_x} (u_{k-1} + u_{k+1}) - \mu_1 \frac{h_x}{h_y} u_{k-n} - \mu_I \frac{h_x}{h_y} u_{k+n} = -\mu_I \frac{h_x}{h_y} \psi_i,$$

où ψ_i est le saut de potentiel à travers l'arête σ_i de l'interface considérée ici. De même, l'équation associée à une maille k avec $k = np + i, i = 2, \ldots, n-1$, située au dessus de l'interface s'écrit :

$$\left(2\mu_1 \frac{h_y}{h_x} + \mu_1 \frac{h_x}{h_y} + \mu_I \frac{h_x}{h_y} \right) u_k - \mu_1 \frac{h_y}{h_x} (u_{k-1} + u_{k+1}) - \mu_1 \frac{h_x}{h_y} u_{k+n} - \mu_I \frac{h_x}{h_y} u_{k-n} = +\mu_I \frac{h_x}{h_y} \psi_i.$$

La discrétisation des conditions aux limites de Neumann sur Γ_2 et Γ_4 est effectuée de la même manière que pour le cas du problème thermique, voir exercice 16.
Il ne reste plus qu'à discrétiser la troisième équation du problème (1.7), qui relie les flux sur la frontière Γ_1 avec les flux sur la frontière Γ_3. En écrivant la même condition avec les flux discrets, on obtient :

$$\mu_1 \sum_{i=1}^{n} \frac{2h_x}{h_y} (u_i - u_{B,i}) + \mu_2 \sum_{i=1}^{n} \frac{2h_x}{h_y} (u_{H,i} - u_{k(i)}) = 0,$$

où : $\mu_{B,i}$ représente l'inconnue discrète sur la i-ème arête de Γ_1 et $\mu_{H,i}$ l'inconnue discrète sur la i-ème arête de Γ_3, et $k(i) = n(p-1) + i$ est le numéro de la maille jouxtant la i-ème arête de Γ_3.
Remarquons que tel qu'il est posé, le système n'est pas inversible : on n'a pas assez d'équations pour éliminer les inconnues $u_{B,i}$ et $u_{H,i}$,$i = 1 \dots N$. On peut par exemple pour les obtenir considérer une différence de potentiel fixée entre Γ_1 et Γ_3, et se donner un potentiel fixé sur Γ_1.

Chapitre 2

Problèmes paraboliques : la discrétisation en temps

On a vu au paragraphe comme exemple type de problème parabolique l'équation de la chaleur instationnaire :

$$u_t - \Delta u = f$$

qui fait intervenir la dérivée en temps d'ordre 1, u_t, ainsi qu'un opérateur différentiel d'ordre 2 en espace. Pour que ce problème soit bien posé, il faut spécifier des conditions aux limites sur la frontière de Ω, et une condition initiale en $t = 0$.

2.1 Le problème continu, et la discrétisation espace-temps

On considère maintenant le même problème en une dimension d'espace. Au temps $t = 0$, on se donne une condition initiale u_0, et on considère des conditions aux limites de type Dirichlet homogène. Le problème unidimensionnel s'écrit :

$$\begin{cases} u_t - u_{xx} = 0, \forall x \in]0,1[, \forall t \in]0,T[\\ u(x,0) = u_0(x), \forall x \in]0,1[, \\ u(0,t) = u(1,t) = 0, \forall t \in]0,T[, \end{cases} \tag{2.1.1}$$

où $u(x,t)$ représente la température au point x et au temps t. On admettra le théorème d'existence et unicité suivant :

Théorème 2.1 (Résultat d'existence et unicité) *Si $u_0 \in C(]0,1[,\mathbb{R})$ alors il existe une unique fonction $u \in C^2(]0,1[\times]0,T[,\mathbb{R}) \cap C([0,1] \times [0,T],\mathbb{R})$ qui vérifie (2.1.1).*
On a même $u \in C^\infty(]0,1[\times]0,T[,\mathbb{R})$. Ceci est appelé, effet "régularisant" de l'équation de la chaleur.

Proposition 2.2 (Principe du maximum) *Sous les hypothèses du théorème 2.1, soit u la solution du problème (2.1.1) ;*

1. *si $u^0(x) \geq 0$ pour tout $x \in [0,1]$, alors $u(x,t) \geq 0$, pour tout $t \geq 0$ pour tout $x \in]0,1[$.*
2. *$\|u\|_{L^\infty([0,1[\times]0,T[)} \leq \|u\|_{L^\infty(]0,1[)}$.*

Ces dernières propriétés peuvent être importantes dans le modèle physique, et il est donc souvent souhaitable que les solutions approchées les vérifient également. Pour calculer une solution approchée, on se donne une discrétisation en temps et en espace, qu'on notera $\mathcal{D}$. On choisit pour l'instant de discrétiser par différences finies en temps et en espace. La discrétisation consiste donc à se donner un ensemble de points t_n, $n = 1,\ldots,M$ de l'intervalle $]0,T[$, et un ensemble de points x_i, $i = 1,\ldots,N$. Pour simplifier, on considère un pas constant en temps et en espace. Soit : $h = \frac{1}{N+1} = \Delta x$ le pas de discrétisation en espace, et $k = \Delta t = \frac{T}{M}$, le pas de discrétisation en temps. On pose alors $t_n = nk$ pour $n = 0,\ldots,M$ et $x_i = ih$ pour $i = 0,\ldots,N+1$. On cherche à calculer une solution approchée $u_{\mathcal{D}}$ du problème (2.1.1) ; plus précisement, on cherche à déterminer $u_{\mathcal{D}}(x_i,t_n)$ pour $i = 1,\ldots,N$, et $n = 1,\ldots,M$. Les inconnues discrètes sont notées $u_i^{(n)}$, $i = 1,\ldots,N$ et $n = 1,\ldots,M$.

2.2 Discrétisation par Euler explicite en temps.

L'approximation en temps par la méthode d'Euler explicite consiste à écrire la première équation de (2.1.1) en chaque point x_i et temps t_n, à approcher $u_t(x_i,t_n)$ par le quotient différentiel :

$$\frac{u(x_i,t_{n+1}) - u(x_i,t_n)}{k},$$

et $-u_{xx}(x_i,t_n)$ par

$$\frac{1}{h^2}(2u(x_i,t_n) - u(x_{i-1},t_n) - u(x_{i+1},t_n)).$$

Remarque 2.3 *On a choisi une discrétisation en espace de type différences finies, mais on aurait aussi bien pu prendre un schéma de volumes finis ou d'éléments finis.*
On approche donc :

$$-u_{xx}(x_i,t_n) \text{ par } \frac{1}{h^2}(2u(x_i,t_n) - u(x_{i-1},t_n) - u(x_{i+1},t_n)).$$

On obtient le schéma suivant :

$$\begin{cases} \dfrac{u_i^{(n+1)} - u_i^{(n)}}{k} + \dfrac{1}{h^2}(2u_i^{(n)} - u_{i-1}^{(n)} - u_{i+1}^{(n)}) = 0, \quad i = 1,\ldots N, \quad n = 1,\ldots,M, \\ u_i^0 = u_0(x_i), \quad i = 1,\ldots,N, \\ u_0^{(n)} = u_{N+1}^{(n)} = 0, \quad \forall n = 1,\ldots,M. \end{cases} \tag{2.2.2}$$

le schéma est dit explicite, car la formule ci-dessus donne $u_i^{(n+1)}$ de manière explicite en fonction des $(u_i^{(n)})_{i=1,\ldots,N}$. En effet on a :

$$u_i^{(n+1)} = u_i^{(n)} - \lambda(2u_i^{(n)} - u_{i-1}^{(n)} - u_{i+1}^{(n)}),$$

avec $\lambda = \dfrac{k}{h^2}$.

2.2.1 Consistance du schéma

Soit $\bar{u}_i^{(n)} = u(x_i,t_n)$ la valeur exacte de la solution en x_i et t_n : L'erreur de consistance R_i en (x_i,t_n) peut s'écrire comme la somme des erreurs de consistance en temps et en espace : $R_i^{(n)} = \tilde{R}_i^{(n)} + \hat{R}_i^n$ avec :

$$\tilde{R}_i^{(n)} = \frac{\bar{u}_i^{(n+1)} - \bar{u}_i^{(n)}}{k} - u_t(x_i,t_n) \text{ et } \hat{R}_i^{(n)} = \frac{1}{h^2}\left(2\bar{u}_i^{(n)} - \bar{u}_{i-1}^{(n)} - \bar{u}_{i+1}^{(n)}\right) - u_{xx}(x_i,t_n).$$

Proposition 2.4 *Le schéma (2.2.2) est consistant d'ordre 1 en temps et d'ordre 2 en espace, c'est à dire qu'il existe $C \in \mathbb{R}_+$ ne dépendant que de u tel que :*

$$|R_i^{(n)}| \leq C(k + h^2). \tag{2.2.3}$$

Démonstration : On a vu lors de l'étude des problèmes elliptiques que l'erreur de consistance en espace $\widetilde{R}_i^{(n)}$ est d'ordre 2 (voir formule (1.2.19) page 16). Un développement de Taylor en temps donne facilement que $\widetilde{R}_i^{(n)}$ est d'ordre 1 en temps. ∎

2.2.2 Stabilité

On a vu à la proposition 2.2 page 69 que la solution exacte vérifie :

$$\|u\|_{L^\infty(]0,1[\times]0,T[)} \leq \|u_0\|_{L^\infty(]0,1[)}$$

Si on choisit correctement les pas de temps et d'espcace, nous allons voir qu'on peut avoir l'équivalent discret sur la solution approchée.

Définition 2.5 *On dit qu'un schéma est L^∞-stable si la solution approchée est bornée dans L^∞ indépendamment du pas du maillage.*

Proposition 2.6 *Si la condition de stabilité*

$$\lambda = \frac{k}{h^2} \leq \frac{1}{2} \tag{2.2.4}$$

est vérifiée, alors le schéma (2.2.2) est L^∞–stable au sens où :

$$\sup_{\substack{i=1,\dots,N \\ n=1,\dots,M}} |u_i^{(n)}| \leq \|u_0\|_\infty$$

Démonstration : On peut écrire le schéma sous la forme

$$u_i^{(n+1)} = u_i^{(n)} - \lambda(2u_i^{(n)} - u_{i-1}^{(n)} - u_{i+1}^{(n)}),$$

soit encore :

$$u_i^{(n+1)} = (1-2\lambda)u_i^{(n)} + \lambda u_{i-1}^{(n)} + \lambda u_{i+1}^{(n)}.$$

Si $0 \leq \lambda \leq \dfrac{1}{2}$, on a $\lambda \geq 0$ et $1 - 2\lambda \geq 0$, et la quantité $u_i^{(n+1)}$ est donc combinaison convexe de $u_i^{(n)}$, $u_{i-1}^{(n)}$ et $u_{i+1}^{(n)}$. Soit $M^{(n)} = \max\limits_{i=1,\dots,N} u_i^{(n)}$, on a alors :

$$u_i^{(n+1)} \leq (1-2\lambda)M^{(n)} + \lambda M^{(n)} + \lambda M^{(n)}, \quad \forall i = 1,\dots,N,$$

et donc $u_i^{(n+1)} \leq M^{(n)}$. On en déduit en passant au maximum que :

$$M^{(n+1)} \leq M^{(n)}.$$

On montre de la même manière que

$$\min_{i=1,\dots,N} u_i^{(n+1)} \geq \min_{i=1,\dots,N} u_i^{(n)}.$$

On en déduit $\max_{i=1,\dots,N}(u_i^{(n+1)}) \leq \max u_i^0$ et $\min_{i=1,\dots,N}(u_i^{(n+1)}) \geq \min u_i^0$ d'où le résultat. ∎

2.2.3 Convergence

Définition 2.7 *Soit u la solution du problème (2.1.1) et* $(u_i^{(n)})_{\substack{i=1,\dots N,\\ n=1,\dots,M}}$ *la solution de (2.2.2). On appelle erreur de discrétisation au point* (x_i,t_n) *la quantité* $e_i^n = u(x_i,t_n) - u_i^n$.

Théorème 2.8 *Sous les hypothèses du théorème 2.1, et sous la condition de stabilité (2.2.4), il existe* $C \in \mathbb{R}_+$ *ne dépendant que de u tel que*

$$\|e_i^{(n+1)}\|_\infty \leq \|e_i^{(0)}\|_\infty + TC(k+h^2), \text{ pour tout } i = 1,\dots,N \text{ et } n = 0,\dots,M-1.$$

Ainsi, si $\|e_i^{(0)}\|_\infty = 0$, *alors* $\max_{i=1,\dots N} \|e_i^{(n)}\|$ *tend vers 0 lorsque k et h tendent vers 0, pour tout* $n = 1,\lambda dots M$ *. Le schéma (2.2.2) est donc convergent.*

Démonstration : On note $\bar{u}_i^{(n)} = u(x_i,t_n)$. On a donc, par définition de l'erreur de consistance,

$$\frac{\bar{u}_i^{(n+1)} - \bar{u}_i^{(n)}}{k} - \frac{1}{h^2}(2\bar{u}_i^{(n)} - \bar{u}_{i-1}^{(n)} - \bar{u}_{i+1}^{(n)}) = R_i^{(n)}. \tag{2.2.5}$$

D'autre part, le schéma numérique s'écrit :

$$\frac{u_i^{(n+1)} - u_i^{(n)}}{k} - \frac{1}{h^2}(2u_i^{(n)} - u_{i-1}^{(n)} - u_{i+1}^{(n)}) = 0. \tag{2.2.6}$$

Retranchons (2.2.6) à (2.2.5), on obtient :

$$\frac{e_i^{(n+1)} - e_i^{(n)}}{k} - \frac{1}{h^2}(2e_i^{(n)} - e_{i+1}^{(n)} - e_{i-1}^{(n)}) = R_i^{(n)},$$

soit encore :

$$e_i^{(n+1)} = (1-2\lambda)e_i^{(n)} + \lambda e_{i-1}^{(n)} + \lambda e_{i+1}^{(n)} + kR_i^{(n)}$$

Or $(1-2\lambda)e_i^{(n)} + \lambda e_{i-1}^{(n)} + \lambda e_{i+1}^{(n)} \leq \|e^{(n)}\|_\infty$, car $\lambda \leq \frac{1}{2}$, et donc comme le schéma est consistant, l'inégalité (2.2.3) entraîne que :

$$|e_i^{(n+1)}| \leq \|e^{(n)}\|_\infty + kC(k+h^2).$$

On a donc par récurrence :

$$\|e_i^{(n+1)}\|_\infty \leq \|e^{(0)}\|_\infty + MkC(k+h^2)$$

ce qui démontre le théorème. ∎

Donnons maintenant un exemple où lorsque la condition (2.2.3) n'est pas vérifiée, le schéma est instable.

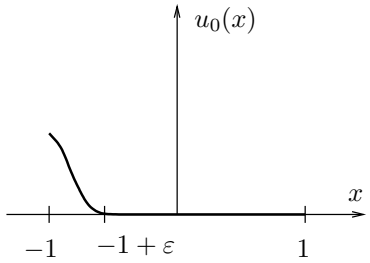

FIG. 2.1 – *Condition initiale pour le contre exemple*

2.2.4 Exemple de non convergence

Montrons que si la condition $\lambda \leq \frac{1}{2}$ n'est pas respectée, on peut construire une condition initiale pour lequel le schéma n'est pas stable. Soit $u_0 \in C([-1,1],\mathbb{R})$ qui vérifie (voir Figure (2.1) :

$$\begin{cases} u_0(x) \geq 0 \\ u_0(x) \neq 0 \text{ si } x \in]-1;-1+\varepsilon[\\ u_0(x) = 0 \text{ si } x > -1+\varepsilon \end{cases}$$

On considère le problème :

$$\begin{cases} u_t - u_{xx} = 0, \forall x \in]-1,1[; \forall t > 0. \\ u(x,0) = u_0(x), \forall x \in]-1,1[\\ u(1,t) = u(-1,t) = 0, \forall t > 0. \end{cases} \tag{2.2.7}$$

On peut montrer que la solution exacte u de (2.2.7) vérifie $u(x,t) > 0, \forall x \in]-1,1[, \forall t > 0$. En particulier, pour un temps $T > 0$ donné, on a $u(0,T) > 0$. Soit $M \in \mathbb{N}$ et $k = T/M$. Soit $u_i^{(n)}$ la solution approchée par (2.2.2), sensée approcher $u(x_i,t_n)$ $(i \in \{-N,\ldots,N\}, n \in N)$. On va montrer que $u_0^M = 0$ pour k et h choisis de manière non admissible ; ceci montre que le schéma ne peut pas converger. Calculons u_0^M :

$$u_0^M = (1-2\lambda)u_0^{M-1} + \lambda u_{-1}^{M-1} + \lambda u_1^{M-1}.$$

Donc u_0^M dépend de

$$\begin{aligned} & u^{(M-1)} \text{ sur } [-h,h] \\ & u^{(M-2)} \text{ sur } [-2h,2h] \\ & \vdots \\ & u^{(0)} \text{ sur } [-Mh,Mh] = [-\frac{T}{k}h, \frac{T}{k}h] \end{aligned}$$

Par exemple, si on prend $\dfrac{h}{k} = \dfrac{1}{2T}$ on obtient : $[-\frac{T}{k}h, \frac{T}{k}h] = [-\frac{1}{2}, \frac{1}{2}]$, et donc, si $\varepsilon < \frac{1}{2}$, on a $u_0^M = 0$. On peut donc remarquer que si $\frac{h}{k} = \frac{1}{2T}$, même si $h \to 0$ et $k \to 0$,

$$u_0^M \not\to u(0,T).$$

Le schéma ne converge pas ; notons que ceci n'est pas en contradiction avec le résultat de convergence 2.8 page 72, puisqu'ici, on n'a pas satisfait à la condition $\frac{k}{h^2} \leq \frac{1}{2}$.

2.2.5 Stabilité au sens des erreurs d'arrondi

On considère le schéma d'Euler explicite pour l'équation (2.1.1). On appelle u la solution exacte de (2.1.1), $u_{\mathcal{D}}$ la solution exacte de (2.2.2), u_{num} la solution effectivement calculée. On peut écrire :

$$u - u_{num} = u - u_{\mathcal{D}} + u_{\mathcal{D}} - u_{num}.$$

On sait que l'erreur de discrétisation $u - u_{\mathcal{D}}$ tend vers 0 lorsque h et k tendent vers 0, sous condition de stabilité (2.2.4), c.à.d.

$$\lambda \leq \frac{1}{2}.$$

Pour contrôler l'erreur entre la solution $u_{\mathcal{D}}$ du schéma (2.2.2) et la solution numérique obtenue u_{num}, on cherche à estimer l'amplification de l'erreur commise sur la donnée initiale. Rappelons que le schéma s'écrit :

$$u_i^{(n+1)} = (1 - 2\lambda)u_i^{(n)} + \lambda u_{i-1}^{(n)} + \lambda u_{i+1}^{(n)},$$

avec $\lambda = \frac{k}{h^2}$. Ce schéma se met sous la forme $u^{(n+1)} = AU^{(n)}$, avec

$$A = \begin{pmatrix} 1-2\lambda & \lambda & 0 & \dots & 0 \\ \lambda & 1-2\lambda & \lambda & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \lambda & 1-2\lambda & \lambda \\ 0 & \dots & 0 & \lambda & 1-2\lambda \end{pmatrix}$$

Définition 2.9 (Stabilité au sens des erreurs d'arrondi) *Supposons que l'on commette une erreur ε^0 sur la condition initiale. La nouvelle condition initiale $\widetilde{u}^0$, s'écrit donc $\widetilde{u}^0 = u^0 + \varepsilon^0$. A cette nouvelle condition initiale correspond une nouvelle solution calculée $\widetilde{u}^{(n)} = u^{(n)} + \varepsilon^{(n)}$. On dit que le schéma est stable au sens des erreurs d'arrondi s'il existe $C > 0$ indépendant de n tel que $\varepsilon^{(n)} \leq C\varepsilon^{(0)}$.*

On peut trouver une condition suffisante pour que le schéma 2.2.2 soit stable au sens des erreurs d'arrondi. En effet, on va démontrer le résultat suivant :

Proposition 2.10 *On suppose que $\lambda = \frac{k}{h^2} < \frac{1}{2}$. Alors le schéma 2.2.2 est stable au sens des erreurs d'arrondi.*

Démonstration : Soit donc une condition initiale perturbée $\widetilde{u}^0 = u^0 + \varepsilon^0$ à laquelle on associe une nouvelle solution calculée $\widetilde{u}^{(n)} = u^{(n)} + \varepsilon^{(n)}$. On a $\varepsilon^{(n)} = A^n\varepsilon^0$. Comme A est symétrique, A est diagonalisable dans $\mathbb{R}$. Soient $\mu_1, \dots, \mu_N$ les valeurs propres de A, et $e_1, \dots, e_N$ les vecteurs propres associés, c'est–à–dire tels que $Ae_i = \mu_i e_i, \forall i = 1, \dots N$. On décompose la perturbation ε^0 sur la base des vecteurs propres :

$$\varepsilon^0 = \sum_{i=1}^{N} a_i e_i. \text{ On a donc } A^n\varepsilon^0 = \sum_{i=1}^{N} a_i \mu_i^n e_i = \varepsilon^{(n)}.$$

Si on prend par exemple : $\varepsilon^0 = a_i e_i$, on obtient $\varepsilon^{(n)} = a_i \mu_i^n e_i$. Il y a diminution de l'erreur d'arrondi sur ε^0 si

$$\sup_{i=1\ldots N} |\mu_i| \leq 1$$

c'est-à-dire si $\rho(A) \leq 1$, où $\rho(A)$ désigne le rayon spectral de A. Calculons $\rho(A)$. On écrit : $A = I + \lambda B$ où B est la matrice symétrique définie négative, définie par :

$$B = \begin{pmatrix} -2 & 1 & 0 & \ldots & 0 \\ 1 & -2 & 1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & -2 & 1 \\ 0 & \ldots & 0 & 1 & -2 \end{pmatrix} \tag{2.2.8}$$

Soit $\mathcal{VP}(A)$ l'ensemble de valeurs propres de A. Alors $\mathcal{VP}(A) = \{1 + \lambda\mu, \mu \in \mathcal{VP}(B))\}$. Or $\mathcal{VP}(B) = \{-4\sin^2 \frac{j\pi}{2(N+1)}, j = 1, \ldots, N\}$ (voir Lemme 2.11 plus loin). Pour que $\varepsilon^{(n)} < \varepsilon^0$, il faut donc que :

$$\sup_{j=1,\ldots,N} |1 - 4\lambda \sin^2 \frac{j\pi}{2(N+1)}| < 1,$$

c.à.d.

$$\lambda \sin^2 \frac{j\pi}{2(N+1)} < \frac{1}{2}.$$

Une condition suffisante pour avoir une diminution de l'erreur est donc que $\lambda < \frac{1}{2}$. ∎

Lemme 2.11 (Valeurs propres de B) *L'ensemble $\mathcal{VP}(B)$ des valeurs propres de la matrice B définie par (2.2.8) est donné par :*

$$\mathcal{VP}(B) = \{-4\sin^2 \frac{j\pi}{2(N+1)}, j = 1, \ldots, N\}.$$

Démonstration : Les valeurs propres B peuvent se calculer à partir des valeurs propres de l'opérateur continu ; on commence donc par chercher u solution de :

$$\begin{cases} -u'' + \alpha u = 0, \\ u(0) = u(1) = 0. \end{cases}$$

Cherchons $u(x)$ sous la forme :

$$u(x) = a\cos\sqrt{\alpha}x + b\sin\sqrt{\alpha}x$$

Comme $u(0) = 0$, on a : a = 0. De même, $u(1) = B\sin\sqrt{\alpha} = 0$, et donc $\sqrt{\alpha} = k\pi$. Les valeurs propres et vecteurs propres associés de l'opérateur continu sont donc : $\left(k^2\pi^2, \sin k\pi x\right) \quad k \in \mathbb{N}^*$. Pour $k = 1, \ldots, N$, soit $v^{(k)} \in \mathbb{R}^N$ tel que $v_i^{(k)} = \sin k\pi i h$. Calculons $Bv^{(k)}$:

$$(Bv^{(k)})_i = v_{i-1}^{(k)} - 2v_i^{(k)} + v_{i+1}^{(k)}$$

et donc

$$(Bv^{(k)})_i = \sin k\pi(i-1)h - 2\sin k\pi i h + \sin k\pi(i+1)h$$

En développant, on obtient :
$(Bv^{(k)})_i = \sin k\pi ih\cos(-k\pi h) + \cos k\pi ih\sin(-k\pi h) - 2\sin k\pi ih + \sin k\pi ih\cos k\pi h + \cos k\pi ih\sin k\pi h.$
Après simplifications, il vient :

$$(Bv^{(k)})_i = 2\sin k\pi ih(-1+\cos k\pi h).$$

Or, $\cos k\pi h = 1 - 2\sin^2\frac{k\pi h}{2}$. On a donc :

$$\begin{aligned}(Bv^{(k)})_i &= 2\sin k\pi ih \times (-2\sin^2\frac{k\pi h}{2}) \\ &= -4\sin^2\frac{k\pi h}{2}(v^{(k)})_i, \quad \forall k = 1\ldots N.\end{aligned}$$

On a $h = \frac{1}{N+1}$, et donc les valeurs propres de B s'écrivent $\mu_k = -4\sin^2\frac{k\pi}{2(N+1)}, \quad k = 1,\ldots,N.$ ∎

2.2.6 Stabilité au sens de Von Neumann

L'analyse de stabilité au sens de Von Neumann consiste à étudier l'impact du schéma sur un mode de Fourier isolé. Pour que le mode de Fourier en question soit solution du problème continu, on remplace les conditions de Dirichlet homogènes du problème (2.1.1) par des conditions périodiques, et pour alléger les notations, on considère l'intervalle]0,2π[comme intervalle détude en espace plutôt que l'intervalle]0,1[.

Problème continu avec conditions aux limites périodiques On considère le problème avec conditions aux limites périodiques

$$\left\{\begin{array}{ll} u_t - u_{(xx)} & = 0, \quad t\in]0,T[,\, x\in]0,2\pi[, \\ u(0,t) & = u(2\pi,t), \forall t\in]0,T[, \\ u(x,0) & = u_0(x). \end{array}\right. \tag{2.2.9}$$

Le problème (2.2.9) est bien posé, au sens où $\forall u_0 \in C([0,2\pi])$, il existe une unique $u \in C^2(]0,2\pi[\times]0,T[,\mathbb{R})$ solution de (2.2.9). On suppose que $u_0 \in L^2(]0,2\pi[)$. On rappelle que L^2 est un espace de Hilbert, et que $\{e^{inx}, n \in \mathbb{Z}\}$ est une base hilbertienne[1] de $L^2(]0,2\pi[)$. On décompose donc la condition initiale dans cette base hilbertienne : $u_0(x) = \sum_{n\in\mathbb{Z}} c_n(0)e^{inx}$ (au sens de la convergence dans L^2). Dans un premier temps, calculons formellement les solutions de (2.2.9) sous la forme d'un développement dans la base hilbertienne :

$$u(x,t) = \sum_{n\in\mathbb{Z}} c_n(t)e^{inx}.$$

En supposant qu'on ait le droit de dériver terme à terme, on a donc :

$$u_t(x,t) = \sum_{n\in\mathbb{Z}} c_n'(t)e^{inx} \text{ et } u_{xx}(x,t) = \sum_{n\in\mathbb{Z}} -c_n(t)n^2e^{inx}.$$

1. Soit H,un espace de Hilbert, $(e_i)_{i\in\mathbb{Z}}$ est une base hilbertienne de H si : $(e_i)_{i\in\mathbb{Z}}$ est une famille orthonormée telle que $\forall x \in H, \exists (x_i)_{i\in\mathbb{Z}} \subset \mathbb{R}$; $x = \sum_{i\in\mathbb{Z}} x_ie_i$ au sens de la convergence dans H, avec $x_i = (x,e_i)$, où $(.,.)$ désigne le produit scalaire sur H.

On obtient, en remplaçant dans l'équation

$$c_n'(t) = -n^2 c_n(t)$$

c'est-à-dire $c_n(t) = c_n(0)e^{-n^2 t}$ en tenant compte de la condition initiale. On a donc finalement :

$$u(x,t) = \sum_{n\in\mathbb{Z}} c_n(0)e^{-n^2 t}e^{inx}. \tag{2.2.10}$$

Justifions maintenant ce calcul formel. On a :

$$\sum_{n\in\mathbb{Z}} |c_n(0)|^2 = \|u^0\|_{L^2}^2 < +\infty$$

De plus, en dérivant (2.2.10) terme à terme, on obtient :

$$u_t - u_{xx} = 0.$$

La condition de périodicité est bien vérifiée par u donnée par (2.2.10). Enfin on a bien : $u(x,t) \to u_0(t)$ lorsque $t \to 0$, donc la condition initiale est vérifiée. On peut remarquer qu'il y a "amortissement" des coefficients de Fourier $c_n(0)$ lorsque t augmente, c.à.d. qu'on a : $c_n(t) \leq c_n(0), \quad \forall\, t > 0$.

Discrétisation du problème (2.2.9) Si on utilise le schéma (2.2.2), pour la discrétisation de (2.2.9) on a :

$$u_j^{(n+1)} = (1-2\lambda)u_j^{(n)} + \lambda u_{j-1}^{(n)} + \lambda u_{j+1}^{(n)}. \tag{2.2.11}$$

On prend comme condition initiale $u^0(x) = a_p e^{ipx}$, pour $p \in \mathbb{Z}$ fixé . En discrétisant, on obtient : $u_j^0(x) = a_p e^{ipjh}$, pour $j = 1,\ldots,N$, avec $h = \frac{2\pi}{N+1}$. On a bien $u_0^0 = u_{N+1}^0 = 0$. Calculons :

$$u_j^{(1)} = (1-2\lambda)a_p e^{ipjh} + \lambda a_p e^{ip(j-1)h} + \lambda a_p e^{ip(j+1)h}$$

donc : $u_j^{(1)} = a_p e^{ipjh}\xi_p$. On appelle ξ_p le facteur d'amplification associé à la fonction e^{ipx} (appelé aussi "p-ième mode"). On a donc :

$$\begin{cases} u_j^{(1)} = \xi_p u_j^{(0)} \\ \vdots \\ u_j^{(n)} = (\xi_p)^n u_j^{(0)} \end{cases}$$

On dit que le schéma est "stable au sens de Von Neumann" : si :

$$|\xi_p| < 1, \quad \forall p.$$

Calculons ξ_p :

$$\begin{aligned} \xi_p &= 1 - 2\lambda + 2\lambda\cos ph \\ &= 1 - 2\lambda + 2\lambda(1 - 2\sin^2\frac{ph}{2}) \\ &= 1 - 4\lambda\sin^2\left(\tfrac{2\pi}{N+1}, \tfrac{p}{2}\right). \end{aligned}$$

Pour avoir $|\xi_p| < 1$, il faut $\lambda\sin^2\left(\dfrac{2\pi}{N+1}, \dfrac{p}{2}\right) < \dfrac{1}{4}$ Une condition suffisante pour que le schéma soit stable au sens de Von Neumann est que : $\lambda < \dfrac{1}{2}$. Remarquons que c'est la même condition que pour la stabilité des erreurs d'arrondis.

Convergence du schéma avec la technique de Von Neumann Soit $u \in C^2(]0,2\pi[\times]0,T[,\mathbb{R})$ la solution exacte de (2.2.9) on a $u(jh,nk) = \sum_{p\in\mathbb{Z}} c_p(0)e^{-p^2nk}e^{ipjh}$ où $h = \dfrac{2\pi}{N+1}$ est le pas de discrétisation en espace et $k = \dfrac{T}{M}$ le pas de discrétisation en temps. Soit $u_\mathcal{D}$ la solution de (2.2.2), et :

$$u_\mathcal{D}(jh,nk) = \sum_{p\in\mathbb{Z}} c_p(0)\xi_p^{(n)}e^{ipjh}.$$

On cherche à montrer la convergence de $u_\mathcal{D}$ vers u au sens suivant :

Proposition 2.12 *Soit $u_0 = \sum_{n\in\mathbb{Z}} c_n(0)e^{inx}$ et u la solution du problème (2.2.9). On note $u_\mathcal{D}$ la solution approchée obtenue par le schéma d'Euler explicite (2.2.11). Alors $\forall\varepsilon > 0$, $\exists\eta \geq 0$ tel que si $k \leq \eta$ et $\frac{k}{h^2} \leq \frac{1}{2}$, alors*

$$|u(jh,nk) - u_\mathcal{D}(jh,nk)| \leq \varepsilon, \forall j = 1\ldots N, n = \frac{T}{k}.$$

Démonstration : On note $(u-u_\mathcal{D})_j^{(n)}$ la quantité $u(jh,nk)-u_\mathcal{D}(jh,nk)$. On fera l'hypothèse supplémentaire :

$$\sum_{p\in\mathbb{Z}} |c_p(0)| < +\infty.$$

Donc pour tout $\varepsilon \in \mathbb{R}+$, il existe $A \in \mathbb{R}$ tel que $2\sum_{|p|\geq A} |c_p(0)| \leq \varepsilon$. On écrit alors :

$$(u-u_\mathcal{D})_j^{(n)} \leq \sum_{|p|\leq A} c_p(0)(e^{-p^2nk} - \xi_p^n)e^{ipjh} + \sum_{|p|\geq A} c_p(0)(e^{-p^2nk} - \xi_p^n)e^{ipjh}$$

On a donc :

$$(u-u_\mathcal{D})_j^{(n)} \leq X + 2\sum_{|p|\geq A} |c_p(0)|, \text{ avec } X = \sum_{|p|\leq A} |c_p(0)|(e^{-p^2nk} - \xi_p^n)$$

et $2\sum_{|p|\geq A} |c_p(0)| \leq 2\varepsilon$. Montrons maintenant que $X \to 0$ lorsque $h \to 0$. Remarquons que

$$e^{-p^2nk} - \xi_p^n = e^{-p^2T} - \xi_p^n, \text{ et } \xi_p = 1 - 4\lambda\sin^2\frac{ph}{2}.$$

Or, $\sin^2\dfrac{ph}{2} = \dfrac{p^2h^2}{4} + O(h^4)$, et $\lambda = \dfrac{k}{h^2}$. Donc : $4\lambda\sin^2\dfrac{ph}{2} = p^2k + O(kh^2)$. On en déduit :

$$(\xi_p)^n = \left(1 - 4\lambda\sin^2\frac{ph}{2}\right)^{T/k} \text{ et donc } \ln\xi_p^n = \frac{T}{k}\ln\left(1 - 4\lambda\sin^2\frac{ph}{2}\right) = -Tp^2 + O(h^2).$$

On en déduit que $\xi_p^n \to e^{-p^2T}$ lorsque $h \to 0$. Tous les termes de X tendent vers 0, et X est une somme finie ; on a donc montré que $(u-u_\mathcal{D})_j^{(n)}$ tend vers 0 lorsque h tend vers 0. ∎

Remarque 2.13 *On peut adapter la technique de Von Neumann au cas Dirichlet homogène sur* [0,1], *en effectuant le développement de u par rapport aux fonctions propres de l'opérateur u'' avec conditions aux limites de Dirichlet :*

$$u(x,t) = \sum c_n(t)\sin(n\pi x).$$

L'avantage du développement en série de Fourier est qu'il marche pour n'importe quel opérateur linéaire à condition d'avoir pris des conditions aux limites périodiques.

2.3 Schéma implicite et schéma de Crank-Nicolson

2.3.1 Le θ-schéma

Commençns par un petit rappel sur les 'équations différentielles (voir aussi polycopié d'analyse numérique de licence, sur le site web http://www.cmi.univ-mrs.fr/~herbin On considère le problème de Cauchy :

$$\begin{cases} y'(t) &= f(y(t)),t > 0. \\ y(0) &= y_0 \end{cases} \tag{2.3.12}$$

Soit k un pas (constant) de discrétisation , on rappelle que les schémas d'Euler explicite et implicite pour la discrétisatision de ce problème s'ecrivent respectivement :

$$\text{Euler explicite : } \frac{y^{(n+1)} - y^{(n)}}{k} = f(y^{(n)}),\ n \geq 0 \tag{2.3.13}$$

$$\text{Euler implicite : } \frac{y^{(n+1)} - y^{(n)}}{k} = f(y^{(n+1)}),\ n \geq 0, \tag{2.3.14}$$

avec $y^{(n)} = y_0$. On rappelle également que le θ-schéma, où θ est un paramètre de l'intervalle [0,1] sécrit :

$$y^{(n+1)} = y^{(n)} + k\theta f(y^{(n+1)}) + k(1-\theta)f(y^{(n)}). \tag{2.3.15}$$

Remarquons que pour $\theta = 0$ on retrouve le schéma (2.3.13) et pour $\theta = 1$ le schéma (2.3.14). On peut facilement adapter le θ schéma à la résolution des équations paraboliques. Par exemple, le θ-schéma pour la discrétisation en temps du problème (2.1.1), avec une discrétisation par différences finies en espace s'écrit :

$$\begin{array}{l} \frac{u_i^{(n+1)} - u_i^{(n)}}{k} = \frac{\theta}{h^2}(-2u_i^{(n+1)} + u_{i-1}^{(n+1)} + u_{i+1}^{(n+1)}) + \frac{1-\theta}{h^2}(-2u_i^{(n)} + u_{i-1}^{(n)} + u_{i+1}^{(n)}), ; n \geq 0,\ i = 1 \\ u_i^{(0)} = u_0(x_i), i = 1, \ldots, N. \end{array} \tag{2.3.16}$$

Si $\theta = 0$, on retrouve le schéma d'Euler explicite ; si $\theta = 1$, celui d'Euler implicite. Dans ce cas où $\theta = \dfrac{1}{2}$ ce schéma s'appelle schéma de Crank-Nicolson. Notons que dès que $\theta > 0$, le schéma est implicite, au sens où on n'a pas d'expression explicite de $u_i^{(n+1)}$ en fonction des $u_j^{(n)}$.

2.3.2 Consistance et stabilité

Proposition 2.14 (Consistance du θ-schéma) *Le θ schéma (2.3.16) pour la discrétisation du problème (2.1.1) est d'ordre 2 en espace. Il est d'ordre 2 en temps si $\theta = \dfrac{1}{2}$, et d'ordre 1 sinon.*

Démonstration : On pose $\bar{u}_j^n = u(x_j, t_n), h = \frac{1}{N+1}$,

$$R_j^{(n)} = \frac{\bar{u}_j^{(n+1)} - \bar{u}_j^{(n)}}{k} + \frac{\theta}{h^2}\left(-2\bar{u}_i^{(n+1)} + \bar{u}_{i-1}^{(n+1)} + \bar{u}_{i+1}^{(n+1)}\right) + \frac{1-\theta}{h^2}\left(-2u_i^{(n)} + u_{i-1}^{(n)} + u_{i+1}^{(n)}\right)$$

On va montrer, en effectuant des développements limités, que : $\left|R_j^{(n)}\right| \leq C(k + h^2)$ si $\theta \neq \frac{1}{2}$ et que $\left|R_j^{(n)}\right| \leq C(k^2 + h^2)$ si $\theta = \frac{1}{2}$. En effet, on décompose

$$R_j^{(n)} = T_j^{(n,1)} + \theta T_j^{(n,2)} + (1-\theta)T_j^{(n,3)}$$

avec :

$$T_j^{(n,1)} = \frac{\bar{u}_j^{(n+1)} - \bar{u}_j^{(n)}}{k},\ T_j^{(n,2)} = \frac{\theta}{h^2}\left(-2\bar{u}_i^{(n+1)} + \bar{u}_{i-1}^{(n+1)} + \bar{u}_{i+1}^{(n+1)}\right)$$
$$T_j^{(n,3)} = \frac{1-\theta}{h^2}\left(-2u_i^{(n)} + u_{i-1}^{(n)} + u_{i+1}^{(n)}\right)$$

Effectuons un développement limité pour calculer $T_j^{(n,1)}$:

$$T_j^{(n,1)} = (\bar{u}_t)(x_j,t_n) + \frac{k}{2}(u_{tt})(x_j,t_n) + R_1 \quad \text{avec } |R_1| \leq Ck^2.$$

Faisons de même pour $T_j^{(n,2)}$:

$$T_j^{(n,2)} = \theta\big(\bar{u}_{xx}(x_j,t_{n+1}) + R_2\big) \text{ avec } |R_2| \leq Ch^2.$$

Or $\bar{u}_{xx}(x_j,t_{n+1}) = \bar{u}_{xx}(x_j,t_n) + k\bar{u}_{xxt}(x_j,t_n) + R_3$ avec $|R_3| \leq Ck^2$, donc:

$$T_j^{(n,2)} = \theta\big(u_{xx}(x_j,t_n) + +ku_{xxt}(x_j,t_n) + R_4\big) \text{ avec } |R_4| \leq C(h^2+k^2).$$

De même pour $T_j^{(n,3)}$, on a :

$$T_j^{(n,3)} = (1-\theta)u_{xx}(x_j,t_n) + R_5, \text{ avec } |R_5| \leq Ck^2.$$

En regroupant, on obtient que

$$R_j^{(n)} = u_t(x_j,t_n) - u_{xx}(x_j,t_n)\frac{k}{2}\frac{\partial}{\partial t}u_t(x_j,t_n) + \theta k(u_{xx})(x_j,t_n) + R$$

avec $R = R_1 + R_4 + R_5$

• Si $\theta = \dfrac{1}{2}$, on a un schéma d'ordre 2 en temps et en espace.

En effet, $\frac{k}{2}(\bar{u}_{tt})(x_j,t_n) - \theta k(\bar{u}_{xxt})(x_j,t_n) = \frac{\partial}{\partial t}\left(k\left[\frac{1}{2}(\bar{u}_t)(x_j,t_n) - \theta(\bar{u}_{xx})(x_j,t_n)\right)\right]$ et $u_t - u_{xx} = 0$.

• Si $\theta \neq \dfrac{1}{2}$: on a un schéma d'ordre 2 en espace et d'ordre 1 en temps. ∎

Proposition 2.15 (Stabilité au sens de Von Neumann) *Si $\theta \geq \dfrac{1}{2}$ le θ-schéma est inconditionnellement stable. En particulier, les schémas d'Euler implicite et de Crank-Nicolson sont inconditionnellement stables. Si $\theta < \dfrac{1}{2}$ le schéma est stable sous condition.*

$$\lambda \leq \frac{1}{2(1-2\theta)}.$$

(On retrouve en particulier que le schéma d'Euler explicite n'est que si $\lambda \leq \frac{1}{2}$).

Démonstration : On remplace les conditions aux limites de Dirichlet sur [0,1] par des conditions périodiques sur [0,2π]. La solution exacte sécrit alors :

$$u = \sum_{p\in\mathbb{Z}} c_p(0)e^{-p^2t}e^{ipx}.$$

Pronons comme condition initiale $u_0(x) = e^{ipx}$. On a :

$$u_j^{(n+1)} - u_j^{(n)} = \frac{k}{h^2}\left[-\theta(2u_j^{(n+1)} - u_{j-1}^{(n+1)} - u_{j+1}^{(n+1)}) - (1-\theta)(2u_j^{(n)} - u_{j-1}^{(n)} - u_{j+1}^{(n)}),\right.$$

ce qui sécrit encore, avec : $\lambda = \frac{k}{h^2}$:

$$(1+2\lambda)u_j^{(n+1)} - \lambda\theta u_{j-1}^{(n+1)} - \lambda\theta u_{j+1}^{(n+1)} = (1-2\lambda(1-\theta))u_j^{(n)} + \lambda(1-\theta)u_{j+1}^{(n)} + \lambda(1-\theta)u_{j-1}^{(n)}. \quad (2.3.17)$$

En discrétisant la condition intiale (mode de Fourier) on obtient $u_j^{(0)} = e^{ipjh}$ et on cherche le facteur d'amplifcation ξ_p tel que $u_j^1 = \xi_p u_j^0 = \xi_p e^{ipjh}$; en appliquant le schéma ci–dessus pour $n = 0$, on obtient :

$$(1+2\lambda\theta)\xi_p - \lambda\theta\xi_p[e^{-iph} + e^{iph}] = [1-2\lambda(1-\theta)] + \lambda(1-\theta)[e^{iph} + e^{iph}]$$

et donc :

$$\xi_p = \frac{1-2\lambda(1-\theta)+2\lambda(1-\theta)\cos ph}{(1+2\lambda\theta)-2\lambda\cos ph} = \frac{1-4\lambda(1-\theta)\sin^2 ph/2}{1+4\lambda\theta\sin^2\frac{ph}{2}}$$

Pour que le schéma soit stable au sens de Von Neumann, il faut que : $|\xi_p| < 1$ pour tout p, soit encore :

$$1 - 4\lambda(1-\theta)\sin^2\frac{ph}{2} < 1 + 4\lambda\theta\sin^2\frac{ph}{2} \quad (2.3.18)$$

et

$$4\lambda(1-\theta)\sin^2\frac{ph}{2} - 1 < 1 + 4\lambda\theta\sin^2\frac{ph}{2} \quad (2.3.19)$$

L'inégalité (2.3.18) est toujours vérifiée. En ce qui concerne l'inégalité (2.3.19), on distingue deux cas :

1. Si $\theta \leq \frac{1}{2}$ alors $0 \leq 1-\theta \leq \theta$ et dans ce cas (2.3.19) est toujours vraie.
2. Si $\theta < \frac{1}{2}$, on veut :

$$4\lambda\left[(1-\theta)\sin^2\frac{ph}{2} - \theta\sin^2\frac{ph}{2}\right] < 2$$

Il faut donc que

$$\lambda < \frac{1}{2}\left\{(1-2\theta)\sin^2\frac{ph}{2}\right\}^{-1}$$

Une condition suffisante est donc :

$$\lambda \leq \frac{1}{2(1-2\theta)} \text{ si } \theta < \frac{1}{2}.$$

■

2.3.3 Convergence du schéma d'Euler implicite.

Prenons $\theta = 1$ dans le θ-schéma : on obtient le schéma d'Euler implicite :

$$(1+2\lambda)u_j^{(n+1)} - \lambda u_{j-1}^{(n+1)} - \lambda u_{j+1}^{(n+1)} = u_j^{(n)} \quad (2.3.20)$$

On rappelle que ce schéma est inconditionnellement stable au sens de Von Neumann. On va montrer de plus qu'il est L^∞–stable :

Proposition 2.16 (Stabilité L^∞ pour Euler implicite) *Si $(u_j^{(n)})_{j=1,\dots,N}$ est solution du schéma (2.3.20), alors :*

$$\max_{j=1,\dots,N} u_j^{(n+1)} \leq \max_{j=1,\dots,N} u_j^{(n)} \leq \max_{j=1,\dots,N} u_j^{(0)} \tag{2.3.21}$$

de même :

$$\min_{j=1,\dots,N} u_j^{(n+1)} \geq \min_{j=1,\dots,N} u_j^{(n)} \geq \min_{j=1,\dots,N} u_j^{(0)} \tag{2.3.22}$$

Le schéma (2.3.20) est donc L^∞ stable.

Démonstration : Prouvons l'estimation (2.3.21), la preuve de (2.3.22) est similaire. Soit j_0 tel que $u_{j_0}^{(n+1)} = \max_{j=1,\dots,N} u_j^{(n+1)}$ Par définition du schéma d'Euler implicite (2.3.20), On a :

$$u_{j_0}^{(n)} = (1+2\lambda)u_{j_0}^{(n+1)} - \lambda u_{j_0-1}^{(n+1)} - \lambda u_{j_0+1}^{(n+1)}.$$

On en déduit : $u_{j_0}^{(n+1)} \leq \max_{j=1,\dots,N} u_j^{(n)}$, ce qui prouve que

$$\max_{j=1,\dots,N} u_j^{(n+1)} \leq \max_{j=1,\dots,N} u_j^{(n)}.$$

Donc le schéma (2.3.20) est L^∞ stable. ∎

Théorème 2.17 *Soit $e^{(n)}$ l'erreur de discrétisation, définie par*

$$e_j^{(n)} = u(x_j,t_n) - u_j^{(n)} \text{ pour } j = 1,\dots,N.$$

Alors $\|e^{(n+1)}\|_\infty \leq \|e^{(0)}\|_\infty + TC(k+h^2)$. Si $\|e^{(0)}\|_\infty = 0$, le schéma est donc convergent (d'ordre 1 en temps et 2 en espace.

Démonstration : En utilisant la définition de l'erreur de consistance, on obtient :

$$(1+2\lambda)e_j^{(n+1)} - \lambda e_{j-1}^{(n)} - \lambda e_{j+1}^{(n)} = e_j^{(n)} + R_j^{(n)}$$

et donc :

$$\|e^{(n+1)}\|_\infty \leq \|e^h\|_\infty + kC(k+h^2)$$

On en déduit, par récurrence sur n, que :

$$\|e^{(n+1)}\|_\infty \leq \|e^0\|_\infty + TC(k+h^2)$$

d'où la convergence du schéma. ∎

On peut montrer que le schéma saute-mouton (ou "Leap-frog")

$$\frac{u_j^{(n+1)} - u_j^{(n-1)}}{2k} = \frac{1}{h^2}(u_{j-1}^{(n)} - 2u_j^{(n)} + u_{j+1}^{(n)})$$

est d'ordre 2 en espace et en temps (voir exercice 24 page 88. Malheureusement il est aussi inconditionnellement instable. On peut le modifier pour le rendre stable, en introduisant le schéma Dufort-Frankel, qui s'écrit :

$$\frac{u_j^{(n+1)} - u_j^{(n-1)}}{2k} = \frac{1}{h^2}(u_{j-1}^{(n)} - (u_j^{(n+1)} + u_j^{(n-1)}) + u_{j+1}^{(n)})$$

Ce schéma est consistant et inconditionnellement stable.

2.4 Cas de la Dimension 2

Soit Ω un ouvert borné de $\mathbb{R}^2$, on considère le problème suivant :

$$\begin{cases} u_t - \Delta u = 0 & x \in \Omega, t \in]0,T[\\ u(x,0) = u_0(x) & x \in \Omega \\ u(x,t) = g(t) & x \in \partial\Omega \quad \forall t \in]0,T[\end{cases}$$

Si le domaine est rectangulaire, ce problème se discrétise facilement à l'aide de θ schéma en temps et de différences finies en espace, en prenant un maillage rectangulaire. On peut montrer, comme dans le cas 1D, la consistance, la stabilité, la L^∞ stabilité, la stabilité au sens de Von Neumann

2.5 Exercices

Exercice 18 (Existence de solutions "presque classiques") *Corrigé en page 93*

Soit $u_0 \in L^2(\Omega)$. On s'intéresse au problème :

$$\begin{array}{l} u_t(x,t) - u_{xx}(x,t) = 0, x \in]0,1[, t \in \mathbb{R}_+^\star, \\ u(0,t) = u(1,t) = 0, t \in \mathbb{R}_+^\star, \\ u(x,0) = u_0(x), x \in]0,1[. \end{array} \tag{2.5.23}$$

1. On définit $u : [0,1] \times \mathbb{R}_+^\star \to \mathbb{R}$ par :

$$u(x,t) = \sum_{n \in \mathbb{N}^\star} e^{-n^2\pi^2 t} a_n \sin(n\pi x), x \in [0,1], t \in \mathbb{R}_+^\star, \tag{2.5.24}$$

avec $a_n = (\int_0^1 u_0(x)\sin(n\pi x)dx)/(\int_0^1 \sin^2(n\pi x)dx)$.

Montrer que u est bien définie de $[0,1] \times \mathbb{R}_+^\star$ dans $\mathbb{R}$ et est solution de (2.5.23) au sens suivant :

$$\begin{array}{l} u \in C^\infty([0,1] \times \mathbb{R}_+^\star, \mathbb{R}), \\ u_t(x,t) - u_{xx}(x,t) = 0, \forall x \in [0,1], \forall t \in \mathbb{R}_+^\star, \\ u(0,t) = u(1,t) = 0, \forall t \in \mathbb{R}_+^\star, \\ \|u(.,t) - u_0\|_{L^2(]0,1[)} \to 0, \text{ quand } t \to 0. \end{array} \tag{2.5.25}$$

2. Montrer qu'il existe une unique fonction u solution de (2.5.25).

Exercice 19 (Exemple de schéma non convergent) *Suggestions en page 92, corrigé en page 95*

Soit $u_0 \in L^2(]-4,4[)$. On note u l'unique solution (au sens vu en cours ou en un sens inspiré de l'exercice précédent) du problème suivant :

$$\begin{array}{l} u_t(x,t) - u_{xx}(x,t) = 0, x \in]-4,4[, t \in]0,1[, \\ u(-4,t) = u(4,t) = 0, t \in]0,1[, \\ u(x,0) = u_0(x), x \in]-4,4[. \end{array} \tag{2.5.26}$$

On sait que la solution de (2.5.26) est de classe C^∞ sur $[-4,4]\times]0,1]$ (voir l'exercice précédent). On admettra que si $u_0 \geq 0$ p.p. sur $]-4,4[$ et $u_0 \neq 0$ (dans $L^2(]-4,4[)$) alors $u(x,t) > 0$ pour tout $x \in]-4,4[$ et tout $t \in]0,1]$.

On suppose maintenant que $u_0 \in C([-4,4],\mathbb{R})$, $u_0(-4) = u_0(4) = 0$, $u_0 \geq 0$ sur $]-4,4[$, u_0 nulle sur $[-3,4]$ et qu'il existe $a \in]-4,-3[$ t.q. $u_0(a) > 0$. On a donc $u(x,t) > 0$ pour tout $x \in]-4,4[$.

Avec les notations du cours, on considère la solution de (2.5.26) donnée par le schéma d'Euler explicite (2.2.2) avec le pas de temps $k = 1/(M+1)$ et le pas d'espace $h = 8/(N+1)$ (M,$N \in \mathbb{N}^\star$, N impair). La solution approchée est définie par les valeurs u_i^n pour $i \in \{-(N+1)/2, \ldots, (N+1)/2\}$ et $n \in \{0, \ldots, M+1\}$. La valeur u_i^n est censée être une valeur approchée de $\overline{u}_i^n = u(ih,nk)$.

1. Donner les équations permettant de calculer u_i^n pour $i \in \{-(N+1)/2, \ldots, (N+1)/2\}$ et $n \in \{0, \ldots, M+1\}$.
2. On suppose maintenant que $k = h$. Montrer que $u_i^n = 0$ pour $i \geq 0$ et $n \in \{0, \ldots, M+1\}$. En déduire que $\max\{|u_i^{M+1} - \overline{u}_i^{M+1}|, i \in \{-(N+1)/2, \ldots, (N+1)/2\}$ ne tends pas vers 0 quand $h \to 0$ (c'est-à-dire quand $N \to \infty$).

Exercice 20 (Schémas explicites centré et décentré) *Corrigé en page 96*

Soient $\alpha > 0$, $\mu > 0$, $T > 0$ et $u_0 : \mathbb{R} \to \mathbb{R}$. On s'intéresse au problème suivant :

$$\begin{array}{l} u_t(x,t) + \alpha u_x(x,t) - \mu u_{xx}(x,t) = 0, x \in]0,1[, t \in]0,T[, \\ u(0,t) = u(1,t) = 0, t \in]0,T[, \\ u(x,0) = u_0(x), x \in]0,1[. \end{array} \tag{2.5.27}$$

On rappelle que $u_t = \frac{\partial u}{\partial t}$, $u_x = \frac{\partial u}{\partial x}$ et $u_{xx} = \frac{\partial^2 u}{\partial x^2}$. On suppose qu'il existe $u \in C^4([0,1] \times [0,T])$ solution (classique) de (2.5.27) (noter que ceci implique $u_0(0) = u_0(1) = 0$). On pose $A = \min\{u_0(x),\ x \in [0,1]\}$ et $B = \max\{u_0(x),\ x \in [0,1]\}$ (noter que $A \leq 0 \leq B$).
On discrétise le problème (2.5.27). On reprend les notations du cours. Soient $h = 1/(N+1)$ et $k = T/M$ (N, $M \in \mathbb{N}^\star$).

1. Schéma explicite décentré. Pour approcher la solution u de (2.5.27), on considère le schéma suivant :

$$\begin{array}{l} \frac{1}{k}(u_i^{n+1} - u_i^n) + \frac{\alpha}{h}(u_i^n - u_{i-1}^n) - \frac{\mu}{h^2}(u_{i+1}^n - 2u_i^n + u_{i-1}^n) = 0, \\ \qquad\qquad i \in \{1,\ldots,N\}, n \in \{0,\ldots,M-1\}, \\ u_0^n = u_{N+1}^n = 0, n \in \{1,\ldots,M\}, \\ u_i^0 = u_0(ih), i \in \{0,\ldots,N+1\}. \end{array} \tag{2.5.28}$$

   On pose $\overline{u}_i^n = u(ih,nk)$ pour $i \in \{0,\ldots,N+1\}$ et $n \in \{0,\ldots,M\}$.

   (a) (Consistance) Montrer que l'erreur de consistance du schéma (2.5.28) est majorée par $C_1(k+h)$, où C_1 ne dépend que de u, T, α et μ.
   (b) (Stabilité) Sous quelle condition sur k et h (cette condition peut dépendre de α et μ) a-t-on $A \leq u_i^n \leq B$ pour tout $i \in \{0,\ldots,N+1\}$ et tout $n \in \{0,\ldots,M\}$? Sous cette condition, en déduire $\|u^n\|_\infty \leq \|u_0\|_{L^\infty(]0,1[)}$ pour tout $n \in \{0,\ldots,M\}$ (avec $\|u^n\|_\infty = \max\{|u_i^n|,\ i \in \{0,\ldots,N+1\}\}$)
   (c) (Estimation d'erreur) On pose $e_i^n = \overline{u}_i^n - u_i^n$.
   Sous la condition sur k et h trouvée précédemment, montrer que $|e_i^n| \leq C_2(k+h)$ pour tout $i \in \{0,\ldots,N+1\}$ et tout $n \in \{0,\ldots,M\}$ avec C_2 ne dépendant que de u, T, α et μ.

2. Schéma explicite centré.
   On change dans le schéma (2.5.28) la quantité $(\alpha/h)(u_i^n - u_{i-1}^n)$ par $(\alpha/2h)(u_{i+1}^n - u_{i-1}^n)$.

   (a) (Consistance) Montrer que l'erreur de consistance est maintenant majorée par $C_3(k+h^2)$, où C_3 ne dépend que de u, T, α et μ.
   (b) Reprendre les questions de stabilité et d'estimation d'erreur du schéma (2.5.28).

Exercice 21 (Schéma implicite et principe du maximum) *Corrigé en page 98*

1. Soit $T > 0$, et $u_0 \in C([0,1])$. On considère le problème d'évolution suivant :

$$\left\{ \begin{array}{l} u_t(x,t) - u_{xx}(x,t) + v(x)u_x(x,t) = 0, x \in]0,1[,, t \in]0,T[, \\ u(0) = a_0, u(1) = a_1, \\ u(x,0) = u_0(x). \end{array} \right. \tag{2.5.29}$$

dont on cherche à approcher la solution par différences finies. On choisit pour cela le schéma de la question 1 de l'exercice **??** pour la discrétisation en espace, et on discrétise par le schéma d'Euler implicite en temps avec un pas de temps uniforme $k = \frac{T}{P}$ où $P \geq 1$.

1.1 Ecrire le schéma ainsi obtenu et montrer qu'il admet une solution qu'on notera $U = (u_i^{(p)})_{\substack{i=1,\dots,N\\p=1,\dots,P}}$, où $u_i^{(p)}$ est censé être une approximation de $u(x_i,t_p)$, où $t_p = pk, p = 0,\dots,P$.
1.2 Montrer que

$$\min(\min_{[0,1]} u_0, \min(a_0,a_1)) \leq u_i^{(p)} \leq \max(\max_{[0,1]} u_0, \max(a_0,a_1)), \text{ pour tout } i = 1,\dots,N \text{ et } p = 1,\dots,P.$$

2. Soit $T > 0$, et $u_0 \in C([0,1])$. On considère maintenant le problème d'évolution suivant :

$$\begin{cases} u_t(x,t) - u_{xx}(x,t) + (vu)_x(x,t) = 0, x \in]0,1[, t \in]0,T[, \\ u(0) = a_0, u(1) = a_1, \\ u(x,0) = u_0(x). \end{cases} \tag{2.5.30}$$

dont on cherche à approcher la solution par différences finies. On choisit pour cela le schéma de la question 1 de l'exercice 6 pour la discrétisation en espace, et on discrétise par le schéma d'Euler implicite en temps avec un pas de temps uniforme $k = \frac{T}{P}$ où $P \geq 1$.

2.1 Ecrire le schéma ainsi obtenu et montrer qu'il admet une solution qu'on notera $U = (u_i^{(p)})_{\substack{i=1,\dots,N\\p=1,\dots,P}}$, où $u_i^{(p)}$ est censé être une approximation de $u(x_i,t_p)$, où $t_p = pk, p = 0,\dots,P$.

2.2 Montrer que si $a_0 \geq 0$, $a_1 \geq 0$ et $u_0 \geq 0$, alors on a : $u_i^{(p)} \geq 0$, pour tout $i = 1,\dots,N$ et $p = 1,\dots,P$.

3. On considère maintenant $\Omega =]0,1[^2$; soient $v \in C^\infty(\Omega,\mathbb{R}_+)$, $a \in C(\partial\Omega,\mathbb{R})$ et $u_0 \in C(\Omega,\mathbb{R}_+)$. En s'inspirant des schémas étudiés aux questions 3 et 4, donner une discrétisation en espace et en temps des deux problèmes suivants (avec pas uniforme) :

$$\begin{cases} u_t - \Delta u + v \cdot \nabla u = 0, \\ u|_{\partial\Omega} = a, \\ u(\cdot,0) = u_0. \end{cases} \tag{2.5.31}$$

$$\begin{cases} u_t - \Delta u + \operatorname{div}(vu) = 0, \\ u|_{\partial\Omega} = a, \\ u(\cdot,0) = u_0. \end{cases} \tag{2.5.32}$$

Exercice 22 (Discrétisation d'un problème parabolique.) *Suggestions en page 92, corrigé en page 98*

Dans cet exercice on s'intéresse à des schémas numériques pour le problème :

$$\begin{cases} u_t + u_x - \varepsilon u_{xx} = 0 & (x,t) \in \mathbb{R}^+ \times]0,T[\\ u(1,t) = u(0,t) = 0 & t \in]0,T[\\ u(x,0) = u_0(x) & x \in]0,1[\end{cases} \tag{2.5.33}$$

où u_0 et ε sont donnés ($\varepsilon > 0$). On reprend dans la suite de l'exercice les notations du cours.
1. Donner un schéma d'approximation de (2.5.33) différences finies à pas constant en espace et Euler explicite à pas constant en temps. Montrer que l'erreur de consistance est majorée par $C(k + h^2)$; avec C dépendant de la solution exacte de (2.5.33). Sous quelle(s) condition(s) sur k et h a-t'on $\|u^n\|_\infty \leq \|u^0\|_\infty, \forall n \leq M$?
Donner un résultat de convergence pour ce schéma.

2. Même question que 1. en remplaçant Euler explicite par Euler implicite.
3. En s'inspirant du schéma de Crank-Nicolson (vu en cours) construire un schéma d'ordre 2 (espace et temps). Sous quelle(s) condition(s) sur k et h a-t'on $\|u^n\|_2 \leq \|u^0\|_2, \forall n \leq M$? Donner un résultat de convergence pour ce schéma.
4. Dans les schémas trouvés aux questions 1., 2. et 3. on remplace l'approximation de u_x par une approximation décentrée à gauche. Quel est l'ordre des schémas obtenus et sous quelle(s) condition(s) sur k et h a-t'on $\|u^n\|_\infty \leq \|u^0\|_\infty$ où $\|u\|u^0\|_2, \forall n \leq M$? Donner un résultat de convergence pour ces schémas.

Exercice 23 (Equation parabolique avec terme source) *Suggestions en page 92, corrigé 101*

Soit u_0 une fonction donnée de [0,1] dans $\mathbb{R}$. On s'intéresse ici à la discrétisation du problème suivant :

$$u_t(t,x) - u_{xx}(t,x) - u(t,x) = 0, t \in \mathbb{R}^+, x \in [0,1], \tag{2.5.34}$$

$$u(t,0) = u(t,1) = 0, t \in \mathbb{R}_+^\star \, ; \, u(0,x) = u_0(x), x \in [0,1]. \tag{2.5.35}$$

On note u la solution de (2.5.34), (2.5.35), et on suppose que u est la restriction à $\mathbb{R}_+ \times [0,1]$ d'une fonction de classe C^∞ de $\mathbb{R}^2$ dans $\mathbb{R}$.
Pour $h = \frac{1}{N+1}$ ($N \in \mathbb{N}^\star$) et $k > 0$, on pose $x_i = ih, i \in \{0, \ldots, N+1\}$, $t_n = nk, n \in \mathbb{N}$, $\overline{u}_i^n = u(x_i, t_n)$, et on note u_i^n la valeur approchée recherchée de $\overline{u}_i^n$.
On considère deux schémas numériques, (2.5.36)–(2.5.38) et (2.5.37)–(2.5.38) définis par les équations suivantes :

$$\frac{u_i^{n+1} - u_i^n}{k} - \frac{(u_{i+1}^{n+1} + u_{i-1}^{n+1} - 2u_i^{n+1})}{h^2} - u_i^{n+1} = 0, n \in \mathbb{N}, i \in \{1, \ldots, N\}, \tag{2.5.36}$$

$$\frac{u_i^{n+1} - u_i^n}{k} - \frac{(u_{i+1}^{n+1} + u_{i-1}^{n+1} - 2u_i^{n+1})}{h^2} - u_i^{n} = 0, n \in \mathbb{N}, i \in \{1, \ldots, N\}, \tag{2.5.37}$$

$$u_0^{n+1} = u_{N+1}^{n+1} = 0, n \in \mathbb{N} \, ; \, u_i^0 = u_0(x_i), = \, i \in \{0, \ldots, N+1\}. \tag{2.5.38}$$

Pour $n \in N$, on note $u^n = (u_1^n, \ldots, u_N^n)^t \in \mathbb{R}^N$.

1. (Consistance) Soit $T > 0$. Pour $n \in \mathbb{N}$, et $i \in \{1, \ldots, N\}$, on note R_i^n l'erreur de consistance (définie en cours) du schéma numérique (2.5.36), (2.5.38) [resp. du schéma numérique (2.5.37), (2.5.38)]. Montrer qu'il existe $C \in \mathbb{R}$, ne dépendant que de u et T, t. q. $|R_i^n| \leq C(k + h^2)$, pour tout $i \in \{1, \ldots, N\}$ et tout $n \in \mathbb{N}$, t.q. $kn \leq T$.
2. Montrer que le schéma (2.5.36), (2.5.38) [resp. (2.5.37), (2.5.38)] demande, à chaque pas de temps, la résolution du système linéaire $Au^{n+1} = a$ [resp. $Bu^{n+1} = b$] avec $A, B \in \mathbb{R}^{N,N}$ et $a, b \in \mathbb{R}^N$ à déterminer.
   Montrer que B est inversible (et même s.d.p.) pour tout $h > 0$ et $k > 0$. Montrer que A est inversible (et même s.d.p.) pour tout $h > 0$ et $k \in]0,1[$.
3. (Stabilité) Pour $n \in \mathbb{N}$, on pose $\|u^n\|_\infty = \sup_{i \in \{1, \ldots, N\}} |u_i^n|$. Soit $T > 0$. On considère le schéma (2.5.37), (2.5.38). Montrer qu'il existe $C_1(T) \in \mathbb{R}$, ne dépendant que de T, t.q. $\|u^n\|_\infty \leq C_1(T)\|u_0\|_\infty$, pour tout $h > 0$, $k > 0$, et $n \in \mathbb{N}$ tel que $kn \leq T$.
   Soit $\alpha \in [0,1]$. On considère le schéma (2.5.36), (2.5.38). Montrer qu'il existe $C_2(T, \alpha) \in \mathbb{R}$, ne dépendant que de T et de α, t.q. $\|u^n\|_\infty \leq C_2(T,\alpha)\|u_0\|_\infty$, pour tout $h > 0$, $k \in]0,\alpha[$, et $n \in \mathbb{N}$ tel que $kn \leq T$.

4. (Estimation d'erreur) Pour $n \in \mathbb{N}$ et $i \in \{1,\ldots,N\}$, on pose $e_i^n = \overline{u}_i^n - u_i^n$. Soit $T > 0$. Donner, pour $kn \leq T$, des majorations de $\|e^n\|_\infty$ en fonction de T, C, $C_1(T)$, $C_2(T,\alpha)$ (définis dans les questions précédentes), k et h pour les deux schémas étudiés.

Exercice 24 (Schéma "saute-mouton") *Corrigé en page 104*

On considère le problème suivant:

$$\left\{\begin{array}{lcl} u_t(x,t) - u_{xx}(x,t) & = & 0, x \in]0,1[, t \in]0,T[, \\ u(0,t) = u(1,t) & = & 0, t \in]0,T[, \\ u(x,0) & = & u_0(x), x \in]0,1[. \end{array}\right. \tag{2.5.39}$$

Pour trouver une solution approchée de ((2.5.39)), on considère le schéma "saute-mouton" :

$$\left\{\begin{array}{l} \dfrac{u_j^{n+1} - u_j^{(n-1)}}{2k} = \dfrac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{h^2}, j = 1,\ldots,N-1, n = 1,\ldots,M-1, \\ u_0^{n+1} = u_{N+1}^{n+1} = 0, n = 1,\ldots,M-1, \end{array}\right. \tag{2.5.40}$$

où $(u_j^0)_{j=1,\ldots,N}$ et $(u_j^1)_{j=1,\ldots,N}$ sont supposés connus, $h = 1/N$, $k = T/M$.

1. Montrer que le schéma (2.5.40) est consistant. Quel est son ordre?
2. Montrer que le schéma (2.5.40) est inconditionnellement instable au sens de Von Neumann.

On modifie "légèrement" le schéma (2.5.40) en prenant

$$\left\{\begin{array}{l} \dfrac{u_j^{n+1} - u_j^{(n-1)}}{2k} = \dfrac{u_{j-1}^n - (u_j^{n+1} + u_j^{(n-1)}) + u_{j+1}^n}{h^2}, j = 1,\ldots,N, n = 1,\ldots,M-1, \\ u_0^{n+1} = u_{N+1}^{n+1} = 0, n = 1,\ldots,M-1, \end{array}\right. \tag{2.5.41}$$

(schéma de Dufort-Frankel).

3. Montrer que le schéma (2.5.41) est consistant avec (2.5.39) quand $h,k \to 0$ sous la condition $\frac{k}{h} \to 0$.

4. Montrer que (2.5.41) est inconditionnellement stable.

Exercice 25 (Problème parabolique non linéaire) *Corrigé en page 106*

On se propose, dans cet exercice, de montrer l'existence d'une solution faible au problème (2.5.42)-(2.5.44), à partir de l'existence de la solution approchée donnée par un schéma numérique. L'inconnue de ce problème est la fonction u de $[0,1] \times [0,T]$ dans $\mathbb{R}$, elle doit être solution des équations suivantes :

$$\frac{\partial u}{\partial t}(x,t) - \frac{\partial^2 \varphi(u)}{\partial x^2}(x,t) = v(x,t), x \in]0,1[, t \in]0,T[, \tag{2.5.42}$$

$$\frac{\partial \varphi(u)}{\partial x}(0,t) = \frac{\partial \varphi(u)}{\partial x}(1,t) = 0, t \in]0,T[, \tag{2.5.43}$$

$$u(x,0) = u_0(x), x \in]0,1[, \tag{2.5.44}$$

où φ, v, T, u_0 sont donnés et sont t.q.

1. $T > 0$, $v \in L^\infty(]0,1[\times]0,T[)$,

2. φ croissante, lipschitzienne de $\mathbb{R}$ dans $\mathbb{R}$,
3. $u_0 \in L^\infty(]0,1[)$ et $\varphi(u_0)$ lipschitzienne de $[0,1]$ dans $\mathbb{R}$.

Un exemple important est donné par $\varphi(s) = \alpha_1 s$ si $s \leq 0$, $\varphi(s) = 0$ si $0 \leq s \leq L$ et $\varphi(s) = \alpha_2(s-L)$ si $s \geq L$, avec α_1,α_2 et L donnés dans $\mathbb{R}_+^\star$. Noter pour cet exemple que $\varphi' = 0$ sur $]0,L[$.

Les ensembles $]0,1[$ et $D =]0,1[\times]0,T[$ sont munis de leur tribu borélienne et de la mesure de Lebesgue sur cette tribu.

On appelle "solution faible" de (2.5.42)-(2.5.44) une solution de :

$$u \in L^\infty(]0,1[\times]0,T[), \tag{2.5.45}$$

$$\int_D (u(x,t)\frac{\partial \psi}{\partial t}(x,t) + \varphi(u(x,t))\frac{\partial^2 \psi}{\partial x^2}(x,t) + v(x,t)\psi(x,t))dxdt + \int_{]0,1[} u_0(x)\psi(x,0)dx = 0, \tag{2.5.46}$$

$\forall \psi \in C_T^{2,1}(\mathbb{R}^2)$,

où $\psi \in C_T^{2,1}(\mathbb{R}^2)$ signifie que ψ est une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$ deux fois continûment dérivable par rapport à x, une fois continûment dérivable par rapport à t et t.q. $\dfrac{\partial \psi}{\partial x}(0,t) = \dfrac{\partial \psi}{\partial x}(1,t) = 0$, pour tout $t \in [0,T]$ et $\psi(x,T) = 0$ pour tout $x \in [0,1]\}$.

Question 1 (Solution classique versus solution faible)

On suppose, dans cette question seulement, que φ est de classe C^2, v est continue sur $[0,1] \times [0,T]$ et u_0 est continue sur $[0,1]$. Soit $u \in C^2(\mathbb{R}^2,\mathbb{R})$. On note encore u la restriction de u à $]0,1[\times]0,T[$. Montrer que u est solution de (2.5.45)-(2.5.46) si et seulement si u vérifie (2.5.42)-(2.5.44) au sens classique (c'est-à-dire pour tout $(x,t) \in [0,1] \times [0,T]$).

On cherche maintenant une solution approchée de (2.5.42)-(2.5.44).

Soient $N,M \in \mathbb{N}^\star$. On pose $h = \dfrac{1}{N}$ et $k = \dfrac{T}{M}$. On va construire une solution approchée de (2.5.42)-(2.5.44) à partir de la famille $\{u_i^n,\ i = 1,\ldots,N,\ n = 0,\ldots,M\}$ (dont on va prouver l'existence et l'unicité) vérifiant les équations suivantes :

$$u_i^0 = \frac{1}{h}\int_{(i-1)h}^{ih} u_0(x)dx, i = 1,\ldots,N, \tag{2.5.47}$$

$$\frac{u_i^{n+1} - u_i^n}{k} - \frac{\varphi(u_{i-1}^{n+1}) - 2\varphi(u_i^{n+1}) + \varphi(u_{i+1}^{n+1})}{h^2} = v_i^n, i = 1,\ldots,N, n = 0,\ldots,M-1, \tag{2.5.48}$$

avec $u_0^{n+1} = u_1^{n+1}$, $u_{N+1}^{n+1} = u_N^{n+1}$, pour tout $n = 0,\ldots,M-1$ et $v_i^n = \dfrac{1}{kh}\displaystyle\int_{nk}^{(n+1)k}\int_{(i-1)h}^{ih} v(x,t)dxdt$, pour tout $i = 1,\ldots,N$, pour tout $n = 0,\ldots,M$.

Question 2 (Existence et unicité de la solution approchée)

Soit $n \in \{0,\ldots,M-1\}$. On suppose connu $\{u_i^n,\ i = 1,\ldots,N\}$. On va prouver dans cette question l'existence et l'unicité de $\{u_i^{n+1},\ i = 1,\ldots,N\}$ vérifiant (2.5.48) (avec $u_0^{n+1} = u_1^{n+1}$, $u_{N+1}^{n+1} = u_N^{n+1}$).

1. Soit $a > 0$, Pour $s \in \mathbb{R}$, on pose $g_a(s) = s + a\varphi(s)$. Montrer que g_a est une application strictement croissante bijective de $\mathbb{R}$ dans $\mathbb{R}$.

2. Soit $\overline{w} = (\overline{w}_i)_{i=1,\dots,N} \in \mathbb{R}^N$. On pose $\overline{w}_0 = \overline{w}_1$ et$\overline{w}_{N+1} = \overline{w}_N$. Montrer qu'il existe un et un seul couple $(u,w) \in \mathbb{R}^N \times \mathbb{R}^N$, $u = (u_i)_{i=1,\dots,N}$, $w = (w_i)_{i=1,\dots,N}$, t.q. :

$$\varphi(u_i) = w_i, \text{ pour tout } i \in \{1,\dots,N\}, \tag{2.5.49}$$

$$u_i + \frac{2k}{h^2} w_i = \frac{k}{h^2}(\overline{w}_{i-1} + \overline{w}_{i+1}) + u_i^n + k v_i^n, \text{ pour tout } i = 1,\dots,N. \tag{2.5.50}$$

On peut donc définir une application F de $\mathbb{R}^N$ dans $\mathbb{R}^N$ par $\overline{w} \mapsto F(\overline{w}) = w$ où w est solution de (2.5.49)–(2.5.50).

3. On munit $\mathbb{R}^N$ de la norme usuelle $\|\cdot\|_\infty$. Montrer que l'application F est strictement contractante. [On pourra utiliser la monotonie de φ et remarquer que, si $a = \varphi(\alpha)$ et $b = \varphi(\beta)$, on a $|\alpha - \beta| \geq (1/L)|a - b|$, où L ne dépend que de φ.]

4. Soit $\{u_i^{n+1},\ i = 1,\dots,N\}$ solution de (2.5.48). On pose $w = (w_i)_{i=1,\dots,N}$, avec $w_i = \varphi(u_i^{n+1})$ pour $i \in \{1,\dots,N\}$. Montrer que $w = F(w)$.

5. Soit $w = (w_i)_{i=1,\dots,N}$ t.q. $w = F(w)$. Montrer que pour tout $i \in \{1,\dots,N\}$ il existe $u_i^{n+1} \in \mathbb{R}$ t.q. $w_i = \varphi(u_i^{n+1}$. Montrer que $\{u_i^{n+1},\ i = 1,\dots,N\}$ est solution de (2.5.48).

6. Montrer qu'il existe une unique famille $\{u_i^{n+1},\ i = 1,\dots,N\}$ solution de (2.5.48).

Question 3 (Estimation $L^\infty(]0,1[\times]0,T[)$ sur u)

On pose $A = \|u_0\|_{L^\infty(]0,1[)}$ et $B = \|v\|_{L^\infty(]0,1[\times]0,T[)}$. Montrer, par récurrence sur n, que $u_i^n \in [-A - nkB, A + nkB]$ pour tout $i = 1,\dots,N$ et tout $n = 0,\dots,M$. [On pourra, par exemple, considérer (2.5.48) avec i t.q. $u_i^{n+1} = \min\{u_j^{n+1},\ j = 1,\dots,N\}$.]

En déduire qu'il existe $c_{u_0,v,T} \in \mathbb{R}_+$ t.q. $\|u^n\|_{L^\infty(]0,1[)} \leq c_{u_0,v,T}$.

Question 4 (Estimation de la dérivée p.r. à x de $\varphi(u)$)

Montrer qu'il existe C_1 (ne dépendant que de T, φ, v et u_0) t.q., pour tout $n = 0,\dots,M-1$,

$$\sum_{n=0}^{M-1} \sum_{i=1}^{N-1} (\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1}))^2 \leq C_1 \frac{h}{k}. \tag{2.5.51}$$

[Multiplier (2.5.48) par u_i^{n+1} et sommer sur i et sur n et utiliser l'inégalité $a^2 - ab \geq \frac{a^2}{2} - \frac{b^2}{2}$.]

Question 5 (Estimation de la dérivée p.r. à t de $\varphi(u)$)

Montrer qu'il existe C_2 (ne dépendant que de T, φ, v et u_0) t.q.

$$\sum_{n=0}^{M-1} h \sum_{i=0}^{N+1} (\varphi(u_{i+1}^n) - \varphi(u_{i+1}^{n+1}))^2 \leq C_2 k. \tag{2.5.52}$$

et

$$\sum_{i=0}^{N+1} (\varphi(u_i^{n+1}) - \varphi(u_{i+1}^{n+1}))^2 \leq C_2 h, \text{ pour tout } n \in \{0,\dots,M\}. \tag{2.5.53}$$

[indication: multiplier (2.5.48) par $\varphi(u_i^{n+1}) - \varphi(u_i^n)$ et sommer sur i et n]

Dans la suite de l'exercice, il s'agit de passer à la limite (quand N,$M \to \infty$) pour trouver une solution de (2.5.42)-(2.5.44).

Pour $M \in \mathbb{N}^\star$ donné, on prend $N = M^2$ (et donc h et k sont donnés et $k = T\sqrt{h}$), on définit (avec les u_i^n trouvés dans les questions précédentes) une fonction, u_h, sur [0,1] $\times$ [0,T] en posant

$$u_h(x,t) = \frac{t-nk}{k} u_h^{(n+1)}(x) + \frac{(n+1)k-t}{k} u_h^{(n)}(x), \text{ si } t \in [nk,(n+1)k]$$

et

$$u_h^{(n)}(x) = u_i^n, \text{ si } x \in](i-1)h,ih[, i = 1,\ldots,N, n = 0,\ldots,M.$$

Enfin, on définit $\varphi(u_h)$ par $\varphi(u_h)(x,t) = \varphi(u_h(x,t))$.

Question 6 Montrer que les suites $(u_h)_{M\in\mathbb{N}^\star}$ et $(\varphi(u_h))_{M\in\mathbb{N}^\star}$ sont bornées dans $L^\infty(]0,1[\times]0,T[)$ (on rappelle que h est donné par M).

Question 7
Montrer qu'il existe C (ne dépendant que de T, φ, v et u_0) t.q. l'on ait, pour tout $M \in \mathbb{N}^\star$:

1. Pour tout $t \in [0,T]$,

$$\int_{\mathbb{R}} |\varphi(u_h)(x+\eta,t) - \varphi(u_h)(x,t)|^2 dx \leq C\eta,$$

   pour tout $\eta \in \mathbb{R}_+^\star$, avec $\varphi(u_h)(\cdot,t)$ prolongée par 0 hors de [0,1].
2. $\|\varphi(u_h)(\cdot,t) - \varphi(u_h)(\cdot,s)\|_{L^2(]0,1[)} \leq C|t-s|$, pour tout t,$s \in [0,T]$.

Une conséquence des questions 6 et 7 (que l'on admet ici est que l'on peut trouver une suite $(h_n)_{n\in\mathbb{N}}$ et $u \in L^\infty(]0,1[\times]0,T[)$ telle que, en posant $u_n = u_{h_n}$ (on rappelle que $k_n = T\sqrt{h_n}$), l'on ait, quand $n \to \infty$,

1. $h_n \to 0$ et $k_n \to 0$,
2. $u_n \to u$ dans $L^\infty(]0,1[\times]0,T[)$ pour la topologie faible-$\star$,
3. $\varphi(u_n) \to \varphi(u)$ dans $L^p(]0,1[\times]0,T[)$, pour tout $p \in [1,\infty[$.

Question 8 Montrer que la fonction u ainsi trouvée est solution de (2.5.45),(2.5.46).

Remarque. On peut aussi montrer l'unicité de la solution de (2.5.45),(2.5.46).

2.6 Suggestions pour les exercices

Exercice 19 (Exemple de schéma non convergent)

1. Ecrire le schéma d'Euler explicite.
2. Démontrer par récurrence que

$$\text{Si } n \in \{0,\dots,M+1\}, i \in \left\{-\frac{N+1}{2},\dots,\frac{N+1}{2}\right\} \text{ et } \quad i \geq -\frac{N+1}{4} + n \text{ alors } u_i^n = 0.$$

En déduire que $u_i^n = 0$ pour $n \in \{0,\dots,M+1\}$ et $i \in \left\{0,\dots,\frac{N+1}{2}\right\}$ et conclure.

Exercice 22 (Discrétisation d'un problème parabolique)

1. Calculer l'erreur de consistance et la majorer par des développements de Taylor.
Chercher ensuite les conditions pour que :

$$\|u^n\|_\infty \leq \|u^0\|_\infty.$$

Pour étudier la convergence du schéma, majorer l'erreur de discrétisation : $e_j^n = \bar{u}_j^n - u_j^n$ où u_j^n est calculé par (2.7.64), et $\bar{u}_j^n$ est la solution du problème (2.5.33) en $x_j = jh$ et $t_n = nk$.

Même chose pour les questions suivantes...

Exercice 23 (Problème parabolique avec terme source)

1. Effectuer des développements de Taylor...

3. Montrer par récurrence que $\max_{j=1,\dots,N} u_j^n \leq (1+k)^n \max_{j=1,\dots,N} u_j^0$. et que $\min_{j=1,\dots,N} u_j^{(n)} \geq (1+k)^n \min_{j=1,\dots,N} u_j^{(0)}$.
4. Utiliser l'équation, le schéma, et l'erreur de consistance.

Exercice 24 (Schémas de Saute-Mouton et Dufort-Frankel)

1. Effectuer des développements de Tyalor pour majorer l'erreur de consistance.
2. Montrer que le facteur d'amplification ξ_n obtenu par l'analyse de stabilité de Von Neumann satisfait :

$$\xi_{n+1} - \alpha\xi_n - \xi_{n-1} = 0,\ n \geq 2.$$

Etudier ensuite les racines de léquation $r^2 - \alpha r - 1 = 0$ et montrer que l'une de ses racines est, en module, supérieure à 1.

4. Reprendre la méthode développée à la question 2, en montrant que l'équation caractéristique pour ξ est maintenant :

$$p(r) = ar^2 + br + c = 0,$$

avec

$$a = \frac{1}{2k} + \frac{1}{h^2}, b = -\frac{2\cos(ph)}{h^2} \text{ et } c = \frac{1}{h^2} - \frac{1}{2h}.$$

Etudier ensuite les racines de cette équation.

2.7 Corrigés des exercices

Corrigé de l'exercice 18 page 84

On note $\|\cdot\|_2 = \|\cdot\|_{L^2(]0,1[)}$.
1) Pour $n \in \mathbb{N}^*$, on a

$$\int_0^1 \sin^2(n\pi x)dx = \int_0^1 \frac{1-\cos(2n\pi x)}{2}dx = \frac{1}{2},$$

et

$$\int_0^1 |u_0(x)\sin(n\pi x)|dx \leq \|u_0\|_2 \left(\int_0^1 \sin^2(n\pi x)dx\right)^{1/2} = \frac{r_2}{2}\|u_0\|_2.$$

La quantité a_n est donc bien définie et

$$|a_n| \leq r_2\|u_0\|_2$$

Pour tout $t > 0$ et $x \in [0,1]$, on a

$$|e^{-n^2\pi^2t^2}a_n\sin(n\pi x)| \leq r_2\|u_0\|_2 e^{-n^2\pi^2t^2} \qquad \forall n \in \mathbb{N}^*.$$

Ceci montre que la série $\sum_{n>0} e^{-n^2\pi^2t^2}a_n\sin(n\pi x)$ est absolument convergente et donc que u est bien définie pour tout $t > 0$ et tout $x \in [0,1]$ et même pour tout $x \in \mathbb{R}$.
On remarque ensuite que u est de classe C^∞ sur $\mathbb{R} \times \mathbb{R}_+^*$, en appliquant les théorèmes classiques de dérivation terme à terme d'une série. En effet, soit $\varepsilon > 0$, pour tout $x \in \mathbb{R}$ et $t > \varepsilon$ on a

$$\left|e^{-n^2\pi^2t^2}a_n\sin(n\pi x)\right| \leq r_2\|u_0\|_2 e^{-n^2\pi^2\varepsilon^2}, \forall n \in \mathbb{N}^*$$

Comme $(x,t) \to e^{-n^2\pi^2t^2}a_n\sin(n\pi t)$ est continue (pour tout $n \in \mathbb{N}^*$), on en déduit que u est continue sur $\mathbb{R}\times]\varepsilon,\infty[$, et finalement sur $\mathbb{R}\times]0,\infty[$ car $\varepsilon > 0$ est arbitraire.
Pour dériver terme à terme la série définissant u, il suffit également d'obtenir sur $]\varepsilon,\infty[\times\mathbb{R}$ (pour tout $\varepsilon > 0$) une majoration du terme général de la série des dérivées par le terme général d'une série convergente (indépendant de $(x,t) \in \mathbb{R}\times]\varepsilon,\infty[$. On obtient cette majoration en remarquant que, pour $(x,t) \in \mathbb{R}\times]\varepsilon,\infty[$,

$$|-n^2\pi^2e^{-n^2\pi^2t^2}a_n\sin(n\pi x)| \leq n^2\pi^2e^{-n^2\pi^2\varepsilon^2}r_2\|u_0\|_2$$

On montre ainsi finalement que u est de classe C^1 par rapport à t et que

$$u_t(x,t) = \sum_{n>0} -n^2\pi^2e^{-n^2\pi^2t^2}a_n\sin(n\pi x), x \in \mathbb{R}, t > 0.$$

En itérant ce raisonnement on montre que u est de classe C^∞ par rapport à t sur $\mathbb{R} \times \mathbb{R}_+^*$.
Un raisonnement similaire montre que u est de classe C^∞ par rapport à x sur $\mathbb{R} \times \mathbb{R}_+^*$ et que l'on peut dériver terme à terme la série définissant u. On obtient donc aussi

$$u_{xx}(xt) = \sum_{n>0} -n^2\pi^2e^{-n^2\pi^2t^2}a_n\sin(n\pi x), x \in \mathbb{R}, t > 0,$$

et ceci donne $u_t = u_{xx}$ sur $\mathbb{R} \times \mathbb{R}_t^*$ et donc aussi un $[0,1] \times \mathbb{R}_+^*$. Le fait que $u(0,t) = u(1,t)$ pour tout $t > 0$ est immédiat car $\sin n\pi t = \sin 0 = 0$, pour tout $n \in \mathbb{N}^*$.
Il reste à montrer que $u(.,t) \to u_0$ dans $L^2(]0,1[)$ quand $t \to 0$.

On définit $e_n \in L^2(]0,1[)$ par $e_n(x) = \sqrt{2}\sin(n\pi x)$. La famille $\{e_n, n \in \mathbb{N}^*\}$ est une base hilbertienne de $L^2(]0,1[)$.
On a donc :

$$\sum_{n=1}^{N} a_n \sin n\pi x \to u_0, \text{ dans } L^2(]0,1[), \text{ quand } n \to \infty,$$

et

$$\sum_{n=1}^{\infty} a_n^2 = 2\|u_0\|_2^2.$$

On remarque maintenant que

$$u(x,t) - u_0(x) = u(x,t) - u^{(N)}(x,t) + u^{(N)}(xt) - u_0^{(N)}(x) - u_0^{(N)}(x) - u_0(x),$$

avec

$$u^{(N)}(x,t) = \sum_{n=1}^{N} a_n e^{-n^2\pi^2 t^2} \sin(n\pi x)$$

$$u_0^{(N)}(x) = \sum_{n=1}^{N} a_n \sin(n\pi x).$$

Il est clair que, pour tout $N \in \mathbb{N}^*$, on a $u^{(N)}(.,t) \to u_0^{(N)}$ uniformément sur $\mathbb{R}$, quand $N \to \infty$, et donc $u^{(N)}(.,t) \to u_0^{(N)}$ dans $L^2(]0,1[)$.
Comme

$$\|u(.,t) - u^{(N)}(.,t)\|_2^2 = \sum_{n=N+1}^{\infty} a_n^2 \frac{1}{2} e^{-2n^2\pi^2 t^2} \leq \sum_{n=N+1}^{\infty} a_n^2 \frac{1}{2} = \|u_0^{(N)} - u_0\|_2^2 \to 0$$

quand $N \to \infty$, on en déduit que $u(.,t) \to u_0, qdt \to 0$, dans $L^2(]0,1[)$.
2) On note w la différence de 2 solutions de (2.5.25). On a donc

$$\begin{array}{l} w \in C^\infty([0,1] \times \mathbb{R}_+^*, \mathbb{R}) \\ w_t - w_{xx} = 0 \text{ sur } [0,1] \times \mathbb{R}_+^* \\ w(0,t) = w(1,t) = 0 \text{ pour } t > 0 \\ w(.,t) \to 0, \text{ dans } L^2(]0,1[), \text{ quand } t \to 0 \end{array}$$

Soit $0 < \varepsilon < T < \infty$. On intègre l'équation $ww_t - ww_{xx} = 0$ sur $]0,1[\times]\varepsilon,T[$. En utilisant une intégration par parties (noter que $w \in C^\infty([0,1] \times [\varepsilon,T])$, on obtient :

$$\frac{1}{2}\int_0^1 w^2(x,T)dx - \frac{1}{2}\int_0^1 w^2(x,\varepsilon).dx + \int_0^1 \int_\varepsilon^T w_x^2(x,t)dxdt = 0.$$

D'où l'on déduit $\|w(.,T)\|_2 \leq \|w(.,\varepsilon)\|_2$. Quand $\varepsilon \to 0$, on a $\|w(.,\varepsilon)\|_2 \to 0$, on a donc $\|w(.,T)\|_2 = 0$ et donc, comme $w(.,t)$ est contenue sur $[0,1], wX \in [0,1]$. Comme $T > 0$ est arbitraire, on a finalement

$$w(x,t) = 0 \qquad \forall t \in [0,1], \forall t > 0$$

Ce qui montre bien l'unicité de la solution de (2.5.25).

Corrigé de l'exercice 19 page 84

1) La formule pour calculer u_i^0 est :

$$u_1^0 = u_0(ih,0), \quad i = -\frac{N+1}{2}, \ldots, \frac{N+1}{2}$$

Soit maintenant $n \in \{0, \ldots, M\}$. On a :

$$u_i^{n+1} = 0 \qquad \text{pour} \qquad i = -\frac{N+1}{2} \quad \text{et} \quad i = \frac{N+1}{2}$$

$$u_i^{n+1} = u_i^n + \frac{k}{h^2}\left(u_{i+1}^n + u_{i-1}^n - 2u_i^n\right), \quad i = -\frac{N+1}{2} + 1, \ldots, \frac{N+1}{2} - 1.$$

2) On va montrer, par récurrence (finie) sur n, que :

$$\text{Si } n \in \{0, \ldots, M+1\}, i \in \left\{-\frac{N+1}{2}, \ldots, \frac{N+1}{2}\right\} \text{ et } \quad i \geq -\frac{N+1}{4} + n \text{ alors } u_i^n = 0. \tag{2.7.54}$$

Pour initialiser la récurrence, on suppose que $n = 0$ et $i \geq -\dfrac{N+1}{4}$. On a alors

$$ih \geq -\frac{N+1}{4} \quad \frac{8}{N+1} = -2 > -3$$

et donc $u_i^0 = 0$.
Soit maintenant $n \in \{0, \ldots, M\}$. On suppose que l'hypothèse de récurrence est vérifiée jusqu'au rang n, et on démontre la propriété au rang $n+1$. Soit donc $i \in \left\{-\frac{N+1}{2}, \ldots, \frac{N+1}{2}\right\}$ tel que $i \geq -\frac{N+1}{4} + (n+1)$. Alors :

- Si $i = \frac{N+1}{2}$ on a bien $u_i^{N+1} = 0$.
- Si $i < \frac{N+1}{2}$, les indices $i-1$, i et $i+1$ sont tous supérieurs ou égaux à $-\frac{N+1}{4} + n$, et donc par hypothèse de récurrence,

$$u_i^{n+1} = u_i^n \left(1 - \frac{2k}{h^2}\right) + \frac{k}{h^2} u_{i+1}^n + \frac{k}{h^2} u_{i-1}^n = 0.$$

On a donc bien démontré (2.7.54). On utilise maintenant l'hypothèse $k = h$, c'est–à–dire $\frac{1}{M+1} = \frac{8}{N+1}$. On a alors

$$-\frac{N+1}{4} + M + 1 = -2(M+1) + M + 1 = -(M+1) < 0.$$

On en déduit que si $n \in \{0, \ldots, M+1\}$ et $i \geq 0$, alors $i \geq -\frac{N+1}{4} + n$. On en déduit que $u_i^n = 0$ pour $n \in \{0, \ldots, M+1\}$ et $i \in \left\{0, \ldots, \frac{N+1}{2}\right\}$. On remarque alors que

$$\begin{aligned} \max\left\{|u_i^{M+1} - \bar{u}_i^{M+1}|, i \in \left\{-\frac{N+1}{2}, \ldots, \frac{N+1}{2}\right\}\right\} \quad &\geq !la \max\left\{|\bar{u}_i^{M+1}|, i \in \left\{0, \ldots, \frac{N+1}{2}\right\}\right\} \\ &\geq \inf_{[0,4]} u(x,1) > 0, \end{aligned}$$

et donc ne tend pas vers 0 quand $h \to 0$.

Corrigé de l'exercice 20 page 85 : schéma implicite et principe du maximum

1. Schéma explicite décentré

   (a) Par définition, l'erreur de consistance en (x_i,t_n) s'écrit : On s'intéresse ici à l'ordre du schéma au sens des différences finies. On suppose que $u \in C^4([0,1] \times [0,T])$ est solution de (2.5.27) et on pose

   $$\bar{u}_i^n = u(ih,nk), i = 0,\dots,N, k = 0,\dots,M.$$

   Pour $i = 1,\dots,N-1$ et $k = 1,\dots,M-1$, l'erreur de consistance en (x_i,t_k) est définie par :

   $$R_i^n = \frac{1}{k}(\bar{u}_i^{n+1} - \bar{u}_i^n) - \frac{\alpha}{h}(\bar{u}_i^n - \bar{u}_{i-1}^n) - \frac{\mu}{h^2}(\bar{u}_{i-1}^n - 2\bar{u}_i^n + \bar{u}_{i+1}^n). \tag{2.7.55}$$

   Soit $i \in \{1,\dots,N-1\}, k \in \{1,\dots,M-1\}$. On cherche une majoration de R_i^n en utilisant des développements de Taylor. En utilisant ces développements, on obtient qu'il existe $(\xi_\ell,t_\ell) \in [0,1] \times [0,T]$, $\ell = 1,\dots,4$, t.q.:

   $$\bar{u}_i^{n+1} = \bar{u}_i^n + ku_t(ih,nk) + \frac{k^2}{2}u_{tt}(\xi_1,t_1), \tag{2.7.56}$$

   $$\bar{u}_{i-1}^n = \bar{u}_i^n - hu_x(ih,nk) + \frac{h^2}{2}u_{xx}(\xi_2,t_2), \tag{2.7.57}$$

   $$\bar{u}_{i-1}^n = \bar{u}_i^n - hu_x(ih,nk) + \frac{h^2}{2}u_{xx}(ih,nk) - \frac{h^3}{6}u_{xxx}(ih,nk) - \frac{h^4}{24}u_{xxxx}(\xi_3,t_3), \tag{2.7.58}$$

   $$\bar{u}_{i+1}^n = \bar{u}_i^n + hu_x(ih,nk) + \frac{h^2}{2}u_{xx}(ih,nk) + \frac{h^3}{6}u_{xxx}(ih,nk) + \frac{h^4}{24}u_{xxxx}(\xi_4,t_4). \tag{2.7.59}$$

   On en déduit :

   $$\begin{aligned} R_i^n &= u_t(ih,nk) + \frac{k}{2}u_{tt}(\xi_1,t_1) + \alpha u_x(ih,nk) + \alpha\frac{h}{2}u_{xx}(\xi_2,t_2) \\ &\quad -\mu u_{xx}(ih,nk) - \mu\frac{h^2}{24}\left(u_{xxxx}(\xi_3,t_3) + \mu u_{xxxx}(\xi_4,t_4)\right), \end{aligned}$$

   et donc, comme u est solution de (2.5.27), pour h assez petit, on a :

   $$|R_i^n| \le C_1(h+k),$$

   où C_1 ne dépend que de u. Le schéma (2.5.28) est donc consistant d'ordre 1 en temps et en espace.

   (b) Cherchons les conditions pour que u_i^{n+1} s'écrive comme combinaison convexe de u_i^n,u_{i-1}^n et u_{i+1}^n. On peut réécrire le schéma (2.5.28) :

   $$u_i^{n+1} = au_i^n + bu_{i+1}^n + cu_{i-1}^n, \text{ avec } a = 1 - \frac{\alpha k}{h} - \frac{2\mu k}{h^2}, b = \frac{\mu k}{h^2} \text{ et } c = \frac{\alpha k}{h} + \frac{\mu k}{h^2}.$$

   Il est facile de voir que $a+b+c = 1$, et que $b \ge 0$, $c \ge 0$. Il reste à vérifier que $a \ge 0$; pour cela, il faut et il suffit que $\frac{\alpha k}{h} + \frac{2\mu k}{h^2} \le 1$. Cette condition sécrit encore :

   $$k \le \frac{h^2}{\alpha h + 2\mu}. \tag{2.7.60}$$

Si h et k vérifient la condition (2.7.60), on pose : $M^n = \max_{i=1\ldots N} u_i^n$ (resp. $m^n = \min_{i=1\ldots N} u_i^n$. Comme u_i^{n+1} est une combinaison convexe de u_i^n,u_{i-1}^n et u_{i+1}^n, on a alors : $u_i^{n+1} \leq M^n \quad \forall i = 1,\ldots,N$ (resp. $u_i^{n+1} \geq m^n \quad \forall i = 1,\ldots,N$) et donc : $M^{n+1} \leq M^n$ (resp. $m^{n+1} \geq m^n$). On a ainsi montré que :

$$\|u^{n+1}\|_\infty \leq \|u^n\|_\infty.$$

On a de même :

$$\|u^n\|_\infty \leq \|u^{n-1}\|_\infty.$$

$$\vdots$$

$$\|u^1\|_\infty \leq \|u^0\|_\infty.$$

En sommant ces inégalités, on obtient :

$$\|u^n\|_\infty \leq \|u^0\|_\infty.$$

Donc, sous la condition (2.7.60), on a $\|u^{n+1}\|_\infty \leq \|u^n\|_\infty$ et donc $\|u^n\|_\infty \leq \|u^0\|_\infty$, pour tout $n = 1,\ldots,N$.

(c) En retranchant l'égalité (2.7.55) au schéma (2.5.28), on obtient l'équation suivante sur e_i^n :

$$\frac{1}{k}(e_i^{n+1} - e_i^n) + \frac{\alpha}{h}(e_i^n - e_{i-1}^n) - \frac{\mu}{h^2}(e_{i-1}^n - 2e_i^n + e_{i+1}^n) = R_i^n.$$

ce qu'on peut encore écrire :

$$e_i^{n+1} = (1 - \frac{k\alpha}{h} - 2\frac{k\mu}{h^2})e_i^n + e_{i-1}^n\frac{k\mu}{h^2} + kR_i^n. \tag{2.7.61}$$

Sous la condition de stabilité (2.7.60), on obtient donc :

$$\begin{array}{rcccl} |e_i^{n+1}| & \leq & \|e^{n+1}\|_\infty & + & C_1(k+h)k, \\ |e_i^n| & \leq & \|e^{n-1}\|_\infty & + & C_1(k+h)k, \\ \vdots & \leq & \vdots & + & \vdots \\ |e_i^1| & \leq & \|e^0\|_\infty & + & C_1(k+h)k, \end{array}$$

Si à $t = 0$, on a $\|e^0\| = 0$, alors on éduit des inégalités précédentes que $|e_i^n| \leq C1T(k+h)$ pour tout $n \in \mathbb{N}$. Le schéma est donc convergent d'ordre 1.

2. Schéma explicite centré.

(a) (Consistance) En utilisant les développements de Taylor (2.7.56) (2.7.58) et (2.7.59), et les développements suivants :

$$\bar{u}_{i-1}^n = \bar{u}_i^n - hu_x(ih,nk) + \frac{h^2}{2}u_{xx}(ih,nk) - \frac{h^3}{6}u_{xxx}(\xi_5,t_5),$$

$$\bar{u}_{i+1}^n = \bar{u}_i^n + hu_x(ih,nk) + \frac{h^2}{2}u_{xx}(ih,nk) + \frac{h^3}{6}u_{xxx}(\xi_6,t_6),$$

on obtient maintenant :

$$\begin{aligned} R_i^n &= u_t(ih,nk) + \frac{k}{2}u_{tt}(\xi_1,t_1) + \alpha u_x(ih,nk) + \alpha\frac{h^2}{12}\left(u_{xxx}(\xi_5,t_5) + \mu u_{xxx}(\xi_6,t_6)\right) \\ &\quad -\mu u_{xx}(ih,nk) - \mu\frac{h^2}{24}\left(u_{xxxx}(\xi_3,t_3) + \mu u_{xxxx}(\xi_4,t_4)\right), \end{aligned}$$

On en déduit que

$$|R_i^n| \leq C_3(k+h^2),$$

où $C_3 = \max(\frac{1}{2}\|u_{tt}\|_\infty, \frac{1}{6}\|u_{xxx}\|_\infty, \frac{1}{12}\|u_{xxxx}\|_\infty)$.

(b) Le schéma s'écrit maintenant :

$$u_i^{n+1} = \tilde{a}u_i^n + \tilde{b}u_{i+1}^n + \tilde{c}u_{i-1}^n, \text{ avec } \tilde{a} = 1 - \frac{2\mu k}{h^2}, \tilde{b} = \frac{\mu k}{h^2} - \frac{\alpha k}{h} \text{ et } \tilde{c} = \frac{\mu k}{h^2} + \frac{\alpha k}{h}.$$

Remarquons que l'on a bien : $\tilde{a} + \tilde{b} + \tilde{c} = 1$. Pour que u_i^{n+1} soit combinaison convexe de u_i^n, u_{i+1}^n et u_{i-1}^n, il faut et il suffit donc que $\tilde{a} \geq 0$, $\tilde{b} \geq 0$, et $\tilde{c} \leq 0$. L'inégalité $\tilde{c} \geq 0$ est toujours vérifiée. Les deux conditions qui doivent être vérifiées par h et k s'écrivent donc :

i. $\tilde{a} \geq 0$, i.e. $1 - \frac{2\mu k}{h^2} \geq 0$, soit encore

$$k \leq \frac{h^2}{2\mu}.$$

ii. $\tilde{b} \geq 0$ i.e. $\frac{\mu k}{h^2} - \frac{\alpha k}{h} \geq 0$, soit encore

$$h \leq \frac{\mu}{2\alpha}.$$

Le schéma centré est donc stable sous les deux conditions suivantes :

$$h \leq \frac{\mu}{2\alpha} \text{ et } k \leq \frac{1}{2\mu}h^2. \tag{2.7.62}$$

Pour obtenir une borne d'erreur, on procède comme pour le schéma (2.5.28) : on soustrait la définition de l'erreur de consistance au schéma numérique, et on obtient :

$$e_i^{n+1} = \tilde{a}e_i^n + \tilde{b}e_{i+1}^n + \tilde{c}e_{i-1}^n + kR_i^n. \tag{2.7.63}$$

Par le même raisonnement que pour le schéma décentré, on obtient donc que si $e_i^0 = 0$, on a $|e_i^n| \leq C_4(k+h^2)$, avec $C_4 = TC_3$.

Corrigé de l'exercice 21 page 85 : schéma implicite et principe du maximum

Corrigé en cours d'élaboration.

Corrigé de l'exercice 22 page 86

1. On admettra que la solution de (2.5.33) existe est qu'elle est assez régulière. Soient $M \in \mathbb{N}^*$ et $N \in \mathbb{N}^*$, et soient k le pas de temps, choisi tel que $Mk = T$ et h le pas espace, choisi tel que $Nh = 1$. On applique un schéma d'Euler explicite en temps, et un schéma de différences finies centré en espace, on obtient donc :

$$u_j^{n+1} = k\left[\frac{1}{k}u_j^n - \frac{1}{2h}\left(u_{j+1}^n - u_{j-1}^n\right) + \frac{\varepsilon}{h^2}\left(u_{j+1}^n + u_{j-1}^n - 2u_j^n\right)\right] \tag{2.7.64}$$

On tient compte des conditions aux limites et des conditions initiales en posant :

$$\left\{ \begin{array}{ll} u_0^n & = u_{N+1}^n = 0, \\ u_j^0 & = u_0(jh). \end{array} \right.$$

On a, par développement de Taylor :

$$u(x+h,t) = u(x,t) + hu_x(x,t) + \frac{h^2}{2}u_{xx}(x,t) + \frac{h^3}{6}u^{(3)}(x,t) + \frac{h^4}{24}u^{(4)}(\alpha,t),$$

$$u(x-h,t) = u(x,t) - hu_x(x,t) + \frac{h^2}{2}u_{xx}(x,t) - \frac{h^3}{6}u'''(x,t) + \frac{h^4}{24}u^{(4)}(\beta,t)$$

et

$$u(x,t+k) = u(x,t) + ku_t(x,t) + \frac{k^2}{2}u_{tt}(x,\tau_k), \tau_k \in [t,t+k].$$

De ces développements de Taylor, il ressort que l'erreur de consistance vérifie $|R| \leq C(k+h^2)$, où C ne dépend que de u. Le schéma est donc explicite d'ordre 1 en temps et 2 en espace.

Cherchons alors les conditions pour que :

$$\|u^n\|_\infty \leq \|u^0\|_\infty.$$

Par définition,

$$\|u^n\|_\infty = \max_{j=1,\ldots,N} |u_j^n|.$$

On essaye d'abord de vérifier que : $\|u^{n+1}\|_\infty \leq \|u^n\|_\infty$, c'est-à-dire :

$$\max_{j=1,\ldots,N} |u_j^{n+1}| \leq \max_{j=1,\ldots,N} |u_j^n|,$$

On veut donc montrer que

$$\left\{ \begin{array}{l} \max\limits_{j=1,\ldots,N} u_j^{n+1} \leq \max\limits_{j=1,\ldots,N} u_j^n, \\ \\ \min\limits_{j=1,\ldots,N} u_j^{n+1} \geq \min\limits_{j=1,\ldots,N} u_j^n. \end{array} \right.$$

On peut réécrire le schéma (2.7.64) :

$$u_j^{n+1} = u_j^n \left(1 - \frac{2\varepsilon k}{h^2}\right) + u_{j+1}^n \left(-\frac{k}{2h} + \frac{k\varepsilon}{h^2}\right) + u_{j-1}^n \left(\frac{\varepsilon k}{h^2} + \frac{k}{2h}\right).$$

Posons :

$$M^n = \max_{j=1\ldots N} u_j^n$$

Supposons que k et h vérifient :

$$1 \geq \frac{2\varepsilon k}{h^2} \text{ et } \frac{k\varepsilon}{h^2} - \frac{k}{2h} \geq 0,$$

ce qui s'écrit encore :

$$\left\{ \begin{array}{l} \dfrac{k}{h^2} \leq \dfrac{1}{2\varepsilon} \\ k \leq \dfrac{2\varepsilon}{h}, \end{array} \right. \tag{2.7.65}$$

on a alors :

$$u_j^{n+1} \leq M^n \left(1 - \frac{2\varepsilon k}{h^2}\right) + M^n \left(-\frac{k}{2h} + \frac{k\varepsilon}{h^2}\right) + M^n \left(\frac{\varepsilon k}{h^2} + \frac{k}{2h}\right) \quad \forall j = 1, \ldots, N,$$

et donc :

$$M^{n+1} \leq M^n.$$

Posons maintenant :

$$m^n = \min_{j=1\ldots N} u_j^n.$$

Si k et h satisfont les conditions (2.7.65), on obtient de la même manière

$$m^{n+1} \geq m^n$$

On a ainsi montré que :

$$\|u^{n+1}\|_\infty \leq \|u^n\|_\infty.$$

On a de même :

$$\|u^n\|_\infty \leq \|u^{n-1}\|_\infty.$$

$$\vdots$$

$$\|u^1\|_\infty \leq \|u^0\|_\infty.$$

En sommant ces inégalités, on obtient :

$$\|u^n\|_\infty \leq \|u^0\|_\infty.$$

Donc, sous les conditions (2.7.65), on a $\|u^{n+1}\|_\infty \leq \|u^n\|_\infty$ et donc $\|u^n\|_\infty \leq \|u^0\|_\infty$, pour tout $n = 1, \ldots, N$.
Pour étudier la convergence du schéma, on va tenter de majorer l'erreur de discrétisation :

$$e_j^n = \bar{u}_j^n - u_j^n,$$

où u_j^n est calculé par (2.7.64), et $\bar{u}_j^n$ est la solution du problème (2.5.33) en $x_j = jh$ et $t_n = nk$.
On a donc, par définition de l'erreur de consistance,

$$\frac{1}{k}(\bar{u}_j^{n+1} - \bar{u}_j^n) + \frac{1}{2h}(\bar{u}_{j+1}^n - \bar{u}_{j-1}^n) - \frac{\varepsilon}{h^2}(-2\bar{u}_j^n + \bar{u}_{j+1}^n + \bar{u}_{j-1}^n) = R_j^n$$

où $|R_i^n| \leq C(k + h^2)$
ce qui entraine :

$$\frac{1}{k}(e_j^{n+1} - e_j^n) + \frac{1}{2h}(e_j^n - e_{j-1}^n) - \frac{\varepsilon}{h^2}(-2e_j^n + e_{j+1}^n + e_{j-1}^n) = R_j^n$$

soit encore :

$$e_j^{n+1} = \left(1 - \frac{2\varepsilon k}{h^2}\right) e_j^n + \left(-\frac{k}{2h} + \frac{k\varepsilon}{h^2}\right) e^{j+1} + \left(\frac{\varepsilon k}{h^2} + \frac{k}{2h}\right) e_{j-1}^n + kR_j^n.$$

de même que précédemment, on obtient sous les conditions (2.7.65)

$$\begin{aligned} |e_j^{n+1}| &\leq \|e^n\|_\infty + C(k+h^2)k \\ &\vdots \\ |e_j^n| &\leq \|e^{n-1}\|_\infty + C(k+h^2)k \\ &\vdots \\ |e_j^1| &\leq \|e^0\|_\infty + C(k+h^2)k. \end{aligned}$$

Et donc en sommant ces inégalités :

$$\|e^n\|_\infty \leq \|e^0\|_\infty + nCk(k+h^2)$$

Si à $t=0$ on a $\|e^0\|_\infty = 0$, alors :

$$\|e^{n+1}\|_\infty \leq CMk(k+h^2) = T(k+h^2).$$

Et donc sous les conditions (2.7.65) on a $\|e^n\|_\infty$ qui tend vers 0 lorsque k,$h \to 0$, ce qui prouve que le schéma est convergent.

Corrigé de l'exercice 23 page 87

1. Notons $R_i^{(n)}$ l'erreur de consistance en (x_i,t_n). Pour le schéma (2.5.36), on a donc par définition :

$$\begin{aligned} R_i^{(n)} &= \tfrac{\bar{u}_i^{(n+1)} - \bar{u}_i^{(n)}}{k} + \tfrac{1}{h^2}(2\bar{u}_i^{(n+1)} - \bar{u}_{i-1}^{(n+1)} - \bar{u}_{i+1}^{(n+1)}) - \bar{u}_i^{n+1} \\ &= \tilde{R}_i^{(n)} + \hat{R}_i^n, \end{aligned}$$

où

$$\tilde{R}_i^n = \frac{\bar{u}_i^{n+1} - \bar{u}_i^n}{k} - u_t(x_i,t_n) \text{ est l'erreur de consistance en temps}$$

et

$$\hat{R}_i^n = \frac{1}{h^2}(2\bar{u}_i^{n+1} - \bar{u}_{i-1}^{n+1} - \bar{u}_{i+1}^{n+1}) - (u_{xx}(x_i,t_n)) \text{ est l'erreur de consistance en espace.}$$

On a vu (voir (1.2.19)) que

$$\left|\hat{R}_i^n\right| \leq \frac{h^2}{12} \sup_{[0,1]} \left|\frac{\partial^4 u}{\partial x^4}(\cdot,t_n)\right|, \forall i \in \{1,\ldots,N\}$$

Effectuons maintenant un développement de Taylor en fonction du temps d'ordre 2 :

$$u(x_i,t_{n+1}) = u(x_i,t_n) + ku_t + \frac{k^2}{2}u_{tt}(x_i,\xi_n)$$

avec $\xi_n \in [t_n,t_{n+1}]$. Donc $\dfrac{u(x_i,t_{n+1}) - u(x_i,t_n)}{k} - u_t = \dfrac{k}{2}u_{tt}(x_i,\xi_n)$. Comme $\xi_n \in [0,T]$, et u_{tt} admet un maximum (à x_i fixé) dans $[0,T]$ (qui est compact), on a donc

$$\left|\tilde{R}_i^n\right| \leq \frac{k}{2} \max_{[0,T]} |u_{tt}(x_i,\cdot)|.$$

Par conséquent,

$$|R_i^n| = \left|\tilde{R}_i^n + \hat{R}_i^n\right| \leq \left|\tilde{R}_i^n\right| + \left|\hat{R}_i^n\right| \leq \frac{k}{2}\max_{[0,T]}|u_{tt}(x_i,\cdot)| + \frac{h^2}{12}\max_{[0,1]}\left|\frac{\partial^4 u}{\partial x^4}(\cdot,t_{n+1})\right|.$$

Donc $|R_i^n| \leq C(k+h^2)$ avec

$$C = \frac{1}{2}\max\left(\|u_{tt}\|_{L^\infty([0,1]\times[0,T]}; \frac{1}{6}\|\frac{\partial^4 u}{\partial x^4}\|_{L^\infty([0,1]\times[0,T]}\right).$$

Le calcul de l'erreur de consistance pour le schéma (2.5.37) s'effectue de manière semblable.

2. Le schéma (2.5.36) est complètement implicite alors que le schéma (2.5.37) ne l'est que partiellement, puisque le terme de réaction est pris à l'instant n. Le schéma (2.5.36) s'écrit : $AU^{n+1} = U^n$ avec $U^{n+1} = (U_1^{n+1}, \dots U_N^{n+1})^t$,$U^n = (U_1^n, \dots, U_N^n)^t$, et

$$A = \begin{pmatrix} 1+2\lambda-k & -\lambda & 0 & \dots & 0 \\ -\lambda & 1+2\lambda-k & -\lambda & \ddots & 0 \\ 0 & & & & \\ & & & & \\ \vdots & & & & 0 \\ 0 & 0 & -\lambda & & 1+2\lambda-k \end{pmatrix}$$

où $\lambda = \dfrac{k}{h^2}$. Notons que par définition, A est symétrique. De même, le schéma (2.5.37) s'écrit : $BU^{n+1} = U^n$ avec

$$B = \frac{1}{1+k}\begin{pmatrix} 1+2\lambda & -1 & 0 & \dots & 0 \\ -1 & 1+2\lambda & & \ddots & 0 \\ 0 & & & & -1 \\ & & & & \\ & & & & \\ 0 & 0 & -1 & & 1+2\lambda \end{pmatrix}$$

On a donc $A = \lambda A_h$, où A_h est définie en (1.2.15) page 13, avec $c_i = \frac{1-k}{\lambda}$, et $B = \frac{\lambda}{k+1}\ A_h$ avec $c_i = \frac{1}{\lambda}$. Dans les deux cas, les matrices sont donc s.d.p. en vertu de la proposition 1.2 page 13. Notons que l'hypothèse $k \in]0,1[$ est nécessaire dans le cas du premier schéma, pour assurer la positivité de c_i.

3. Le schéma (2.5.37) s'écrit

$$(1+k)u_i^n = u_i^{n+1} + \lambda(2u_i^{n+1} - u_{i+1}^{n+1} - u_{i-1}^{n+1}).$$

On montre facilement par récurrence que $\max_{j=1,\dots,N} u_j^n \leq (1+k)^n \max_{j=1,\dots,N} u_j^0$, (voir preuve de la stabilité L^∞ d'Euler implicite page 82) et que $\min\limits_{j=1,\dots,N} u_j^{(n)} \geq (1+k)^n \min\limits_{j=1,\dots,N} u_j^{(0)}$. On en déduit que

$$\|u^{(n)}\|_\infty \leq (1+k)^n \|u_0\|_\infty$$

Or $(1+k)^n \leq (1+k)^{T/k}$ car $kn \leq T$. Or

$$\begin{aligned}(1+k)^{T/k} &= \exp\left(\frac{T}{k}\ell n(1+k)\right) \\ &\leq \exp(\frac{T}{k}k) = e^T\end{aligned}$$

On en déduit le résultat, avec $C_1(T) = e^T$. De même, pour le schéma (2.5.36), on montre par récurrence que :

$$\|u^{(n)}\|_\infty \leq \frac{1}{(1-k)^n}\|u^{(0)}\|.$$

Mais pour $k \in]0,\alpha[$, avec $\alpha \in]0,1[$, on a :

$$\frac{1}{(1-k)} \leq 1 + \beta k, = \text{ avec } \beta = \frac{1}{(1-\alpha)}.$$

On en déduit par un calcul similaire au précédent que

$$(1-k)^{T/k} \leq e^{\beta T},$$

d'où le résultat avec $C_2(T,\alpha) = e^{\beta T}$.

4. Par définition de l'erreur de consistance, on a pour le schéma (2.5.36)

$$\frac{\bar{u}_j^{(n+1)} - \bar{u}_j^n}{k} - \frac{\bar{u}_{j+1}^{(n+1)} + \bar{u}_{j-1}^{n+1} - 2\bar{u}_j^{(n+1)}}{h^2} - \bar{u}_j^{n+1} = R_i^{(n,1)}$$

et donc, en notant $e_j^{(n)} = \bar{u}_j^n - u_j^{(n)}$ l'erreur de discrétisation en (x_j,t_n), on a :

$$e_j^{(n+1)}(1 + 2\lambda - k) - \lambda e_{j-1}^{(n+1)} - \lambda e_{j+1}^{(n+1)} = e_j^{(n)} + kR_j^{(n,1)}$$

On obtient donc, de manière similaire à la question 3 :
(en considérant $e_{j_0}^{(n+1)} = \max e_j^{(n+1)}$ puis $e_{j_0}^{(n+1)} = \min e_j^{(n+1)}$)

$$\frac{1}{1-k}\|e^{(n+1)}\| \leq \|e^{(n)}\| + kC(k+h^2).$$

Par récurrence sur n, on obtient alors

$$\|e^{(n)}\|_\infty \leq \left(\frac{1}{1-k}\right)^n [kC(k+h^2) + \|e^0\|_\infty]$$

d'où

$$\|e^{(n)}\|_\infty \leq C_2(T,\alpha)(TC(k+h^2) + \|e^0\|_\infty).$$

De même, pour le schéma (2.5.37), on écrit l'erreur de consistance :

$$\frac{\bar{u}_j^{(n+1)} - \bar{u}_j^n}{k} - \frac{u_{j+1}^{(n+1)} + \bar{u}_{j-1}^{n+1} - 2\bar{u}_j^{(n+1)}}{h^2} - \bar{u}_j^n = R_j^{(n,2)}$$

et donc :

$$e_j^{(n+1)}(1 + 2\lambda) - \lambda e_{j-1}^{(n+1)} - \lambda e_{j+1}^{(n+1)} = e_j^{(n)}(1+k) + kR_j^{(n,2)}.$$

Par des raisonnements similaires à ceux de la question 3 on obtient alors :

$$\|e_j^{(n)}\| \leq (1+k)^n(\|e^{(0)}\| + kC(k+h^2))$$

d'où

$$\|e^{(n)}\|_\infty \leq C_1(T)(\|e^{(0)}\| + kC(k+h^2)).$$

Corrigé de l'exercice 24 page 88

1. On s'intéresse ici à l'ordre du schéma au sens des différences finies. On suppose que $u \in C^4([0,1]\times[0,T])$ est solution de (2.5.39) et on pose

$$\bar{u}_j^n = u(jh,nk), j = 0,\ldots,N, k = 0,\ldots,M.$$

L'erreur de consistance est définie par :

$$R_j^n = \frac{\bar{u}_j^{n+1} - \bar{u}_j^{n-1}}{2k} - \frac{\bar{u}_{j-1}^n - 2\bar{u}_j^n + \bar{u}_{j+1}^n}{h^2}, j = 1,\ldots,N-1, k = 1,\ldots,M-1.$$

On cherche une majoration de R_j^n en utilisant des développement de Taylor. Soit $j \in \{1,\ldots,N-1\}, k \in \{1,\ldots,M-1\}$. Il existe $(\xi_i,t_i) \in [0,1]\times[0,T]$, $i = 1,\ldots,4$, t.q.:

$$\bar{u}_j^{n+1} = \bar{u}_j^n + ku_t(jh,nk) + \frac{k^2}{2}u_{tt}(jh,nk) + \frac{k^3}{6}u_{ttt}(\xi_1,t_1),$$

$$\bar{u}_j^{n-1} = \bar{u}_j^n - ku_t(jh,nk) + \frac{k^2}{2}u_{tt}(jh,nk) - \frac{k^3}{6}u_{ttt}(\xi_2,t_2),$$

$$\bar{u}_{j-1}^n = \bar{u}_j^n - hu_x(jh,nk) + \frac{h^2}{2}u_{xx}(jh,nk) - \frac{h^3}{6}u_{xxx}(jh,nk) - \frac{h^4}{24}u_{xxxx}(\xi_3,t_3),$$

$$\bar{u}_{j+1}^n = \bar{u}_j^n + hu_x(jh,nk) + \frac{h^2}{2}u_{xx}(jh,nk) + \frac{h^3}{6}u_{xxx}(jh,nk) + \frac{h^4}{24}u_{xxxx}(\xi_4,t_4).$$

On en déduit :

$$R_j^n = u_t(jh,nk) + \frac{k^2}{12}\left(u_{ttt}(\xi_1,t_1) + u_{ttt}(\xi_2,t_2)\right) - u_{xx}(jh,nk) - \frac{h^2}{24}\left(u_{xxxx}(\xi_3,t_3) + u_{xxxx}(\xi_4,t_4)\right),$$

et donc, comme u est solution de (2.5.39),

$$|R_j^n| \leq C_1(k^2+h^2),$$

où C_1 ne dépend que de u. Le schéma (2.5.40) est donc consistant d'ordre 2.

2. Pour étudier la stabilité au sens de Von Neumann, on "oublie" les conditions aux limites dans (2.5.39). Plus précisément, on s'intéresse à (2.5.39)avec $x \in \mathbb{R}$ (au lieu de $x \in]0,1[$) et on remplace les conditions aux limites par des conditions de périodicité (exactement comme on l'a vu au paragraphe 2.2.6 page 76). Enfin, on prend une condition initiale de type "mode de Fourier", avec $p \in \mathbb{R}$ arbitraire, et u_0 défini par :

$$u_0(x) = e^{ipx}, x \in \mathbb{R}.$$

La solution exacte est alors :

$$u(x,t) = e^{-p^2t}e^{ipx}, x \in \mathbb{R},\ t \in \mathbb{R}_+,$$

c'est–à–dire

$$u(\cdot,t) = e^{-p^2t}u_0, t \in \mathbb{R}_+.$$

Le facteur d'amplification est donc, pour tout $t \in \mathbb{R}_+$, le nombre e^{-p^2t}. Ce facteur est toujours, en module, inférieur à 1. On va maintenant chercher la solution du schéma numérique sous la forme :

$$u_j^n = \xi_n e^{ipjh}, j \in \mathbb{Z}, n \in \mathbb{N}, \tag{2.7.66}$$

où ξ_0 et $\xi_1 \in \mathbb{R}$ sont donnés (ils donnent u_j^0 et u_j^1 pour tout $j \in \mathbb{Z}$) et $\xi_n \in \mathbb{R}$ est à déterminer de manière à ce que la première équation de (2.5.40)) soit satisfaite.
Ce facteur ξ_n va dépendre de k,h et p. Pour k et h donnés, le schéma est stable au sens de Von Neumann si, pour tout $p \in \mathbb{R}$, la suite $(\xi_n)_{n\in\mathbb{N}}$ est bornée. Dans le cas contraire, le schéma est (pour ces valeurs de k et h) dit instable au sens de Von Neumann.
Un calcul immédiat donne que la famille des u_j^n, définie par (2.7.66), est solution de la première équation si et seulement si la suite $(\xi_n)_{n\in\mathbb{N}}$ vérifie (on rappelle que ξ_0 et ξ_1 sont donnés) :

$$\frac{\xi_{n+1}-\xi_{n-1}}{2k} = \frac{2}{h^2}(\cos ph - 1)\xi_n, n \geq 2,$$

ou encore, en posant

$$\alpha = \frac{4k}{h^2}(\cos ph - 1)\ (\leq 0),$$

$$\xi_{n+1} - \alpha\xi_n - \xi_{n-1} = 0,\ n \geq 2 \tag{2.7.67}$$

En excluant le cas $\alpha = -2$ (qui correspond, pour k et h donnés, à des valeurs de p très particulières), la solution de (2.7.67) est

$$\xi_n = Ar_1^n + Br_2^n, A \geq 0, \tag{2.7.68}$$

où A et B sont déterminés par ξ_0 et ξ_1 (de sorte que $\xi_0 = A + B$,$\xi_1 = Ar_1 + Br_2$) et r_1,r_2 sont les deux racines distinctes de :

$$r^2 - \alpha r - 1 = 0. \tag{2.7.69}$$

Les nombres r_1 et r_2 sont réels et comme $r_1 r_2 = 1$, l'un de ces nombres est, en module, supérieur à 1. Ceci montre que $(\xi_n)_n$ est une suite non bornée (sauf pour des choix très particulier de ξ_0 et ξ_1, ceux pour les quelles $\xi_1 = \xi_0 r_2$ où r_2 est la racine de (2.7.69) de module inférieur à 1). Ce schéma est donc instable au sens de Von Neumann, pour tout $k > 0$ et $h > 0$.

3. On reprend les notations de la question 1. On s'intéresse maintenant à la quantité S_j^n (qui est toujours l'erreur de consistance) :

$$S_j^n = \frac{\bar{u}_j^{n+1} - u_j^{n-1}}{2k} - \frac{\bar{u}_{j-1}^n - (\bar{u}_j^{n+1} + \bar{u}_j^{n-1}) + \bar{u}_{j+1}}{h^2}, j = 1,\ldots,N-1, \quad k = 0,\ldots,M-1.$$

En reprenant la technique de la question 1, il existe $(\xi_i,t_i), i = 1,\ldots,6 \qquad t.q.$

$$S_j^n = \frac{h^2}{12}\left(u_{ttt}(\xi_1,t_1) + u_{ttt}(\xi_2,t_2)\right) - \frac{h^2}{24}\left(u_{xxxx}(\xi_3,t_3) - u_{xxxx}(\xi_4,t_4)\right) + \frac{k^2}{2h^2}h_{tt}(\xi_5,t_5) + \frac{k^2}{2h^2}u_{tt}(\xi_6,t_6).$$

Ce qui donne, avec C_2 ne dépendant que de u,

$$|S_j^n| \leq C_2\left(h^2 + k^2 + \frac{k^2}{h^2}\right), j = 1,\ldots,N-1, \quad k = 0,\ldots,M-1.$$

Le schéma est donc consistant quand $h \to 0$ avec $\dfrac{k}{h} \to 0$.
4. On reprend la méthode développée à la question 2, la suite $(\xi_n)_n$ doit maintenant vérifier la relation suivante (avec ξ_0,ξ_1 donnés).

$$\frac{\xi_{n+1}-\xi_{n-1}}{2k} = \frac{2\cos(ph)}{h^2}\xi_n - \frac{\xi_{n-1}+\xi_{n+1}}{h^2}, n \geq 2$$

c'est à dire :

$$\xi_{n+1}\left(\frac{1}{2k}+\frac{1}{h^2}\right)-\frac{2\cos(ph)}{h^2}\xi_n+\xi_{n-1}\left(\frac{1}{h^2}-\frac{1}{2k}\right)=0, n\geq 2.$$

L'équation caractéristique est maintenant :

$$p(r)=ar^2+br+c=0,$$

avec

$$a=\frac{1}{2k}+\frac{1}{h^2}, b=-\frac{2\cos(ph)}{h^2} \text{ et } c=\frac{1}{h^2}-\frac{1}{2h}.$$

Pour montrer la stabilité au sens de Von Neumann, il suffit d'après (2.7.68) de montrer que les deux racines du polynôme p sont de module inférieur ou égal à 1. On note r_1 et r_2 ces deux racines (qui peuvent être confondues) et on distingue 2 cas :

1. 1er cas : Les racines de p ne sont pas réelles. Dans ce cas, on a $|r_1|=|r_2|=\gamma$ et

   $$\gamma=|\frac{c}{a}|<1,$$

   car $k>0$.
2. 2ème cas : Les racines de p sont réelles. Dans ce cas, on remarque que

   $$r_1r_2=\frac{c}{a}<1,$$

   et l'une des racines, au moins, est donc entre -1 et 1 (strictement). De plus on a $p(1)=\frac{2}{h^2}-\frac{2\cos ph}{h^2}\geq 0$ et $p(-1)=\frac{2}{h^2}+\frac{2\cos ph}{h^2}\geq 0$, l'autre racine est donc aussi entre -1 et 1 (au sens large).

On en déduit que le schéma (2.5.41) est stable au sens de Von Neumann.

Corrigé de l'exercice 25 page 88 : Discrétisation d'un problème parabolique non linéaire

On se propose, dans cet exercice, de montrer l'existence d'une solution faible au problème (2.5.42)-(2.5.44), à partir de l'existence de la solution approchée donnée par un schéma numérique. L'inconnue de ce problème est la fonction u de $[0,1]\times[0,T]$ dans $\mathbb{R}$, elle doit être solution des équations suivantes :

$$\frac{\partial u}{\partial t}(x,t)-\frac{\partial^2\varphi(u)}{\partial x^2}(x,t)=v(x,t), x\in]0,1[, t\in]0,T[, \tag{2.7.70}$$

$$\frac{\partial\varphi(u)}{\partial x}(0,t)=\frac{\partial\varphi(u)}{\partial x}(1,t)=0, t\in]0,T[, \tag{2.7.71}$$

$$u(x,0)=u_0(x), x\in]0,1[, \tag{2.7.72}$$

où φ, v, T, u_0 sont donnés et sont t.q.

1. $T>0$, $v\in L^\infty(]0,1[\times]0,T[)$,
2. φ croissante, lipschitzienne de $\mathbb{R}$ dans $\mathbb{R}$,
3. $u_0\in L^\infty(]0,1[)$ et $\varphi(u_0)$ lipschitzienne de $[0,1]$ dans $\mathbb{R}$.

Un exemple important est donné par $\varphi(s) = \alpha_1 s$ si $s \leq 0$, $\varphi(s) = 0$ si $0 \leq s \leq L$ et $\varphi(s) = \alpha_2(s-L)$ si $s \geq L$, avec α_1,α_2 et L donnés dans $\mathbb{R}_+^\star$. Noter pour cet exemple que $\varphi' = 0$ sur $]0,L[$.

Les ensembles $]0,1[$ et $D =]0,1[\times]0,T[$ sont munis de leur tribu borélienne et de la mesure de Lebesgue sur cette tribu.

On appelle "solution faible" de (2.5.42)-(2.5.44) une solution de :

$$u \in L^\infty(]0,1[\times]0,T[), \tag{2.7.73}$$

$$\int_D (u(x,t)\frac{\partial \psi}{\partial t}(x,t) + \varphi(u(x,t))\frac{\partial^2 \psi}{\partial x^2}(x,t) + v(x,t)\psi(x,t))dxdt + \int_{]0,1[} u_0(x)\psi(x,0)dx = 0, \tag{2.7.74}$$

$\forall \psi \in C_T^{2,1}(\mathbb{R}^2)$,

où $\psi \in C_T^{2,1}(\mathbb{R}^2)$ signifie que ψ est une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$ deux fois continûment dérivable par rapport à x, une fois continûment dérivable par rapport à t et t.q. $\dfrac{\partial \psi}{\partial x}(0,t) = \dfrac{\partial \psi}{\partial x}(1,t) = 0$, pour tout $t \in [0,T]$ et $\psi(x,T) = 0$ pour tout $x \in [0,1]\}$.

Question 1 (Solution classique versus solution faible)

Soit $u \in C^2(\mathbb{R}^2,\mathbb{R})$; notons u sa restriction à $D =]0,1[\times]0,T[$; notons que l'on a bien $u \in L^\infty(]0,1[\times]0,T[)$. Supposons que u satisfait (2.5.42)-(2.5.44), et montrons qu'alors u vérifie (2.5.46). Soit $\psi \in C_T^{2,1}(\mathbb{R}^2)$. Multiplions (2.5.42) par ψ et intégrons sur D. On obtient :

$$\int_D \frac{\partial u}{\partial t}(x,t)\psi(x,t)dxdt - \int_D \frac{\partial^2 \varphi(u)}{\partial x^2}(x,t)\psi(x,t)dxdt = \int_D v(x,t)\psi(x,t)dxdt. \tag{2.7.75}$$

Par intégration par parties, il vient :

$$\int_D \frac{\partial u}{\partial t}(x,t)\psi(x,t)dxdt = \int_0^1 u(x,T)\psi(x,T)dx - \int_0^1 u(x,0)\psi(x,0)dx - \int_D u(x,t)\frac{\partial \psi}{\partial t}(x,t).$$

Comme $\psi \in C_T^{2,1}(\mathbb{R}^2)$ on a donc $\psi(x,T) = 0$ pour tout $x \in [0,1]$ et comme u vérifie (2.5.44), on a $u(x,0) = u_0(x)$. On en déduit que

$$\int_D \frac{\partial u}{\partial t}(x,t)\psi(x,t)dxdt = -\int_0^1 u_0(x)\psi(x,0)dx - \int_D \frac{\partial \psi}{\partial t}(x,t)u(x,t). \tag{2.7.76}$$

Intégrons par parties le deuxième terme de (2.7.75) :

$$\begin{aligned}\int_D \frac{\partial^2 \varphi(u)}{\partial x^2}(x,t)\psi(x,t)dxdt = &\int_0^T [\frac{\partial \varphi(u)}{\partial x}(1,t)\psi(1,t) - \frac{\partial \varphi(u)}{\partial x}(0,t)\psi(0,t)]dt \\ &- \int_D \frac{\partial \varphi(u)}{\partial x}(x,t)\frac{\partial \psi(u)}{\partial x}(x,t)dxdt.\end{aligned}$$

et comme u vérifie (2.5.43), on a

$$\frac{\partial \varphi(u)}{\partial x}(0,t) = \frac{\partial \varphi(u)}{\partial x}(1,t) = 0, t \in]0,T[.$$

En tenant compte de ces relations et en ré–intégrant par parties, on obtient :

$$\int_D \frac{\partial^2 \varphi(u)}{\partial x^2}(x,t)\psi(x,t)dxdt = -\int_D \varphi(u)(x,t)\frac{\partial^2 \psi(u)}{\partial x^2}(x,t)dxdt. \tag{2.7.77}$$

En remplaçant dans (2.7.76) et (2.7.77) dans (2.7.75), on obtient (2.5.42).
Réciproquement, supposons que u satisfait (2.5.46), et soit ψ continûment différentiable à support compact dans D. En intégrant (2.5.46) par parties et en tenant compte que ψ et toutes ses dérivées sont nulles au bord de D, on obtient :

$$\int_D \big[-\frac{\partial u}{\partial t}(x,t) + \frac{\partial^2 \varphi(u)}{\partial x^2}(x,t) - v(x,t)\big]\psi(x,t)dxdt = 0, \forall \psi \in C_c^\infty(D).$$

Comme u est régulière, ceci entraîne que l'équation (2.5.42) est donc satisfaite par u.
On prend ensuite $\psi \in C_T^{2,1}(\mathbb{R}^2)$, et on intègre (2.5.46) par parties. En tenant compte du fait que $\psi(x,T) = 0$, pour tout x et $\dfrac{\partial \psi}{\partial x}(0,t) = \dfrac{\partial \psi}{\partial x}(1,t) = 0$, pour tout t, on obtient :

$$\begin{aligned} &-\int_0^1 u(x,0)\psi(x,0)dx - \int_D \frac{\partial u}{\partial t}(x,t)\psi(x,t)dxdt + \int_0^T \frac{\partial \varphi(u)}{\partial x}(1,t)\psi(1,t)dt \\ &-\int_0^T \frac{\partial \varphi(u)}{\partial x}(0,t)\psi(0,t)dt + \int_D \frac{\partial^2 \varphi(u)}{\partial x^2}\psi(x,t)dxdt + \int_0^1 u_0(x)\psi(x,0)dx = 0. \end{aligned}$$

En regroupant et en utilisant le fait que u satisfait (2.5.42), on obtient :

$$\int_0^1 (u_0(x) - u(x,0))\psi(x,0)dx + \int_0^T \frac{\partial \varphi}{\partial x}(1,t)\psi(1,t)dt - \int_0^T \frac{\partial \varphi}{\partial x}(0,t)\psi(0,t)dt = 0.$$

En choisissant successivement une fonction ψ nulle en $x = 0$ et $x = 1$ puis nulle en $x = 1$ et $t = T$ et enfin nulle en $x = 0$ et $t = T$, on obtient que u satisfait la condition initiale (2.5.44) et les conditions aux limites (2.5.43), ce qui conclut la question.

Question 2 (Existence et unicité de la solution approchée)
Soit $n \in \{0,\ldots,M-1\}$. On suppose connu $\{u_i^n,\ i = 1,\ldots,N\}$. On va prouver dans cette question l'existence et l'unicité de $\{u_i^{n+1},\ i = 1,\ldots,N\}$ vérifiant (2.5.48) (avec $u_0^{n+1} = u_1^{n+1}$, $u_{N+1}^{n+1} = u_N^{n+1}$).

1. L'application $s \mapsto s$ est strictement croissante, et par hypothèse sur φ, l'application $s \mapsto a\varphi(s)$ est croissante. La somme d'une fonction strictement croissante et d'une fonction croissante est strictement croissante. D'autre part, comme φ est croissante, pour tout $\varphi(s) \leq \varphi(0), \forall s \leq 0$, et donc $\lim_{s\to-\infty} g_a(s) = -\infty$. De même, $\varphi(s) \geq \varphi(0), \forall s \geq 0$, et donc $\lim_{s\to+\infty} g_a(s) = +\infty$. La fonction g_a est continue et prend donc toutes les valeurs de l'intervalle $]-\infty,+\infty[$. Comme elle est strictement croissante, elle est bijective.
2. L'équation (4.2.5) s'écrit encore :

   $$g_a(u_i) = \frac{k}{h^2}(\overline{w}_{i-1} + \overline{w}_{i+1}) + u_i^n + kv_i^n, \text{ pour tout } i = 1,\ldots,N,$$

   avec $a = \frac{k}{h^2}$. Par la question précédente, il existe donc un unique u_i qui vérifie cette équation ; il suffit alors de poser $\varphi(u_i) = w_i$ pour déterminer de manière unique la solution de (2.5.49)–(2.5.50).

3. Soit $\overline{w}^1$ et $\overline{w}^2 \in \mathbb{R}^N$ et soit $w^1 = F(\overline{w}^1)$ et $w^2 = F(\overline{w}^2)$. Par définition de F, on a :

$$u_i^1 - u_i^2 + \frac{2k}{h^2}(w_i^1 - w_i^2) = \frac{k}{h^2}\big((\overline{w}_{i-1}^1 + \overline{w}_{i+1}^1) - (\overline{w}_{i-1}^2 + \overline{w}_{i+1}^2)\big), \text{ pour tout } i = 1,\dots,N. \quad (2.7.78)$$

Comme φ est monotone, le signe de $w_i^1 - w_i^2 = \varphi(u_i^1) - \varphi(u_i^2)$ est le même que celui de $u_i^1 - u_i^2$, et donc

$$|u_i^1 - u_i^2 + \frac{2k}{h^2}(w_i^1 - w_i^2)| = |u_i^1 - u_i^2| + \frac{2k}{h^2}|w_i^1 - w_i^2|. \quad (2.7.79)$$

Et comme φ est lipschitzienne de rapport L, on a

$$|w_i^1 - w_i^2| = |\varphi(u_i^1) - \varphi(u_i^2)| \leq L|u_i^1 - u_i^2|,$$

d'où :

$$|u_i^1 - u_i^2| \geq \frac{1}{L}|w_i^1 - w_i^2|. \quad (2.7.80)$$

On déduit donc de (2.7.78),(2.7.79) et(2.7.80) que

$$\frac{1}{L}|w_i^1 - w_i^2| + \frac{2k}{h^2}|w_i^1 - w_i^2| \leq \frac{k}{h^2}\big(|\overline{w}_{i-1}^1 - \overline{w}_{i-1}^2| + |\overline{w}_{i+1}^1 - \overline{w}_{i+1}^2|\big), \text{ pour tout } i = 1,\dots,N.$$

On a donc

$$|w_i^1 - w_i^2| \leq \frac{1}{1 + \frac{h^2}{2kL}} \max_{i=1,\dots,N} |\overline{w}_i^1 - \overline{w}_i^2|, \text{ pour tout } i = 1,\dots,N.$$

d'où on déduit que $\|w^1 - w^2\|_\infty \leq C\|\overline{w}^1 - \overline{w}^2\|_\infty$ avec $C = \frac{1}{1+\frac{h^2}{2kL}} < 1$. L'application F est donc bien strictement contractante.

4. Soit $\{u_i^{n+1}$, Si $\{u_i^{n+1}$, $i = 1,\dots,N\}$ est solution de (2.5.48) et $w_i = \varphi(u_i^{n+1})$ pour $i \in \{1,\dots,N\}$, alors on remarque que $(u_i^{n+1})_{i=1,\dots,N}$ et $(w_i)_{i=1,\dots,N}$ vérifient (2.5.49)–(2.5.50) avec $\overline{w}_i = w_i$ pour $i = 1,\dots,N$. On en déduit que $w = F(w)$.
5. Soit $w = (w_i)_{i=1,\dots,N}$ t.q. $w = F(w)$. Montrer que pour tout
   Par définition de F, on a $F(w) = \tilde{w}$ avec $(\tilde{u},\tilde{w}) \in \mathbb{R}^N \times \mathbb{R}^N$, $\tilde{u} = (\tilde{u}_i)_{i=1,\dots,N}$, $\tilde{w} = (\tilde{w}_i)_{i=1,\dots,N}$, t.q. :

$$\varphi(\tilde{u}_i) = \tilde{w}_i, \text{ pour tout } i \in \{1,\dots,N\},$$

$$\tilde{u}_i + \frac{2k}{h^2}\tilde{w}_i = \frac{k}{h^2}(w_{i-1} + w_{i+1}) + u_i^n + kv_i^n, \text{ pour tout } i = 1,\dots,N. \quad (2.7.81)$$

   Comme $F(w) = w$, on a donc $\tilde{w}_i = w_i$ et on obtient l'existence de $u_i^{n+1} = \tilde{u}_i$ tel que $w_i = \varphi(u_i^{n+1})$ pour $i = 1,\dots,N$. Il suffit alors de remplacer w_i et $\tilde{w}_i$ par $\varphi(u_i^{n+1}$ dans (2.7.81) pour conclure que $\{u_i^{n+1}$, $i = 1,\dots,N\}$ est solution de (2.5.48).
6. On vient de montrer dans les questions précédentes que $\{u_i^{n+1}$, $i = 1,\dots,N\}$ est solution de (2.5.48) si et seulement si w défini par $w_i = \varphi(u_i^{n+1}$ est solution de $w = F(w)$, où F est définie par (2.5.49) $-$ $-$(2.5.50). Comme F est une application strictement croissante, il existe un unique point fixe $w = F(w)$. Donc par définition de F il existe une unique famille $\{u_i^{n+1}$, $i = 1,\dots,N\}$ solution de (2.5.48).

Question 3 (Estimation $L^\infty(]0,1[\times]0,T[)$ sur u)

La relation à démontrer par récurrence est clairement vérifiée au rang $n = 0$, par définition de A. Supposons qu'elle soit vraie jusqu'au rang n, et démontrons–la au rang $n+1$. La relation (2.5.48) s'écrit encore :

$$u_i^{n+1} = u_i^n + \frac{k}{h^2}(\varphi(u_{i-1}^{n+1}) - \varphi(u_i^{n+1})) + \frac{k}{h^2}(\varphi(u_{i+1}^{n+1} - \varphi(u_i^{n+1})) + kv_i^n, i = 1, \ldots, N, n = 0, \ldots, M-1,$$

Supposons que i est tel que $u_i^{n+1} = \min_{j=1,\ldots,N} u_j^{n+1}$. Comme φ est croissante, on a dans ce cas : $\varphi(u_{i-1}^{n+1}) - \varphi(u_i^{n+1}) \geq 0$ et $\varphi(u_{i+1}^{n+1} - \varphi(u_i^{n+1}) \geq 0$, et on en déduit que $\min_{j=1,\ldots,N} u_j^{n+1} \geq u_i^n - kB$ d'où, par hypothèse de récurrence, $\min_{j=1,\ldots,N} u_j^{n+1} \geq -A - nkB - kB$. Un raisonnement similaire en considérant maintenant i tel que $u_i^{n+1} = \max_{j=1,\ldots,N} u_j^{n+1}$ conduit à : $\max_{j=1,\ldots,N} u_j^{n+1} \leq u_i^n + kB \leq A + nkB + kB$. On a donc bien : $-A - (n+1)kB \leq u_i^{n+1} \leq A + (n+1)kB$, pour tout $i = 1, \ldots, N$ et tout $n = 0, \ldots, M$.
On en déduit alors que $\|u^n\|_{L^\infty(]0,1[)} \leq c_{u_0,v,T}$, avec $c_{u_0,v,T} = A + BT$.

Question 4 (Estimation de la dérivée p.r. à x de $\varphi(u)$)
En multipliant (2.5.48) par u_i^{n+1} et en sommant sur i, on obtient $A_n + B_n = C_n$, avec

$$A_n = \sum_{i=1}^N \frac{u_i^{n+1} - u_i^n}{k} u_i^{n+1},\ B_n = -\sum_{i=1}^N \frac{\varphi(u_{i-1}^{n+1}) - 2\varphi(u_i^{n+1}) + \varphi(u_{i+1}^{n+1})}{h^2} u_i^{n+1} \text{ et } C_n = \sum_{i=1}^N v_i^n u_i^{n+1}.$$

En utilisant l'inégalité $a^2 - ab = \frac{a^2}{2} - \frac{b^2}{2}$, on obtient :

$$A_n \geq \alpha_{n+1} - \alpha_n, \text{ avec } \alpha_n = \frac{1}{2k} \sum_{i=1}^N (u_i^n)^2.$$

En développant B_n, on obtient :

$$B_n = -\frac{1}{h^2}\Big(\sum_{i=1}^N (\varphi(u_{i-1}^{n+1}) - \varphi(u_i^{n+1}))u_i^{n+1} + \sum_{i=1}^N (-\varphi(u_i^{n+1}) + \varphi(u_{i+1}^{n+1}))u_i^{n+1})\Big).$$

Par un changement d'indice sur les sommes, on obtient alors :

$$B_n = -\frac{1}{h^2}\Big(\sum_{i=0}^{N-1} (\varphi(u_i^{n+1}) - \varphi(u_{i+1}^{n+1}))u_{i+1}^{n+1} - \sum_{i=1}^N (-\varphi(u_i^{n+1}) + \varphi(u_{i+1}^{n+1}))u_i^{n+1})\Big).$$

En tenant compte du fait que $u_0^{n+1} = u_1^{n+1}$, $u_{N+1}^{n+1} = u_N^{n+1}$, pour tout $n = 0, \ldots, M-1$, on obtient alors que :

$$B_n = \frac{1}{h^2}\Big(\sum_{i=1}^{N-1} (\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1}))(u_{i+1}^{n+1} - u_i^{n+1})\Big).$$

En utilisant le caractère lipschitzien de φ, on obtient la minoration suivante :

$$B_n \geq \frac{1}{Lh^2} \sum_{i=1}^{N-1} (\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1}))^2.$$

Enfin, on majore C_n :

$$C_n \leq \frac{Bc_{u_0,v,T}}{h}.$$

L'égalité $A_n + B_n = C_n$ entraîne donc :

$$\alpha_{n+1} - \alpha_n + \frac{1}{Lh^2}\sum_{i=1}^{N-1}(\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1}))^2 \leq \frac{Bc_{u_0,v,T}}{h}.$$

En sommant pour $n = 0$ à $M-1$, et en notant que $\alpha_M \geq 0$,on obtient alors :

$$\frac{1}{Lh^2}\sum_{n=0}^{M-1}\sum_{i=1}^{N-1}(\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1}))^2 \leq \frac{Bc_{u_0,v,T}}{h} + \alpha_0.$$

Il reste à remarquer que $\alpha_0 \leq \frac{h}{2k}c_{u_0,v,T}^2$ pour conclure que :

$$\sum_{n=0}^{M-1}\sum_{i=1}^{N-1}(\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1}))^2 \leq C_1\frac{h}{k}, \text{ avec } C_1 = Lc_{u_0,v,T}(B + \frac{1}{2}c_{u_0,v,T}).$$

Question 5 (Estimation de la dérivée p.r. à t de $\varphi(u)$)
Multiplions (2.5.48) par $\varphi(u_i^{n+1}) - \varphi(u_i^n)$ et sommons pour $i = 1, \ldots, N$. On obtient :

$$A_n + B_n = C_n, \tag{2.7.82}$$

avec $A_n = \sum_{i=1}^{N}\frac{u_i^{n+1} - u_i^n}{k}(\varphi(u_i^{n+1}) - \varphi(u_i^n))$, $B_n = -\sum_{i=1}^{N}\frac{\varphi(u_{i-1}^{n+1}) - 2\varphi(u_i^{n+1}) + \varphi(u_{i+1}^{n+1})}{h^2}(\varphi(u_i^{n+1}) - \varphi(u_i^n))$ et $C_n = \sum_{i=1}^{N} v_i^n(\varphi(u_i^{n+1}) - \varphi(u_i^n))$.
En utilisant le caractère lipschitzien de φ, on obtient la minoration suivante :

$$A_n \geq \frac{1}{Lk}\sum_{i=1}^{N-1}(\varphi(u_i^{n+1}) - \varphi(u_i^n))^2. \tag{2.7.83}$$

En développant B_n, on obtient :

$$B_n = -\frac{1}{h^2}\Big(\sum_{i=1}^{N}(\varphi(u_{i-1}^{n+1}) - \varphi(u_i^{n+1}))(\varphi(u_i^{n+1}) - \varphi(u_i^n)) + \sum_{i=1}^{N}(-\varphi(u_i^{n+1}) + \varphi(u_{i+1}^{n+1}))(\varphi(u_i^{n+1}) - \varphi(u_i^n))\Big).$$

Par un changement d'indice sur les sommes, on obtient alors :

$$B_n = -\frac{1}{h^2}\Big(\sum_{i=0}^{N-1}(\varphi(u_i^{n+1}) - \varphi(u_{i+1}^{n+1}))(\varphi(u_{i+1}^{n+1}) - \varphi(u_{i+1}^n)) + \sum_{i=1}^{N}(-\varphi(u_i^{n+1}) + \varphi(u_{i+1}^{n+1}))(\varphi(u_i^{n+1}) - \varphi(u_i^n)\Big).$$

En tenant compte du fait que $u_0^{n+1} = u_1^{n+1}$, $u_{N+1}^{n+1} = u_N^{n+1}$, pour tout $n = 0, \ldots, M-1$, on obtient alors que :

$$B_n = \frac{1}{h^2}\Big(\sum_{i=1}^{N-1}(\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1}))((\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1})) - (\varphi(u_{i+1}^n) - \varphi(u_i^n))\Big).$$

En utilisant à nouveau la relation $a(a-b) \geq \frac{a^2}{2} - \frac{b^2}{2}$, on obtient :

$$B_n \geq \beta_{n+1} - \beta_n, \text{ avec } \beta_n = \frac{1}{2h^2} \sum_{i=1}^{N-1} (\varphi(u_{i+1}^n) - \varphi(u_i^n))^2 \tag{2.7.84}$$

Enfin, on majore C_n par :

$$C_n \leq \frac{1}{2Lk} \sum_{i=1}^{N} (\varphi(u_i^{n+1}) - \varphi(u_i^n))^2 + C \sum_{i=1}^{N} k \geq C\frac{k}{h}. \tag{2.7.85}$$

En utilisant (2.7.82), (2.7.83), (2.7.84) et (2.7.85), on obtient :

$$\frac{1}{2Lk} \sum_{i=1}^{N} (\varphi(u_i^{n+1}) - \varphi(u_i^n))^2 + \beta_{n+1} - \beta_n \leq C\frac{k}{h}. \tag{2.7.86}$$

En sommant sur n, on obtient d'une part, en utilisant le fait que $\beta_n \geq 0$:

$$\sum_{n=0}^{M-1} \sum_{i=1}^{N} (\varphi(u_i^{n+1}) - \varphi(u_i^n))^2 \leq 2LC\frac{k}{h} + 2L\beta_0 k. \tag{2.7.87}$$

d'autre part, en utilisant que le fait que le premier terme est positif, on obtient par (2.7.86) une majoration sur β_M, et donc sur β_n pour tout $n \leq M$:

$$\beta_n \leq \frac{C}{h} + \beta_0. \tag{2.7.88}$$

Il ne reste donc plus qu'à majorer β_0 pour obtenir (2.5.52) et (2.5.53). Par définition, on a

$$\beta_0 = \sum_{i=1}^{N-1} \frac{\varphi(u_i^0) - \varphi(u_{i+1}^0)}{2h^2}.$$

En utilisant le fait que φ est lipschitzienne et que la différence entre u_i^0 et u_{i+1}^0 est en h, on obtient (2.5.53) à partir de (2.7.87) et (2.5.52) à partir de (2.7.88).

Question 6 Par définition de la fonction u_h, et grâce au résultat de la question 3, on a :

$$\begin{aligned} \sup_{\substack{x\in](i-1)h,ih[\\ t\in[nk,(n+1)k]}} u_h(x,t) &\leq \frac{t-nk}{k}\|u_h^{(n+1)}\|_\infty + \frac{(n+1)k-t}{k}\|u_h^{(n)}\|_\infty \\ &\leq c_{u_0,v,T}. \end{aligned}$$

ce qui prouve que la suite $(u_h)_{M\in\mathbb{N}^\star}$ est bornée dans $L^\infty(]0,1[\times]0,T[)$. Comme φ est continue, on en déduit immédiatement que $(\varphi(u_h))_{M\in\mathbb{N}^\star}$ est bornée dans $L^\infty(]0,1[\times]0,T[)$

Chapitre 3

Méthodes variationnelles

3.1 Exemple de problèmes variationnels

3.1.1 Le problème de Dirichlet

Soit Ω un ouvert borné de $\mathbb{R}^d$, $d \geq 1$. On considère le problème suivant :

$$\begin{cases} -\Delta u = f, \text{ dans } \Omega, \\ u = 0 \quad \text{sur } \partial\Omega, \end{cases} \tag{3.1.1}$$

où $f \in C(\bar{\Omega})$ et $\Delta u = \partial_1^2 u + \partial_2^2 u$, où l'on désigne par $\partial_i^2 u$ la dérivée partielle d'ordre 2 par raport à la i-ème variable.

Définition 3.1 *On appelle solution classique de (3.1.1) une fonction $u \in C^2(\bar{\Omega})$ qui vérifie (3.1.1).*

Soit $u \in C^2(\bar{\Omega})$ une solution classique de (3.1.1), et soit $\varphi \in C_c^\infty(\Omega)$, où $C_c^\infty(\Omega)$ désigne l'ensemble des fonctions de classe C^∞ à support compact dans Ω. On multiplie (3.1.1) par φ et on intègre sur Ω (on appellera par la suite φ "fonction test") : on a donc :

$$\int_\Omega -\Delta u(x)\varphi(x)dx = \int_\Omega f(x)\varphi(x)dx.$$

Notons que ces intégrales sont bien définies, puisque $\Delta u \in C(\Omega)$ et $f \in C(\Omega)$. Par intégration par parties (formule de Green), on a :

$$\begin{aligned} \int_\Omega -\Delta u(x)\varphi(x)dx &= -\sum_{i=1}^d \int_\Omega \partial_i^2 u(x)\varphi(x)dx \\ &= \sum_{i=1}^d \int_\Omega \partial_i u(x)\varphi(x)dx + \sum_{i=1}^d \int_{\partial\Omega} \partial_i u \cdot n_i(s)\varphi(s)d\gamma(s) \end{aligned}$$

où n_i désigne la i-ème composante du vecteur unitaire normal à la frontière $\partial\Omega$ de Ω, et extérieur à Ω, et $d\gamma$ désigne le symbole d'intégration sur $\partial\Omega$. Comme φ est nulle sur $\partial\Omega$, on obtient:

$$\sum_{i=1}^d \int_\Omega \partial_i u(x)\partial_i\varphi(x)dx = \int_\Omega f(x)\varphi(x)dx.$$

ce qui s'écrit encore:

$$\int_\Omega \nabla u(x)\cdot\nabla\varphi(x)dx = \int_\Omega f(x)\varphi(x)dx. \tag{3.1.2}$$

Donc toute solution classique de (3.1.1) satisfait (3.1.2)
Prenons maintenant comme fonction test φ, non plus une fonction de $C_c^\infty(\Omega)$, mais une fonction de $H_0^1(\Omega)$. On rappelle que l'espace $H_0^1(\Omega)$ est défini comme l'adhérence de $C_c^\infty(\Omega)$ dans $H^1(\Omega) = \{u \in L^2(\Omega); Du \in L^2(\Omega)\}$, où Du désigne la dérivée faible de u, voir par exemple [1]. On rappelle que l'espace $H^1(\Omega)$ muni du produit scalaire

$$(u,v)_{H^1} = \int_\Omega u(x)v(x)dx + \sum_{i=1}^d \int_\Omega D_iu(x)D_iv(x)dx \tag{3.1.3}$$

est un espace de Hilbert. Les espaces $H^1(\Omega)$ et $H_0^1(\Omega)$ font partie des espaces dits "de Sobolev" (voir [1] pour une introduction).
Si $\varphi \in H_0^1(\Omega)$, par définition, il existe $(\varphi_n)_{n\in\mathbb{N}} \subset C_c^\infty(\Omega)$ telle que

$$\varphi_n \to \varphi \text{ dans } H^1 \text{ lorsque } n \to +\infty,$$

Soit encore

$$\|\varphi_n - \varphi\|_{H^1} = \|\varphi_n - \varphi\|_{L^2}^2 + \sum \|D_i\varphi_n - D_i\varphi\|_{L^2}^2 \to 0 \text{ lorsque } n \to +\infty.$$

Pour chaque fonction $\varphi_n \in C_c^\infty(\Omega)$ on a par (3.1.2) :

$$\sum_{i=1}^N \int_\Omega \partial_i u(x)\partial_i\varphi_n(x)dx = \int_\Omega f(x)\varphi_n(x)dx,\ \forall n \in \mathbb{N}.$$

Or la i-ème dérivée partielle $\partial_i\varphi_n = \dfrac{\partial\varphi_n}{\partial x_i}$ converge vers $D_i\varphi$ dans L^2 donc dans L^2 faible lorsque n tend vers ∞, et φ_n tend vers φ dans $L^2(\Omega)$. On a donc :

$$\int_\Omega \partial_i u(x)\partial_i\varphi_n(x)dx \to \int_\Omega \partial_i u(x)D_i\varphi(x)dx \text{ lorsque } n \to +\infty$$

et

$$\int_\Omega f(x)\varphi_n(x)dx \to \int_\Omega f(x)\varphi(x)dx \text{ lorsque } n \to +\infty.$$

L'égalité (3.1.1) est donc vérifiée pour toute fonction $\varphi \in H_0^1(\Omega)$. Montrons maintenant que si u est solution classique (3.1.1) alors $u \in H_0^1(\Omega)$. En effet, si $u \in C^2(\Omega)$, alors $u \in C(\bar\Omega)$ et donc $u \in L^2(\Omega)$; de plus $\partial_i u \in C(\bar\Omega)$ donc $\partial_i u \in L^2(\Omega)$. On a donc bien $u \in H^1(\Omega)$. Il reste à montrer que $u \in H_0^1(\Omega)$. Pour cela on rappelle (ou on admet ...) les théorèmes de trace suivant :

Théorème 3.2 (Existence de l'opérateur trace) *Soit Ω un ouvert borné de $\mathbb{R}^d$, $d \geq 1$, de frontière $\partial\Omega$ lipschitzienne, alors l'espace $C_i^\infty(\bar\Omega)$ des fonctions de classe C^∞ et à support compact dans $\bar\Omega$ est dense dans $H^1(\Omega)$. On peut donc définir par continuité l'application "trace", qui est linéaire continue de $H^1(\Omega)$ dans $L^2(\partial\Omega)$, définie par :*

$$\gamma(u) = u|_{\partial\Omega} \text{ si } u \in C_c^\infty(\bar\Omega)$$

et par

$$\gamma(u) = \lim_{n\to+\infty} \gamma(u_n) \text{ si } u \in H^1(\Omega),\ u = \lim_{n\to+\infty} u_n, \text{ où } (u_n)_{n\in\mathbb{N}} \subset C_c^\infty(\bar{\Omega}).$$

Dire que l'application (linéaire) γ est continue est équivalent à dire qu'il existe $C \in \mathbb{R}_+$ tel que

$$\|\gamma(u)\|_{L^2(\partial\Omega)} \leq C\|u\|_{H^1(\Omega)} \text{ pour tout } u \in H^1(\Omega). \tag{3.1.4}$$

Notons que $\gamma(H^1(\Omega)) \subset L^2(\Omega)$, mais $\gamma(H^1(\Omega)) \neq L^2(\partial\Omega)$. On note $H^{1/2}(\Omega) = \gamma(H^1(\Omega))$.

Théorème 3.3 (Noyau de l'opérateur trace) *Soit Ω un ouvert borné de $\mathbb{R}^d$ de frontière $\partial\Omega$ lipschitzienne, et γ l'opérateur trace défini par le théorème (3.2). Alors*

$$Ker\gamma = H_0^1(\Omega).$$

Si $u \in C^2(\bar{\Omega})$ est une solution classique de (3.1.1), alors $\gamma(u) = u|_{\partial\Omega} = 0$ donc $u \in Ker\gamma$, et par le théorème 3.3, ceci prouve que $u \in H_0^1(\Omega)$.
Nous avons ainsi montré que toute solution classique de (3.1.1) vérifie $u \in H_0^1(\Omega)$ et l'egalité (3.1.2). Cette remarque motive l'introduction de solutions plus gńérales, qui permettent de s'affranchir de la régularité C^2, et qu'on appellera "solutions faibles".

Définition 3.4 (Formulation faible) *Soit $f \in L^2(\Omega)$, on dit que u est solution faible de (3.1.1) si u est solution de*

$$\begin{cases} u \in H_0^1(\Omega), \\ \displaystyle\sum_{i=1}^{N} \int_\Omega D_i u(x) D_i\varphi(x)dx = \int_\Omega f(x)\varphi(x)dx, \forall\varphi \in H_0^1(\Omega). \end{cases} \tag{3.1.5}$$

Définition 3.5 (Formulation variationnelle) *Soit $f \in L^2(\Omega)$; on dit que u est solution variationnelle de (3.1.1) si u est solution du problème de minimisation suivant :*

$$\begin{cases} u \in H_0^1(\Omega) \\ \\ J(u) \leq J(v) \quad \forall v \in H_0^1(\Omega) \end{cases} \tag{3.1.6}$$
$$\text{avec } J(v) = \frac{1}{2}\int_\Omega \nabla v(x)\cdot\nabla v(x)dx - \int_\Omega f(x)v(x)dx,$$

où on a noté :

$$\int_\Omega \nabla u(x)\cdot\nabla\varphi(x)dx = \sum_{i=1}^{d}\int_\Omega D_i u(x) D_i\varphi(x)dx.$$

On cherche à montrer l'existence et l'unicité de la solution de (3.1.5) et (3.1.6). Pour cela, on utilise le théorème de Lax-Milgram, qu'on rappelle ici :

Théorème 3.6 (Lax-Milgram) *Soit H un espace de Hilbert, soit a une forme bilinéaire continue coercive sur H et $T \in H'$. Il existe un unique élément $u \in H$ tel que $a(u,v) = T(v)$ pour tout $v \in H$. De plus, si a est symétrique, u est l'unique solution du problème de minimisation suivant :*

$$\begin{cases} u \in H, \\ J(u) \leq J(v), \end{cases} \tag{3.1.7}$$

où J est définie de H dans $\mathbb{R}^N$ par :

$$J(v) = \frac{1}{2}a(v,v) - T(v). \tag{3.1.8}$$

Démonstration :

- Si a est symétrique l'existence et l'unicité de u est immédiate par le théorème de représentation de Riesz (car dans ce cas a est un produit scalaire, et la forme linéaire définie par $\varphi \mapsto \int_\Omega f(x)\varphi(x)dx$ est continue pour la norme associée à ce produit scalaire.).
- Si a est non symétrique, on considère l'application de H dans H, qui à u associe Au, défini par :

$$(Au,v) = a(u,v) \quad \forall v \in H.$$

L'application qui à u associe Au est linéaire continue, et

$$(Au,v) \leq a(u,v) \leq M\|u\|\|v\|$$

car a est continue. D'autre part, par le théorème de représentation de Riesz, on a existence et unicité de $\psi \in H$ tel que $T(v) = (\psi,v)$, pour tout $v \in H$. Donc u est solution de $a(u,v) = T(v), \forall v \in H$ si et seulement si $Au = \psi$. Pour montrer l'existence et l'unicité de u, il faut donc montrer que A est bijectif.

Montrons d'abord que A est injectif. On suppose que $Au = 0$. On a $(Au,u) \geq \alpha\|u\|^2$ par coercitivité de a et comme $\|Au\|\,\|v\| \geq (Au,v)$, on a donc :

$$\|Au\| \geq \alpha\|u\|,$$

En conclusion, si $Au = 0 \Rightarrow u = 0$.

Montrons maintenant que A est surjectif. On veut montrer que $AH = H$. Pour cela, on va montrer que AH est fermé et $AH^\top = \{0\}$. Soit $w \in \overline{AH}$; il existe alors une suite $(v_n)_{n\in\mathbb{N}} \subset H$ telle que $Av_n \to w$ dans H. Montrons que la suite $(v_n)_{n\in\mathbb{N}}$ converge dans H. On a :

$$\|Av_n - Av_m\| = \|A(v_n - v_m)\| \geq \alpha\|v_n - v_m\|_H$$

donc la suite $(v_n)_{n\in\mathbb{N}}$ est de Cauchy. On en déduit qu'elle converge vers un certain $v \in H$. Comme A est continue, on a donc : $Av_n \to Av$ dans H, et donc $w = Av \in AH$.
Montrons maintenant que

$$AH^\top = \{0\}$$

Soit $v_0 \in AH^\top$, comme a est coercive, on a :

$$\alpha\|v_0\|^2 \leq a(v_0,v_0) = (Av_0,v_0) = 0,$$

on en déduit que $v_0 = 0$, ce qui prouve que $AH^\top = \{0\}$.

Pour conclure la preuve du théorème, il reste à montrer que si a est symétrique, le problème (3.1.7), qui s'écrit :

$$\begin{cases} u \in H, \\ J(u) \leq J(v), \ \forall v \in H, \end{cases} \tag{3.1.9}$$

est équivalent au problème

$$\begin{cases} u \in H, \\ a(u,v) = T(v), \quad \forall v \in H. \end{cases} \tag{3.1.10}$$

Soit $u \in H$ solution unique de (3.1.10). Soit $w \in H$, on va montrer que $J(u+w) \geq J(u)$.

$$\begin{aligned} J(u+w) &= \frac{1}{2}a(u+w,u+w) - T(u+w) \\ &= \frac{1}{2}a(u,u) + \frac{1}{2}\left[a(u,w) + a(w,u)\right] + \frac{1}{2}a(w,w) - T(u) - T(w) \\ &= \frac{1}{2}a(u,u) + \frac{1}{2}a(w,w) + a(u,w) - T(u) - T(w) \\ &= J(u) + \frac{1}{2}a(w,w) \geq J(u) + \frac{\alpha}{2}\|w\|^2 \end{aligned}$$

Donc $J(u+w) > J(u)$ sauf si $w = 0$. ∎

Montrons qu'on peut appliquer le théorème de Lax Milgram pour les problèmes (3.1.5) et (3.1.6).

Proposition 3.7 (Existence et unicité de la solution de (3.1.1)) *Si $f \in L^2(\Omega)$, il existe un unique $u \in H_0^1(\Omega)$ solution de (3.1.5) et (3.1.6).*

Démonstration : Montrons que les hypothèses du théorème de Lax Milgram sont vérifiées. L'espace $H = H_0^1(\Omega)$ est un espace de Hilbert. La forme bilinéaire a est définie par :

$$a(u,v) = \int_\Omega \nabla u(x) \cdot \nabla v(x) dx \left(= \sum_{i=1}^N \int_\Omega D_i u(x) D_i v(x) dx \right),$$

et la forme linéaire T par :

$$T(v) = \int_\Omega f(x)v(x)dx.$$

Montrons que $T \in H'$; en effet, la forme T est linéaire, et on a :

$$T(v) \leq \|f\|_{L^2}\|v\|_{L^2} \leq \|f\|_{L^2}\|v\|_{H^1}.$$

On en déduit que T est une forme linéaire continue sur $H_0^1(\Omega)$, ce qui est équivalent à dire que $T \in H^{-1}(\Omega)$ (dual topologique de $H_0^1(\Omega)$).

Montrons maintenant que a est bilinéaire, continue et symétrique. La continuité de a se démontre en écrivant que

$$\begin{aligned} a(u,v) = \int_\Omega \nabla u(x) \cdot \nabla v(x) dx &\leq \|\nabla u\|_{L^2}\|\nabla v\|_{L^2} \\ &\leq \|u\|_{H^1}\|v\|_{H^1} \end{aligned}$$

Les caractères bilinéaire et symétrique sont évidents. Montrons maintenant que a est coercitive : en effet,

$$a(v,v) = \int_\Omega \nabla v(x) \cdot \nabla v(x) dx = \sum_{i=1}^N \int_\Omega D_i v(x) D_i v(x) dx \geq \frac{1}{\mathrm{diam}(\Omega)^2 + 1}\|u\|_{H^1}^2,$$

par l'inégalité de Poincaré (voir note page 1 page 24. Comme $T \in H'$ et comme a est linéaire, continue, coercitive donc le théorème de Lax Milgram s'applique : on en conclut qu'il existe une unique fonction $u \in H_0^1(\Omega)$ solution de (3.1.5) et comme a est symétrique, u est l'unique solution du problème de minimisation associée. ∎

Définition 3.8 (Solution forte dans H^2) *Soit $f \in L^2(\Omega)$, on dit que u est solution forte de (3.1.1) dans H^2 si $u \in H^2(\Omega) \cap H_0^1(\Omega)$ vérifie - $\Delta u = f$ dans $L^2(\Omega)$.*

Remarquons que si u est solution forte C^2 de (3.1.1), alors u est solution forte H^2. De même, si u est solution forte H^2 de (3.1.1) alors u est solution faible de (3.1.1). Les réciproques sont fausses. On admettra le théorème (difficile) de régularité, qui s'énonce de la manière suivante :

Théorème 3.9 (Régularité) *Soit Ω un ouvert borné de $\mathbb{R}^d$. On suppose que Ω a une frontière de classe C^2, ou que Ω est convexe à frontière lipschitzienne. Si $f \in L^2(\Omega)$ et si $u \in H_0^1(\Omega)$ est solution faible de (3.1.1), alors $u \in H^2(\Omega)$. De plus, si $f \in H^m(\Omega)$ alors $u \in H^{m+2}(\Omega)$*

Remarque 3.10 (Différences entre les méthodes de discrétisation) *Lorsqu'on adopte une discrétisation par différences finies, on a directement le problème (3.1.1). Lorsqu'on adopte une méthode de volumes finis, on discrétise le "bilan" obtenu en intégrant (3.1.1) sur chaque maille. Lorsqu'on utilise une méthode variationnelle, on discrétise la formulation faible (3.1.5) dans le cas de la méthode de Galerkin, et la formulation variationnelle (3.1.6) dans le cas de la méthode de Ritz.*
Remarquons également que dans la formulation faible, (3.1.5), les conditions aux limites de Dirichlet homogènes $u = 0$ sont prises en compte dans l'espace $u \in H_0^1(\Omega)$, et donc également dans l'espace d'approximation H_N.

3.1.2 Problème de Dirichlet non homogène

On se place ici en dimension 1 d'espace, $d = 1$, et on considère le problème suivant :

$$\left\{ \begin{array}{ll} u'' = f & \text{sur }]0,1[\\ u(0) = a, & \\ u(1) = b, & \end{array} \right. \tag{3.1.11}$$

où a et b sont des réels donnés. Ces conditions aux limites sont dites de type Dirichlet non homogène ; comme a et b ne sont pas forcément nuls, on cherche une solution dans $H^1(\Omega)$ et non plus dans $H_0^1((\Omega))$. Cependant, pour se ramener à l'espace $H_0^1(\Omega)$ (en particulier pour obtenir que le problème est bien posé grâce au théorème de Lax Milgram et à la coercivité de la forme bilinéaire $a(u,v) = \displaystyle\int_\Omega \nabla u(x)\nabla v(x)dx$ sur $H_0^1(\Omega)$, on va utiliser une technique dite de "relèvement". On pose : $u = u_0 + \widetilde{u}$ où u_0 est définie par :

$$u_0(x) = a + (b-a)x.$$

On a en particulier $u_0(0) = a$ et $u_0(1) = b$. On a alors $\widetilde{u}(0) = 0$ et $\widetilde{u}(1) = 0$. La fonction $\widetilde{u}$ vérifie donc le système :

$$\left\{ \begin{array}{l} -\widetilde{u}'' = f, \\ \widetilde{u}(0) = 0, \\ \widetilde{u}(1) = 0, \end{array} \right.$$

dont on connait la formulation faible, et dont on sait qu'il est bien posé (voir paragraphe 3.1.1 page 113). Donc il existe un unique $u \in H^1(\Omega)$ vérifiant $u = u_0 + \widetilde{u}$, où $\widetilde{u} \in H_0^1(\Omega)$ est l'unique solution du problème

$$\int_0^1 \widetilde{u}'v' = \int_0^1 fv \quad \forall v \in H_0^1(]0,1[)$$

De manière plus générale, soit $u_1 \in H^1_{a,b}(]0,1[) = \{v \in H^1\,;\, v(0) = a \text{ et } v(1) = b\}$, et soit $\bar{u} \in H_0^1(]0,1[)$ l'unique solution faible du problème :

$$\left\{ \begin{array}{l} -\bar{u}'' = u_1'' + f, \\ \bar{u}(0) = 0, \\ \bar{u}(1) = 0. \end{array} \right.$$

Alors $\bar{u} + u_1$ est l'unique solution faible de (3.1.11), c'est–à–dire la solution du problème

$$\left\{ \begin{array}{l} u \in H^1_{a,b}(]0,1[), \\ \displaystyle\int_0^1 u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx, \forall v \in H^1_0(]0,1[). \end{array} \right.$$

Remarque 3.11 *Il est facile de montrer que u ne dépend pas du relèvement choisi (voir exercice 30 page 138).*

Considérons maintenant le cas de la dimension 2 d'espace : $d = 2$.
Soit Ω un ouvert borné de $\mathbb{R}^d$, considère le problème :

$$\left\{ \begin{array}{ll} -\Delta u = f & \text{dans } \Omega \\ u = g & \text{sur } \partial\Omega \end{array} \right. \tag{3.1.12}$$

Pour se ramener au problème de Dirichlet homogène, on veut construire un relèvement, c'est à dire une fonction $u_0 \in H^1(\Omega)$ t.q. $\gamma(u_0) = g$ où γ est l'application trace. On ne peut plus le faire de manière explicite comme en dimension 1. En particulier, on rappelle qu'en dimension 2, l'espace $H^1(\Omega)$ n'est pas inclus dans l'espace $C(\bar{\Omega})$ des fonctions continues, contrairement au cas de la dimension 1. Mais si on a $g \in H^{1/2}(\partial\Omega)$, on sait qu'il existe $u_0 \in H^1(\Omega)$ tel que $g = \gamma(u_0)$. On cherche donc u sous la forme $u = \widetilde{u} + u_0$ avec $\widetilde{u} \in H^1_0(\Omega)$ et $u_0 \in H^1(\Omega)$ telle que $\gamma(u_0) = g$ Soit $v \in H^1_0(\Omega)$; on multiplie (3.1.12) par v et on intègre sur Ω :

$$\int_\Omega -\Delta u(x)v(x)dx = \int_\Omega f(x)v(x)dx,$$

c'est–à–dire :

$$\int_\Omega \nabla u(x)\nabla v(x)dx = \int_\Omega f(x)v(x)dx.$$

Comme $u = u_0 + \widetilde{u}$, on a donc :

$$\left\{ \begin{array}{l} \widetilde{u} \in H^1_0(\Omega), \\ \displaystyle\int_\Omega \nabla\widetilde{u}(x)\nabla v(x)dx = \int_\Omega f(x)v(x)dx - \int \nabla u_0(x)\nabla v(x)dx, \forall v \in H^1_0(\Omega) \end{array} \right. \tag{3.1.13}$$

En dimension 2, il n'est pas toujours facile de construire le relèvement u_0. Il est donc usuel, dans la mise en oeuvre des méthodes d'approximation (par exemple par éléments finis), de servir de de la formulation suivante, qui est équivalente à la formulation (3.1.13) :

$$\left\{ \begin{array}{l} u \in \{v \in H^1(\Omega); \gamma(v) = g \text{ sur } \partial\Omega\} \\ \displaystyle\int_\Omega \nabla u(x)\nabla v(x)dx = \int_\Omega f(x)v(x)dx \quad \forall\ v \in H^1_0(\Omega). \end{array} \right. \tag{3.1.14}$$

3.1.3 Problème avec conditions aux limites de Fourier

On considère ici le problème de diffusion avec conditions aux limites de type "Fourier" (ou "Robin" dans la littérature anglo-saxonne).

$$\left\{ \begin{array}{l} -\Delta u = f \text{ dans } \Omega, \\ \nabla u \cdot \mathbf{n} + \lambda u = 0 \text{ sur } \partial\Omega, \end{array} \right. \tag{3.1.15}$$

où :

1. Ω est un ouvert borné de $\mathbb{R}^d$, $d = 1$, 2 ou 3, et $\partial\Omega$ sa frontière,

2. $f \in C^2(\bar{\Omega})$,
3. $\mathbf{n}$ est le vecteur unitaire normal à $\partial\Omega$, extérieur à Ω,
4. $\lambda(x) > 0, \forall x \in \partial\Omega$, est un coefficient qui modélise par exemple un transfert thermique à la paroi.

Supposons qu'il existe $u \in C^2(\bar{\Omega})$ vérifiant (3.1.15). Soit $\varphi \in C^\infty(\bar{\Omega})$ une "fonction test". On multiplie formellement (3.1.15) par φ et on intègre sur Ω.
On obtient :

$$-\int_\Omega \Delta u(x)\varphi(x)dx = \int_\Omega f(x)\varphi(x)dx.$$

Par intégration par parties, on a alors

$$\int_\Omega \nabla u(x)\nabla\varphi(x)dx - \int_{\partial\Omega} \nabla u(x)\cdot\mathbf{n}(x)\varphi(x)d\gamma(x) = \int_\Omega f(x)\varphi(x)dx.$$

Notons que la fonction φ qui n'est pas à support compact, et que la condition aux limites :

$$\nabla u \cdot \mathbf{n} = -\lambda u$$

va donc intervenir dans cette formulation. En remplaçant on obtient :

$$\int_\Omega \nabla u(x)\cdot\nabla\varphi(x)dx + \int_{\partial\Omega} \lambda u(x)\varphi(x)d\gamma(x) = \int_\Omega f(x)\varphi(x)dx, \forall\ \varphi \in C^\infty(\bar{\Omega}).$$

Par densité de $C^\infty(\bar{\Omega})$ dans $H^1(\Omega)$, on a donc également

$$\int_\Omega \nabla u(x)\cdot\nabla\varphi(x)dx + \int_{\partial\Omega} \lambda u(x)\varphi(x) = \int_\Omega f(x)\varphi(x)dx, \forall\ \varphi \in H^1(\Omega).$$

Définition 3.12 (Solution faible) *On dit que u est solution faible de (3.1.15) si u est solution de :*

$$\begin{cases} u \in H^1(\Omega), \\ \displaystyle\int_\Omega \nabla u(x)\cdot\nabla v(x) + dx\int_{\partial\Omega} \lambda(x)u(x)v(x)dx = \int_\Omega f(x)v(x)dx \quad \forall v \in H^1(\Omega). \end{cases} \tag{3.1.16}$$

On peut remarquer que sous les hypothèses :

$$f \in L^2(\Omega),\ \lambda \in L^\infty(\partial\Omega),$$

toutes les intégrales de (3.1.16) sont bien définies. (On rappelle que si $\varphi \in L^2(\Omega)$ et $\psi \in L^2(\Omega)$, alors $\varphi\psi \in L^1(\Omega)$).
Pour vérifier que le problème (3.1.16) est bien posé, on a envie d'appliquer le théorème de Lax-Milgram. Définissons pour cela $a : H^1(\Omega)\times H^1(\Omega) \to \mathbb{R}$ par :

$$a(u,v) = \int_\Omega \nabla u(x)\cdot\nabla v(x)dx + \int \lambda(x)u(x)v(x)dx. \tag{3.1.17}$$

Il est facile de voir que a est une forme bilinéaire symétrique. On peut donc lui associer une forme quadratique définie par :

$$E(v) = \frac{1}{2}\int_\Omega \nabla v(x)\cdot\nabla v(x)dx + \int_{\partial\Omega} \lambda(x)v^2(x)d\gamma(x) - \int_\Omega f(x)v(x)dx. \tag{3.1.18}$$

Définition 3.13 (Solution variationnelle) *On dit que u est solution variationnelle de (3.1.15) si u vérifie :*

$$\begin{cases} u \in H^1(\Omega), \\ E(u) \leq E(v), \forall v \in H^1(\Omega), \end{cases} \tag{3.1.19}$$

où E est défini par (3.1.18).

Lemme 3.14 *On suppose que $\lambda \in L^\infty(\partial\Omega)$. Alors la forme bilinéaire définie par (3.1.17) est continue sur $H^1(\Omega) \times H^1(\Omega)$.*

Démonstration : On a :

$$\begin{aligned} a(u,v) &= \int_\Omega \nabla u(x)\nabla v(x)dx + \int_{\partial\Omega} \lambda(x)u(x)v(x)d\gamma(x) \\ &\leq \|\nabla u\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} + \|\lambda\|_{L^\infty(\partial\Omega)}\|u\|_{L^2(\partial\Omega)}\|v\|_{L^2(\partial\Omega)}. \end{aligned}$$

Or par le théorème de trace (théorème 3.2), et plus particulièrement grâce à la continuité de la trace (3.1.4), on a

$$\|u\|_{L^2(\partial\Omega)} \leq C\|u\|_{H^1(\Omega)}.$$

On en déduit que

$$a(u,v) \leq M\|u\|_{H^1}\|v\|_{H^1}$$

avec $M = 1 + C^2\|\lambda\|_{L^\infty}(\partial\Omega)$. Donc a est bilinéaire continue ∎

Lemme 3.15 *Soit $\lambda \in L^\infty(\partial\Omega)$ tel qu'il existe $\underline{\lambda} > 0$ tel que $\lambda(x) \geq \underline{\lambda}$ p.p. sur $\partial\Omega$. Alors la forme bilinéaire a définie par (3.1.17) est coercitive :*

Montrons qu'il existe $\alpha > 0$ tel que $a(v,v) \geq \alpha\|v\|^2$, pour tout $v \in H^1$ où

$$a(v,v) = \int_\Omega \nabla v(x) \cdot \nabla v(x)dx + \int_\Omega \alpha(x)v^2(x)d\gamma(x).$$

Attention, comme $v \in H^1(\Omega)$ et non $H^1_0(\Omega)$, on ne peut pas écrire l'inégalité de Poincaré, qui nous permettrait de minorer $\int_\Omega \nabla v(x).\nabla v(x)dx$. On va montrer l'existence de α par l'absurde. On suppose que a n'est pas coercive. Dans ce cas : c'est-à-dire que :

$$\forall\ \alpha > 0, \exists\ v \in H^1(\Omega); a(v,v) < \alpha\|v\|^2.$$

On a donc en particulier, en prenant $\alpha = \frac{1}{n}$:

$$\forall\ n \in \mathbb{N}, \exists\ v_n \in H^1(\Omega); a(v_n,v_n) < \frac{1}{n}\|v_n\|^2_{H^1}.$$

Dans cette dernière assertion, on peut prendre v_n de norme 1, puisque l'inégalité est homogène de degré 2. On a donc :

$$\forall n \in \mathbb{N}, \exists\ v_n \in H^1(\Omega); \|v_n\|_{H^1(\Omega)} = 1; a(v_n,v_n) < \frac{1}{n}.$$

Or, par le théorème de Rellich, toute suite bornée $(v_n)_{n\in\mathbb{N}}$ de $H^1(\Omega)$, est relativement compacte dans $L^2(\Omega)$. Comme on a $\|v_n\|_{H^1(\Omega)} = 1$, il existe donc une sous-suite encore notée $(v_n)_{n\in\mathbb{N}} \subset H^1(\Omega)$ telle que v_n converge vers v dans $L^2(\Omega)$ lorsque n tend vers $+\infty$.

De plus, comme :

$$a(v_n,v_n) = \int_\Omega \nabla v_n(x).\nabla v_n(x)dx + \int_{\partial\Omega} v_n(x)v_n(x)dx < \frac{1}{n} \to 00 \text{ lorsque } n \to +\infty,$$

On en déduit que, chaque terme étant positif :

$$\int_\Omega \nabla v_n(x).\nabla v_n(x)dx \to_{n\to+\infty} 0 \tag{3.1.20}$$

et

$$\int_{\partial\Omega} v_n(x)v_n(x)dx \to_{n\to+\infty} 0 \tag{3.1.21}$$

On a donc : $\nabla v_n \to 0$ dans $L^2(\Omega)$ lorsque $n \to +\infty$. On en déduit que

$$\int_\Omega \partial_i v_n(x)\varphi dx \to 0 \text{ lorsque } n \to +\infty, \text{ pour } i = 1,\dots,d.$$

Donc par définition de la dérivée faible (voir note page 23), on a aussi

$$\int_\Omega v_n(x)\partial_i\varphi(x)dx \to 0 \text{ lorsque } n \to +\infty.$$

Comme $v_n \to v$ dans $L^2(\Omega)$ lorsque $n \to +\infty$, on peut passer à la limite ci-dessus et écrire que $\int_\Omega v(x)\partial_i\varphi(x) = 0$. On en déduit que la dérivée faible $D_i v$ existe et est nulle dans Ω. La fonction v est donc constante par composante connexe. Mais par (3.1.21), on a $v = 0$ sur $\partial\Omega$, et la trace d'une fonction constante est la constante elle-même. On a donc

$$v = 0 \text{ dans } \Omega.$$

On a ainsi montré que

$$D_i v_n \to D_i v \text{ et } v_n \to v = 0 \text{ dans } L^2(\Omega) \text{ lorsque } n \to +\infty.$$

Donc $v_n \to 0$ dans $H^1(\Omega)$ lorsque $n \to +\infty$, ce qui contredit le fait que $\|v_n\|_{H^1(\Omega)} = 1$. On a ainsi montré la coercivité de a. ∎

Proposition 3.16 *Soit $f \in L^2(\Omega)$ et $\lambda \in L^\infty(\Omega)$ t.q. $\lambda \geq \underline{\lambda}$ p.p. avec $\underline{\lambda} > 0$ alors il existe un unique u solution de (3.1.16) qui est aussi l'unique solution de (3.1.19).*

3.1.4 Condition de Neumann

Considérons maintenant le problème (3.1.15) avec $\lambda = 0$, on obtient le problème :

$$\begin{cases} -\Delta u = f, \text{ dans } \Omega \\ \\ \dfrac{\partial u}{\partial n} = 0 \text{ sur } \partial\Omega \end{cases}$$

qu'on appelle problème de Dirichlet avec conditions de Neumann homogènes. En intégrant la première équation du système, il est facile de voir qu'une condition nécessaire d'existence d'une solution de (3.1.4) est que :

$$\int_\Omega -\Delta u(x)dx = \int_{\partial\Omega} \frac{\partial u}{\partial n}(x)dx = \int_\Omega f(x)dx = 0$$

Si la condition aux limites de Neumann est non-homogène : $\dfrac{\partial u}{\partial n} = g$, la condition de compatibilité devient

$$\int_\Omega f(x)dx + \int_{\partial\Omega} g(x)d\gamma(x) = 0.$$

Remarquons que si $\alpha = 0$, la forme bilinéaire est

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

et que celle-ci n'est pas coercive sur $H^1(\Omega)$. De fait, il est clair que la solution de (3.1.4) n'est pas unique, puisque si u est solution de (3.1.4) alors $u+c$ est aussi solution, pour tout $c \in \mathbb{R}$. Pour éviter ce problème on va chercher les solutions de (3.1.4) à moyenne nulle. On cherche donc à résoudre (3.1.4) dans l'espace

$$H = \{v \in H^1(\Omega); \int_\Omega v(x)dx = 0\}$$

On admettra que a est coercive sur H (ceci est vrai grâce à l'inégalité de Poincaré–Wirtinger [1]. Le problème

$$\begin{cases} u \in H, \\ a(u,v) = \int fv \quad \forall v \in H, \end{cases}$$

admet donc une unique solution.

3.1.5 Formulation faible et formulation variationnelle.

Nous donnons ici un exemple de problème pour lequel on peut établir une formulation faible, mais pas variationnelle. On se place en une dimension l'espace $N = 1$, et on considère $\Omega =]0,1[$ et $f \in L^2(]0,1[)$. On s'intéresse au problème suivant (dit "d'advection diffusion") :

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

Cherchons une formulation faible. Par la même méthode qu'au paragraphe 3.1.1, on choisit $v \in H_0^1(\Omega)$, on multiplie (3.1.5) par v et on intègre par parties :

$$\int_\Omega u'(x)v'(x)dx + \int_\Omega u'(x)v(x)dx = \int_\Omega f(x)v(x)dx.$$

1. L'inégalité de Poincaré–Wirtinger s'énonce de la faqn suivante : soit Ω un ouvert borné de $\mathbb{R}^d$ de frontière lipschitzienne, alors il existe $C \in \mathbb{R}_+$, ne dépendant que de Ω, Ω, tel que pour tout $u \in H^1(\Omega)$, on a :

$$\|u\|^2_{L^2(\Omega)} \leq C|u|^2_{H^1(\Omega)} + 2(\mathrm{m}(\Omega))^{-1}(\int_\Omega u(x)dx)^2$$

Il est donc naturel de poser :

$$a(u,v) = \int_\Omega u'(x)v'(x)dx + \int_\Omega u'(x)v(x)dx, \text{ et } T(v) = \int_\Omega f(x)v(x)dx.$$

Il est évident que T est une forme linéaire continue sur $H_0^1(\Omega)$ (c'est à dire $T \in H^{-1}(\Omega)$) et que la forme a est bilinéaire continue, mais pas symétrique. De plus elle est coercive : En effet, on a :

$$\begin{aligned} a(u,u) &= \int_\Omega u'^2(x)dx + \int_\Omega u'(x)u(x)dx \\ &= \int_\Omega u'^2(x)dx + \int_\Omega \frac{1}{2}(u^2)'(x)dx \end{aligned}$$

Or, comme $u \in H_0^1(\Omega)$, on a $u = 0$ sur $\partial\Omega$ et donc $\displaystyle\int_\Omega (u^2)'(x)dx = u^2(1) - u^2(0) = 0$. On en déduit que : $a(u,u) = \displaystyle\int_0^1 (u')^2$, et par l'inégalité de Poincaré (voir page 24), on conclut que a est coercive sur $H_0^1(\Omega)$. On en déduit par le théorème de Lax Milgram, l'existence et l'unicité de u solution du problème :

$$\left\{ \begin{array}{l} u \in H_0^1(]0,1[) \\ \displaystyle\int_0^1 (u'(x)v'(x) + u'(x)v(x))dx = \int_0^1 f(x)v(x)dx. \end{array} \right.$$

3.2 Méthodes de Ritz et Galerkin

3.2.1 Principe général de la méthode de Ritz

On se place sous les hypothèses suivantes :

$$\left\{ \begin{array}{l} H \text{ est un espace Hilbert} \\ a \text{ est une forme bilinéaire continue coercitive et symétrique} \\ T \in H' \end{array} \right. \tag{3.2.22}$$

On cherche à calculer $u \in H$ telle que :

$$a(u,v) = T(v), \qquad \forall v \in H,$$

ce qui revient à calculer $u \in H$ solution du problème du problème de minimisation (3.1.7), avec J définie par (3.1.8). L'idée de la méthode de Ritz est de remplacer H par un espace $H_N \subset H$ de dimension finie (où $\dim H_N = N$), et de calculer u_N solution de

$$\left\{ \begin{array}{l} u_N \in H_N \\ J(u_N) \leq J(v), \qquad \forall v \in H_N, \end{array} \right. \tag{3.2.23}$$

en espérant que u_N soit "proche" (en un sens à définir) de u.

Théorème 3.17 *Sous les hypothèses (3.2.22), si H_N est un s.e.v. de H et $\dim H_N < +\infty$ alors le problème (3.2.23) admet une unique solution.*

Démonstration : Puisque H_N est un espace de dimension finie inclus dans H, c'est donc aussi un Hilbert. On peut donc appliquer le théorème de Lax Milgram, et on en déduit l'existence et l'unicité de $u_N \in H_N$ solution de (3.2.23), qui est aussi solution de :

$$\begin{cases} u_N \in H_N, \\ a(u_N,v) = T(v), \quad \forall v \in H_N. \end{cases}$$

■

Nous allons maintenant exposer une autre méthode de démonstration du théorème 3.17, qui a l'avantage d'être constructive, et qui nous permet d'introduire les idées principales des méthodes numériques envisagées plus loin. Comme l'espace H_N considéré dans le théorème 3.2.23 est de dimension N, il existe une base $(\phi_1,\ldots,\phi_N)$ de H_N. Si $u \in H_N$, on peut donc développer $u = \sum_{i=1}^{N} u_i\phi_i$. On note :

$$U = (u_1,\ldots,u_N)^t \in \mathbb{R}^N$$

L'application ξ qui à u associe U est une bijection de H_N dans $\mathbb{R}^N$. Posons $j = J \circ \xi^{-1}$. On a donc :

$$j(U) = J(u).$$

Or :

$$\begin{aligned} J(u) &= \frac{1}{2}a\left(\sum_{i=1}^{N} u_i\phi_i, \sum_{i=1}^{N} u_i\phi_i\right) - T\left(\sum_{i=1}^{N} u_i\phi_i\right) \\ &= \frac{1}{2}\sum_{i=1}^{N} \quad \sum_{j=1}^{N} u_iu_ja(\phi_i,\phi_j) - \sum_{i=1}^{N} u_iT(\phi_i). \end{aligned}$$

On peut donc écrire $J(u)$ sous la forme :

$$J(u) = \frac{1}{2}U^t\mathcal{K}U - U^t\mathcal{G} = j(U),$$

où $\mathcal{K} \in M^{N,N}(\mathbb{R})$ est définie par $\mathcal{K}_{ij} = a(\phi_i,\phi_j)$, et où $\mathcal{G}_i = T(\phi_i)$. Chercher u_N solution de (3.2.23) est donc équivalent à chercher U solution de :

$$\begin{cases} U \in \mathbb{R}^N, \\ j(U) \leq j(V), \quad \forall V \in \mathbb{R}^N. \end{cases} \tag{3.2.24}$$

où

$$j(V) = \frac{1}{2}V^t\mathcal{K}V - V^t\mathcal{G}. \tag{3.2.25}$$

Il est facile de vérifier que la matrice $\mathcal{K}$ est symétrique définie positive. Donc j est une fonctionnelle quadratique sur $\mathbb{R}^N$, et on a donc existence et unicité de $U \in \mathbb{R}^N$ tel que $j(U) \leq j(V) \quad \forall V \in \mathbb{R}^N$. La solution du problème de minimisation (3.2.24) est aussi la solution du système linéaire $\mathcal{K}U = \mathcal{G}$; on appelle souvent $\mathcal{K}$ la matrice de rigidité.

Proposition 3.18 (Existence et unicité de la solution du problème de minimisation) *. Soit $j = \mathbb{R}^N \to \mathbb{R}$ définie par (3.2.25). Il existe un unique $u \in \mathbb{R}^N$ solution du problème de minimisation (3.2.24).*

Démonstration : Ceci est une conséquence du résultat général de minimisation dans $\mathbb{R}^N$ (voir cours de licence). ∎

Résumé sur la technique de Ritz.

1. On se donne $H_N \subset H$.
2. On trouve une base de H_N.
3. On calcule la matrice de rigidité $\mathcal{K}$ et le second membre $\mathcal{G}$. Les coefficients de $\mathcal{K}$ sont donnés par $\mathcal{K}_{ij} = a(\phi_i,\phi_j)$.
4. On minimise j par la résolution de $\mathcal{K}V = \mathcal{G}$.
5. On calcule la solution approchée : $u^{(N)} = \sum_{i=1}^{N} u_i\phi_i$.

On appelle H_N l'espace d'approximation. Le choix de cet espace sera fondamental pour le développement de la méthode d'approximation. Le choix de H_N est formellement équivalent au choix de la base $(\phi_i)_{i=1\ldots N}$. Pourtant, le choix de cette base est capital même si $u^{(N)}$ ne dépend que du choix de H_N et pas de la base.

Choix de la base Un premier choix consiste à choisir des bases indépendantes de N c'est à dire $\{$ base de $H_{N+1}\} = \{$ base de $H_N\} \cup \{\phi_{N+1}\}$. Les bases sont donc emboitées les unes dans les autres. Considérons par exemple $H = H^1(]0,1[)$, et l'espace d'approximation :

$$H_N = Vect\{1,X\ldots,X^{N-1}\}$$

Les fonctions de base sont donc $\phi_i = X^{i-1}$, $i = 1,\ldots,N$. On peut remarquer que ce choix de base amène à une méthode d'approximation qui donne des matrices pleines. Or, on veut justement éviter les matrices pleines, car les systèmes linéaires associés sont coûteux (en temps et mémoire) à résoudre.

Le choix idéal serait de choisir une base $(\phi_i)i = 1,\ldots,N$ de telle sorte que

$$a(\phi_i,\phi_j) = \lambda_i\delta_{ij}$$

$$\text{où } \delta_{ij} = \begin{cases} 1 \text{ si } i = j \\ 0 \text{ sinon} \end{cases} \tag{3.2.26}$$

On a alors $\mathcal{K} = diag(\lambda_1,\ldots,\lambda_N)$, et on a explicitement : $u^{(N)} = \sum_{i=1}^{N} \frac{T(\phi_i)}{a(\varphi_i,\varphi_i)}\phi_i$. Considérons par exemple le problème de Dirichlet (3.1.1) Si ϕ_i est la i-ème fonction propre de l'opération $-\Delta$ avec conditions aux limites de Dirichlet associée à λ_i, on obtient bien la propriété souhaitée. Malheureusement, il est rare que l'on puisse connaître explicitement les fonctions de base ϕ_i.

Un deuxième choix consiste à choisir des bases dépendantes de N. Mais dans ce cas, la base de H_N n'est pas incluse dans celle de H_{N+1}. La technique des éléments finis qu'on verra au chapitre suivant, est un exemple de ce choix. Dans la matrice $\mathcal{K}$ obtenue est creuse (c'est à dire qu'un grand nombre de ses coefficients sont nuls). Par exemple, pour des éléments finis appliqués à un opérateur du second ordre, on peut avoir un nombre de coefficients non nuls de l'ordre de $0(N)$.

Convergence de l'approximation de Ritz Une fois qu'on a calculé u_N solution de (3.2.24), il faut se préocupper de savoir si $u^{(N)}$ est une bonne approximation de u solution de (3.2.1), c'est à dire de savoir si

$$u^{(N)} \to u \text{ lorsque } N \to +\infty$$

Pour vérifier cette convergence, on va se servir de la notion de consistance.

Définition 3.19 (Consistance) *Sous les hypothèses (3.2.22), on dit que l'approximation de Ritz définie par l'espace $H_N \subset H$ avec $\dim H_N = N < +\infty$ est consistante si $d(H,H_N)$ tend vers 0 lorsque $N \to +\infty$, c'est à dire $d(u,H_N) \to_{N\to+\infty} 0$, $\forall u \in N$ ou encore $\inf_{v\in H_N} \|u-v\| \to_{N\to+\infty} 0$, $\forall u \in H$.*

L'autre notion fondamentale pour prouver la convergence est la stabilité, elle même obtenue grâce à la propriété de coercivité de a. Par stabilité, on entend estimation a priori sur la solution approchée $u^{(N)}$ (avant même de savoir si elle existe), où $u^{(N)}$ est solution de (3.2.24) ou encore de :

$$\begin{cases} a(u^{(N)},v) = T(v) \quad \forall v \in H_N \\ \\ u^{(N)} \in H_N \end{cases} \tag{3.2.27}$$

On a l'estimation a priori suivante sur u_N :

Proposition 3.20 (Stabilité) *Sous les hypothèses du théorème 3.2.22, on a :*

$$\|u^{(N)}\|_H \leq \frac{\|T\|_{H'}}{\alpha}.$$

Démonstration :
Le caractère coercif de a nous permet d'écrire :

$$\alpha\|u^{(N)}\|^2 \leq a(u^{(N)},u^{(N)}).$$

Or comme $u^{(N)}$ est solution de (3.2.27), on a :

$$a(u^{(N)},u^{(N)}) = T(u^{(N)}).$$

Comme T est linéaire continue, on obtient

$$T(u^{(N)}) \leq \|T\|_{H'}\|u^{(N)}\|_H.$$

Théorème 3.21 (Lemme de Céa) *Soit H un espace de Hilbert réel, et a une forme bilinéaire continue sumétrique coercive. Soit T une forme linéaire continue et $T \in H'$, et soit $M > 0$ et $\alpha > 0$ tels que $a(u,v) \leq M\|u\|_H\|v\|_H$ et $a(u,u) \geq \alpha\|u\|_H^2$. Soit $u \in H$ l'unique solution du problème suivant :*

$$\begin{cases} u \in H, \\ a(u,v) = T(v), \forall v \in H. \end{cases} \tag{3.2.28}$$

Soit $H_N \subset H$ tel que $\dim H_N = N$, et soit $u^{(N)} \in H_N$ l'unique solution de

$$\begin{cases} u^{(N)} \in H_N, \\ a(u^{(N)},v) = T(v), \forall v \in H_N. \end{cases} \tag{3.2.29}$$

Alors

$$\|u - u^{(N)}\| \leq \sqrt{\frac{M}{\alpha}} d(u,H_N) \tag{3.2.30}$$

où

$$d(u,H_N) = \inf_{v\in H_N} d(u,v).$$

Démonstration :
Etape 1 : On va montrer que $u^{(N)}$ est la projection de u sur H_N pour le produit scalaire $(\cdot,\cdot)_a$ induit par a, défini de $H \times H$ $(u,v)_a = a(u,v)$. On note $\|u\|_a = \sqrt{a(u,u)}$, la norme induite par le produit scalaire a. La norme $\|.\|_a$ est équivalente à la norme $\|.\|_H$, en effet, grâce à la coercivité et la continuité de la forme bilinéaire a, on peut écrire :

$$\alpha\|u\|_H^2 \leq \|u\|_a^2 \leq M\|u\|_H^2$$

Donc $(H,\|.\|_a)$ est un espace de Hilbert. Soit u la solution de (3.2.28), et soit $v = P_{H_N}u$ la projection orthogonale de u sur H_N relative au produit scalaire $a(.,.)$. Par définition de la projection orthogonale, on a donc

$$v - u \in H_N^{\perp}$$

Soit encore $a(v-u,w) = 0, \forall w \in H_N$. On en déduit que $a(v,w) = a(u,w) = T(w), \forall w \in H$, et donc que $v = u^{(N)}$. On a donc montré que $u^{(N)}$ est la projection orthogonale de v sur H_N, c'est–à–dire $u^{(N)} = P_{H_N}u$.
Etape 2 : On va établir une estimation de la norme de la différence entre u et u_N ; par définition de P_{H_N}, on a :

$$\|u - P_{H_N}u\|_a^2 \leq \|u - v\|_a^2, \forall v \in H_N,$$

ce qui s'écrit (puisque $P_{H_N}u = u^{(N)}$) :

$$a(u-u^{(N)},u-u^{(N)}) \leq a(u-v,u-v), \quad \forall v \in H_N$$

Par coercivité et continuité de la forme bilinéaire a, on a donc :

$$\alpha\|u-u^{(N)}\|_H^2 \leq a(u-u^{(N)},u-u^{(N)}) \leq a(u-v,u-v) \leq M\|u-v\|_H^2, \forall\ v \in H_N.$$

On en déduit que :

$$\|u-u^{(N)}\| \leq \sqrt{\frac{M}{\alpha}}\|u-v\|, \forall v \in H_N.$$

En passant à l'inf sur v, on obtient alors :

$$\|u-u^{(N)}\| \leq \sqrt{\frac{M}{\alpha}} \inf_{v\in H_N} \|u-v\|$$

Ce qui est exactement (3.2.30). ∎

3.2.2 Méthode de Galerkin

On se place maintenant sous les hypothèses suivantes :

$$\begin{cases} H \text{ espace de Hilbert,} \\ a \ : \text{ forme bilinéaire continue et coercive, } T \in H'. \end{cases} \tag{3.2.31}$$

Remarquons que maintenant, a n'est pas nécessairement symétrique, les hypothèses (3.2.31) sont donc plus générales que les hypothèses (3.2.22). On considère le problème

$$\begin{cases} u \in H \\ \\ a(u,v) = T(v),\ v \in H. \end{cases} \tag{3.2.32}$$

Par le théorème de Lax Milgram, il y a existence et unicité de $u \in H$ solution de (3.2.32).
Le principe de la méthode de Galerkin est similaire à celui de la méthode de Ritz. On se donne $H_N \subset H$, tel que dim $H_N < +\infty$, et on cherche à résoudre le problème approché :

$$(P_N) \left\{ \begin{array}{l} u^{(N)} \in H_N, \\ a(u^{(N)},v) = T(v), \forall v \in H_N. \end{array} \right. \tag{3.2.33}$$

Par le théorème de Lax-Milgram, on a immédiatement :

Théorème 3.22 *Sous les hypothèses , si $H_N \subset H$ et dim $H_N = N$, il existe un unique $u^{(N)} \in H_N$ solution de (3.2.33).*

Comme dans le cas de la méthode de Ritz, on va donner une autre méthode, constructive, de démonstration de l'existence et unicité de u_N qui permettra d'introduire la méthode de Galerkin. Comme dim $H_N = N$, il existe une base $(\phi_1 \dots \phi_N)$ de H_N. Soit $v \in H_N$, on peut donc développer v sur la base :

$$v = \sum_{i=1}^{N} v_i \phi_i,$$

et identifier v au vecteur $(v_1 \dots v_N)^t \in \mathbb{R}^N$. En écrivant que u satisfait (3.2.33) pour tout $v = \phi_i = 1,N$:

$$a(u,\phi_i) = T(\phi_i), \forall i = 1, \dots, N,$$

et en développant u sur la base $(\phi_i)_{i=1,\dots,N}$, on obtient :

$$\sum_{j=1}^{N} a(\phi_j,\phi_i) u_j = T(\phi_i), \forall i = 1, \dots, N.$$

On peut écrire cette dernière égalité sous forme d'un système linéaire : $\mathcal{K}U = \mathcal{G}$,

$$\mathcal{K}_{ij} = a(\phi_j,\phi_i) \text{ et } \mathcal{G}_i = T(\phi_i), \text{ pour } i,j = 1, \dots,,N.$$

La matrice $\mathcal{K}$ n'est pas en général symétrique.
Proposition 3.23 *Sous les hypothèses du théorème 3.22 le système linéaire (3.2.2) admet une solution .*
Démonstration : On va montrer que $\mathcal{K}$ est inversible en vérifiant que son noyau est réduit à $\{0\}$. Soit $w \in \mathbb{R}^N$ tel que $\mathcal{K}w = 0$. Décomposons w sur le N base $(\phi_1, \dots, \phi_N)$ de H_N : On a donc : $\sum_{j=1}^{N} a(\phi_j,\phi_i) w_j = 0$. Multiplions cette relation par w_i et sommons pour $i = 1$ à N, on obtient :

$$\sum_{i=1}^{N} \quad \sum_{j=1}^{N} a(\phi_j,\phi_i) w_j w_i = 0.$$

Soit encore : $a(w,w) = 0$. Par coercitivité de a, ceci entraine que $w = 0$. On en déduit que $w_i = 0, \forall_i = 1, \dots, N$, ce qui achève la preuve. ∎

Remarque 3.24 *Si a est symétrique, la méthode de Galerkin est équivalente à celle de Ritz.*

En résumé, la méthode de Galerkin comporte les mêmes étapes que la méthode de Ritz, c'est à dire :

1. On se donne $H_N \subset H$
2. On trouve une base de H_N
3. On calcule $\mathcal{K}$ et $\mathcal{G}$
4. On résout $\mathcal{K}U = \mathcal{G}$
5. On écrit $u^{(N)} = \sum_{i=1}^{N} u_i \phi_i$.

La seule différence est que l'étape 4 n'est pas issue d'un problème de minimisation. Comme pour la méthode de Ritz, il faut se poser la question du choix du sous espace H_N et de sa base, ainsi que de la convergence de l'approximation de u solution de (3.2.32) par $u^{(N)}$ obtenue par la technique de Galerkin. En ce qui concerne le choix de la base $\{\phi_1,\ldots,\phi_N\}$, les possibilités sont les mêmes que pour la méthode de Ritz, voir paragraphe 3.2.1. De même, la notion de consistance est identique à celle donnée pour la méthode de Ritz (voir définition 3.19) et la démonstration de stabilité est identique à celles effectuée pour la méthode de Ritz ; voir proposition 3.20 page 127. On peut alors établir le théorème de convergence :

Théorème 3.25 *Sous les hypothèses du théorème (3.22), si u est la solution de (3.2.32) et u_N la solution de (3.2.33), alors*

$$\|u - u^{(N)}\|_H \leq \frac{M}{\alpha} d(u,H_N), \tag{3.2.34}$$

où M et α sont tels que : $\alpha\|v\|^2 \leq a(v,u) \leq M\|v\|^2$ pour tout v dans H (les réels M et α existent en vertu de la continuité et de la coercivité de a).

Démonstration : Comme la forme bilinéaire a est coercive de constante α, on a :

$$\alpha\|u - u^{(N)}\|_H^2 \leq a(u - u^{(N)},u - u^{(N)})$$

On a donc, pour tout $v \in H$:

$$\alpha\|u - u^{(N)}\|_H^2 \leq a(u - u^{(N)},u - v) + a(u - u^{(N)},v - u^{(N)})$$

Or $a(u - u^{(N)},v - u^{(N)}) = a(u,v - u^{(N)}) - a(u^{(N)},v - u^{(N)})$ et par définition de u et $u^{(N)}$, on a :

$$a(u,v - u^{(N)}) = T(v - u^{(N)})$$

$$a(u^{(N)},v - u^{(N)}) = T(v - u^{(N)})$$

On en déduit que :

$$\alpha\|u - u^{(N)}\|_H^2 \leq a(u - u^{(N)},u - v),\forall v \in H_N,$$

et donc, par continuité de la forme bilinéaire a :

$$\alpha\|u - u^{(N)}\|_H^2 \leq M\|u - u^{(N)}\|_H\|u - v\|_H.$$

On obtient donc :

$$\|u - u^{(N)}\|_H \leq \frac{M}{\alpha}\|u - v\|_H,\forall v \in H_N,$$

ce qui entraine (3.2.34). ∎

Remarque 3.26 *On peut remarquer que l'estimation (3.2.34) obtenue dans le cadre de la métode de Galerkin est moins bonne que l'estimation (3.2.30) obtenue dans le cadre de la méthode de Ritz. Ceci est moral, puisque la méthode de Ritz est un cas particulier de la méthode de Galerkin.*

Grâce au théorème 3.25, on peut remarquer que $u^{(N)}$ converge vers u dans H lorsque N tend vers $+\infty$ dès que $d(u,H_N) \to 0$ lorsque $N \to +\infty$. C'est donc là encore une propriété de consistance dont nous avons besoin. La propriété de consistance n'est pas toujours facile à montrer directement. On utilise alors la caractérisation suivante :

Proposition 3.27 (Caractérisation de la consistance) *Soit V un sous espace vectoriel de H dense dans H On suppose qu'il existe une fonction $r_N : V \to H_N$ telle que*

$$\|v - r_N(v)\|_H \to_{N\to+\infty} 0,$$

alors

$$d(u,H_N) \to_{N\to+\infty} 0$$

Démonstration : Soit $v \in V$, et $w = r_N(v)$. Par définition, on a

$$\begin{aligned} d(u,H_N) &\leq \|u - r_N(v)\|_H \\ &\leq \|u - v\|_H + \|v - r_N(v)\| \end{aligned}$$

Comme V est dense dans H, pour tout $\varepsilon > 0$, il existe $v \in V$, tel que $\|u - v\|_H \leq \varepsilon$. Choisissons v qui vérifie cette dernière inégalité. Par hypothèse sur r_N :

$$\forall \varepsilon > 0, \exists N_0 / N \geq N_0 \text{ alors } \|v - r_N(v)\| \leq \varepsilon.$$

Donc si $N \geq N_0$, on a $d(u,H_N) \leq 2\varepsilon$. On en déduit que $d(u,H_N) \to 0$ quand $N \to +\infty$. ∎

3.2.3 Méthode de Petrov-Galerkin

La méthode de Petrov-Galerkin s'apparente à la méthode de Galerkin. On cherche toujours à résoudre :

$$\begin{cases} a(u,v) = T(v), & \forall v \in H, \\ u \in H \end{cases}$$

Mais on choisit maintenant deux sous-espaces H_N et V_N de H, tous deux de même dimension finie :

$$\dim H_N = \dim V_N = N.$$

On cherche une approximation de la solution du problème dans l'espace H_N, et on choisit comme fonction test les fonctions de base de V_N. On obtient donc le système :

$$\begin{cases} u \in H_N \\ a(u,v) = T(v) \quad \forall v \in V_N \end{cases}$$

On appelle H_N l'espace d'approximation, et V_N l'espace des fonctions test. Si $(\phi_1, \ldots, \phi_N)$ est une base de H_N et $(\psi_1, \ldots, \psi_N)$ une base de V_N, en développant $u^{(N)}$ sur la base de $(\phi_1, \ldots, \phi_N)$. $u^{(N)} = \sum u_j \phi_j$, et en écrivant (3.2.3) pour $v = \phi_j$, on obtient :

$$\begin{cases} u \in H_N \\ a(u,\psi_i) = T(\psi_i), \quad \forall i = 1, \ldots, N. \end{cases}$$

Le système à résoudre est donc :

$$\begin{cases} u^{(N)} = \sum_{i=1}^{N} u_i \phi_i, \\ \mathcal{K}U = \mathcal{G}, \end{cases}$$

avec $\mathcal{K}_j = a(\phi_j, d_i)$ et $\mathcal{G}_i = T(\psi_i)$, pour $i = 1, \ldots, N$.

3.3 La méthode des éléments finis

La méthode des éléments finis est une façon de choisir les bases des espaces d'approximation pour les méthodes de Ritz et Galerkin.

3.3.1 Principe de la méthode

On se limitera dans le cadre de ce cours à des problèmes du second ordre. L'exemple type sera le problème de Dirichlet (3.1.1), qu'on rappelle ici :

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

et l'espace de Hilbert sera l'espace de Sobolev $H^1(\Omega)$ ou $H_0^1(\Omega)$.
On se limitera à un certain type d'éléments finis, dits "de Lagrange". Donnons les principes généraux de la méthode.

Eléments finis de Lagrange Soit $\Omega \subset \mathbb{R}^2$ (ou $\mathbb{R}^3$), Soit H l'espace fonctionnel dans lequel on recherche la solution (par exemple $H_0^1(\Omega)$ s'il s'agit du problème de Dirichlet (3.1.1)). On cherche $H_N \subset H = H_0^1(\Omega)$ et les fonctions de base $\phi_1, \ldots, \phi_N$. On va déterminer ces fonctions de base à partir d'un découpage de Ω en un nombre fini de cellules, appelés, "éléments". la procédure est la suivante :

1. On construit un "maillage" $\mathcal{T}$ de Ω (en triangles ou rectangles) que l'on appelle <u>éléments</u> $\mathcal{K}$.
2. Dans chaque élément, on se donne des points que l'on appelle "noeuds".
3. On définit H_N par :
$$H_N = \left\{ u \ : \Omega \to \mathbb{R} / u_{|K} \in P_k, \forall K \in \mathcal{T} \right\} \cap H$$
où P_k désigne l'ensemble des polynômes de degré inférieur ou égal à k. Le degré des polynômes est choisi de manière à ce que u soit entièrement déterminée par ses valeurs aux noeuds. Pour une méthode d'éléments finis de type Lagrange, les valeurs aux noeuds sont également les "degrés de liberté", c.à.d. les valeurs qui déterminent entièrement la fonction recherchée.
4. On construit une base $\{\phi_i \ldots \phi_N\}$ de H_N tel que le support de ϕ_i soit "le plus petit possible". Les fonctions ϕ_i sont aussi appelées fonctions de forme.

Remarque 3.28 (Eléments finis non conformes) *Notons qu'on a introduit ici une méthode déléments finis conforme, c'est–à–dire que l'espace d'approximation H_N est inclus dans l'espace H. Dans une méthode non conforme, on n'aura plus $H_N \subset H$, et par conséquence, on devra aussi construire une forme bilinéaire approchée $a_{\mathcal{T}}$; on pourra voir à ce sujet l'exercice 37 page 141 où on exprime la méthode des volumes finis comme une méthode déléments finis non conformes.*

Exemple en dimension 1 Soit $\Omega =]0,1[\subset \mathbb{R}$ et soit $H = H_0^1([0,1[)$; on cherche un espace H_N d'approximation de H. Pour cela, on divise l'intervalle $]0,1[$ en N intervalles de longueur $h = \frac{1}{N+1}$. On pose $x_i = i$, $i = 0,N+1$. Les étapes 1. à 4. décrites précédemment donnent dans ce cas :

1. <u>Construction des éléments</u> On a construit $n+1$ éléments $K_i =]x_i,x_{i+1}[, i = 0, \ldots, N$.
2. <u>Noeuds</u> : On a deux noeuds par élément, (x_i et x_{i+1} sont les noeuds de K_i,$i = 0, \ldots, N$) Le fait que $H_N \subset H_0^1(]0,1[)$ impose que les fonctions de H_N soient nulles en $x_0 = 0$ et $x_{N+1} = 1$. On appelle $x_1, \ldots, x_N$ les noeuds libres et x_0,x_{N+1} les noeuds liés. Les degrés de liberté sont donc les valeurs de u en $x_1, \ldots, x_N$. Aux noeuds liés, on a $u(x_0) = u(x_{N+1}) = 0$

3. Choix de l'espace On choisit comme espace de polynôme : $P_1 = \{ax + b, a, b \in \mathbb{R}\}$ et on pose :

$$H_N = \{u \ : \Omega \to \mathbb{R} \text{ t.q. } u_{|K_i} \in P_1, \forall i = 1 \dots N, u \in C(\bar{\Omega}) = C([0,1]) \text{ et } u(0) = u(1) = 0\}.$$

Rappelons que $H = H_0^1(]0,1[) \subset C([0,1])$. Avec le choix de H_N, on a bien $H_N \subset H$.

4. Choix de la base de H_N.
Si on prend les fonctions de "type 1" de la méthode de Ritz, on choisit les fonctions décrites sur la figure 3.1. On a donc $H_1 = Vect\{\phi_1\}$, $H_3 = Vect\{\phi_1, \phi_2, \phi_3\}$, et $H_7 = Vect\{\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6, \phi_7\}$, où $Vect$ désigne le sous espace engendré par la famille considérée. Avec ce choix, on a donc $H_1 \subset H_3 \subset H_7$.

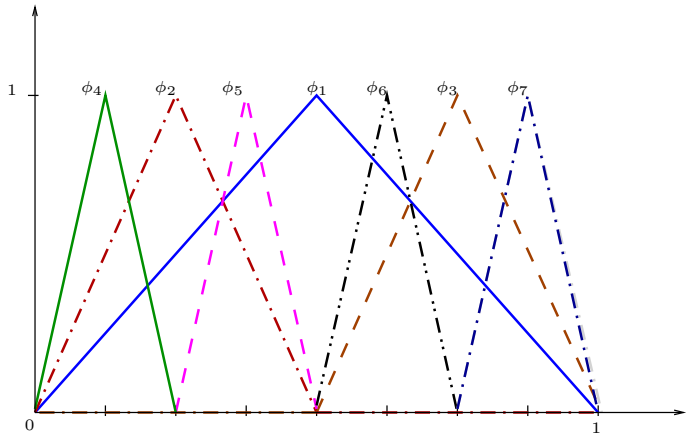

FIG. 3.1 – *Fonctions de forme de type 1 (espaces emboîtés)*

Si maintenant on choisit des fonctions de forme de "type 2" on peut définir ϕ_i pour $i = 1$ à N par :

$$\begin{cases} \phi_i \ : \text{ affine par morceaux, continue} \\ \text{supp}(\phi_i) = [x_{i-1}, x_{i+1}] \\ \phi_i(x_i) = 1 \\ \phi_i(x_{i-1}) = \phi_i(x_{i+1}) = 0 \end{cases} \tag{3.3.35}$$

Il est facile de voir que $\phi_i \in H_N$ et que $\{\phi_1, \dots, \phi_N\}$ engendre H_N, c'est à dire que pour tout $u \in H_N$, il existe $(u_1, \dots, u_N) \in \mathbb{R}^N$ tel que $u = \sum_{i=1}^{N} u_i \phi_i$. On a représenté sur la figure 3.2 les fonctions de base obtenue pour H_3 (à gauche) et H_7 (à droite). On peut remarquer que dans ce cas, les espaces d'approximation ne sont plus inclus les uns dans les autres.

Exemple en dimension 2 Soit Ω un ouvert polygonal de $\mathbb{R}^2$, et $H = H_0^1(\Omega)$. Les étapes de construction de la méthode des éléments finis sont encore les mêmes.

1. Eléments : on choisit des triangles.

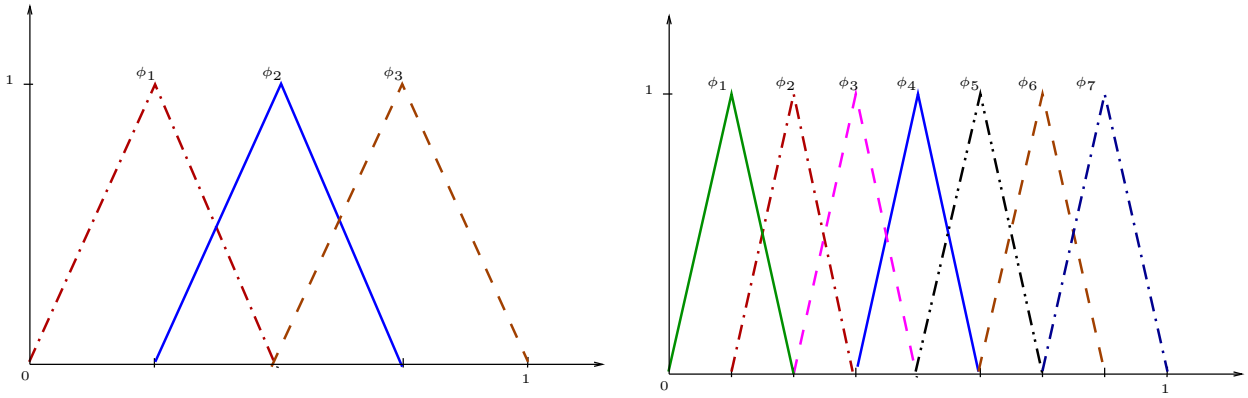

FIG. 3.2 – *Fonctions de forme de type 2 (fonction P1) en une dimension d'espace*

2. Noeuds : on les place aux sommets des triangles. Les noeuds $x_i \in \Omega$ (intérieurs à Ω) sont libres, et les noeuds $x_i \in \partial\Omega$ (sur la frontière de Ω sont liés. On notera Σ l'ensemble des noeuds libres, Σ_F l'ensemble des noeuds liés, et, $\Sigma = \Sigma_I \cup \Sigma_F$.
3. Espace d'approximation L'espace des polynômes est l'ensemble des fonctions affines, noté P_1. Une fonction $p \in P_1$ est de la forme :

$$\begin{aligned} p &: \mathbb{R}^2 \to \mathbb{R}, \\ x = (x_1,x_2)^t &\mapsto a_1x_1 + a_2x_2 + b, \end{aligned}$$

avec $(a_1,a_2,b) \in \mathbb{R}^3$. L'espace d'approximation H_N est donc défini par :

$$H_N = \{u \in C(\bar{\Omega}); u|_K \in P_1, \forall K, \text{ et } u(x_i) = 0, \forall x_i \in \Sigma_F\}$$

4. Base de H_N : On choisit comme base de H_N la famille de fonctions $\{\phi_i\} i = 1,\ldots,N$, où $N = \text{card}(\Sigma_I)$, où ϕ_i est définie, pour $i = 1$ à N, par :

$$\begin{cases} \phi_i \text{ est affine par morceaux,} \\ \phi_i(x_i) = 1, \\ \phi_i(x_j) = 0, \quad \forall j \neq 1. \end{cases} \tag{3.3.36}$$

La fonction ϕ_i associée au noeud x_i a donc l'allure présentée sur la figure 3.3. Le support de chaque fonction ϕ_i (c'est à dire l'ensemble des points où ϕ_i est non nulle), est constitué de l'ensemble des triangles dont x_i est un sommet.

En résumé Les questions à se poser pour construire une méthode d'éléments finis sont donc :

1. La construction du maillage.
2. Un choix cohérent entre éléments, noeuds et espace des polynômes.
3. La construction de l'espace d'apporximation H_N et de sa base $\{\phi_i\}_{i=1\ldots N}$.
4. La construction de la matrice de rigidité $\mathcal{K}$ et du second membre $\mathcal{G}$.
5. L'évaluation de $d(u,H_N)$ en vue de l'analyse de convergence.

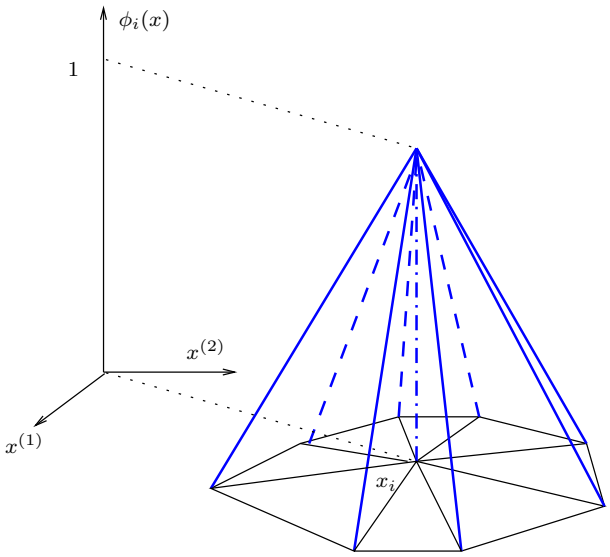

FIG. 3.3 – *Fonction de forme de type 2 (fonction P1) en deux dimensions d'espace*

3.3.2 Construction du maillage, de l'espace H_N et de sa base ϕ_N

Construction des éléments

Soit $\Omega \in \mathbb{R}^2$ un ouvert borné polygonal. On construit un maillage de Ω en divisant $\bar{\Omega}$ en parties fermées $\{K_\ell\}_{\ell=1,\ldots,L}$ où L est le nombre d'éléments.
Les principes pour la construction du maillage sont :

– Eviter les angles trop grands ou trop petits. On préfèrera par exemple les triangles de gauche plutôt que ceux de droite dans la figure 3.4.

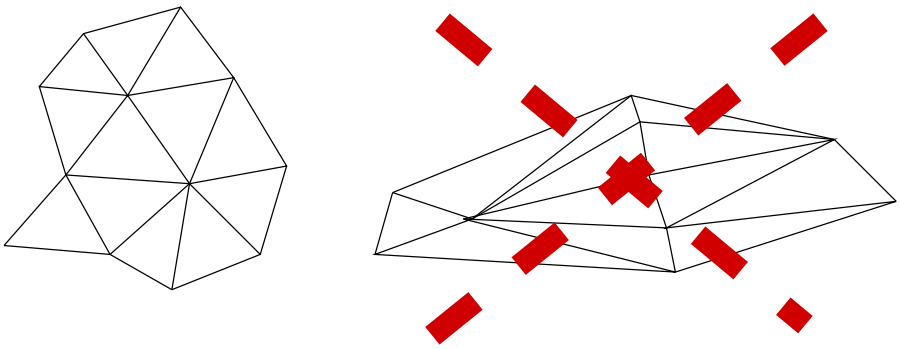

FIG. 3.4 – *Exemple de triangles "bons" (à gauche) et "mauvais" (à droite)*

– Mettre beaucoup d'éléments là où u varie rapidement (ceci ne peut se faire que si on connait a priori les zones de de variation rapide, ou si on a les moyens d'évaluer l'erreur entre la solution exacte du

problème et la solution calculée et de remailler les zones ou celle–ci est jugée trop grande.

– On peut éventuellement mélanger des triangles et des rectangles, mais ceci n'est pas toujours facile.

Il existe un très grand nombre de logiciels de maillages en deux ou trois dimensions d'espace. On pourra pour s'en convaincre utiliser le moteur de recherche google sur internet avec les mots clés : "mesh 2D structured", "mesh 2D unstructured", "mesh 3D structured", "mesh 3D unstructured". Le mot "mesh" est le terme anglais pour maillage, les termes 2D et 3D réfèrent à la dimension de l'espace physique. Le terme "structured" (structuré en français) désigne des maillages que dont on peut numéroter les éléments de façon cartésienne, le terme "unstructured" (non structuré) désigne tous les autres maillages. L'avantage des maillages "structurés" est qu'ils nécessitent une base de données beaucoup plus simple que les maillages non structurés, car on peut connaitre tous les noeuds voisins à partir du numéro global d'un noeud d'un maillage structuré, ce qui n'est pas le cas dans un maillage non structuré (voir paragraphe suivant pour la numérotation des noeuds). La figure 3.5 montre un exemple de maillage de surface

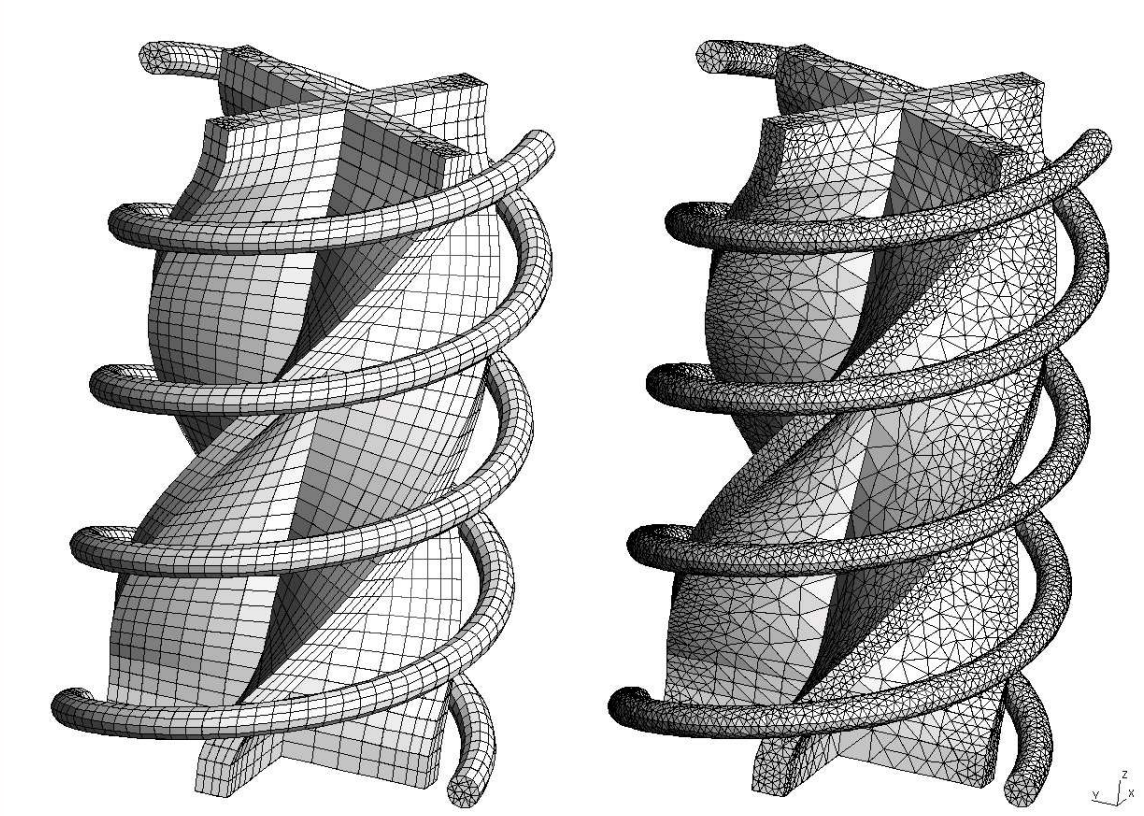

FIG. 3.5 – *Exemple de maillage structuré (à gauche) et non-structuré (à droite) d'une surface surface*

structuré ou non-structuré, pris sur le site web du logiciel Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities, dévelippé par C. Geuzaine and J.-F. Remacle (http://www.geuz.org/gmsh/).

Choix des noeuds

On se donne une famille $\{S_i\}_{i=1,\ldots,M}$ de M points de $\bar{\Omega}$, de composantes (x_i,y_i), pour $i = 1,\ldots,M$.
Le maillage éléments finis est défini par éléments $\{K_\ell\}_{\ell=1\ldots L}$ et les noeuds $\{S_i\}_{i=1\ldots M}$. Ces éléments et noeuds ne peuvent bien sûr pas être choisis indépendamment. Dans le cas général, on choisit tous les éléments de même type (par exemple, des triangles) et on se donne un nombre fixe de noeuds par élément, ce qui détermine le nombre total de noeuds. Chaque noeud appartient donc à plusieurs éléments. Dans le cas d'un maillage structuré tel que celui qu'on a décrit dans la figure 1.3 page 28, une numérotation globale des noeuds est suffisante pour retrouver les éléments dont font partie ce noeud, ainsi que tous les voisins du noeud. Par contre, dans le cas d'un maillage non structuré (un maillage en triangles, par exemple), on aura besoin d'une numérotation locale des noeuds c'est à dire une numérotation des noeuds de chaque élément, pour $k = 1,\ldots,N_\ell$, où N_ℓ est le nombre de noeuds par élément ; on aura également besoin d'une numérotation globale des noeuds, et d'une table de correspondance, l'une qui donne pour chaque élément, les numéros dans la numérotation globale des noeuds qui lui appartiennent.

$$i_r^\ell = \text{ notation globale } (\ell,r) \text{ r-ième noeud de l'élément } \ell$$

Amélioration de la précision

On a vu aux paragraphes précédents que l'erreur entre la solution exacte u recherchée et la solution $u(N)$ obtenue par la méthode de Ritz ou de Galerkin est majorée par une constante fois la distance entre H et H_N. On a donc intérêt à ce que cette distance soit petite. Pour ce faire, il paraît raisonnable d'augmenter la dimension de l'espace H_N. Pour cela, on a deux possibilités :

- augmenter le nombre d'éléments : on augmente alors aussi le nombre global de noeuds, mais pas le nombre local.
- augmenter le degré des polynômes : on augmente alors le nombre de noeuds local, donc on augmente aussi le nombre global de noeuds, mais pas le nombre d'éléments. Ce deuxième choix (augmentation du degré des polynômes) ne peut se faire que si la solution est suffisamment régulière ; si la solution n'est pas régulière, on n'arrivera pas à diminuer $d(H,H_N)$ en augmentant le degré des polynômes.

3.4 Exercices

Exercice 26 (Fonctions H^1 en une dimension d'espace)

Montrer que si $u \in H^1(]0,1[)$ alors u est continue. En déduire que $H^2(]0,1[) \subset C^1([0,1])$.

Exercice 27 (Minimisation de la semi-norme) *Suggestions en page 142, corrigé en page 142*

Soit Ω un ouvert borné de $\mathbb{R}^n$. On suppose que sa frontière est de classe C^1 par morceaux. Etant donné une fonction $u_0 \in H^1(\Omega)$,

1. Montrer qu'il existe une unique fonction $u \in u_0 + H_0^1(\Omega)$ tel que :

$$|u|_{1,\Omega} = \inf_{v \in u_o + H_0^1(\Omega)} |v|_{1,\Omega}.$$

2. Caractériser u comme étant la solution d'un problème aux limites.

Exercice 28 (Formulation faible pour le problème de Dirichlet en 1D) *Corrigé en page 144*

Soit $f \in L^2(]0,1[)$. On s'intéresse au problème suivant:

$$-u_{xx}(x) = f(x),\ x \in]0,1[, \tag{3.4.37}$$
$$u(0) = 0,\ u(1) = 0. \tag{3.4.38}$$

Donner une formulation faible et une formulation variationnelle de (3.4.38).

Exercice 29 (Relèvement en une dimension d'espace) *Suggestions en page 142*

Ecrire une formulation faible pour laquelle on puisse appliquer le théorème de Lax Milgram, dans le cas du problème suivant :

$$\begin{cases} -u''(x) = f(x), \quad x \in [0,1] \\ u'(0) = 0 \\ u(1) = 1. \end{cases} \tag{3.4.39}$$

Exercice 30 (Relèvement) *Corrigé en page 146*

Soient a et $b \in \mathbb{R}$, et $f \in C(\mathbb{R},\mathbb{R})$.

1. Soient u_0 et u_1 définies de $[0,1]$ dans $\mathbb{R}$ par $u_0(x) = a + (b-a)x$ et $u_1(x) = a + (b-a)x^2$. Montrer qu'il existe un unique $\tilde{u}$ (resp. $\bar{u}$) tel que $u = u_0 + \tilde{u}$ (resp. $v = u_1 + \bar{u}$) soit solution de (3.1.11). Montrer que $u = v$.
2. Mêmes questions en supposant maintenant que u_0 et u_1 sont des fonctions de $C^2([0,1])$ telles que $u_0(0) = u_1(0) = a$ et $u_0(1) = u_1(1) = b$.

Exercice 31 (Conditions aux limites de Fourier et Neumann) *Corrigé en page 146*

Soit $f \in L^2(]0,1[)$. On s'intéresse au problème suivant:

$$\begin{array}{l} -u_{xx}(x) + u(x) = f(x),\ x \in]0,1[, \\ u'(0) - u(0) = 0,\ u'(1) = -1. \end{array} \tag{3.4.40}$$

Donner une formulation faible et une formulation variationnelle de (3.4.40); y-a-t-il existence et unicité des solutions faibles de (3.4.40)?

Exercice 32 (Conditions aux limites de Fourier et Neumann, bis) *Corrigé en page 148*

Soit $f \in L^2(]0,1[)$. On s'intéresse au problème suivant:

$$\begin{cases} -u''(x) - u'(x) + u(x) = f(x),\ x \in]0,1[, \\ u(0) + u'(0) = 0,\ u(1) = 1 \end{cases} \tag{3.4.41}$$

1. Donner une formulation faible du problème de la forme

$$\begin{cases} \text{Trouver } u \in H^1(]0,1[); u(1) = 1, \\ a(u,v) = T(v), \forall v \in H. \end{cases} \tag{3.4.42}$$

où $H = \{v \in H^1(]0,1[); v(1) = 0\}$, a et T sont respectivement une forme bilinéaire sur $H^1(]0,1[)$ et une forme linéaire sur $H^1(]0,1[)$, à déterminer.

2. Y-a-t-il existence et unicité de solutions de cette formulation faible?

Exercice 33 (Conditions mêlées) *Suggestions en page 142, corrigé en page 150*

Soit Ω un ouvert borné $\mathbb{R}^d$, $d = 1ou2$, de frontière $\partial\Omega = \Gamma_0 \cup \Gamma_1$, avec $\Gamma_0 \cap \Gamma_1 = \emptyset$; on suppose que la mesure $d-1$ dimensionnelle de Γ_0 est non nulle, et soit $f \in L^2(\Omega)$. On s'intéresse ici au problème suivant:

$$\begin{array}{l} -\Delta u(x) = f(x), x \in \Omega, \\ u(x) = 0, x \in \Gamma_0, \\ \nabla u(x) \cdot \mathbf{n}(x) = 0, x \in \Gamma_1, \end{array} \tag{3.4.43}$$

où $\mathbf{n}$ est la normale unitaire à $\partial\Omega$ extérieure à Ω.

Donner une formulation faible et une formulation variationnelle de (3.4.43) telle qu'on puisse appliquer le lemme de Lax-Milgram. (On rappelle que l'inégalité de Poincaré donnée en bas de page 1 page 24 pour les fonctions de $H_0^1(\Omega)$ est encore valable pour les fonctions de $H^1(\Omega)$ dont la trace est nulle sur un sous-ensemble de $\partial\Omega$ de mesure ($(d-1)$–dimensionnelle) non nulle.)

Exercice 34 (Problème elliptique pour un problème avec conditions mixtes) *Corrigé en page 150*

Soit Ω un ouvert borné $\mathbb{R}^d$, $d = 1ou2$, de frontière $\partial\Omega = \Gamma_0 \cup \Gamma_1$, avec $\Gamma_0 \cap \Gamma_1 = \emptyset$; on suppose que la mesure $d-1$ dimensionnelle de Γ_0 est non nulle. On s'intéresse ici au problème suivant:

$$\begin{array}{l} -\mathrm{div}(p(x)\nabla u(x)) + q(x)u(x) = f(x), x \in \Omega, \\ u(x) = g_0(x), x \in \Gamma_0, \\ p(x)\nabla u(x).\mathbf{n}(x) + \sigma u(x) = g_1(x), x \in \Gamma_1, \end{array} \tag{3.4.44}$$

où:
$f \in L^2(\Omega)$,
$p \in L^\infty(\Omega)$, est telle qu'il existe $\alpha > 0$ t.q. $p(x) \geq \alpha$ p.p.
$q \in L^\infty(\Omega)$, $q \geq 0$,
$\sigma \in \mathbb{R}_+$,
$g_0 \in L^2(\Gamma_0)$ est telle qu'il existe $\tilde{g} \in H^1(\Omega)$ t.q. $\gamma(\tilde{g})|_{\Gamma_0} = g_0$
$g_1 \in L^2(\Gamma_1)$,
$\mathbf{n}$ est la normale unitaire à $\partial\Omega$ extérieure à Ω.

1. Donner une formulation faible et une formulation variationnelle de (3.4.44) telle qu'on puisse appliquer le lemme de Lax-Milgram.

2. On suppose dans cette question que $p \in C^1(\overline{\Omega})$, $q \in C(\overline{\Omega})$ $g_0 \in C(\Gamma_0)$ et $g_1 \in C(\Gamma_1)$. Soit $u \in C^2(\overline{\Omega})$. Montrer que u est solution faible si et seulement si u est une solution classique de (3.4.44).

Exercice 35 (Condition inf-sup) *Corrigé en page 152*

Soit V un espace de Hilbert réel de produit scalaire $(\cdot;\cdot)$ induisant une norme $||\cdot||$. On se donne $a(\cdot;\cdot)$ une forme bilinéaire continue sur $V \times V$, avec M comme constante de continuité. Soit L une forme linéaire continue sur V. On suppose de plus qu'il existe une solution $u \in V$ au problème suivant :

$$a(u,v) = L(v),\ \forall v \in V. \tag{3.4.45}$$

Soit V_h un sous-espace de V de dimension finie. On suppose qu'il existe $\beta_h \in R_+$ telle que :

$$\inf_{(v_h \in V_h, \|v_h\|=1)} \left(\sup_{(w_h \in V_h, \|w_h\|=1)} (a(v_h; w_h)) \right) \geq \beta_h \tag{3.4.46}$$

On cherche alors u_h solution de :

$$\begin{array}{l} u_h \in V_h, \\ a(u_h,v_h) = L(v_h),\ \forall v_h \in V_h \end{array} \tag{3.4.47}$$

1. Montrer que le problème (3.4.47) admet une unique solution.
2. Soit u la solution de (3.4.45) et u_h la solution de (3.4.47). Montrer que :

$$\|u - u_h\| \leq \left(1 + \frac{M}{\beta_h}\right) \inf_{v_h \in V_h} \|u - v_h\|. \tag{3.4.48}$$

Exercice 36 (Condition inf-sup pour un problème mixte) *Corrigé en page 153*

Soient V et Q deux espaces de Hilbert, on note $(\cdot,\cdot)_V$, $\|\cdot\|_V$ et $(\cdot,\cdot)_Q$, $\|\cdot\|_Q$ leurs produits scalaires et normes respectives, et on considère le problème suivant:

$$\left\{ \begin{array}{l} \text{Trouver } u \in V, p \in Q, \text{ tels que} \\ a(u,v) + b(v,p) = (f,v)_H, \forall v \in V, \\ b(u,q) = (g,q)_Q, \forall q \in Q. \end{array} \right. \tag{3.4.49}$$

où a est une forme bilinéaire continue et coercive sur V et b est une application bilinéaire continue de $V \times Q$ dans $\mathbb{R}$.
Pour (u,p) et (v,q) éléments de $V \times Q$, on pose :

$$\begin{array}{rcl} B(u,p;v,q) & = & a(u,v) + b(v,p) + b(u,q), \\ F(v,q) & = & (f,v)_H + (g,q)_Q. \end{array}$$

et on munit $V \times Q$ d'une norme notée $\|(\cdot,\cdot)\|$, définie par $\|(v,q)\| = \|v\|_V + \|q\|_Q$ pour $(v,q) \in V \times Q$.

1. Montrer que B est une forme bilinéaire continue sur $V \times Q$.
2. Montrer que le problème (3.4.49) est équivalent au problème :

$$\left\{ \begin{array}{l} \text{Trouver } (u,p) \in V \times Q, \text{ tels que} \\ B(u,p;v,q) = F(v,q), \forall (v,q) \in V \times Q. \end{array} \right. \tag{3.4.50}$$

On considère maintenant des espaces d'approximation (par exemple construits par élements finis). Soient donc $(V_n)_{n\in\mathbb{N}}$ et $(Q_n)_{n\in\mathbb{N}}$ des espaces de Hilbert de dimension finie tels que $V_n \subset V$ et $Q_n \subset Q$, pour tout $n \in \mathbb{N}$.

3. On suppose dans cette question que la condition suivante (dite condition "inf-sup") est satisfaite :

$$\text{Il existe } \beta \in \mathbb{R}_+^* \text{ (indépendant de } n) \text{ tel que } \inf_{q \in Q_n} \sup_{\substack{w \in V_n \\ \|w\|_V \neq 0}} \frac{b(w,q)}{\|w\|_V} \geq \beta \|q\|_Q. \tag{3.4.51}$$

   (a) Montrer qu'il existe $\alpha \in \mathbb{R}_+^*$ tel que :

$$\text{Pour tout } q \in Q_n \text{ et } v \in V_n, B(v,q;v,-q) \geq \alpha \|v\|_V^2. \tag{3.4.52}$$

   (b) Soit $(v,q) \in V_n \times Q_n$, montrer qu'il existe $w \in V_n$ tel que $\|w\|_V = \|q\|_Q$ et $b(w,q) \geq \beta \|q\|_Q^2$. Montrer que pour ce choix de w, on a :

$$B(v,q;w,0) \geq -M\|v\|_V \|w\|_V + \beta \|q\|_Q^2,$$

   où M est la constante de continuité de a.

(c) En déduire qu'il existe des réels positifs C_1 et C_2 indépendants de n tels que

$$B(v,q;w,0) \geq -C_1\|v\|_V^2 + C_2\|q\|_Q^2.$$

(On pourra utiliser, en le démontrant, le fait que pour tout $a_1 \geq 0, a_2 \geq 0$ et $\epsilon > 0$, on a $a_1 a_2 \leq \frac{1}{\varepsilon} a_1^2 + \varepsilon a_2^2$.)

(d) Soit $\gamma \in \mathbb{R}_+^*$. Montrer que si γ est suffisamment petit, on a :

$$B(v,q;v+\gamma w,-q) \geq C_3[\|v\|_V^2 + \|q\|_Q^2].$$

et

$$\|(v+\gamma w,-q)\| \leq C_4\|(v,q)\|,$$

où C_3 et C_4 sont deux réels positifs qui ne dépendent pas de n.

(e) En déduire que la condition suivante (dite de stabilité) est satisfaite :

$$\begin{gathered}\text{Il existe } \delta \in \mathbb{R}_+^* \text{ (indépendant de } n) \text{ tel que pour tout } (u,p) \in V_n \times Q_n, \\ \sup_{\substack{(v,q)\in V_n\times Q_n \\ \|(v,q)\|\neq 0}} \frac{B((u,p);(v,q))}{\|(v,q)\|} \geq \delta\|(u,p)\|.\end{gathered} \tag{3.4.53}$$

4. On suppose maintenant que la condition (3.4.53) est satisfaite.

(a) Montrer que pour tout $p \in Q$,

$$\sup_{\substack{(v,q)\in V_n\times Q_n \\ \|(v,q)\|\neq 0}} \frac{b(v,p)}{\|(v,q)\|} \geq \delta\|p\|_Q.$$

(b) En déduire que pour tout $p \in Q$,

$$\sup_{\substack{v\in V_n \\ \|v\|_V\neq 0}} \frac{b(v,p)}{\|v\|_V} \geq \delta\|p\|_Q.$$

5. Déduire des questions précédentes que la condition (3.4.51) est satisfaite si et seulement si la condition (3.4.53) est satisfaite.

Exercice 37 (Volumes finis vus comme des éléments finis non conformes) *Suggestions en page 142, corrections en page 155*

Soit un ouvert borné polygonal de $\mathbb{R}^2$, et $\mathcal{T}$ un maillage admissible au sens des volumes finis (voir page 1.4.2 page 28) de Ω.

1. Montrer que la discrétisation par volumes finis de (3.1.1) se ramène à chercher $(u_K)_{K\in\mathcal{T}}$, qui vérifie :

$$\sum_{\sigma\in\mathcal{E}_{\rm int},\ \sigma=K|L} \tau_\sigma(u_L - u_K) + \sum_{\sigma\in\mathcal{E}_{\rm ext},\ \sigma\in\mathcal{E}_K} \tau_\sigma\ u_K = m(K) f_K \tag{3.4.54}$$

où $\mathcal{E}_{\rm int}$ représente l'ensemble des arêtes internes (celles qui ne sont pas sur le bord) $\mathcal{E}_{\rm ext}$ l'ensemble des arêtes externes (celles qui sont sur le bord), et

$$\tau_\sigma = \begin{cases} \dfrac{m(\sigma)}{d_{K,\sigma} + d_{L,\sigma}} \text{ si } \sigma \in \mathcal{E}_{\rm int},\ \sigma = K|L, \\ \dfrac{m(\sigma)}{d_{K,\sigma}} \text{ si } \sigma \in \mathcal{E}_{\rm ext},\ \sigma \in \mathcal{E}_K, \end{cases} \tag{3.4.55}$$

(voir figure 1.4 page 29).

2. On note $H_{\mathcal{T}}(\Omega)$ le sous-espace de $L^2(\Omega)$ formé des fonctions constantes par maille (c.à.d. constantes sur chaque élément de $\mathcal{T}$). Pour $u \in H_{\mathcal{T}}(\Omega)$, on note u_K la valeur de u sur K. Montrer que $(u_K)_{K\in\mathcal{T}}$ est solution de (3.4.55) si et seulement si $u \in H_{\mathcal{T}}(\Omega)$ est solution de :

$$\left\{ \begin{array}{l} u \in H_{\mathcal{T}}(\Omega), \\ a_{\mathcal{T}}(u,v) = T_{\mathcal{T}}(v), \forall v \in H_{\mathcal{T}}(\Omega), \end{array} \right. \tag{3.4.56}$$

où $a_{\mathcal{T}}$ est une forme bilinéaire sur $H_{\mathcal{T}}(\Omega)$ (à déterminer), et $T_{\mathcal{T}}$ est une forme linéaire sur $H_{\mathcal{T}}(\Omega)$ (à déterminer).

3.5 Suggestions pour les exercices

Exercice 27 page 137

Rappel ; par définition, l'ensemble $u_0 + H_0^1$ est égal à l'ensemble $\{v = u_0 + w, w \in 1H_0^1\}$

1. Montrer que le problème s'écrit sous la forme: $J(u) \leq J(v), \forall v \in H$, ou H est un espace de Hilbert, avec $J(v) = a(v,v)$, où a est une forme bilinéaire symétrique définie positive.

2. Prendre une fonction test à support compact dans la formulation faible.

Exercice 29 page 138

Considérer les espaces

$$H^1_{1,1} = \{v \in H^1(]0,1[); v(1) = 1\} \quad \text{et}$$
$$H^1_{1,0} = \{v \in H^1(]0,1[); v(1) = 0\}.$$

Exercice 33 page 138

Considérer comme espace de Hilbert l'ensemble $\{u \in H^1(\Omega); u = 0 \text{ sur } \Gamma_0\}$.

Exercice 37 page 141

1. Intégrer l'équation sur la maille et approcher les flux sur les arêtes par des quotients différentiels.

2. Pour montrer que (3.4.54) entraîne (3.4.56), multiplier par v_K, où $v \in H_{\mathcal{T}}(\Omega)$, et développer. Pour montrer la réciproque, écrire u comme combinaison linéaire des fonctions de base de $H_{\mathcal{T}}(\Omega)$, et prendre pour v la fonction caractéristique de la maille K. (3.4.54)

3.6 Corrigés des exercices

Corrigé de l'exercice 27 page 137

1. Par définition, on sait que $|u|_{1,\Omega} = (\int_\Omega \sum_{i=1}^N |\partial_i u(x)|^2 dx)^{\frac{1}{2}}$, où $\partial_i u$ désigne la dérivée partielle de u par rapports à sa i–ème variable. Attention ceci $|\cdot|_{1,\Omega}$ définit une semi-norme et non une norme sur l'espace $H^1(\Omega)$. Cependant sur $H_0^1(\Omega)$ c'est bien une norme, grâce à l'inégalité de Poincaré. On rappelle que $H_0^1(\Omega) = \mathrm{Ker}(\gamma) = \left\{u \in H^1(\Omega) \textit{ tel que } \gamma(u) = 0\right\}$, où γ est l'opérateur de trace linéaire et continu de

$H^1(\Omega)$ dans $L^2(\Omega)$ (voir théorème 3.2 page 114). Le problème consiste à minimiser $\left(\int_\Omega \sum_{i=1}^N |\partial_i u|^2\, dx\right)^{\frac{1}{2}}$ sur $u_0 + H_0^1(\Omega)$. Tentons de nous ramener à minimiser une certaine fonctionnelle sur $H_0^1(\Omega)$. Soit $v \in u_0 + H_0^1(\Omega)$. Alors $v = u_0 + w$ avec $w \in H_0^1(\Omega)$, et donc :

$$\begin{aligned}
|v|_{1,\Omega}^2 &= |u_0 + w|_{1,\Omega}^2 \\
&= \int_\Omega \sum_{i=1}^N |\partial_i(u_0 + w)|^2\, dx \\
&= \int_\Omega \sum_{i=1}^N \left|(\partial_i u_0)^2 + (\partial_i w)^2 + 2\,(\partial_i u_0)\,(\partial_i w)\right| dx \\
&= |u_0|_{1,\Omega}^2 + |w|_{1,\Omega}^2 + 2\int_\Omega \sum_{i=1}^N |\partial_i u_0 \partial_i w|\, dx
\end{aligned}$$

Ainsi chercher à mimimiser $|v|_{1,\Omega}$ sur $u_0 + H_0^1(\Omega)$ revient à minimiser J sur $H_0^1(\Omega)$, où J est définie par :

$$J(w) = \inf_{H_0^1(\Omega)} 2\left(\int_\Omega \sum_{i=1}^N |\partial_i u_0 \partial_i w|\, dx + \frac{1}{2}\int_\Omega \sum_{i=1}^N (\partial_i w)^2\right).$$

Pour montrer l'existence et l'unicité du minimum de J, nous allons mettre ce problème sous une forme faible, puis utiliser le théorème de Lax-Milgram pour en déduire l'existence et l'unicité d'une solution faible, et finalement conclure que la fonctionnelle $J(w)$ admet un unique inf. On pose :

$$a(w,v) = \int_\Omega \sum_{i=1}^N (\partial_i w\, \partial_i v)\, dx, \;\; \forall w,\, v \in H_0^1(\Omega)$$

et

$$L(v) = -\int_\Omega \sum_{i=1}^N (\partial_i u_0\, \partial_i v)\, dx, \;\; \forall v \in H_0^1(\Omega)$$

Voyons si les hypothèses de Lax-Milgram sont vérifiées. La forme $a(w,v)$ est clairement symétrique, on peut changer l'ordre de w et de v dans l'expression sans changer la valeur de l'intégrale. La forme $a(w,v)$ est bilinéaire. En effet, elle est linéaire par rapport au premier argument, puisque : $\forall u,v,w \in H_0^1(\Omega)$ et $\forall \lambda,\mu \in \mathbb{R}$, on a: $a(\lambda w + \mu u, v) = \lambda a(w,v) + \mu a(u,v)$. Ainsi par symétrie, elle est aussi linéaire par rapport au second argument. Donc elle est bien bilinéaire. Pour montrer que la forme $a(w,v)$ est continue, on utilise la caractérisation de la continuité des applications bilinéaires. On va donc montrer l'existence de $C \in \mathbb{R}_+$ tel que $|a(u,v)| \leq C\|u\|_{H^1}\|v\|_{H^1}$ pour tous u, $v \in H_0^1(\Omega)$. Or, par l'inégalité de Cauchy-Schwarz, on a :

$$\begin{aligned}
|a(u,v)| &= \left|\int_\Omega \sum_{i=1}^N (\partial_i u \partial_i v)\, dx\right| \\
&\leq \left(\int_\Omega \sum_{i=1}^N (\partial_i u)^2\, dx\right)^{\frac{1}{2}} \left(\int_\Omega \sum_{i=1}^N (\partial_i v)^2\, dx\right)^{\frac{1}{2}} \\
&\leq \|u\|_{H^1}\|v\|_{H^1}.
\end{aligned}$$

La forme a est donc bien continue. Montrons alors qu'elle est coercive, c'est–à–dire qu'il existe $\alpha > 0$ tel que $a(v,v) \geq \alpha \|v\|^2_{H^1}$ pour tout $v \in H^1_0(\Omega)$.

$$\begin{aligned} a(v,v) &= \int_\Omega \sum_{i=1}^N (\partial_i v(x))^2 \, dx \\ &= \int_\Omega \nabla v(x) \cdot \nabla v(x) dx \\ &\geq \frac{1}{1+\mathrm{diam}(\Omega)^2} \|v\|^2_{H^1}, \end{aligned}$$

grâce à l'inégalite de Poincaré, qu'on rappelle ici :

$$\|v\|_{L^2(\Omega)} \leq c(\Omega) \|\nabla v\|_{L^2(\Omega)}, \, \forall v \in H^1_0(\Omega). \tag{3.6.57}$$

Donc a est bien une forme bilinéaire, symétrique, continue et coercive. Par le même genre de raisonnement, on montre facilement que L est linéaire et continue. Ainsi toutes les hypothèses de Lax-Milgram sont vérifiées, donc le problème :

$$\text{Trouver } u \in H^1_0(\Omega) \text{ tel que } a(u,v) = L(v) \text{ pour tout } v \in H^1_0(\Omega)$$

a une unique solution dans $H^1_0(\Omega)$. De plus, comme a est symétrique, la fonctionnelle J admet un unique minimum.

2. On va maintenant caractériser u comme étant la solution d'un problème aux limites. Soit $\varphi \in D(\Omega)$, donc φ est à support compact dans Ω. On a :

$$a(u,\varphi) = L(\varphi) \quad \forall \varphi \in D(\Omega),$$

et donc :

$$\int_\Omega \nabla u(x) \nabla(x) \varphi dx = -\int_\Omega \nabla u_0(x) \nabla(x) \varphi(x) dx.$$

Comme u et $u_0 \in H^1(\Omega)$, et comme φ est réguliére, on peut intégrer par parties ; en remarquant que φ est nulle sur $\partial\Omega$, on a donc :

$$-\int_\Omega \Delta u(x) \phi(x) dx = \int_\Omega \Delta u_0(x) \phi(x) dx.$$

On en déduit que $-\Delta u = \Delta u_0$. Comme $u \in H^1_0(\Omega)$, ceci revient à résoudre le problème aux limites $\tilde{u} = u - u_0 \in H^1(\Omega)$, tel que $-\Delta \tilde{u} = 0$ dans Ω et $\tilde{u} = u_0$ sur $\partial\Omega$.

Corrigé de l'exercice 28 page 137 (Formulation faible du problème de Dirichlet)

Soit $\varphi \in C^\infty_c([0,1])$, on multiplie la première équation de (3.4.38), on intègre par parties et on obtient :

$$\int_0^1 u'(x) \varphi'(x) dx = \int_0^1 f(x) \varphi(x) dx. \tag{3.6.58}$$

Pour trouver une formulation faible (ou variationnelle) il faut commencer par trouver un espace de Hilbert pour les fonctions duquel (3.6.58) ait un sens, et qui soit compatible avec les conditions aux limites. Comme $f \in L^2(]0,1[)$, le second membre de (3.6.58) est bien défini dès que $\varphi \in L^2(]0,1[)$.

De même, le premier membre de (3.6.58) est bien défini dès que $u' \in L^2(]0,1[$ et $\varphi' \in L^2(]0,1[)$.
Comme de plus, on doit avoir $u = 0$ en 0 et en 1, il est naturel de choisir $H = H_0^1(]0,1[) \stackrel{def}{=} \{u \in L^2(]0,1[; Du \in L^2(]0,1[)$ et $u(0) = u(1) = 0\}$
(Rappelons qu'en une dimension d'espace $H^1(]0,1[) \subset C([0,1])$ et donc $u(0)$ et $u(1)$ sont bien définis).
Une formulation faible naturelle est donc :

$$\begin{cases} u \in H = \{u \in H_0^1(\Omega); v(0) = v(1) = 0\}, \\ a(u,v) = T(v), \forall v \in H, \end{cases}$$

où $a(u,v) = \displaystyle\int_0^1 u'(x)v'(x)dx$ et $T(v) = \displaystyle\int_0^1 f(x)v(x)dx$.
La formulation variationnelle associée (notons que a est clairement symétrique), s'écrit :

$$\begin{cases} \text{Trouver } u \in H, \\ J(u) = \min\limits_{v \in H} J(v) \end{cases}$$

avec $J(v) = \dfrac{1}{2}a(u,v) - T(v)$
Le fait que a soit une forme bilinéaire continue symétrique et coercive etque $T \in H'$ a été prouvé (dans le cas plus général de la dimension quelconque) lors de la démonstration de la proposition 3.7 page 117.

Corrigé de l'exercice 29 page 138

On introduit les espaces :

$$H_{1,1}^1 = \{v \in H^1(]0,1[); v(1) = 1\}$$

$$H_{1,0}^1 = \{v \in H^1(]0,1[); v(1) = 0\}$$

Soit $u_0 :]0,1[\to \mathbb{R}$, définie par $u_0(x) = x$. On a bien $u_0(1) = 1$, et $u_0 \in H_{1,1}^1$.
Cherchons alors u sous la forme $u = u_0 + \widetilde{u}$, avec $\widetilde{u} \in H_{1,0}^1$.

$$\int_0^1 u'(x)v'(x)dx - u'(1)v(1) + u'(0)v(0) = \int_0^1 f(x)vdx, \forall \in H_{1,0}^1.$$

Comme $v(1) = 0$ et $u'(0) = 0$, on obtient donc :

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

ou encore :

$$\int_0^1 \widetilde{u}'(x)v'(x)dx = \int_0^1 f(x)v(x)dx - \int_0^1 u_0'(x)v'(x)dx = \int_0^1 f(x)v(x)dx - \int_0^1 v'(x)dx.$$

car $u_0' = 1$.

Corrigé de l'exercice 30 page 138

1. Comme $f \in C(\mathbb{R},\mathbb{R})$, et comme $-u'' = f$, on a $u \in C^2(\mathbb{R},\mathbb{R})$. Or $u_0 \in C^2(\mathbb{R},\mathbb{R})$ et $u_0'' = 0$; de même, $u_1 \in C^2(\mathbb{R},\mathbb{R})$ et $u_1'' = 2(b-a)$.
Les fonctions $\widetilde{u}$ et $\bar{u}$ doivent donc vérifier :

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

et

$$\begin{cases} -\bar{u}'' = f + 2(b-a) \\ \bar{u}(0) = 0 \\ \bar{u}(1) = 0. \end{cases}$$

Donc $\widetilde{u}$ est l'unique solution du problème

$$\begin{cases} \widetilde{u} \in H_0^1(\Omega) \\ a(u,\varphi) = \widetilde{T}(\varphi), \forall \varphi \in H_0^1(\Omega), \end{cases}$$

avec $a(u,\varphi) = \displaystyle\int_0^1 u'(x)\varphi'(x)dx$ et $\widetilde{T}(\varphi) = \displaystyle\int_0^1 f(x)\varphi(x)dx$, et $\bar{u}$ est l'unique solution du problème.

$$\begin{cases} \bar{u} \in H_0^1(\Omega) \\ a(\bar{u},\varphi) = \bar{T}(\varphi), \forall \varphi \in H_0^1(\Omega), \end{cases}$$

avec $\bar{T}(\varphi) = \displaystyle\int_0^1 (f(x) + 2(b-a))\varphi(x)dx$.
Montrons maintenant que $u = v$. Remarquons que $w = u - v$ vérifie

$$\begin{cases} w'' = 0 \\ w(0) = w(1) = 0 \end{cases}$$

ce qui prouve que w est solution de

$$\begin{cases} w \in H_0^1(\Omega) \\ a(w,\varphi) = 0, \forall \varphi \in H_0^1(\Omega), \end{cases}$$

ce qui prouve que $w = 0$.
2. Le même raisonnement s'applique pour u_0 et $u_1 \in C^2([0,1])$ tel que

$$u_0(0) = u_1(0) = a \text{ et } u_1(0) = u_1(1) = b.$$

Corrigé de l'exercice 31 page 138

1. Soit $v \in C_c^\infty([0,1])$, on multiplie la première équation de (3.4.40), on intègre par parties et on obtient :

$$\int_0^1 u'(x)v'(x)dx - u'(1)v(1) + u'(0)v(0) + \int_0^1 u(x)v(x)dx = \int_0^1 f(x)v(x)dx.$$

En tenant compte des conditions aux limites sur u en 0 et en 1, on obtient :

$$\int_0^1 u'(x)v'(x)dx + \int_0^1 u(x)v(x)dx + u(0)v(0) = \int_0^1 f(x)\varphi(x)dx - v(1). \tag{3.6.59}$$

Pour trouver une formulation faible (ou variationnelle) il faut commencer par trouver un espace de Hilbert pour les fonctions duquel (3.6.59) ait un sens, et qui soit compatible avec les conditions aux limites. Comme $f \in L^2(]0,1[)$, le second membre de (3.6.59) est bien défini dès que $v \in L^2(]0,1[)$.
De même, le premier membre de (3.6.59) est bien défini dès que $u \in H^1(]0,1[$ et $v \in H^1(]0,1[) \stackrel{def}{=} \{u \in L^2(]0,1[; Du \in L^2(]0,1[)$. Il est donc naturel de choisir $H = H(]0,1[)$. On obtient ainsi la formulation faible suivante :

$$\begin{cases} u \in H = \{u \in H(\Omega)\}, \\ a(u,v) = T(v), \forall v \in H, \end{cases}$$

où $a(u,v) = \displaystyle\int_0^1 u'(x)v'(x)dx + \int_0^1 u(x)v(x)dx + u(0)v(0)$ et $T(v) = \displaystyle\int_0^1 f(x)v(x)dx - v(1)$.
La formulation variationnelle associée (notons que a est clairement symétrique), s'écrit :

$$\begin{cases} \text{Trouver } u \in H, \\ J(u) = \min\limits_{v \in H} J(v) \end{cases}$$

avec $J(v) = \dfrac{1}{2}a(u,v) - T(v)$

Pour montrer l'existence et l'unicité des solutions de (3.6.59), on cherche à appliquer le théorème de Lax–Milgram. On remarque d'abord que T est bien une forme linéaire sur H, et que de plus, par l'inégalité de Cauchy–Schwarz, :

$$|T(v)| = |\int_0^1 f(x)v(x)dx| + |v(1)| \leq \|f\|_{L^2(]0,1[)}\|v\|_{L^2(]0,1[)} + |v(1)|. \tag{3.6.60}$$

Montrons maintenant que $|v(1)| \leq 2\|v\|_{H^1(]0,1[)}$. Ce résultat est une conséquence du théorème de trace, voir cours d'EDP. Dans le cas présent, comme l'espace est de dimension 1, la démonstration est assez simple en remarquant que comme $v \in H^1(]0,1[)$, on peut écrire que v est intégrale de sa dérivée. On a en particulier :

$$v(1) = v(x) + \int_x^1 v'(t)dt,$$

et donc par l'inégalité de Cauchy–Schwarz,

$$|v(1)| = |v(x)| + \int_x^1 |v'(t)|dt \leq |v(x)| + \|v'\|_{L^2(]0,1[)}.$$

En intégrant cette inégalité entre 0 et 1 on obtient :

$$|v(1)| \leq \|v(x)\|_{L^1(]0,1[)} + \|v'\|_{L^2(]0,1[)}.$$

Or $\|v\|_{L^1(]0,1[)} \leq \|v(x)\|_{L^2(]0,1[)}$. De plus

$$\|v\|_{L^2(]0,1[)} + \|v'\|_{L^2(]0,1[)} \leq 2\max(\|v(x)\|_{L^2(]0,1[)}, \|v'\|_{L^2(]0,1[)})$$

on a donc

$$\begin{aligned}\Big(\|v\|_{L^2(]0,1[)}+\|v'\|_{L^2(]0,1[)}\Big)^2 &\leq 4\max(\|v(x)\|^2_{L^2(]0,1[)},\|v'\|^2_{L^2(]0,1[)})\\ &\leq 4(\|v\|^2_{L^2(]0,1[)}+\|v'\|^2_{L^2(]0,1[)}).\end{aligned}$$

On en déduit que

$$|v(1)|\leq\|v\|_{L^2(]0,1[)}+\|v'\|_{L^2(]0,1[)}\leq 2\|v\|_{H^1(]0,1[)}.$$

En reportant dans (3.6.60), on obtient :

$$|T(v)|\leq(\|f\|_{L^2(]0,1[)}+2)\|v\|_{H^1(]0,1[)}$$

ce qui montre que T est bien continue.
Remarquons que le raisonnement effectué ci–dessus pour montrer que $|v(1)|\leq 2\|v\|_{H^1(]0,1[)}$ s'applique de la même manière pour montrer que

$$|v(a)|\leq 2\|v\|_{H^1(]0,1[)}\text{ pour tout } a\in[0,1]. \tag{3.6.61}$$

Ceci est une conséquence du fait que $H^1(]0,1[)$ s'injecte continûment dans $C([0,1])$.
Il est clair que a est une forme bilinéaire symétrique (notons que le caractére symétrique n'est pas nécessaire pour l'application du théorème de Lax–Milgram). Montrons que a est continue. On a :

$$\begin{aligned}|a(u,v)| &\leq \int_0^1|u'(x)v'(x)|dx+\int_0^1|u(x)||v(x)|dx+|u(0)||v(0)|\\ &\leq \|u'\|_{L^2(]0,1[)}\|v'\|_{L^2(]0,1[)}+\|u\|_{L^2(]0,1[)}\|v\|_{L^2(]0,1[)}+|u(0)||v(0)|\end{aligned}$$

Grâce à (3.6.61), on en déduit que

$$\begin{aligned}|a(u,v)| &\leq \|u'\|_{L^2(]0,1[)}\|v'\|_{L^2(]0,1[)}+\|u\|_{L^2(]0,1[)}\|v\|_{L^2(]0,1[)}+4\|u\|_{H^1(]0,1[)}\|v\|_{H^1(]0,1[)}\\ &\leq 6\|u\|_{H^1(]0,1[)}\|v\|_{H^1(]0,1[)}.\end{aligned}$$

On en déduit que a est continue. Soit $u\in H^1(]0,1[)$, Par définition de a, on a :

$$\begin{aligned}a(u,u) &= \int_0^1(u'(x))^2dx+\int_0^1(u(x))^2dx+u(0)^2\\ &\geq \|u\|_{H^1(]0,1[)}.\end{aligned}$$

Ceci prouve que la forme a est coercive. Par le théorème de Lax–Milgram, on en déduit l'existence et l'unicité des solutions faibles de (3.6.59).

Corrigé de l'exercice 32 page 138

1. Soit φ une fonction régulière de [0,1] dans $\mathbb{R}$. Multiplions la première équation de (3.4.41) par φ et intégrons entre 0 et 1 :

$$\int_0^1-u''(x)\varphi(x)dx-\int_0^1u'(x)\varphi(x)dx+\int_0^1u(x)\varphi(x)dx=\int_0^1f(x)\varphi(x)dx$$

Deux intégrations par parties donnent alors :

$$-u'(1)\varphi(1)+u'(0)\varphi(0)+\int_0^1 u'(x)\varphi'(x)dx-u(1)\varphi(1)+u(0)\varphi(0)+\int_0^1 u(x)\varphi'(x)dx \\ +\int_0^1 u(x)\varphi(x)dx=\int_0^1 f(x)\varphi(x)dx.$$

En choisissant φ telle que $\varphi(1)=0$, on obtient alors :

$$u'(0)\varphi(0)+\int_0^1 u'(x)\varphi'(x)dx+u(0)\varphi(0)+\int_0^1 u(x)\varphi'(x)dx+\int_0^1 u(x)\varphi(x)dx=\int_0^1 f(x)\varphi(x)dx.$$

En tenant compte de la condition à la limite $u(0)+u'(0)=0$, on a :

$$\int_0^1 u'(x)\varphi'(x)dx+\int_0^1 u(x)\varphi'(x)dx+\int_0^1 u(x)\varphi(x)dx=\int_0^1 f(x)\varphi(x)dx.$$

Posons alors, pour u, $v\in H^1(]0,1[)$,

$$a(u,v)=\int_0^1 u'(x)\varphi'(x)dx+\int_0^1 u(x)\varphi'(x)dx+\int_0^1 u(x)\varphi(x)dx \text{ et } T(v)=\int_0^1 f(x)\varphi(x)dx.$$

Soit $u\in C^2(]0,1[)\cap C(]0,1[)$ solution de (4.7.31). Montrons que u est solution de (3.4.41). Soit $v\in C^\infty(]0,1[)\cap C(]0,1[)$. On a donc :

$$\int_0^1 u'(x)v'(x)dx+\int_0^1 u(x)v'(x)dx+\int_0^1 u(x)v(x)dx=\int_0^1 f(x)v(x)dx.$$

En intégrant par parties, ceci donne :

$$u'(1)v(1)-u'(0)v(0)-\int_0^1 u''(x)v(x)dx+u(1)v(1)-u(0)v(0)-\int_0^1 u'(x)v(x)dx \\ +\int_0^1 u(x)v(x)dx=\int_0^1 f(x)v(x)dx \tag{3.6.62}$$

En prenant v telle que $v(0)=v(1)=0$, on obtient :

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

Ceci étant vrai pour toute fonction $v\in C^\infty(]0,1[)\cap C(]0,1[)$ nulle sur le bord, on en déduit que $-u''-u'+u=f$ sur $]0,1[$. En prenant alors une fonction $v\in C^\infty(]0,1[)\cap C(]0,1[)\cap H$, on obtient à partir de (3.6.62) que : $-u'(0)v(0)-u(0)v(0)-=0$. Comme ceci est vrai pour toute valeur de $v(0)$, on en déduit que $u'(0)+u(0)=0$.

2. Soit $u_0\in H_0^1(]0,1[)$ telle que $u_0(1)=1$. On peut réecrire (4.7.31) sous la forme :

$$\begin{array}{l} u=u_0+\tilde{u},\tilde{u}\in H; \\ a(\tilde{u},v)=T(v)-a(u_0,v),\forall v\in H. \end{array} \tag{3.6.63}$$

Par le lemme de Lax Milgram, il existe une unique solution $\tilde{u}$ à (4.7.31) si a est une forme bilinéaire continue coercive sur $H_0^1(]0,1[)$, et si $\tilde{T}=T-a(u_0,\cdot)$ est une forme linéaire continue. Montrons que a est coercive (les autres propriétés sont faciles à vérifier). On a ;

$$a(u,u)=\int_0^1 (u'(x))^2dx+\int_0^1 u(x)u'(x)dx+\int_0^1 (u(x))^2dx\geq\frac{1}{2}(\int_0^1 (u'(x))^2dx+\int_0^1 u(x)u'(x)dx)$$

Corrigé de l'exercice 33 page 138

Soit $\varphi \in H = \{v \in H^1(\Omega) : u = 0 \text{ sur } \Gamma_0\}$.
Multiplions la première équation de (3.4.43) par $\varphi \in H$. On obtient :

$$\int_\Omega \nabla u(x).\nabla\varphi(x)dx + \int_{\Gamma_0\cup\Gamma_1} \nabla u.n(x)\varphi(x)dx = \int_\Omega f(x)\varphi(x)dx$$

et comme $\nabla u.n = 0$ sur Γ_1 et $\varphi = 0$ sur Γ_0, on obtient donc

$$\int_\Omega \nabla u(x).\nabla\varphi(x) = \int_\Omega f(x)\varphi(x)dx.$$

On obtient donc la formulation faible.

$$\begin{cases} \text{Trouver } u \in H; \\ \int_\Omega \nabla u(x).\nabla v(x)dx = \int_\Omega f(x)u(x)dx. \end{cases}$$

Notons que cette formulation ne diffère de la formulation faible du problème (3.1.1) que par la donnée de la condition aux limites de Dirichlet sur Γ_0 et non $\partial\Omega$ dans l'espace H. La condition de Neumann homogène est implicitement prise en compte dans la formulation faible.
La démonstration du fait que cette formulation satisfait les hypothèses du théorème de Lax-Milgram est similaire à celle de la proposition 3.7 en utilisant, pour la coercivité, le fait que les fonctions à trace nulle sur une partie du bord de Ω (de mesure non nulle) vérifient encore l'inégalité de Poincaré.

Corrigé de l'exercice 34 page 139

1. Multiplions la première équation de (3.4.44) par $\varphi \in C^\infty(\Omega)$ et intégrons sur Ω. Par la formule de Green, on obtient :

$$\int_\Omega p(x)\nabla u(x).\nabla\varphi(x)dx - \int_{\partial\Omega} p(x)\nabla u(x).n(x)\varphi(x)dx + \int_\Omega q(x)u(x)\varphi(x)dx = \int_\Omega f(x)\varphi(x)dx.$$

En tenant compte des conditions aux limites sur u et en prenant φ nulle sur Γ_0, on obtient alors :

$$a(u,\varphi) = T(\varphi)$$

avec :

$$a(u,\varphi) = \int_\Omega (p(x)\nabla u(x).\nabla\varphi(x) + q(x)u(x)\varphi(x))dx + \int_{\Gamma_1} \sigma(x)u(x)\varphi(x)d\gamma(x), \tag{3.6.64}$$

et

$$T(\varphi) = \int_\Omega f(x)\varphi(x)dx + \int_{\Gamma_1} g_1(x)\varphi(x)d\gamma(x). \tag{3.6.65}$$

Pour assurer la condition aux limites de type Dirichlet non homogène, on choisit donc $u \in H^1_{\Gamma_0,g_0}(\Omega) = \{u \in H^1(\Omega); u = g_0 \text{ sur } \Gamma_0\}$, qu'on peut aussi décomposer en : $u = \widetilde{u}+u_0$ avec $\widetilde{u} \in H^1_{\Gamma_0,g_0}(\Omega)$ ("relèvement" de u) et $u_0 \in H^1_{\Gamma_0}(\Omega) = \{u \in H^1(\Omega); u = 0 \text{ sur } \Gamma_0\}$, Une formulation faible naturelle est alors :

$$\begin{cases} u \in H^1_{\Gamma_0,g_0}(\Omega) \\ a(u,v) = T(v), \forall v \in H^1_{\Gamma_0,0}(\Omega), \end{cases}$$

ou encore :

$$\begin{cases} u = u_0 + \widetilde{u} \\ \widetilde{u} \in H^1_{\Gamma_0,0}(\Omega) \\ a(\widetilde{u},v) = T(v) - T(u_0), \forall v \in H^1_{\Gamma_0,0}(\Omega), \end{cases} \tag{3.6.66}$$

L'espace $H = H^1_{\Gamma_0,0}(\Omega)$ muni de la norme H^1 est un espace de Hilbert. Il est facile de montrer que l'application a définie de $H \times H$ dans $\mathbb{R}$ est bilinéaire. Montrons qu'elle est continue ; soient $(u,v) \in H \times H$, alors

$$a(u,v) \leq \|p\|_{L^\infty(\Omega)} \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + \|q\|_{L^\infty(\Omega)} \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \sigma \|\gamma(u)\|_{L^2(\Omega)} \|\gamma(v)\|_{L^2(\partial\Omega)}.$$

Par le théorème de trace, il existe C_Ω ne dépendant que de Ω tel que

$$\|\gamma(u)\|_{L^2(\partial\Omega)} \leq C_\Omega \|u\|_{H^1(\Omega)} \text{ et } \|\gamma(v)\|_{L^2(\partial\Omega)} \leq C_\Omega \|v\|_{H^1(\Omega)}.$$

On en déduit que

$$a(u,v) \leq \left(\|p\|_{L^\infty(\Omega)} + \|q\|_{L^\infty(\Omega)} + \sigma C_\Omega^2 \right) \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)},$$

ce qui montre que a est continue. La démonstration de la coercivité de a est similaire à la démonstration du lemme 3.15 page 121. Enfin, il est facile de voir que T définie par (3.6.65) est une forme linéaire. On en déduit que le théorème de Lax-Milgram s'applique.

2. On a déjà vu à la question précédente que si u est solution de (3.6.66), alors u est solution de (3.4.44). Il reste à démontrer la réciproque. Soit donc u solution de (3.6.66), et soit $\varphi \in C_c^\infty(\Omega) (\subset H)$. En utilisant la formule de Green, et en notant que φ est nulle sur $\partial\Omega$, on obtient :

$$\int_\Omega (-div(p\nabla u)(x) + q(x)u(x) - f(x))\varphi(x)dx = 0, \forall \varphi \in C_0^\infty(\Omega).$$

Comme $u \in C^2(\bar{\Omega})$, on en déduit que :

$$-div(p\nabla u)(x) + q(x)u(x) - f(x) = 0, \forall x \in \Omega.$$

Comme $u \in H^1_{\Gamma_0,g_0}$ et $u \in C^2(\bar{\Omega})$, on a aussi $u = g_0$ sur Γ_0. Prenons maintenant $\varphi \in H^1_{\Gamma_0,0}$ on a :

$$\int_\Omega p(x)\nabla u(x)\nabla\varphi(x)dx + \int_\Omega q(x)u(x)\varphi(x)dx + \int_{\Gamma_1} \sigma(x)u(x)\varphi(x)d\gamma(x) = \int_\Omega f(x)dx + \int_{\Gamma_1} g(x)\varphi(x)d\gamma(x).$$

Par intégration par parties, il vient donc :

$$\int_\Omega -\mathrm{div}(p(x)\nabla u(x))\varphi(x)dx + \int_{\Gamma_1} p(x)\nabla u(x)\cdot n(x)\varphi(x)dx + \int_{\Gamma_1} \sigma(x)u(x)\varphi(x)d\gamma(x) + \int_\Omega q(x)u(x)\varphi(x)dx =$$

$$= \int_\Omega f(x)\varphi(x)dx + \int_{\Gamma_1} g_1(x)\varphi(x)d\gamma(x).$$

Or on a montré que $-div(p\nabla u) + qu = 0$. On a donc :

$$\int_{\Gamma_1} (p(x)\nabla u(x)\cdot n(x) + \sigma u(x) - g_1(x))\varphi(x)d\gamma(x) = 0, \quad \forall \varphi \in H^1_{\Gamma_0,g_0}.$$

On en déduit que :

$$p\nabla u \cdot n + \sigma u - g_1 = 0 \text{ sur } \Gamma_1.$$

Donc u vérifie bien (3.4.44).

Corrigé de l'exercice 35 page 139

1. Pour montrer que le problème (3.4.47) admet une unique solution, on aimerait utiliser le théorème de Lax-Milgram. Comme $V_h \subset V$ un Hilbert, que a une forme bilinéaire continue sur $V \times V$, et que L est une forme linéaire continue sur V, il ne reste qu'à montrer la coercivité de a sur V_h. Mais la condition (3.4.46) page 139 n'entraîne pas la coercivité de a sur V_h. Il suffit pour s'en convaincre de considérer la forme bilinéaire $a(u,v) = u_1u_2 - v_1v_2$ sur $V_h = \mathbb{R}^2$, et de vérifier que celle-ci vérifie la condition (3.4.46) sans être pour autant coercive. Il faut trouver autre chose...
On utilise le théorème représentation de F. Riesz, que l'on rappelle ici : Soit H un espace de Hilbert et T une une forme lineaire continue sur H, alors il existe un unique $u_T \in H$ tel que $T(v) = (u_T,v)\ \forall v \in H$. Soit A l'opérateur de V_h dans V_h défini par $a(u,v) = (Au,v)$ pour tout $v \in V_h$. Comme L est une forme linéaire continue sur $V_h \subset V$, par le théorème de Riesz, il existe un unique $\psi \in V_h$ tel que $L(v) = (\psi,v)$, pour tout $v \in V_h$. Le problème (3.4.47) s'écrit donc

$$\text{Trouver } u \in V_h \text{ tel que } (Au,v) = (\psi,v), \text{ pour tout } v \in V_h.$$

Si A est bijectif de V_h dans V_h, alors $u = A^{-1}\psi_u$ est donc la solution unique de (3.4.47). Comme V_h est de dimension finie, il suffit de montrer que A est injectif. Soit donc $w \in V_h$ tel que $Aw = 0$, on a dans ce cas $\|Aw\| = 0$ et donc

$$\sup_{(v\in V_h,,\|v\|=1)} a(w,v) = 0.$$

Or par la condition (3.4.46), on a

$$\inf_{w\in V_h, w\neq 0} \sup_{(v\in V_h,\|v\|=1)} a(w,v) \geq \beta_h > 0.$$

On en déduit que $w = 0$, donc que A est bijectif et que le problème (3.4.47) admet une unique solution. On peut remarquer de plus que si A est inversible,

$$\inf_{v\in V_h,\|v\|=1} \sup_{v\in V_h,\|v\|=1} a(w,v) = \|A^{-1}\|^{-1}, \tag{3.6.67}$$

et donc si (3.4.46) est vérifiée, alors

$$\|A^{-1}\| \leq \frac{1}{\beta_h} \tag{3.6.68}$$

En effet, par définition,

$$\begin{aligned}
\|A^{-1}\|^{-1} &= \left(\sup_{v\in V_h, v\neq 0} \frac{\|A^{-1}v\|}{\|v\|}\right)^{-1} \\
&= \inf_{v\in V_h, v\neq 0} \frac{\|v\|}{\|A^{-1}v\|} \\
&= \inf_{f\in V_h, v\neq 0} \frac{\|Af\|}{\|f\|} \\
&= \inf_{f\in V_h, \|f\|=1} \|Af\| \\
&= \inf_{f\in V_h, \|f\|=1} \sup_{w\in V_h, \|w\|=1} (Af,w).
\end{aligned}$$

2. Soit $v \in V_h$, $v \neq 0$; par l'inégalité triangulaire, on a :

$$\|u - u_h\| \leq \|u - v\| + \|v - u_h\|. \tag{3.6.69}$$

Mais grâce à (3.6.68), on a :

$$\begin{aligned}\|v-u_h\| &= \|A^{-1}A(v-u_h)\| \\ &\leq \frac{1}{\beta_h}\|A(v-u_h)\| \\ &\leq \frac{1}{\beta_h}\sup_{w\in V_h,\|w\|=1)} a(v-u_h,w) \\ &\leq \frac{1}{\beta_h}\sup_{w\in V_h,\|w\|=1)} (a(v,w)-a(u_h,w)) \\ &\leq \frac{1}{\beta_h}\sup_{w\in V_h,\|w\|=1)} (a(v,w)-a(u,w)),\end{aligned}$$

car $a(u_h,w)=L(w)=a(u,w)$. On a donc

$$\begin{aligned}\|v-u_h\| &\leq \frac{1}{\beta_h}\sup_{w\in V_h,\|w\|=1)} a(v-u,w) \\ &\leq \frac{1}{\beta_h}\sup_{w\in V_h,\|w\|=1)} M\|v-u\|\,\|w\| \\ &\leq \frac{M}{\beta_h}\|v-u\|.\end{aligned}$$

En reportant dans (3.6.69), il vient alors :

$$\|u-u_h\| \leq \|u-v\| + \frac{M}{\beta_h}\|v-u\|,\ \forall v\in V_h,$$

et donc

$$\|u-u_h\| \leq \left(1+\frac{M}{\beta_h}\right)\inf_{v\in V_h}\|u-v\|.$$

Corrigé de l'exercice 36 page 140 (Condition inf-sup pour un problème mixte)

1. Il est immédiat de vérifier que B est une forme bilinéaire sur $V\times Q$. Montrons que B est continue. Soit $(u,p;v,q)\in(V\times Q)^2$. Comme a et b sont des formes bilinéaires continues, on a, en notant M_a et M_b les constantes de continuité de a et b,

$$\begin{aligned}|B(u,p;v,q)| &\leq |a(u,v)|+|b(v,p)|+|b(u,q)| \\ &\leq M_a\|u\|_V\|v\|_V + M_b\|v\|_V\|p\|_Q + M_b\|u\|_V\|q\|_Q \\ &\leq \max(M_a,M_b)\|(u,p)\|_{V\times Q}\|(v,q)\|_{V\times Q},\end{aligned}$$

avec $\|(u,p)\|_{V\times Q}=(\|u\|_V+\|p\|_Q)$.

2. En additionnant membre à membre les des équations de (3.4.49), on obtient (3.4.50). Réciproquement, en prenant $q=0$ dans (3.4.50) on obtient la première équation de (3.4.49), et en prenant $v=0$ on obtient la deuxième. Les deux formulations sont donc équivalentes.

3.a Pour $q\in Q_n$ et $v\in V_n$, on a $B(v,q;v,-q)=a(v,v)\geq\alpha\|v\|_V^2$ où α est la constante de coercivité de a, ce qui montre que (3.4.52) est vérifiée.

3.b Si $q = 0$, $w = 0$ convient.
Soit maintenant $q \in Q_n$, $q \neq 0$; grâce à la condition (3.4.51) on a :

$$\sup_{\substack{w \in V_n \\ \|w\|_V \neq 0}} \frac{b(w,q)}{\|w\|_V} \geq \beta \|q\|_Q,$$

ou encore, par homogénéité,

$$\sup_{\substack{w \in V_n \\ \|w\|_V = \|q\|_Q}} \frac{b(w,q)}{\|w\|_V} \geq \beta \|q\|_Q.$$

Comme V_n est de dimension finie, il existe donc $w \in V_n$ tel que

$$\|w\|_V = \|q\|_Q \text{ et } b(w,q) \geq \beta \|q\|_Q^2. \tag{3.6.70}$$

Dans le cas où $q = 0$ et $w = 0$, la relation $B(v,q;w,0) \geq -M\|v\|_V\|w\|_V + \beta\|q\|_Q^2$, est évidemment vérifiée. Si $q \neq 0$, la relation est vérifiée car $B(v,q,w,0) = a(v,w) \geq -M\|v\|_V\|w\|_V$ où M est la constante de continuité de a. Et donc, pour w vérifiant (3.6.70), on a :

$$B(v,q,w,0) = a(v,w) + b(w,q) \geq -M\|q\|_Q\|v\|_V + \beta\|q\|_Q^2. \tag{3.6.71}$$

3.c On remarque d'abord que pour tout $a_1 \geq 0$,$a_2 \geq 0$ et $\varepsilon > 0$, on a soit $a_1 \leq \varepsilon a_2$, auquel cas $a_1 a_2 \leq \varepsilon a_2^2$, soit $a_1 > \varepsilon a_2$, auqel cas $a_2 \leq \frac{1}{\varepsilon} a_1$ et donc $a_1 a_2 \leq \frac{1}{\varepsilon} a_1^2$. On a donc bien dans tous les cas : $a_1 a_2 \leq \frac{1}{\varepsilon} a_1^2 + \varepsilon a_2^2$.)
Avec w choisi à la question précédente, et en prenant $a_1 = \|q\|_Q$, $a_2 = M\|v\|_V$ et $\varepsilon = \frac{\beta}{2}$, on obtient :

$$M\|q\|_Q\|v\|_V \leq \frac{\beta}{2}\|q\|_Q^2 + \frac{2M^2}{\beta}\|v\|_V^2,$$

d'où l'on déduit par (3.6.71) que

$$B(v,q,w,0) \geq \frac{\beta}{2}\|q\|_Q^2 - \frac{2M^2}{\beta}\|v\|_V^2.$$

ce qui termine la question.
3.d Soit $v \in V_n$ et $\gamma \in \mathbb{R}_+^*$. Par linéarité, on a :

$$B(v,q;v+\gamma w,-q) = B(v,q;v,-q) + B(v,q;\gamma w,0) = B(v,q;v,-q) + \gamma B(v,q;w,0).$$

Par la question 3.a, on sait que $B(v,q;v,-q) \geq \alpha\|v\|_V^2$ où α est la constante de coercivité de a, et donc, d'après la question précédente,

$$B(v,q;v+\gamma w,-q) \geq (\alpha - C_1\gamma)\|v\|_V^2 + C_2\gamma\|q\|_Q^2$$

Pour $\gamma = \frac{\alpha}{2C_1}$, on obtient donc :

$$B(v,q;v+\gamma w,-q) \geq \frac{\alpha}{2}\|v\|_V^2 + \frac{\alpha C_2}{2C_1}\|q\|_Q^2.$$

et donc

$$B(v,q;v+\gamma w,-q) \geq C_3(\|v\|_V^2 + \|q\|_Q^2) \text{ avec } C_3 = \min(\frac{\alpha}{2}, \frac{\alpha C_2}{2C_1}).$$

De plus,

$$\begin{aligned}\|(v+\gamma w,-q)\|_{V\times Q} &= \|v+\gamma w\|_V + \|q\|_Q \\ &\leq (\|v\| + \|\gamma w\|_V) + \|q\|_Q \\ &\leq (\|v\| + \gamma\|w\|_V) + \|q\|_Q\end{aligned}$$

Or $\|w\|_V = \|q\|_Q$, et donc

$$\|(v+\gamma w,-q)\| \leq C_4\|(v,q)\|_{V\times Q}, \text{ avec } C_4 = \max(1,1+\gamma).$$

3.e
Soit $(u,p) \in V_n \times Q_n$, alors :

$$\sup_{\substack{(v,q)\in V_n\times Q_n \\ \|(v,q)\|_{V\times Q}\neq 0}} \frac{B((u,p);(v,q))}{\|(v,q)\|_{V\times Q}} \geq \frac{B((u,p);(u+\gamma w,-p))}{\|(u+\gamma w,-p)\|_{V\times Q}},$$

où w est le w des questions 3.b à 3.d pour u au lieu de v. On en déduit que

$$\sup_{\substack{(v,q)\in V_n\times Q_n \\ \|(v,q)\|\neq 0}} \frac{B((u,p);(v,q))}{\|(v,q)\|_{V\times Q}} \geq \frac{C_3(\|u\|_V^2 + \|p\|_Q^2)}{C_4\|(u,p)\|_{V\times Q}}.$$

Or $\|u\|_V^2 + \|p\|_Q^2 \geq \frac{1}{2}\|(u,p)\|_{V\times Q}^2$, d'où le résultat, avec $\delta = \frac{C_3}{2C_4}$.

4.1. Le résultat s'obtient immédiatement en prenant $u = 0$ dans (3.4.53).

4.2 On a clairement $\frac{b(v,p)}{\|v\|_V} \geq \frac{b(v,p)}{\|(v,q)\|_{V\times Q}}$ pour tous v et q. On en déduit que

$$\sup_{\substack{v\in V_n \\ \|v\|_V\neq 0}} \frac{b(v,p)}{\|v\|_V} \geq \sup_{\substack{(v,q)\in V_n\times Q_n \\ \|(v,q)\|_{V\times Q}\neq 0}} \frac{b(v,p)}{\|(v,q)\|_{V\times Q}}.$$

D'où le résultat.

5. L'équivalence se déduit immédiatement des deux questions pécédentes.

Corrigé de l'exercice 37 page 141

1. Soit K un volume de contrôle du maillage volumes finis. On intègre (3.1.1) sur K et en utilisant la formule de Stokes, on obtient :

$$\sum_{\sigma\in\mathcal{E}_K} \int_\sigma \nabla u(x)\cdot n_{K,\sigma} d\gamma(x) = m(K) f_K,$$

avec les notations du paragraphe 1.1.2 page 10.
On approche cette équation par :

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} = m(K) f_K,$$

où $F_{K,\sigma}$ est le flux numérique à travers σ, qu'on approche par :

$$F_{K,\sigma} = \begin{cases} \dfrac{m(\sigma)}{d_{K,\sigma}+d_{L,\sigma}}(u_K - u_L) \text{ si } \sigma \in \mathcal{E}_{int} \cap \mathcal{E}_K, \\ \dfrac{m(\sigma)}{d_{K,\sigma}} u_K \text{ si } \sigma \in \mathcal{E}_{ext} \cap \mathcal{E}_K. \end{cases}$$

On obtient donc bien le schéma (3.4.54) - (3.4.55)
2. Soit $v = (v_K)_{K\in\mathcal{T}} \in H_\tau(\Omega)$ une fonction constante par volumes de contrôle.
On multiplie l'équation (3.4.54) par V_K et on somme sur K. On obtient :

$$\sum_{K\in\mathcal{T}} \left(\sum_{\substack{\sigma\in\mathcal{E}_{int} \\ \sigma=K|L}} \tau_\sigma(u_K - u_L)v_K + \sum_{\sigma\in\mathcal{E}_{ext}\cap\mathcal{E}_K} \tau_\sigma u_K V_K \right) = \sum_K m(K) f_K v_K.$$

Remarquons maintenant que le premier membre de cette égalité est aussi égal, en sommant sur les arêtes du maillage à :

$$\sum_{\substack{\sigma\in\mathcal{E} \\ \sigma=K|L}} (\tau_\sigma(u_K - u_L)v_K) + \tau_\sigma(u_L - u_K)v_L) + \sum_{\tau\in\mathcal{E}_{ext}} \tau_\sigma u_{K_\sigma} v_{K_\sigma}$$

où K_σ désigne le volume de contrôle dont σ est une arête (du bord) dans la deuxième sommation. On obtient donc :

$$a_\tau(u,v) = T_\tau(V), \forall v \in H_\tau(\Omega), \tag{3.6.72}$$

avec :

$$a_\tau(u,v) = \sum_{\substack{\sigma\in\mathcal{E} \\ \sigma=K|L}} \tau_\sigma(u_K - u_L)(v_K - v_L) + \sum_{\sigma\in\mathcal{E}xt} u_{K_\sigma} v_{K,\sigma} \text{et } T_\tau(v) = \sum_K m(K) f_K v_K.$$

On a donc montré que si $u = (u_K)_{K\in\mathcal{T}}$ la solution de (3.4.54) - (3.4.55), alors u est solution de (3.6.72).
Montrons maintenant la réciproque. Soit 1_K la solution caractéristique du volume de contrôle K, définie par

$$1_K(x) = \begin{cases} 1 \text{ si } x \in K \\ 0 \text{ sinon }, \end{cases}$$

Prenons $v = 1_K$ dans (3.6.72), on obtient alors

$$\sum_{\sigma\in\mathcal{E}_{int}} \tau_\sigma(u_K - u_L) + \sum_{\sigma\in\mathcal{E}_{ext}\cap\mathcal{E}_K} \tau_\sigma u_K = m(K) f_K.$$

On retrouve donc bien (3.4.54).
Notons qu'en faisant ceci, on a introduit une discrétisation de la formulation faible (3.1.5) page 115 par une méthode de discrétisation non conforme, puisque $H_\tau \not\subset H^1(\Omega)$.

Chapitre 4

Eléments finis de type Lagrange

On a vu au paragraphe précédent que la construction d' une méthode d'éléments finis, nécessite la donnée d'un maillage, de noeuds et d'un espace de polynômes, qui doivent être choisis de manière cohérente. Nous allons ici introduire une famille d'éléments finis, dits de Lagrange, pour lesquels nous établirons des règles qui permettront de s'assurer la cohérence des choix. Les éléments finis de type Lagrange, qui font intervenir comme "degrés de liberté" (c.à.d. les valeurs qui permettent de déterminer entièrement une fonction) les valeurs de la fonction aux noeuds, sont très largement utilisés dans les applications. Il existe d'autres familles d'éléments finis, comme par exemple les éléments finis de type Hermite qui font également intervenir les valeurs des dérivées directionnelles. Dans le cadre de ce cours, nous n'aborderons que les éléments finis de type Lagrange, et nous renvoyons aux ouvrages cités en introduction pour d'autres éléments.

4.1 Définition et cohérence "locale"

Soit $\mathcal{T}$ un maillage de Ω, pour tout élément K de $\mathcal{T}$, on note Σ_K l'ensemble des noeuds de l'élément. On suppose que chaque élément a N_ℓ noeuds $\mathrm{K} : \Sigma_K = \{a_1, \ldots, a_{N_\ell}\}$ (qui ne sont pas forcément ses sommets). On note P un espace de dimension finie constitué de polynômes (à choisir pour définir la méthode).

Définition 4.1 (Unisolvance, élément fini de Lagrange) *Soit K un élément et $\Sigma_K = (a_i)_{i=1,\ldots,N_\ell}$ un ensemble de noeuds de K. Soit P un espace de polynômes de dimension finie. On dit que le triplet (K,Σ_K,P) est un élément fini de Lagrange si Σ_K est P-unisolvant, c'est à dire si pour tout $(\alpha_1, \ldots, \alpha_{N_\ell}) \in \mathbb{R}^{N_\ell}$, il existe un unique élément $f \in P$ tel que $f(a_i) = \alpha_i \quad \forall i = 1 \ldots N_\ell$. Pour $i = 1, \ldots, N_\ell$, on appelle degré de liberté la forme linéaire ζ_i définie par $\zeta_i(p) = p(a_i)$, pour tout $p \in P$. La propriété d'unisolvance équivaut à dire que la famille $(\zeta_i)_{i=1,\ldots,N_\ell}$ forme une base de P' (espace dual de P).*

La P-unisolvance revient à dire que toute fonction de P est entièrement déterminée par ses valeurs aux noeuds.

Exemple: l'élément fini de Lagrange P_1 Prenons par exemple, en dimension 1, l'élément $K = [a_1,a_2]$, avec $\Sigma_K = \{a_1,a_2\}$, et $P = P_1$ (ensemble des polynômes de degré inférieur ou égal à 1). Le triplet (K,Σ_K,P) est unisolvant s'il existe une unique fonction f de P telle que :

$$\begin{cases} f(a_1) = \alpha_1 \\ f(a_2) = \alpha_2 \end{cases}$$

Or toute fonction f de P s'exprime sous la forme $f(x) = \lambda x + \mu$ et le système

$$\begin{cases} \lambda a_1 + \mu = \alpha_1 \\ \lambda a_2 + \mu = \alpha_2 \end{cases}$$

détermine λ et μ de manière unique.

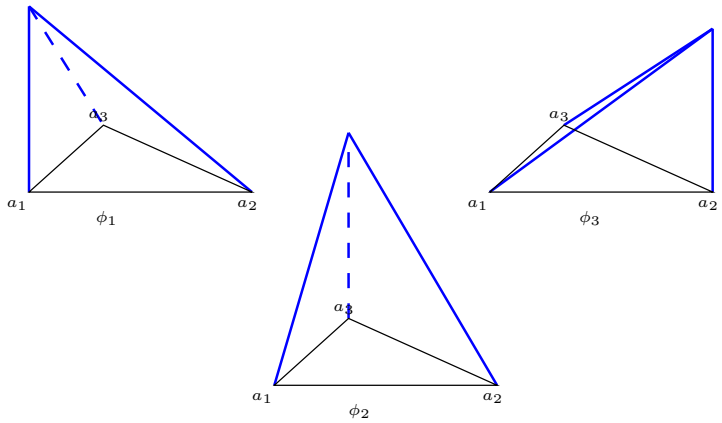

FIG. 4.1 – *fonctions de base locales pour l'élément fini de Lagrange P_1 en dimension 2*

De même si on considère le cas $d = 2$. On prend comme élément K un triangle et comme noeuds les trois sommets, a_1,a_2,a_3 du triangle. Soit $P = P_1 = \{f \ : \mathbb{R} \to \mathbb{R}; f(x) = \lambda x_1 + \mu x_2 + \nu\}$ l'ensemble des fonctions affines. Alors le triplet (K,Σ_K,P) est un élément fini de Lagrange car $f \in P$ est entièrement déterminée par $f(a_1)$,$f(a_2)$ et $f(a_3)$.

Définition 4.2 (Fonctions de base locales) *Si (K,Σ_K,P) est un élément fini de Lagrange, alors toute fonction f de P peut s'écrire :*

$$f = \sum_{i=1}^{N_\ell} f(a_i) f_i$$

avec $f_i \in P$ *et* $f_i(a_j) = \delta_{ij}$. *Les fonctions f_i sont appelées fonctions de base locales.*

Pour l'élément fini de Lagrange P_1 en dimension 2 considéré plus haut, les fonctions de base locales sont décrites sur la figure 4.1

Définition 4.3 (Interpolée) *Soit (K,Σ_K,P) un élément fini de Lagrange, et soit $v \in C(K,\mathbb{R})$. L'interpolée de v est la fonction $\Pi v \in P$ définie par :*

$$\Pi v = \sum_{i=1}^{N_\ell} v(a_i) f_i$$

On montre sur la figure 4.2 un exemple d'interpolée pour l'élément fini de Lagrange P_1 en dimension 1. L'étude de $\|v - \Pi v\|$ va nous permettre d'établir une majoration de l'erreur de consistance $d(u,H_N)$.

Remarque 4.4 *Pour que le triplet (K,Σ_K,P) soit un élément fini de Lagrange, ilfaut, mais il ne suffit pas, que* $\dim P = card\Sigma_K$. *Par exemple si $P = P_1$ et qu'on prend comme noeuds du triangle deux sommets*

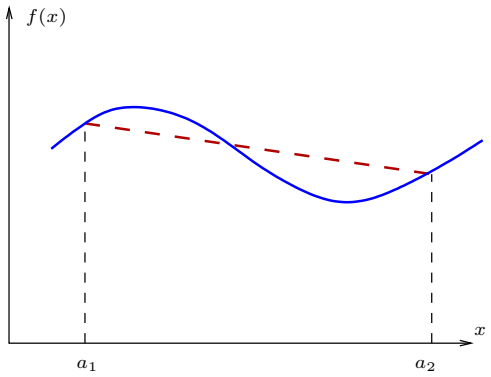

FIG. 4.2 – *Interpolée $P1$ sur $[a_1,a_2]$ (en trait pointillé) d'une fonction régulière (en trait continu)*

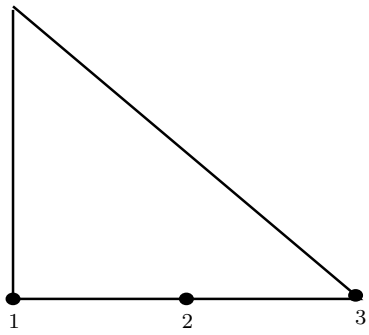

FIG. 4.3 – *Exemple de triangle à trois noeuds qui n'est pas un élément fini de Lagrange)*

et le milieu de l'arête joignant les deux sommets, (voir figure 4.3), (K,Σ_K,P) n'est pas un élément fini de Lagrange.

Proposition 4.5 (Critère de détermination) *Soit (K,Σ,P) un triplet constitué d'un élément, d'un ensemble de noeuds et d'un espace de polynômes, tel que :*

$$\dim P = card\Sigma = N_\ell \tag{4.1.1}$$

Alors

$$si\ \exists! f \in P; f = 0\ sur\ \Sigma \tag{4.1.2}$$

ou si

$$\forall i \in \{1 \dots N_\ell\} \exists f_i \in P \quad f_i(a_j) = \delta_{ij} \tag{4.1.3}$$

alors (K,Σ,P) est un élément fini de Lagrange.

Démonstration : Soit :

$$\phi \ : P \to \mathbb{R}^{N_\ell}$$

$$f \mapsto (f(a_i))^t_{i=1,N_\ell}\,.$$

L'application ϕ est linéaire de P dans $\mathbb{R}^{N_\ell}$, et, par hypothèse $card\Sigma = \dim P$. Donc ϕ est une application linéaire continue de P dans $\mathbb{R}^{N_\ell}$, avec $dimP = dim(\mathbb{R}^{N_\ell}) = N_\ell$. Si (K,Σ,P) vérifie la condition (4.1.2) alors ϕ est injective. En effet, si $\phi(f) = 0$, alors $f(a_i) = 0, \forall i = 1,\ldots,N_\ell$, et donc par hypothèse, $f = 0$. Donc ϕ est une application linéaire, ϕ est injective de P dans $\mathbb{R}^{N_\ell}$ avec $dimP = N_\ell$. On en déduit que ϕ est bijective. Donc toute fonction de P est entièrement déterminée par ses valeurs aux noeuds : (K,Σ,P) est donc un élément fini de Lagrange.

On montre facilement que si la condition (4.1.3) est vérifiée alors ϕ est surjective. Donc ϕ est bijective, et (K,Σ,P) est un élément fini de Lagrange. ∎

Proposition 4.6 *Soit $(\bar{K},\bar{\Sigma},\bar{P})$, un élément fini de Lagrange, où $\bar{\Sigma}$ est l'ensemble des noeuds de $\bar{K}$ et $\bar{P}$ un espace de fonctions de dimension finie, et soit F une bijection de $\bar{K}$ dans K, où K est une maille d'un maillage éléments finis. On pose $\Sigma = F(\bar{\Sigma})$ et $P = \{f : K \to \mathbb{R}; f \circ F \in \bar{P}\}$ (voir figure 4.4). Alors le triplet (K,Σ,P) est un élément fini de Lagrange.*

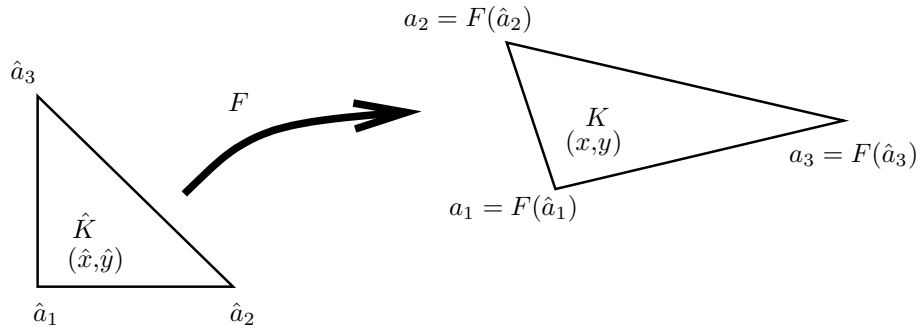

FIG. 4.4 – *Transformation F*

Démonstration : Supposons que les hypothèses de la proposition sont réalisées. On veut donc montrer que (Σ,P) est unisolvant. Soit $\Sigma = (a_1,\ldots,a_{N_\ell})$, et soit $(\alpha_1,\ldots,\alpha_{N_\ell}) \in \mathbb{R}^{N_\ell}$. On veut montrer qu'il existe une unique fonction $f \in P$ telle que

$$f(a_i) = \alpha_i, \quad \forall i = 1,\ldots,N_\ell.$$

Or par hypothèse, $(\bar{\Sigma},\bar{P})$ est unisolvant. Donc il existe une unique fonction $\bar{f} \in \bar{P}$ telle que

$$\bar{f}(\bar{a}_i) = \alpha_i, \quad \forall i = 1,\ldots,N_\ell,$$

(où $(\bar{a}_i)_{i=1,\ldots,N_\ell}$ désignent les noeuds de $\bar{K}$). Soit F la bijection de $\bar{K}$ sur K, on pose $f = \bar{f} \circ F^{-1}$. Or par hypothèse, $a_i = F(\bar{a}_i)$. On a donc : $f(a_i) = \bar{f} \circ F^{-1}(a_i) = \bar{f}(\bar{a}_i) = \alpha_i$. On a ainsi montré l'existence de f telle que $f(a_i) = \alpha_i$.
Montrons maintenant que f est unique. Supposons qu'il existe f et $g \in P$ telles que :

$$f(a_i) = g(a_i) = \alpha_i, \quad \forall i = 1,\ldots,N_\ell.$$

Soit $h = f - g$ on a donc :

$$h(a_i) = 0 \quad \forall i = 1 \ldots N_\ell.$$

On a donc $h \circ F(\bar{a}_i) = h(a_i) = 0$. Or $h \circ F \in \bar{P}$, et comme $(\bar{\Sigma},\bar{P})$ est unisolvant, on en déduit que $h \circ F = 0$. Comme, pour tout $x \in K$, on a $h(x) = h \circ F \circ F^{-1}(x) = h \circ F(F^{-1}(x)) = 0$, on en conclut que $h = 0$. ∎

Définition 4.7 (Eléments affine-équivalents) *. Sous les hypothèses de la proposition 4.6, si la bijection F est affine, on dit que les éléments finis $(\bar{K},\bar{\Sigma},\bar{P})$ et (K,Σ,P) sont affine–équivalents.*

Remarque 4.8 *Soient $(\bar{K},\bar{\Sigma},\bar{P})$ et (K,Σ,P) deux éléments finis afffine–équivalents. Si les fonctions de base locales de $(\bar{K},\bar{\Sigma},\bar{P})$. (resp. de (K,Σ,P)) sont affines, alors celles de K (resp. $\bar{K}$) le sont aussi, et on a :*

$$\left\{ \begin{array}{ll} \bar{f}_i & = f_i \circ F, \\ & \qquad i = 1,\ldots,\mathrm{card}\Sigma \\ f_i & = \bar{f}_i \circ F^{-1}, \end{array} \right.$$

La preuve de cette remarque fait l'objet de l'exercice 43.

Proposition 4.9 (Interpolation) *Sous les hypothèses de la proposition 4.10 page 162, soient $\Pi_{\bar{K}}$ et Π_K les opérateurs d'interpolation respectifs sur $\bar{K}$ et K, voir définition 4.3 page 158. Soient $v \in C(K,\mathbb{R})$, $\Pi_{\bar{K}}v$ et $\Pi_K v$ les interpolées respectives de v sur $(\bar{K},\bar{P})$ et (K,P), alors on a :*

$$\Pi_K v \circ F = \Pi_{\bar{K}}(v \circ F)$$

Démonstration : Remarquons tout d'abord que $\Pi_K v \circ F$ et $\Pi_{\bar{K}}(v \circ F)$ sont toutes deux des fonctions définies de $\bar{K}$ à valeurs dans $\mathbb{R}$, voir figure 4.5. Remarquons ensuite que, par définition de l'interpolée,

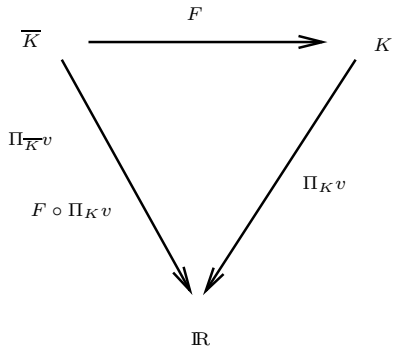

FIG. 4.5 – *Opérateurs d'interpolation $\Pi_{\bar{K}}$ etΠ_K*

$\Pi_K v \in P$. Comme $(\bar{K},\bar{\Sigma},\bar{P})$ est l'élément de référence, on a donc :

$$\Pi_K v \circ F \in \bar{P}$$

On a aussi, par définition de l'interpolée : $\Pi_{\bar{K}}(v \circ F) \in \bar{P}$. On en déduit que $\Pi_K v \circ F$ et $\Pi_{\bar{K}}(v \circ F)$ sont toutes deux des fonctions de $\bar{P}$. Comme l'élément $(\bar{K},\bar{P},\bar{\Sigma})$ est unisolvant (car c'est un élément fini de Lagrange), toute fonction de $\bar{P}$ est uniquement déterminée par ses valeurs aux noeuds de $\bar{\Sigma}$. Pour montrer l'égalité de $\Pi_K v \circ F$ et $\Pi_{\bar{K}}(v \circ F)$, il suffit donc de montrer que :

$$\Pi_{\bar{K}}(v \circ F)(\bar{a}_i) = \Pi_K v \circ F(\bar{a}_i), \quad i = 1,\ldots,N_\ell,$$

où $N_\ell = card\bar{\Sigma}$. Décomposons $\Pi_{\bar{K}}(v \circ F)$ sur les fonctions de base locales $(\bar{f}_j)$,$j = 1,\ldots,N_\ell$. On obtient :

$$\Pi_{\bar{K}}(v \circ F)(\bar{a}_i) = \sum_{j=1}^{N_\ell} v \circ F(\bar{a}_j)\bar{f}_j(\bar{a}_i).$$

On a donc :

$$\Pi_{\bar{K}}(v \circ F)(\bar{a}_i) = v \circ \left(\sum_{j=1}^{N_\ell} F(\bar{a}_j)\bar{f}_j \right)(\bar{a}_i) = v \circ F(\bar{a}_i) = v(a_i).$$

Mais on a aussi :

$$\Pi_K v \circ F(\bar{a}_i) = \Pi_K v(F(\bar{a}_i)) = \Pi_K v(a_i) = v(a_i).$$

D'où l'égalité. ∎

4.2 Construction de H_N et conformité

Nous allons considérer deux cas : le cas où l'espace H est l'espace H^1 tout entier, et le cas où l'espace H est l'espace H_0^1

4.2.1 Cas $H = H^1(\Omega)$

Plaçons-nous ici dans le cas où $H = H^1(\Omega)$, où $\Omega \subset \mathbb{R}^d$ est un ouvert borné polygonal (si $d = 2$, polyèdrique si $d = 3$). Soit $\mathcal{T}$ un maillage éléments finis, avec $\mathcal{T} = (K_\ell)_{\ell=1,\ldots,L}$, où les éléments finis K_ℓ sont fermés et tels que $\cup_{\ell=1}^{L} K_\ell = \bar{\Omega}$. Soit $\mathcal{S} = (S_i)_{i=1,\ldots,M}$ l'ensemble des noeuds du maillage éléments finis, avec $S_i \in \bar{\Omega}$, $\forall i = 1,\ldots,M$.. On cherche à construire une méthode d'éléments finis de Lagrange ; donc à chaque élément K_ℓ, $\ell = 1,\ldots,L$, est associé un ensemble de noeuds $\Sigma_\ell = \mathcal{S} \cap K_\ell$, et un espace P_ℓ de polynômes. On veut que chaque triplet $(K_\ell,\Sigma_\ell,P_\ell)$ soit un élément fini de Lagrange. On définit les fonctions de base globales $(\phi_i)_{i=1,\ldots,M}$, par :

$$\phi_i\,|_{K_\ell} \in P_\ell \qquad \forall i = 1,\ldots,M; \qquad \forall \ell = 1;\ldots,L, \tag{4.2.4}$$

et

$$\phi_i(S_j) = \delta_{ij} \qquad \forall i = 1,\ldots,M, \qquad \forall j = 1,\ldots,M. \tag{4.2.5}$$

Chaque fonction ϕ_i est définie de manière unique, grâce au caractère unisolvant de $(K_\ell,\Sigma_\ell,P_\ell)$, $\ell = 1,\ldots,M$. On pose $H_N = Vect(\phi_1,\ldots,\phi_M)$. Pour obtenir une méthode d'éléments finis conforme, il reste à s'assurer que $H_N \subset H^1$.
Une manière de construire l'espace H_N est de construire un maillage à partir d'un élément de référence, grâce à la proposition suivante, qui se déduit facilement de la proposition 4.6 page 160

Proposition 4.10 (Elément fini de référence) *Soit $\mathcal{T}$ un maillage constitué d'éléments K. On appelle élément fini de référence un élément fini de Lagrange $(\bar{K},\bar{\Sigma},\bar{P})$, où $\bar{\Sigma}$ est l'ensemble des noeuds de $\bar{K}$ et $\bar{P}$ un espace de fonctions, de dimension finie, tel que, pour tout autre élément $K \in \mathcal{T}$, il existe une bijection $F : \bar{K} \to K$ telle que $\Sigma = F(\bar{\Sigma})$ et $P = \{f : K \to \mathbb{R}; f \circ F \in \bar{P}\}$ (voir figure 4.4). Le triplet (K,Σ,P) est un élément fini de Lagrange.*

Proposition 4.11 (Critère de conformité, cas H^1) *Soit Ω un ouvert polygonal (ou polyèdrique) de $\mathbb{R}^d$,$d = 2$ ou 3. Soit $\mathcal{T} = (K_\ell)_{\ell=1,\dots,L}$, un maillage éléments finis de Ω,$\mathcal{S} = (S_i)_{i=1,\dots,M}$ l'ensemble des noeuds de maillage. On se place sous les hypothèses de la proposition 4.10 ; soient $(\phi_i)_{i=1,\dots,M}$ les fonctions de base globales, vérifiant (4.2.4) et (4.2.5), et on suppose de plus que les hypothèses suivantes sont vérifiées :*

$$\text{Pour toute arête (ou face si } d = 3) \ \epsilon = K_{\ell_1} \cap K_{\ell_2}, \text{ on a : } \Sigma_{\ell_1} \cap \epsilon = \Sigma_{\ell_2} \cap \epsilon \text{ et } P_{\ell_1}|_\epsilon = P_{\ell_2}|_\epsilon, \quad (4.2.6)$$

où $P_{\ell_1}|_\epsilon$ (resp. $P_{\ell_2}|_\epsilon$) désigne l'ensemble des restrictions des fonctions de P_{ℓ_1} (resp. P_{ℓ_2}) à ϵ),

$$\text{Si } \epsilon \text{ est un côté de } K_\ell, (\Sigma_\ell \cap \epsilon, P_\ell|_\epsilon) \text{ est unisolvant.} \quad (4.2.7)$$

Alors on a : $H_N \subset C(\bar{\Omega})$ et $H_N \subset H^1(\Omega)$. On a donc ainsi construit une méthode d'éléments finis conformes. (Notons que les côtés de K_ℓ sont des arêtes en 2D et des faces en 3D.)

Démonstration : Pour montrer que $H_N \subset C(\bar{\Omega})$ et $H_N \subset H^1(\Omega)$, il suffit de montrer que pour chaque fonction de base globale ϕ_i, on a $\phi_i \in C(\bar{\Omega})$ et $\phi_i \in H^1(\Omega)$. Or par hypothèse, (4.2.4), chaque fonction ϕ_i est polynômiale par morceaux. De plus, grâce à l'hypothèse (4.2.6), on a raccord des polynômes sur les interfaces des éléments, ce qui assure la continuité de ϕ_i. Il reste à montrer que $\phi_i \in H^1(\Omega)$ pour tout $i = 1,\dots,M$. Comme $\phi_i \in C(\bar{\Omega})$, il est évident que $\phi_i \in L^2(\Omega)$ (car Ω est un ouvert borné, donc $\phi_i \in L^\infty(\Omega) \subset L^2(\Omega)$.
Montrons maintenant que les dérivées faibles $D_j\phi_i, \quad j = 1,\dots,d$, appartiennent à $L^2(\Omega)$. Par définition, la fonction ϕ_i admet une dérivée faible dans $L^2(\Omega)$ s'il existe une fonction $\psi_{i,j} \in L^2(\Omega)$ telle que :

$$\int_\Omega \phi_i(x)\partial_j\varphi(x)dx = -\int_\Omega \psi_{ij}(x)\varphi(x)dx, \quad (4.2.8)$$

pour toute fonction $\varphi \in C_c^1(\Omega)$ (on rappelle que $C_c^1(\Omega)$ désigne l'ensemble des fonctions de classe C^1 à support compact, et que ∂_j désigne la dérivée classique par rapport à la j-ème variable). Or, comme $\bar{\Omega} = \bigcup_{\ell=1}^{L} K_\ell$, on a :

$$\int_\Omega \phi_i(x)D_j\varphi(x)dx = \sum_{\ell=1}^{L} \int_{K_\ell} \phi_i(x)D_j\varphi(x)dx.$$

Sur chaque élément K_ℓ, la fonction ϕ_i est polynômiale. On peut donc appliquer la formule de Green, et on a :

$$\int_{K_L} \phi_i(x)\partial_j\varphi(x)dx = \int_{\partial K_\ell} \phi_i(x)\varphi(x)n_j(x)d\gamma(x) - \int_{K_\ell} \partial_j\phi_i(x)\varphi(x)dx,$$

où $n_j(x)$ est la j-ième composante du vecteur unitaire normal à ∂K_ℓ en x, extérieur à K_ℓ. Mais, si on note $\mathcal{E}_{int}$ l'ensemble des arêtes intérieures du maillage (i.e. celles qui ne sont pas sur le bord), on a :

$$X = \sum_{\ell=1}^{L} \int_{\partial K_\ell} \phi_i(x)\varphi(x)n_j(x)d\gamma(x) = \int_{\partial\bar{\Omega}} \phi_i(x)\varphi(x)n_j(x)d\gamma(x)$$

$$+ \sum_{\epsilon\in\mathcal{E}_{int}} \int \left[(\phi_i(x)\varphi(x)n_j(x))\,\big|_{K_{\ell_1}} + (\phi_i(x)\varphi(x)n_j(x))_{K_{\ell_2}}\right] d\gamma(x).$$

où K_{ℓ_1} et K_{ℓ_2} désignent les deux éléments dont ϵ est l'interface.

Comme φ est à support compact,

$$\int_{\partial\bar{\Omega}} \phi_i(x)\varphi(x)n_j(x)d\gamma(x) = 0.$$

Comme ϕ_i et φ sont continues et comme $n_j(x)\,\big|_{K_{\ell_1}} = -n_j(x)\big|_{K_{\ell_2}}$ pour tout $x \in \epsilon$, on en déduit que $X = 0$. En reportant dans (4.2.1), on obtient donc que :

$$\int_\Omega \phi_i(x)\partial_j\varphi(x)dx = -\sum_{\ell=1}^{L}\int_{K_\ell} \partial_j\phi_i(x)\varphi(x)dx.$$

Soit $\psi_{i,j}$ la fonction de Ω dans $\mathbb{R}$ définie presque partout par

$$\psi_{ij}\,\Big|_{\overset{\circ}{K}_\ell} = -\partial_j\phi_i.$$

Comme $\partial_j\phi_i$ est une fonction polynômiale par morceaux, on a $\psi_{i,j} \in L^2(\Omega)$ qui vérifie (4.2.8), ce qui termine la démonstration. ∎

4.2.2 Cas $H = H_0^1(\Omega)$

Plaçons-nous mainteant dans le cas où $H = H_0^1(\Omega)$. On décompose alors l'ensemble $\mathcal{S}$ des noeuds du maillage :

$$\mathcal{S} = \mathcal{S}_{int} \cup \mathcal{S}_{ext}$$

où

$$\mathcal{S}_{int} = \{S_i, i = 1, \ldots, N\} \subset \Omega$$

est l'ensemble des noeuds intérieurs à Ω et

$$\mathcal{S}_{ext} = \{S_i, i = N+1, \ldots, M\} \subset \partial\Omega$$

est l'ensemble des noeuds de la frontière. Les fonctions de base globales sont alors les fonctions $\phi_i, i = 1, \ldots, N$ telles que

$$phi_i\,|_{K_\ell} \in P_\ell, \forall i = 1, \ldots, N, \quad \forall \ell = 1, \ldots, L \tag{4.2.9}$$

$$\phi_i(S_j) = \delta_{ij}, \forall j = 1, \ldots, N, \tag{4.2.10}$$

et on pose là encore $H_N = Vect\{\phi_1, \ldots, \phi_N\}$. On a alors encore le résultat suivant :

Proposition 4.12 (Critère de conformité, cas H_0^1) *Soit Ω un ouvert polygonal (ou polyèdrique) de $\mathbb{R}^d, d = 2$ ou 3. Soit $\mathcal{T} = (K_\ell)_{\ell=1,\ldots,L}$ un maillage éléments finis de Ω, $\mathcal{S} = (S_i)_{i=1,\ldots,M} = \mathcal{S}_{int} \cup \mathcal{S}_{ext}$ l'ensemble des noeuds du maillage. On se place sous les hypothèses de la proposition 4.6. On suppose que les fonctions de base globale $(\phi_i)_{i=1,\ldots,M}$ vérifient (4.2.9) et (4.2.10), et que les conditions (4.2.6) et (4.2.7) sont vérifiées. Alors on a : $H_N \subset C(\bar{\Omega})$ et $H_N \subset H_0^1(\Omega)$*

Démonstration : La preuve de cette proposition est laissée à titre d'exercice. ∎

Remarque 4.13 (Eléments finis conformes dans $H^2(\Omega)$) *On a construit un espace d'approximation H_N inclus dans $C(\bar{\Omega})$. En général, on n'a pas $H_N \subset C^1(\bar{\Omega})$, et donc on n'a pas non plus $H_N \subset H^2(\Omega)$ (en dimension 1 d'espace, $H^2(\Omega) \subset C^1(\Omega)$). Même si on augmente le degré de l'espace des polynômes,*

on n'obtiendra pas l'inclusion $H_N \subset C^1(\bar{\Omega})$. Si on prend par exemple les polynômes de degré 2 sur les éléments, on n'a pas de condition pour assurer le raccord, des dérivées aux interfaces. Pour obtenir ce raccord, les éléments finis de Lagrange ne suffisent pas : il faut prendre des éléments de type Hermite, pour lesquels les degrés de liberté ne sont plus seulement les valeurs de la fonction aux noeuds, mais aussi les valeurs de ses dérivées aux noeuds. Les éléments finis de Hermite seront par exemple bien adaptés à l'approximation des problèmes elliptiques d'ordre 4, dont un exemple est l'équation :

$$\Delta^2 u = f \text{ dans } \Omega$$

où Ω est un ouvert borné de $\mathbb{R}^2$,$\Delta^2 u = \Delta(\Delta u)$, et avec des conditions aux limites adéquates, que nous ne détaillerons pas ici. On peut, en fonction de ces conditions aux limites, trouver un espace de Hilbert H et une formulation faible de (4.13), qui s'écrit :

$$\begin{cases} \displaystyle\int_\Omega \Delta u(x)\Delta\varphi(x)dx = \int_\Omega f(x)\varphi(x)dx \\ u \in H, \forall \varphi \in H. \end{cases}$$

Pour que cette formulation ait un sens, il faut que $\Delta u \in L^2(\Omega)$ et $\Delta\varphi \in L^2(\Omega)$, et donc que $H \subset H^2(\Omega)$. Pour construire une approximation par éléments finis conforme de ce problème, il faut donc choisir $H_N \subset H^2(\Omega)$, et le choix des éléments finis de Hermite semble donc indiqué.

4.3 Exemples d'éléments finis de Lagrange

Pour chaque méthode d'élément fini de Lagrange, on définit :

1. un élément de référence $\bar{K}$
2. des fonctions de base locales sur $\bar{K}$
3. une bijection F_ℓ de K sur K_ℓ, pour $\ell = 1,\ldots,L$, où L est le nombre déléments du maillage.

4.3.1 Elément fini de Lagrange $P1$ sur triangle $(d = 2)$

Le maillage du domaine est constitué de L triangles $(K_\ell)_{\ell=1,\ldots,L}$, et les polynômes d'approximation sont de degré 1.

Elément fini de référence : on choisit le triangle $\bar{K}$ de sommets (0,0), (1,0) et (0,1), et $\bar{P} = \{\psi : K \to \mathbb{R}(x,y) \mapsto ax + by + c,(a,b,c) \in \mathbb{R}^3\}$.

Proposition 4.14 (Unisolvance) *Soit $\bar{\Sigma} = (\bar{a}_i)_{i=1,2,3}$ avec $\bar{a}_1 = (0,0),\bar{a}_2 = (1,0)$ et $\bar{a}_3 = (0,1)$, et*

$$\bar{P} = \{\psi; K \to \mathbb{R}; (x,y) \mapsto a + bx + cy,(a,b,c) \in \mathbb{R}^3\}$$

Alors le couple $(\bar{\Sigma},\bar{P})$ est unisolvant.

Démonstration : Soit $(\alpha_1,\alpha_2,\alpha_3) \in \mathbb{R}^3$, et $\psi \in \bar{P}$. On suppose que $\psi(\bar{a}_i) = \alpha_i$, $i = 1,2,3$. La fonction ψ est de la forme $\psi(x,y) = a + bx + cy$ et on a donc :

$$\begin{cases} a = \alpha_1 \\ a + b = \alpha_2 \\ a + c = \alpha_3 \end{cases}$$

d'où $c = \alpha_1, b_1 = \alpha_2 - \alpha_1$ et $b_2 = \alpha_3 - \alpha_2$. La connaissance de ψ aux noeuds $(\bar{a}_i)_{i=1,2,3}$ détermine donc entièrement la fonction ψ. ∎

Fonctions de bases locales.
Les fonctions de base locales sur l'élément fini de référence $\bar{K}$ sont définies par $\bar{\phi}_i \in \bar{P}\bar{\phi}_i(\bar{a}_j) = \delta_{ij}$, ce qui détermine les $\bar{\phi}$; de manière unique, comme on vient de le voir. Et on a donc

$$\left\{ \begin{array}{l} \bar{\phi}_1(\bar{x},\bar{y}) = 1 - \bar{x} - \bar{y} \\ \bar{\phi}_2(\bar{x},\bar{y}) = \bar{x} \\ \bar{\phi}_3(\bar{x},\bar{y}) = \bar{y}. \end{array} \right.$$

Transformation F_ℓ
On construit ici une bijection affine qui transforme $\bar{K}$ le triangle de référence en un autre triangle K du maillage. On cherche donc $\ell : \bar{K} \to K$, telle que

$$F_\ell(\bar{a}_i) = a_i \qquad i = 1, \ldots, 3$$

où $\Sigma = (a_i)_{i=1,2,3}$ est l'ensemble des sommets de K. Notons (x_i,y_i) les coordonnées de $a_i, i = 1,2,3$. Comme F_ℓ est une fonction affine de $\mathbb{R}^2$ dans $\mathbb{R}^2$, elle s'écrit sous la forme.

$$F_\ell(\bar{x},\bar{y}) = (\beta_1 + \gamma_1\bar{x} + \delta_1\bar{y}, \beta_2 + \gamma_2\bar{x} + \delta_2\bar{y})^t$$

et on cherche $\beta_i,\gamma_i,\delta_i, \quad i = 1,2$ tels que :

$$\left\{ \begin{array}{l} F_\ell((0,0)) = (x_1,y_1) \\ F_\ell((1,0)) = (x_2,y_2) \\ F_\ell((0,1)) = (x_3,y_3). \end{array} \right.$$

Une résolution de système élémentaire amène alors à :

$$F_\ell(\bar{x},\bar{y}) = \left(\begin{array}{l} x_1 + (x_2 - x_1)\bar{x} + (x_3 - x_1)\bar{y} \\ y_1 + (y_2 - y_1)\bar{x} + (y_3 - y_1)\bar{y} \end{array} \right)$$

D'après la remarque 4.8 page 161, si on note $\bar{\phi}_k, k = 1,2,3$ les fonctions de base locales de l'élément de référence $(\bar{K},\bar{\Sigma},\bar{P})$, et $\phi_k^{(\ell)}, k = 1,2,3$ les fonctions de base locales de l'élément $(K_\ell,\Sigma_\ell,P_\ell)$, on a $\phi_k^{(\ell)} = \bar{\phi}_k \circ F_\ell^{-1}$
Si on note maintenant $(\phi_i)_{i=1,\ldots,N}$ les fonctions de base globales, on a :

$$\phi_i \Big|_{K_\ell} = \phi_k^{(\ell)},$$

où $i = ng(\ell,k)$ est l'indice du k-ième noeud de l'élément ℓ dans la numérotation globale. Notons que l'élément fini de Lagrange ainsi défini vérifie les critères de cohérence 4.2.6 page 163 et (4.2.7) page 163. Pour compléter la définition de l'espace d'approximation H_N, il ne reste qu'à déterminer les "noeuds liés", de la façon dont on a traité le cas de l'espace $H_0^1(\Omega)$.
Il faut également insister sur le fait que cet élément est très souvent utilisé, en raison de sa facilité d'implantation et de la structure creuse des systèmes linéaires qu'il génère. Il est particulièrement bien adapté lorsqu'on cherche des solutions dans l'espace $H^1(\Omega)$. Il se généralise facilement en trois dimensions d'espace, où on utilise alors des tétraèdres, avec toujours comme espace de polynôme l'espace des fonctions affines.

4.3.2 Elément fini triangulaire $P2$

Comme le titre du paragraphe l'indique, on considère un maillage triangulaire, et un espace de polynômes de degré 2 pour construire l'espace d'approximation.

Elément fini de référence On choisit comme élément fini de référence le triangle de sommets (0,0), (1,0) et (0,1), voir Figure 4.6 et on prend pour Σ :

$$\bar{\Sigma} = \{(0,0),(1,0),(0,1),(\frac{1}{2},\frac{1}{2}),(0,\frac{1}{2}),(\frac{1}{2},0)\}$$

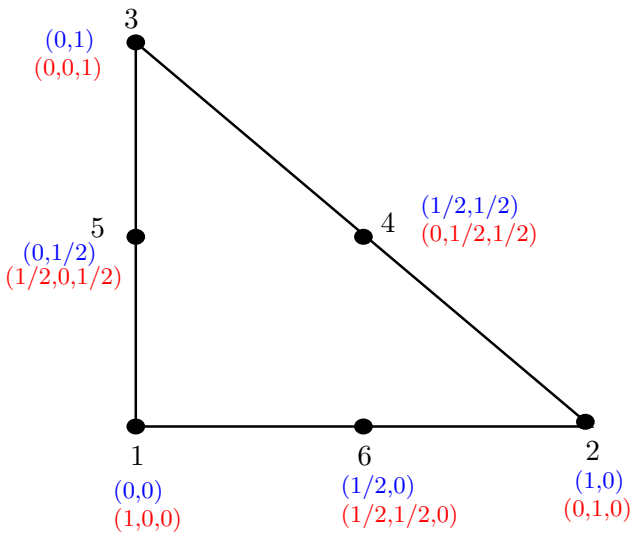

FIG. 4.6 – *Elément de référence pour les éléments finis P2, avec coordonnées cartésiennes et barycentriques des noeuds*

Fonctions de base locales Les fonctions de base locales sont définies à partir des coordonnées barycentriques. On rappelle que les coordonnées barycentriques d'un point $\mathbf{x}$ du triangle K de sommets a_1,a_2 et a_3 sont les réels λ_1,λ_2,λ_3 tels que :

$$\mathbf{x} = \lambda_1 a_1 + \lambda_2 a_2 + \lambda_3 a_3.$$

Dans le cas du triangle de référence $\bar{K}$ de sommets (0,0), (1,0) et (0,1), les coordonnées barycentriques d'un point $\{\mathbf{x}\}$ de coordonnées cartésiennes x et y sont donc : $\lambda_1 = 1 - x - y$, $\lambda_2 = x$, $\lambda_3 = y$. Par définition, on a $\sum_{i=1}^{3} \lambda_i = 1$ et $\lambda_i \geq 0$ (car le triangle K est l'enveloppe convexe de l'ensemble de ses sommets). On peut alors déterminer les fonctions de base en fonction des coordonnées barycentriques des six noeuds de $\bar{K}$ exprimés par leurs coordonnées barycentriques : $a_1 = (1,0,0)$, $a_2 = (0,1,0)$, $a_3 = (0,0,1)$, $a_4 = (0,\frac{1}{2},\frac{1}{2})$, $a_5 = (\frac{1}{2},0,\frac{1}{2})$, $a_6 = (\frac{1}{2},\frac{1}{2},0)$. Les fonctions de base sont telles que $\phi_i \in P_2$, et $\phi_i(a_j) = \delta_{ij}$, $\forall i = 1,\ldots,6$, $forall j = 1,\ldots,6$. Commençons par ϕ_1 ; on veut $\phi_1(a_1) = 1$, et $\phi_i(a_i) = 0$, $\forall i = 2,\ldots,6$. La fonction ϕ_1 définie par

$$\phi_1(x,y) = 2\lambda_1(\lambda_1 - \frac{1}{2})$$

convient, et comme le couple ($\bar{\Sigma}$,$P2$) est unisolvant, c'est la seule fonction qui convient. Par symétrie, on définit

$$\phi_2(x,y) = 2\lambda_2(\lambda_2 - \frac{1}{2}),$$

et

$$\phi_3(x,y) = 2\lambda_3(\lambda_3 - \frac{1}{2}).$$

Les fonctions de base associées aux noeuds a_4,a_5,a_6 sont alors

$$\phi_4(x,y) = 4\lambda_2\lambda_3,$$

$$\phi_5(x,y) = 4\lambda_1\lambda_3,$$

$$\text{et } \phi_6(x,y) = 4\lambda_1\lambda_2.$$

Il est facile de voir que ces fonctions forment une famille libre d'éléments de $P2$ et comme card $\bar{\Sigma}$ = card $P2$, le couple ($\bar{\Sigma}$,$P2$) est bien unisolvant.

Transformation F_ℓ La bijection F_ℓ qui permet de passer de l'élément fini de référence $\bar{K}$ à l'élément K_ℓ a déjà été vue dans le cas de l'élément fini P_1 c'est la fonction affine définie par :

$$F_\ell(x,y) = \begin{pmatrix} x_1 + (x_2 - x_1)x + (x_3 - x_1)y \\ y_1 + (y_2 - y_1)x + (y_3 - y_1)y \end{pmatrix}$$

où (x_i,y_i),$i = 1,2,3$ sont les coordonnées respectives des trois sommets du triangle K_ℓ. Comme cette transformation est affine, les coordonnées barycentriques restent inchangées par cette transformation.

On peut montrer (ce n'est pas facile) que l'erreur d'interpolation $\|u - u_N\|_{H^1}$ est contrôlée, en éléments finis P_1 et P_2 par les inégalités suivantes :

$$P1 : \text{ si } u \in H^2(\Omega), \text{ on a } \|u - u_N\|_{H^1(\Omega)} \leq Ch\|u\|_{H^2(\Omega)}$$

$$P2 : \text{ si } u \in H^3(\Omega), \text{ on a } \|u - u_N\|_{H^1(\Omega)} \leq Ch^2\|u\|_{H^3(\Omega)}.$$

On peut généraliser les éléments finis P_1 et P_2 aux éléments finis P_k sur triangles, pour $k \geq 1$. On prend toujours le même élément de référence, dont on divise chaque côté en k intervalles. Les extrémités de ces intervalles sont les noeuds du mailage. On a donc $3k$ noeuds, qu'on peut repérer par leurs coordonnés barycentriques, qui prennent les valeurs $0,\frac{1}{k},\frac{2}{k},\ldots,1$. On peut montrer que si $u \in H^{k+1}$, alors

$$\|u_N - u\|_{H^1(\Omega)} \leq Ch^k\|u\|_{H^{k+1}(\Omega)}$$

4.3.3 Eléments finis sur quadrangles

Le cas rectangulaire

On prend comme élément fini de référence le carré $\bar{K} = [-1,1] \times [-1,1]$, et comme noeuds les coins de ce carré :

$$a_1 = (1,-1), a_2 = (1,1), a_3 = (-1,1), \text{ et } a_4 = (-1,-1).$$

On prend comme espace de polynômes

$$P = \{f : \bar{K} \to \mathbb{R}; f \in Q_1\}$$

où $Q_1 = \{f : \mathbb{R}^2 \to \mathbb{R}; f(x,y) = a + bx + cy + dxy, (a,b,c,d) \in \mathbb{R}^4\}$ Le couple (Σ,P) est unisolvant. Les fonctions de base locales sont les fonctions :

$$\phi_1(x,y) = -\frac{1}{4}(x+1)(y-1)$$

$$\phi_2(x,y) = \frac{1}{4}(x+1)(y+1)$$

$$\phi_3(x,y) = -\frac{1}{4}(x-1)(y+1)$$

$$\phi_4(x,y) = \frac{1}{4}(x-1)(y-1).$$

La transformation F_ℓ permet de passer de l'élément de référence carré $\bar{K}$ à un rectangle quelconque du maillage K_ℓ. Si on considère un rectangle K_ℓ parallèle aux axes, dont les noeuds sont notés (x_1,y_1), (x_2,y_1), (x_2,y_2), (x_1,y_2), les noeuds du rectangle K_ℓ, la bijection F_ℓ s'écrit :

$$F_\ell(x,y) = \frac{1}{2}\begin{pmatrix} (x_2-x_1)x + x_2 + x_1 \\ (y_2-y_1)y + y_2 + y_1 \end{pmatrix}.$$

Considérons maintenant le cas d'un maillage quadrangulaire quelconque. Dans ce cas, on choisit toujours comme élément de référence le carré unité. La transformation F_ℓ qui transforme l'élément de référence en un quadrangle K_ℓ est toujours affine, mais par contre, les composantes de $F_\ell((x,y))$ dépendent maintenant de x et de y voir exercice 42 page 189. En conséquence, le fait que $f \in Q_1$ n'entraîne plus que $f \circ F_\ell \in Q_1$. Les fonctions de base seront donc des polynômes Q_1 sur l'élément de référence $\bar{K}$, mais pas sur les éléments "courants" K_ℓ.

Eléments finis d'ordre supérieur Comme dans le cas d'un maillage triangulaire, on peut choisir un espace de polynômes d'ordre supérieur, Q_k, pour les fonctions de base de l'élément de référence $\bar{K} = [-1,1] \times [-1,1]$. On choisit alors comme ensemble de noeuds : $\bar{\Sigma} = \bar{\Sigma}_k = \{(x,y) \in \bar{K}, (x,y) \in \{-1, -1+\frac{1}{k}, -1+\frac{1}{k}, \ldots, 1\}^2$. On peut montrer facilement que $(\bar{\Sigma}_k \qquad Q_k)$ est unisolvant. Là encore, si la solution exacte de problème continu est suffisamment régulière, on peut démontrer l'estimation d'erreur suivante (voir [3]) :

$$\|u - u_N\|_{H'(\Omega)} \geq C\|u\|_{H^{k+1}(\Omega)} h^k.$$

Exprimons par exemple l'espace des polynômes Q_2. On a :

$$Q_2 = \{f : \mathbb{R} \to \mathbb{R}; f(x) = a_1 + a_2x + a_3y + a_4xy + a_5x^2 + a_6y^2 + a_7xy^2 + a_8x^2y + a_9x^2y^2, a_i \in \mathbb{R}, i = 1, \ldots, 9\}$$

L'espace Q_2 comporte donc neuf degrés de liberté. On a donc besoin de neuf noeuds dans $\bar{\Sigma}$ pour que le couple $(\bar{\Sigma},Q_2)$ soit unisolvant (voir exercice 47 page 190). On peut alors utiliser comme noeuds sur le carré de référence $[-1,1] \times [-1,1]$:

$$\bar{\Sigma} = \{(-1,-1), (-1,0), (-1,-1), (0,-1), (0,0), (0,1), (1,-1), (1,0), (1,1)\}$$

En général, on préfère pourtant supprimer le noeud central (0,0)et choisir :

$$\Sigma^* = \bar{\Sigma} \setminus \{(0,0)\}.$$

Il faut donc un degré de liberté en moins pour l'espace des polynômes. On définit alors:

$$Q_2^* = \{f : \mathbb{R} \to \mathbb{R}; f(x) = a_1 + a_2x + a_3y + a_4xy + a_5x^2 + a_6y^2 + a_7xy^2 + a_8x^2y, a_i \in \mathbb{R}, i = 1, \ldots, 8\}$$

Le couple (Σ^*,Q_2^*) est unisolvant (voir exercice 48 page 191), et on peut montrer que l'élément Q_2^* est aussi précis (et plus facile à mettre en oeuvre que l'élément Q_2).

4.4 Construction du système linéaire

Pour effectuer la construction du système linéaire obtenu par une discrétisation par éléments finis, on va considérer le problème modèle suivant, qui contient déjà plusieurs difficultés. Soit Ω un ouvert polygonal [1], on suppose que $\partial\Omega : \Gamma_0 \cup \Gamma_1$ avec $\text{mes}(\Gamma_0) \neq 0$. On va imposer des conditions de Dirichlet sur Γ_0 et des conditions de Fourier sur Γ_1 (on dit qu'on a des conditions "mixtes"). On se donne donc des fonctions $p : \Omega \to \mathbb{R}$, $g_0 : \Gamma_0 \to \mathbb{R}$ et $g_1 : \Gamma_1 \to \mathbb{R}$, et on considère le problème :

$$\begin{cases} -div(p(x)\nabla u(x)) + q(x)u(x) = f(x), x \in \Omega, \\ u = g_0 \text{ sur } \Gamma_0, \\ p(x)\nabla u(x).\mathbf{n}(x) + \sigma u(x) = g_1(x), x \in \Gamma_1, \end{cases} \tag{4.4.11}$$

où $\mathbf{n}$ désigne le vecteur unitaire normal à $\partial\Omega$ extérieure à Ω. Pour assurer l'existence et unicité du problème (4.4.11), (voir exercice 34), on se place sous les hypothèses suivantes :

$$\begin{cases} p(x) \geq \alpha > 0, p.p.\, x \in \Omega \\ q \geq 0 \\ \sigma \geq 0 \\ mes(\Gamma_0) > 0. \end{cases} \tag{4.4.12}$$

Pour obtenir une formulation variationnelle, on introduit l'espace

$$H^1_{\Gamma_0,g_0} = \{u \in H^1(\Omega); u = g_0 \text{ sur } \Gamma_0\}$$

et l'espace vectoriel associé :

$$H = H^1_{\Gamma_0,0} = \{u \in H^1(\Omega); u = 0 \text{ sur } \Gamma_0\}$$

Notons que H est un espace de Hilbert. Par contre, attention, l'espace $H^1_{\Gamma_0,g_0}$ n'est pas un espace vectoriel. On va chercher u solution de (4.4.11) sous la forme $u = \widetilde{u} + u_0$, avec $u_0 \in H^1_{\Gamma_0,g_0}$ et $\widetilde{u} \in H^1_{\Gamma_0,0}$. Soit $v \in H$, on multiplie (4.4.11) par v et on intègre sur Ω. On obtient :

$$\int_\Omega -div(p(x)\nabla u(x))v(x)dx + \int_\Omega q(x)u(x)v(x)dx = \int_\Omega f(x)v(x)dx, \quad \forall v \in H.$$

En appliquant la formule de Green, il vient alors :

$$\int_\Omega p(x)\nabla u(x)\nabla v(x)dx - \int_{\partial\Omega} p(x)\nabla u(x)nv(x)d\gamma(x) + \int_\Omega qu(x)v(x)dx = \int_\Omega f(x)v(x)dx, \quad \forall v \in H.$$

Comme $v = 0$ sur Γ_0 on a :

$$\int_{\partial\Omega} p(x)\nabla u(x)nv(x)d\gamma(x) = \int_{\Gamma_1} p\nabla u(x)nv(x)d\gamma(x).$$

Mais sur Γ_1, la condition de Fourier s'écrit : $\nabla u.n = -\sigma u + g_1$, et on a donc

$$\int_\Omega p(x)\nabla u(x)\nabla v(x)dx + \int_{\Gamma_1} p(x)\sigma u(x)v(x)d\gamma(x) + \int_\Omega qu(x)v(x)dx = \int_\Omega f(x)v(x)dx + \int_{\Gamma_1} g_1(x)v(x)d\gamma(x).$$

1. Dans le cas où la frontière $\partial\Omega$ de Ω n'est pas polymale mais courbe, il faut considérer des éléments finis dits "isoparamétriques" que nous verrons plus loin

On peut écrire cette égalité sous la forme : $a(u,v) = \tilde{T}(v)$, avec $a(u,v) = a_\Omega(u,v) + a_{\Gamma_1}(u,v)$, où :

$$\begin{cases} a_\Omega(u,v) = \displaystyle\int_\Omega p(x)\nabla u(x)\nabla v(x)dx + \int_\Omega q(x)u(x)v(x)dx, \\ a_\Gamma(u,v) = \displaystyle\int_{\Gamma_1} p(x)\sigma(x)u(x)v(x)d\gamma(x), \end{cases}$$

et $\tilde{T}(v) = T_\Omega(v) + T_{\Gamma_1}(v)$, avec

$$T_\Omega(v) = \int_\Omega f(x)v(x)dx \text{ et } T_{\Gamma_1} = \int_{\Gamma_1} g_1(x)v(x)d\gamma(x).$$

On en déduit une formulation faible associée à (4.4.11):

$$\begin{cases} \text{chercher } \widetilde{u} \in H \\ a(u_0 + \widetilde{u},v) = \tilde{T}(v), \forall v \in H, \end{cases} \tag{4.4.13}$$

où $u_0 \in H^1(\Omega)$ est un révèlement de g_0, c'est à dire une fonction de $H^1(\Omega)$ telle que $u_0 = g_0$ sur Γ. Le problème (4.4.13) peut aussi s'écrire sous la forme :

$$\begin{cases} \widetilde{u} \in H \\ a(\widetilde{u},v) = T(v), \forall v \in H. \end{cases} \tag{4.4.14}$$

où $T(v) = \tilde{T}(v) - a(u_0,v)$. Sous les hypothèses (4.4.12), on peut alors appliquer le théorème de Lax Milgram (voir théorème 3.6 page 115) au problème (4.4.14) pour déduire l'existence et l'unicité de la solution de (4.4.13) ; notons que, comme la forme bilinéaire a est symétrique, ce problème admet aussi une formulation variationnelle :

$$\begin{cases} J(u) = \displaystyle\min_{v \in H^1_{\Gamma_0,g_0}} J(v), \\ \\ J(v) = \dfrac{1}{2}a(v,v) + T(v), \forall v \in H^1_{\Gamma_0,g_0}. \end{cases} \tag{4.4.15}$$

Dans ce cas, les méthodes de Ritz et Galerkin sont équivalentes. Remarquons que l'on peut choisir u_0 de manière abstraite, tant que u_0 vérifie $u_0 = g_0$ sur Γ_0 et $u_0 \in H^1$. Intéressons nous maintenant à la méthode d'approximation variationnelle. On approche l'espace H par $H_N = Vect\{\phi_1, \ldots, \phi_N\}$ et on remplace (4.4.14) par :

$$\begin{cases} \widetilde{u}_N \in H_N \\ a(\widetilde{u}_N,\phi_i) = T(\phi_i) - a(u_0,\phi_i), \forall i = 1, \ldots, N. \end{cases} \tag{4.4.16}$$

On pose maintenant $\widetilde{u}_N = \displaystyle\sum_{j=1}^{N} \widetilde{u}_j\phi_j$. Le problème (4.4.16) est alors équivalent au système linéaire :

$$\mathcal{K}\widetilde{U} = \mathcal{G},$$

avec

$$\begin{cases} \mathcal{K}_{ij} = a(\phi_j,\phi_i), i,j = 1, \ldots, N, \\ \widetilde{U} = (\widetilde{u}_1 \ldots \widetilde{u}_N)^t, \\ \mathcal{G}_i = T(\phi_i) - a(u_0,\phi_i), i = 1, \ldots, N. \end{cases}$$

L'implantation numérique de la méthode d'approximation nécessite donc de :

1. construire $\mathcal{K}$ et $\mathcal{G}$
2. résoudre $\mathcal{K}\widetilde{U} = \mathcal{G}$.

Commençons par la construction de l'espace H_N et des fonctions de base pour une discrétisation par éléments finis de Lagrange du problème (4.4.14).

4.4.1 Construction de H_N et Φ_i

On considère une discrétisation à l'aide déléments finis de Lagrange, qu'on note : $(K_\ell,\Sigma_\ell,P_\ell)$ $\quad \ell = 1,\ldots,L$, où L est le nombre d'éléments. On note S_i, $i = 1,\ldots,M$, les noeuds du maillage, et $\phi_1,\ldots,\phi_N$, les fonctions de base, avec $N \leq M$. On peut avoir deux types de noeuds :

- les noeuds libres : $S_i \notin \Gamma_0$. On a N noeuds libres
- les noeuds liés : $S_i \in \Gamma_0$. On a $M - N$ noeuds liés.

Notons qu'on a intérêt à mettre des noeuds à l'intérsection de Γ_0 et Γ_1 (ce seront des noeuds liés). Grâce à ceci, et à la cohérence globale et locale des éléments finis de Lagrange, on a $H_N \subset H$. On a donc bien des éléments finis conformes. Récapitulons alors les notations :

- M : nombre de noeuds total
- N : nombre de noeuds libres
- $M_0 = M - N$: nombre de noeuds liés
- $J_0 = \{$ indices des noeuds liés$\} \subset \{1,\ldots,M\}$. On a $cardJ_0 = M_0$
- $J = \{$ indices des noeuds libres $\} = \{1 \ldots M\}\ J_0$. On a $cardJ = N$.

Pour la programmation des éléments finis, on a besoin de connaître, pour chaque noeud (local) de chaque élément, son numéro dans la numérotation globale. Pour cela on introduit un tableau $ng(L,N_\ell)$, où L est le nombre d'éléments et N_ℓ est le nombre de noeuds par élément. (on le suppose constant par souci de simplicité, N_ℓ peut en fait dépendre de L. Exemple : triangle - quadrangle). Pour tout $\ell \in \{1,\ldots,L\}$ et tout $r \in \{1,\ldots,N_\ell\}$, $ng(\ell,r)$ est alors le numéro global du r-ième noeud du ℓ-ième élément. On a également besoin de connaître les coordonnées de chaque noeud. On a donc deux tableaux $x(M)$ et $y(M)$, où $x(i)$,$y(i)$ représentent les coordonnées du i-ème noeud. Notons que les tableaux ng,x et y sont des données du mailleur (qui est un module externe par rapport au calcul éléments finis proprement dit). Pour les conditions aux limites, on se donne deux tableaux :

- CF : conditions de Fourier
- CD : conditions de Dirichlet

(on verra plus tard le format de ces deux tableaux)

4.4.2 Construction de $\mathcal{K}$ et $\mathcal{G}$

On cherche à construire la matrice $\mathcal{K}$ d'ordre $(N \times N)$, définie par :

$$\mathcal{K}_{ij} = a(\phi_j,\phi_i) \qquad i,j \in J$$

Ainsi que le vecteur $\mathcal{G}$, défini par :

$$\mathcal{G}_i = T(\phi_i) - a(u_0,\phi_i) \qquad i \in J \quad cardJ = N$$

La première question à résoudre est le choix de u_0. En effet, contrairement au cas unidimensionnel (voir exercice 29 page 138), il n'est pas toujours évident de trouver $u_0 \in H^1_{\Gamma_0,g_0}$. Pour se faciliter la tâche, on

commet un "crime variationnel", en remplaçant u_0 par

$$u_{0,N} = \sum_{j \in J_0}^{N} u_0(S_j)\phi_j.$$

Notons qu'on a pas forcément :

$$u_{0,N} \in H^1_{\Gamma_0,g_0}$$

(en ce sens, c'est un "crime"). Mais on a $u_{0,N}(S_j) = u_0(S_j), \forall j \in J$. On peut voir la fonction $u_{0,N}$ comme une approximation non conforme de $u_0 \in H^1_{\Gamma_0,g_0}$ On remplace donc $\mathcal{G}_i$ par :

$$\mathcal{G}_i = T(\phi_i) - \sum_{j \in J_0} g_0(S_j)a(\phi_j,\phi_i).$$

Calculons maintenant $a(\phi_j,\phi_i)$ pour $j = 1,\ldots,M$, et $i = 1,\ldots,M$.

Calcul de $\mathcal{K}$ et $\mathcal{G}$

1. Calcul des contributions intérieures : on initialise les coefficients de la matrice $\mathcal{K}$ et les composantes par les contributions provenant de a_Ω et T_Ω.

$$\left.\begin{array}{l} \mathcal{K}_{ij} = a_\Omega(\phi_j,\phi_i) \\ \mathcal{G}_i = T_\Omega(\phi_i) \end{array}\right\} \begin{array}{l} i = 1,\ldots,N, \\ j = 1,\ldots,N. \end{array}$$

2. Calcul des termes de bord de Fourier. On ajoute maintenant à la matrice $\mathcal{K}$ les contributions de bord :

$$\begin{array}{l} \mathcal{K}_{ij} \longleftarrow \mathcal{K}_{ij} + a_{\Gamma_i}(\phi_j,\phi_i), i = 1,\ldots,N, j = 1,\ldots,N. \\ \mathcal{G}_i \longleftarrow \mathcal{G}_i + T_{\Gamma_i}(\phi_i) \quad i = 1 \ldots M. \end{array}$$

3. Calcul des termes de bord de Dirichlet. On doit tenir compte ici du relèvement de la condition de bord :

$$\mathcal{G}_i \longleftarrow \mathcal{G}_i - \sum_{j \in J_0} g_0(N_i)\mathcal{K}_{ij} \qquad \forall i \in J$$

   Après cette affectation, les égalités suivantes sont vérifiées :

$$\begin{array}{l} \mathcal{K}_{ij} = a(\phi_j,\phi_i) \qquad i,j \in J(\cup J_0) \\ \mathcal{G}_i = T(\phi_i) - a(u_{0,N},\phi_i). \end{array}$$

   Il ne reste plus qu'à résoudre le système linéaire

$$\sum_{j \in J} \mathcal{K}_{ij}\alpha_j = \mathcal{G}_i, \forall i \in J. \tag{4.4.17}$$

4. Prise en compte des noeuds liés. Pour des questions de structure de données, on d'inclut en général les noeuds liés dans la résolution du système, et on résout donc le d'ordre $M \geq N$ suivant :

$$\sum_{j=1,\ldots,N} \tilde{\mathcal{K}}_{ij}\alpha_j = \mathcal{G}_i. \forall i = 1,\ldots,N. \tag{4.4.18}$$

   avec $\tilde{\mathcal{K}}_{ij} = \mathcal{K}_{ij}$ pour $i,j \in J$, $\tilde{\mathcal{K}}_{ij} = 0$ si $(i,j) \notin J^2$, et $i \neq j$, et $\tilde{\mathcal{K}}_{ii} = 1$ si $i \notin J$. Ces deux systèmes sont équivalents, puisque les valeurs aux noeuds liées sont fixées.

Si par chance on a numéroté les noeuds de manière à ce que tous les noeuds liés soient en fin de numérotation, c.à.d. si $J = \{1, \ldots, N\}$ et $J_0 = \{N+1, \ldots, M\}$, le système (4.4.18) est de la forme :

$$\left(\begin{array}{c|c} & \\ \mathcal{K} & 0 \\ & \\ \hline & \\ - - - & - - - \\ & \\ \hline & \\ 0 & Id_M \\ & \end{array}\right), U = \left(\begin{array}{c} \alpha_1 \\ \vdots \\ \alpha u_N \\ -- \\ \alpha_{N+1} \\ \vdots \\ \alpha_M \end{array}\right), \text{ et } \mathcal{G} = \left(\begin{array}{c} \mathcal{G}_1 \\ \vdots \\ \mathcal{G}_N \\ -- \\ \mathcal{G}_{N+1} \\ \vdots \\ \mathcal{G}_M \end{array}\right)$$

Dans le cas où la numérotation est quelconque, les noeuds liés ne sont pas forcément à la fin, et pour obtenir le système linéaire d'ordre M (4.4.18) (donc incluant les inconnues α_i, $i \in J_0$, qui n'en sont pas vraiment) on peut adopter deux méthodes :

(a) Première méthode : on force les valeurs aux noeuds liés de la manière suivante :
$\mathcal{K}_{ii} \longleftarrow 1$ pour tout $i \in J_0$
$\mathcal{K}_{ij} \longleftarrow 0$ pour tout $i \in J_0 \qquad j \in \{1 \ldots M\} \quad i \neq j$
$\mathcal{G}_i \longleftarrow g_0(S_i)$ pour tout $i \in J_0$

(b) Deuxième méthode : on force les valeurs aux noeuds liés de la manière suivante :
$\mathcal{K}_{ii} \longleftarrow 10^{20} \qquad \forall i \in J_0$
$\mathcal{G}_i \longleftarrow 10^{20} g_0(S_i) \qquad \forall i \in J_0$

La deuxième méthode permet d'éviter l'affectation à 0 de coefficients extra-diagonaux de la matrice. Elle est donc un peu moins chère en temps de calcul.

Conclusion Après les calculs 1, 2, 3, 4, on a obtenu une matrice $\mathcal{K}$ d'ordre $M \times M$ et le vecteur $\mathcal{G}$ de $\mathbb{R}^M$. Soit $\alpha \in \mathbb{R}^M$ la solution du système $\mathcal{K}\alpha = \mathcal{G}$. Rappelons qu'on a alors :

$$\begin{aligned} u_N &= \sum_{i=1}^{N} \alpha_i \phi_i, \\ &= \sum_{i \in J} \alpha_i \phi_i + \sum_{i \in J_0} \alpha_i \phi_i \end{aligned}$$

$$u_N = \widetilde{u}_N + u_0$$

Remarque 4.15 (Numérotation des noeuds) *Si on utilise une méthode itérative sans préconditionnement, la numérotation des noeuds n'est pas cruciale. Elle l'est par contre dans le cas d'une méthode directe et si on utilise une méthode itérative avec préconditionnement. Le choix de la numérotation s'effectue pour essayer de minimiser la largeur de bande. On pourra à ce sujet étudier l'influence de la numérotation sur deux cas simples sur la structure de la matrice.*

4.4.3 Calcul de a_Ω et T_Ω, matrices élémentaires.

Détaillons maintenant le calcul des contributions intérieures, c'est à dire $a_\Omega(\phi_i,\phi_j)$ $i = 1, \ldots M, j = 1, \ldots, M$ et $T_\Omega(\phi_i)$ $i = 1, \ldots, M$. Par définition,

$$a_\Omega(\phi_i,\phi_j) = \int_\Omega p(x) \nabla\phi_i(x) \nabla\phi_j(x) dx + \int_\Omega q(x) \phi_i(x) \phi_j(x) dx.$$

Décomposons Ω à l'aide du maillage éléments finis.

$$\Omega = \bigcup_{\ell=1}^{L} K_\ell.$$

En notant $\theta(\phi_i,\phi_j)(x) = p(x)\nabla\phi_i(x)\nabla\phi_j(x) + q(x)\phi_i(x)\phi_j(x)$,
On a donc :

$$a_\Omega(\phi_i,\phi_j) = \sum_{\ell=1}^{L} \int_{K_\ell} \theta(\phi_i,\phi_j)dx.$$

Pour r et s numéros locaux de l'élément K_ℓ, on pose :

$$k^\ell_{r,s} = \int_\ell \theta(\phi_s,\phi_r)dx.$$

On va calculer $k^\ell_{r,s}$ puis on calcule $a_\Omega(\phi_i,\phi_j)$, en effectuant un parcours sur les éléments, ce qui s'exprime par l'algorithme suivant :

Initialisation : $\mathcal{K}_{ij} \longleftarrow 0$, $i = 1,\dots M$, $j \leq i$.
Boucle sur les éléments
Pour $\ell = 1$ à L faire
  Pour $r = 1$ à N_ℓ faire
    $i = ng(\ell,r)$ numéro global du noeud r de l'élément ℓ
      Pour $s = 1$ à r faire
        calcul de $k^\ell_{r,s}$
          $j = ng(\ell,s)$
        si $i \geq j$
            $\mathcal{K}_{ij} \longleftarrow \mathcal{K}_{ij} + k^\ell_{r,s}$
        sinon
            $\mathcal{K}_{ji} \longleftarrow \mathcal{K}_{ji} + k^\ell_{rs}$
      Fin pour
  Fin pour

On a ainsi construit complètement la matrice de rigidité $\mathcal{K}$. Il reste à savoir comment calculer

$$k^\ell_{r,s} = \int_{\ell e} \theta(\phi_s,\phi_r)(x)dx.$$

Ce calcul s'effectue sur l'élément de référence, et non sur les éléments K_ℓ. On calcule ensuite la valeur de $k^\ell_{r,s}$ par des changements de variable à l'aide de la transformation F_ℓ (voir Figure 4.4 page 160). Notons :

$$F_\ell(\bar{x},\bar{y}) = (x,y) = (a^\ell_0 + a^\ell_1\bar{x} + a^\ell_2\bar{y}, b^\ell_0 + b^\ell_1\bar{x} + b^\ell_2 y) \tag{4.4.19}$$

Notons que les coefficients a^ℓ_i et b^ℓ_i sont déterminés à partir des la connaissances des coordonnées $(x(i),y(i))$ où $i = ng(\ell,r)$. En effet, on peut déduire les coordonnées locales $x(r),y(r),r = 1,N_\ell$, des noeuds de l'élément ℓ, à partir des coordonnées globales des noeuds $(x(i),y(i))$, et du tableau $ng(\ell,r) = i$. Sur l'élément courant K_ℓ, le terme élémentaire $k^\ell_{r,s}$ s'écrit donc

$$k^\ell_{r,s} = \int_\ell \theta(\phi_s(x,y),\phi_r(x,y))dxdy.$$

Or, $(x,y) = F_\ell(\bar{x},\bar{y})$; donc par changement de variables , on a :

$$k^\ell_{r,s} = \int_{\bar{e}} \theta(\phi_s \circ F_\ell(\bar{x},\bar{y}),\phi_r \circ F_\ell(\bar{x},\bar{y}))\, Jac_{\bar{x},\bar{y}}(F_\ell)\, d\bar{x}d\bar{y}$$

ou $Jac_{\bar{x},\bar{y}}(F_\ell)$ désigne le Jacobien de F_ℓ en $(\bar{x},\bar{y})$. Or, $\phi_s \circ F_\ell = \bar{\phi}_s$, et, puisque F_ℓ est définie par (4.4.19), on a :

$$Jac(F_\ell) = Det(DF_\ell) = \begin{vmatrix} a^\ell_1 & b^\ell_1 \\ a^\ell_2 & b^\ell_2 \end{vmatrix} = \left|a^\ell_1 b^\ell_2 - a^\ell_2 b^\ell_1\right|$$

donc $k^\ell_{r,s} = Jac(F_\ell)\bar{k}_{r,s}$, où

$$\bar{k}_{r,s} = \int_{\bar{e}} \theta(\bar{\phi}_s(\bar{x},\bar{y}),\bar{\phi}_r(\bar{x},\bar{y}))d\bar{x}d\bar{y}$$

Etudions maintenant ce qu'on obtient pour $\bar{k}_{r,s}$ dans le cas du problème modèle (4.4.11), on a :

$$\bar{k}_{r,s} = \int_{\bar{\ell}} \left[p(\bar{x},\bar{y})\nabla\bar{\phi}_s(\bar{x},\bar{y})\nabla\bar{\phi}_r(\bar{x},\bar{y}) + q(\bar{x},\bar{y})\bar{\phi}_s(\bar{x},\bar{y})\bar{\phi}_r(\bar{x},\bar{y})\right] d\bar{x}d\bar{y}.$$

Les fonctions de base $\bar{\phi}_s$ et $\bar{\phi}_r$ sont connues ; on peut donc calculer $\bar{k}_{r,s}$ explicitement si p et q sont faciles à intégrer. Si les fonctions p et q ou les fonctions de base $\bar{\phi}$, sont plus compliquées, on calcule $\bar{k}_{r,s}$ en effectuant une intégration numérique. Rappelons que le principe d'une intégration numérique est d'approcher l'intégrale d 'une fonction continue donnée ψ,

$$I = \int_{\bar{\ell}} \psi(\bar{x},\bar{y})d\bar{x}d\bar{y}, \text{ par } \widetilde{I} = \sum_{i=1}^{NP_I} \omega_i(P_i)\psi(P_i),$$

où NP_I est le nombre de points d'intégration, notés P_i, qu'on appelle souvent points d'intégration de Gauss, et les coefficients ω_i sont les poids associés. Notons que les points P_i et les poids ω_i sont indépendants de ψ. Prenons par exemple, dans le cas unidimensionnel, $\bar{K} = [0,1]$, $p_1 = 0$, $p_2 = 1$, et $\omega_1 = \omega_2 = \frac{1}{2}$. On approche alors

$$I = \int_0^1 \psi(x)dx \text{ par } \widetilde{I} = \frac{1}{2}(\psi(0) + \psi(1)).$$

C'est la formule (bien connue) des trapèzes. Notons que dans le cadre d'une méthode, il est nécessaire de s'assurer que la méthode d'intégration numérique choisie soit suffisamment précise pour que :

1. le système $\mathcal{K}\alpha = \mathcal{G}(N \times N)$ reste inversible,
2. l'ordre de convergence de la méthode reste le même.

Examinons maintenant des éléments en deux dimensions d'espace.

1. Elément fini P_1 sur triangle Prenons $NP_I = 1$ (on a donc un seul point de Gauss), choisissons $p_1 = \left(\frac{1}{3},\frac{1}{3}\right)$, le centre de gravité du triangle $\bar{K}$, et $\omega_1 = 1$. On approche alors

$$I = \int_{\bar{\ell}} \psi(\bar{x})d\bar{x} \text{ par } \psi(p_1).$$

On vérifiera que cette intégration numérique est exacte pour les polynômes d'ordre 1 (exercice 46 page 190).

2. P_2 sur triangles. On prend maintenant $NP_I = 3$, et on choisit comme points de Gauss :

$$p_1 = \left(\frac{1}{2},0\right), p_2 = \left(\frac{1}{2},\frac{1}{2}\right), p_3 = \left(0,\frac{1}{2}\right)$$

et les poids d'intégration $\omega_1 = \omega_2 = \omega_3 = \frac{1}{6}$. On peut montrer que cette intégration numérique est exacte pour les polynômes d'ordre 2 (voir exercice 46 page 190).

Remarquons que, lors de l'intégration numérique du terme élémentaire

$$k^{\ell}_{r,s} = \int_{\bar{\ell}} \left[p(\bar{x},\bar{y})(F_\ell(\bar{x},\bar{y}))\nabla\bar{\phi}_r(\bar{x},\bar{y}) \cdot \nabla\bar{\phi}_s(\bar{x},\bar{y}) + q(\bar{x},\bar{y})(F_\ell(\bar{x},\bar{y}))\bar{\phi}_r(\bar{x},\bar{y})\bar{\phi}_s\right] d\bar{x}d\bar{y},$$

on approche $k^{\ell}_{r,s}$ par

$$\bar{k}_{r,s} \simeq \sum_{i=1}^{NP_I} \omega_i \left[p(F_\ell(P_i))\nabla\bar{\phi}_r(P_i) \cdot \nabla\bar{\phi}_s(P_i) + q(F_\ell(P_i))\bar{\phi}_r(P_i)\bar{\phi}_s(P_i)\right].$$

Les valeurs $\nabla\bar{\phi}_r(P_i)$, $\nabla\bar{\phi}_s(P_i)$, $\bar{\phi}_r(P_i)$ et $\bar{\phi}_s(P_i)$ sont calculées une fois pour toutes, et dans la boucle sur ℓ, il ne reste donc plus qu'à évaluer les fonctions p et q aux points $F_\ell(P_i)$. Donnons maintenant un résumé de la mise en oeuvre de la procédure d'intégration numérique (indépendante de ℓ). Les données de la procédure sont :

- les coefficients ω_i,$i = 1,\ldots,NP_I$,
- les coordonnées $(xpg(i),ypg(i))$, $i = 1,\ldots,NP_I$ des points de Gauss,
- les valeurs de ϕ_r,$\frac{\partial\phi}{\partial x}$ et $\frac{\partial\phi_r}{\partial y}$ aux points de Gauss, notées $\phi(r,i)$,$\phi_x(r,i)$ et $\phi_y(r,i)$, $r = 1\ldots N_\ell$, $i = 1,\ldots,NP_I$.

Pour ℓ donné, on cherche à calculer :

$$I = \int_{\bar{K}} p(F_\ell(\bar{x},\bar{y}))\frac{\partial\phi_r}{\partial\bar{x}}(\bar{x},\bar{y})\frac{\partial\phi_s}{\partial\bar{y}}(\bar{x},y)d\bar{x}d\bar{y} + \int_e q(F_\ell(\bar{x},\bar{y}))\phi_r(\bar{x},y)d\bar{x}d\bar{y}\phi_s(\bar{x},\bar{y}).$$

On propose l'algorithme suivant :

Initialisation : $I \longleftarrow 0$
Pour $i = 1$ à NP_I, faire :
$p_i = p(F_e(P_i))$
$q_i = q(F_e(P_i))$
$I \longleftarrow I + \omega_i(p_i\phi_x(r,i)\phi_y(s,i) + q_i\phi(r,i)\phi(s,i))$
Fin pour

On procède de même pour le calcul du second membre

$$T_\Omega(\phi_i) = \int_\Omega f(x,y)\phi_i'x,y)dxdy = \sum_{\ell=1}^{L} g_\ell, \text{ où } g_\ell = \int_{\ell_e} f(x,y)\phi_i(x,y)dxdy.$$

L'algorithme sécrit :

Initialisation de $\mathcal{G}$ à 0 : $\mathcal{G}_i \longleftarrow 0 \qquad i = 1$ à M
Pour $\ell = 1$ à L
Pour $r = 1$ à N_ℓ
Calcul de $g^r_\ell = \int_{\ell_e} f(x,y)\phi_r(x,y)dxdy$
$i = ng(\ell,r)$

$\mathcal{G}_i \longleftarrow \mathcal{G}_i + g_\ell^r$

Fin pour

Fin pour

Il reste le calcul de g_ℓ^r qui se ramène au calcul de l'élément de référence par changement de variable. On a :

$$g_\ell^r = \int_{K_\ell} f(x,y)\phi_r(x,y)dxdy = \int_{\bar{K}} f \circ F_\ell(\bar{x},\bar{y})\bar{\phi}_r(\bar{x},\bar{y})Jac_{\bar{x},\bar{y}}(F_\ell)d\bar{x}d\bar{y}.$$

L'intégration numérique est identique à celle effectuée pour $\bar{k}_{r,s}$.

4.4.4 Calcul de a_{Γ_1} et T_{Γ_1} (contributions des arêtes de bord "Fourier".

Détaillons maintenant le calcul des contributions des bords où s'applique la condition de Fourier, c'est à dire $a_{\Gamma_1}(\phi_i,\phi_j)$ $i = 1,\dots M, j = 1,\dots,M$ et $T_{\Gamma_1}(\phi_i)$ $i = 1,\dots,M$. Par définition,

$$a_{\Gamma_1}(\phi_i,\phi_j) = \int_{\Gamma_1} p(x)\nabla\phi_i(x)\cdot\nabla\phi_j(x)dx + \int_{\Gamma_1} q(x)\phi_i(x)\phi_j(x)dx.$$

Notons que $a_{\Gamma_1}(\phi_i,\phi_j) = 0$ si ϕ_i et ϕ_j sont associées à des noeuds S_i, S_j de d'un élément sans arête commune avec les arêtes de la frontière. Soit $L1$ le nombre d'arêtes ϵ_k,$k = 1,\dots,L1$ du maillage incluses dans Γ_1. Rappelons que les noeuds soumis aux conditions de Fourier sont repertoriés dans un tableau CF, de dimensions ($L1$,2), qui donne les informations suivantes

1. $CF(k,1)$ contient le numéro ℓ de l'élément K_ℓ auquel appartient l'arête ϵ_k.
2. $CF(k,2)$ contient le premier numéro des noeuds de l'arête ϵ_k dans l'élément K_ℓ. On suppose que la numérotation des noeuds locaux a été effectuée de manière "adroite", par exemple dans le sens trigonométrique. Dans ce cas, $CF(k,2)$ détermine tous les noeuds de l'arête ϵ_k dans l'ordre, puisqu'on connait le nombre de noeuds par arête et le sens de numérotation des noeuds. Donnons des exemples pour trois cas différents, représentés sur la figure 4.7.
   (a) Dans le premier cas (à droite sur la figure), qui représente un élément fini $P1$, on a $CF(k,2) = 3$ et le noeud suivant sur l'arête est 1.
   (b) Dans le second cas (au centre sur la figure), qui représente un élément fini $P2$, on a $CF(k,2) = 3$ et les noeuds suivants sur l'arête sont 4 et 5.
   (c) Enfin dans l'élément $P1$ "de coin" représenté à gauche sur la figure, on a $CF(k,1) = \ell$, $CF(k',1) = \ell$, $CF(k,2) = 1$, $CF(k',2) = 2$.

Pour $k = 1,\dots,L1$, on note $\hat{S}_k$ l'ensemble des noeuds locaux de ϵ_k, donnés par $CF(k,2)$ en appliquant la règle *ad hoc* (par exemple le sens trigonomtrique). On peut alors définir :

$$S_k = \{(r,s) \in (\hat{S}_k)^2 / r < s\}$$

L'algorithme de prise en compte des conditions de Fourier s'écrit alors :

Pour $k = 1 \dots L1$

$\ell = CF(k,1)$.

Pour chaque $(r,s) \in S_k$ faire

calcul de $I_{rs}^\ell = \int_{C_k} p(x)\sigma(x)\phi_r^\ell(x)\phi_s^\ell(x)dx$ (éventuellement avec intégration numérique)

$i = ng(\ell,r)$

$j = ng(\ell,s)$

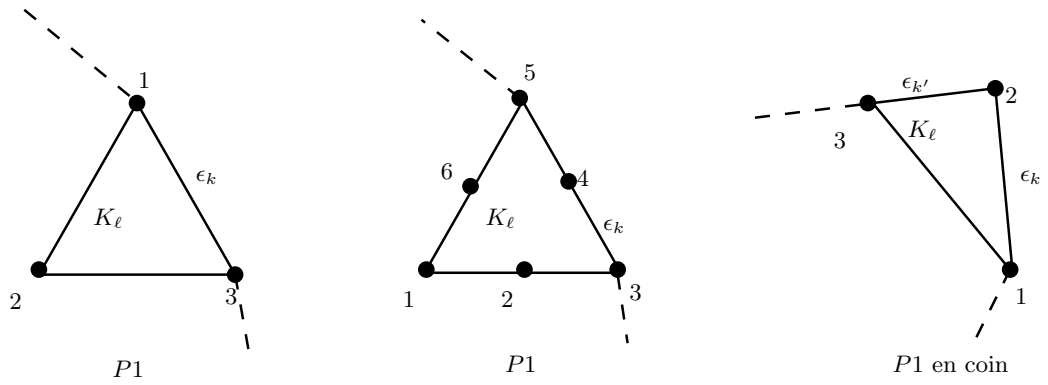

FIG. 4.7 – *Exemples de numérotation d'arête du bord*

si $j \leq i$
$\mathcal{K}ij \longleftarrow \mathcal{K}ij + I^\ell_{rs}$
sinon
$\mathcal{K}ij \longleftarrow \mathcal{K}ji + I^\ell_{rs}$
Fin si
Fin pour
Fin pour

Le calcul de I^ℓ_{rs} s'effectue sur l'élément de référence (avec éventuellement intégration numérique). De même, on a une procédure similaire pour le calcul de $T_{\Gamma_1} = \int_{\Gamma_1} p(x)g_1(x)v(x)d\gamma(x)$.

$$\mathcal{G}_i \longleftarrow \mathcal{G}_i + \int_{\Gamma_2} p(x)g_1(x)\phi_i(x)d\gamma(x)$$

4.4.5 Prise en compte des noeuds liés dans le second membre

Aprés les calculs précédents, on a maintenant dans $\mathcal{G}_i$:

$$\mathcal{G}_i = \int_\Omega f(x)\phi_i(x)dx + \int_{\Gamma_1} p(x)g_1(x)\phi_i(x)d\gamma(x)$$

Il faut maintenant retirer du second membre, les combinaisons venant des noeuds liés :

$$\mathcal{G}_i \longleftarrow \mathcal{G}_i - \sum_{j\in J_0} g_0(S_j)a(\phi_j,\phi_i)$$

où J_0 est l'ensemble des indices des noeuds liés. On utilise pour cela le tableau CD qui donne les conditions, de Dirichlet, de dimension M_0 où $M_0 = cardJ_0$. Pour $i_0 = 1,\ldots,M_0$, $CD(i_0) = j_0 \in J_0$ est le numéro du noeud lié dans la numérotation globale. La procédure est donc la suivante.

Pour $i_0 = 1,\ldots,M_0$, faire
$j = CD(i_0)$
$a = g_0(S_j)$
si $(i \leq j)$ $\mathcal{G}_i \longleftarrow \mathcal{G}_i - a\mathcal{K}_{ij}$ sinon $\mathcal{G}_i \longleftarrow \mathcal{G}_i - a\mathcal{K}_{ji}$ sinon Fin si Fin pour

4.4.6 Stockage de la matrice $\mathcal{K}$

Remarquons que la matrice $\mathcal{K}$ est creuse (et même très creuse), en effet $a(\phi_j,\phi_i) = 0$ dès que

$$supp(\phi_i) \cap supp(\phi_j) = \phi$$

Examinons une possibilité de stockage de la matrice $\mathcal{K}$. Soit NK le nombre d'éléments non nuls de la matrice $\mathcal{K}$ On peut stocker la matrice dans un seul tableau $KMAT$ en mettant bout à bout les coefficients non nuls de la première ligne, puis ceux de la deuxième ligne, etc... jusqu'à ceux de la dernieère ligne. Pour repérer les éléments de $\mathcal{K}$ dans le tableau $KMAT$, on a alors besoin de pointeurs. Le premier pointeur, nommé, IC est de dimension NK. La valeur de $IC(k)$ est le numéro de la colonne de $K(k)$. On introduit alors le pointeur $IL(\ell)$,$\ell = 1,\ldots,NL$, où NL est le nombre de lignes, où $IL(\ell)$ est l'indice dans $KMAT$ du début de la ℓ-ième ligne. L'identification entre $KMAT$ et $\mathcal{K}$ se fait alors par la procédure suivante :

Pour $k = 1 \ldots NK$
  si $IL(m) \leq k < IL(m+1)$ alors
  $KMAT(k) = \mathcal{K}_{m,IC(k)}$
  Fin si
Fin pour

La matrice $\mathcal{K}$ est symétrique définie positive, on peut donc utiliser une méthode de type gradient conjugué préconditionné (voir cours de Licence). Notons que la structure de la matrice dépend de la numérotation des noeuds. Il est donc important d'utiliser des algorithmes performants de maillage et de numérotation.

4.5 Eléments finis isoparamétriques

Dans le cas ou Ω est polygonal, si on utilise des éléments finis de type P_2, les noeuds de la frontière sont effectivement sur la frontière même si on les calcule à partir de l'élément fini de référence. Par contre, si le bord est courbe, ce n'est plus vrai. L'utilisation d'éléments finis "isoparamétriques" va permettre de faire en sorte que tous les noeuds frontières soient effectivement sur le bord, comme sur la figure 4.8. Pour obtenir une transformation isoparamétrique, on définit

$$\begin{aligned} F_\ell : K \quad &\to K_\ell \\ (\bar{x},\bar{y}) \quad &\mapsto (x,y) \end{aligned}$$

à partir des fonctions de base de l'élément fini de référence :

$$x = \sum_{r=1}^{N_\ell} x_r \bar{\phi}_r(\bar{x},\bar{y}),\ y = \sum_{r=1}^{N_\ell} y_r \bar{\phi}_r(\bar{x},\bar{y}),$$

où N_ℓ est le nombre de noeuds de l'élément et (x_r,y_r) sont les coordonnées du r-ième noeud de K_ℓ. Remarquons que la transformation F_ℓ isoparamétrique P_1 est identique à celle des éléments finis classiques. Par contre, la transformation isoparamétrique P_2 n'est plus affine, alors qu'elle l'est en éléments finis classiques. Notons que les fonctions de base locales vérifient toujours

$$\phi_r^\ell \circ F_\ell = \phi_r, \forall \ell = 1,\ldots,L, \quad \forall r = 1,\ldots,N_\ell.$$

On peut alors se poser le problème de l'inversibilité de F_ℓ. On ne peut pas malheureusement démontrer que F_ℓ est inversible dans tous les cas, toutefois, cela s'avère être le cas dans la plupart des cas pratiques. L'intérêt de la transformation isoparamétrique est de pouvoir traiter les bords courbes, ainsi que les éléments finis $Q1$ sur quadrilatères. Notons que le calcul de ϕ_r^ℓ est toujours inutile, car on se ramène encore à l'élément de référence.

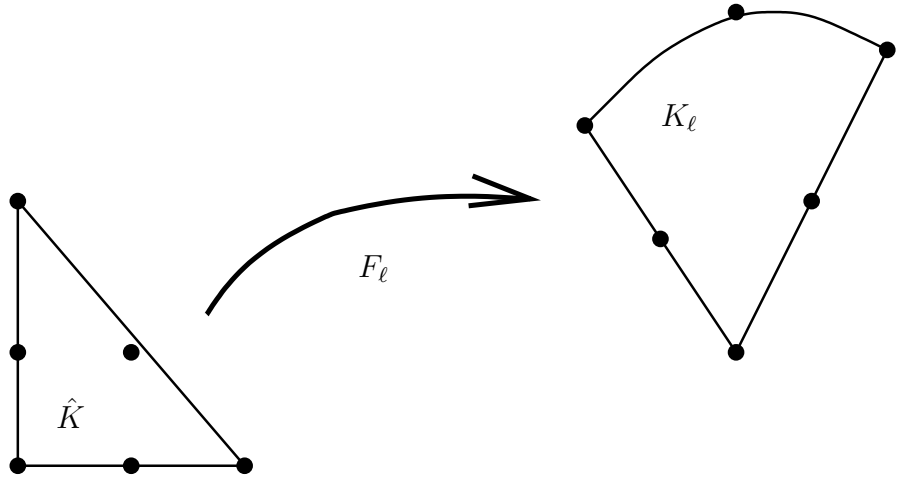

FIG. 4.8 – *Transformation isoparamétrique*

4.6 Analyse de l'erreur pour l'élément fini P1 en une dimension d'espace

4.6.1 Erreur de discrétisation et erreur d'interpolation

On considère toujours le problème modèle (4.4.11) page 170 sur lequel on a étudié la mise en oeuvre de la méthode des éléments finis. On rappelle que la formulation faible de ce problème est donnée en (4.4.14) page 171, et que, sous les hypothèses (4.4.12) page 170, le problème (4.4.14) admet une unique solution $\widetilde{u} \in H = H^1_{\Gamma_0 n} = \{u \in H^1(\Omega); u = 0 \text{ sur } \Gamma_0\}$. La méthode d'approximation variationnelle du problème (4.4.14) consiste à chercher $\widetilde{u}_N \in H_N = Vect\{\phi_1, \ldots, \phi_N\}$ solution de (4.4.16) page 171, où les fonctions $\phi_1, \ldots, \phi_N$ sont les fonctions de base éléments finis associés aux noeuds $x_1, \ldots, x_N$. Comme les hypothèses (3.2.22) page 124 sont vérifiées, l'estimation (3.2.30) page 127 entre $\widetilde{u}$ solution de (4.4.14) et $\widetilde{u}^{(N)}$ solution de (4.4.16) est donc vérifiée. On a donc :

$$\|\widetilde{u} - \widetilde{u}_N\|_{H^1} \leq \sqrt{\frac{M}{\alpha}} d(\widetilde{u}, H_N),$$

où M(resp.α) est la constante de continuité (resp. de coercivité) de a. Comme $u = u_0 + \widetilde{u}$, on a, en posant $c = \sqrt{\frac{M}{\alpha}}$,

$$\|u - u_N\| \leq C\|u - w\| \forall w \in H_N, \tag{4.6.20}$$

où $u_N = \widetilde{u}_N + u_0$. Notons que dans l'implantation pratique de la méthode d'éléments finis, lorsqu'on calcule $T(v) = T(v) - a(u_0, v)$, on remplace u_0 par $u_{0,N} \in H_N$, donc on commet une légère erreur sur T. De plus, on calcule $a(\phi_i, \phi_j)$ à l'aide d'intégrations numériques : l'inégalité (4.6.20) n'est donc vérifiée en pratique que de manière approchée. On supposera cependant, dans la suite de ce paragraphe, que les erreurs commises sont négligeables et que l'inégalité (4.6.20) est bien vérifiée. De la même manière qu'on a défini l'interpolée sur un élément K, (voir définition 4.3 page 158, on va maintenant définir l'interpolée sur $H^1(\Omega)$ tout entier, de manière à établir une majoration de l'erreur de discrétisation grâce à (4.6.20).

Définition 4.16 (Interpolée dans H_N) . *Soit $u \in H^1(\Omega)$ et $H_N = Vect\{\phi_1, \ldots, \phi_N\}$ où les fonctions*

$\phi_1 \dots \phi_N$ sont des fonctions de base éléments finis associées aux noeuds $S_1 \dots S_N$ d'un maillage éléments finis de Ω. Alors on définit l'interpolée de u dans H_N, $u_I \in H_N$ par :

$$u_I = \sum_{i=1}^{N} u(S_i)\phi_i.$$

Comme $u_I \in H_N$, on peut prendre $W = u_I$ dans (4.6.20), ce qui fournit un majorant de l'erreur de discrétisation :

$$\|u - u_N\|_{H^1} \leq C\|u - u_I\|_{H^1}$$

On appelle erreur d'interpolation le terme $\|u - u_I\|_{H^1}$

4.6.2 Etude de l'erreur d'interpolation en dimension 1

Soit $\Omega =]0,1[$, on considère un maillage classique, défini par les $N+2$ points $(x_i)_{i=0\dots N+1}$, avec $x_0 = 0$ et $x_{N+1} = 1$, et on note

$$h_i = x_{i+1} - x_i, i = 0, \dots, N+1, \text{ et } h = \max\{|h_i|,\ i = 0, \dots, N+1\}$$

On va montrer que si $u \in H^2(]0,1[)$, alors on peut obtenir une majoration de l'erreur d'interpolation $\|u - u_I\|_{H^1}$. On admettra le lemme suivant(voir exercice 26 page 137) :

Lemme 4.17 *Si $u \in H^1(]0,1[)$ alors u est continue.*

En particulier, on a donc $H^2(]0,1[) \subset C^1([0,1])$. Remarquons que ce résultat est lié à la dimension 1, voir injection de Sobolev, cours d'analyse fonctionnelle ou [1]. On va démontrer le résultat suivant sur l'erreur d'interpolation.

Théorème 4.18 (majoration de l'erreur d'interpolation, dimension 1) *Soit $u \in H^2(]0,1[)$, et soit u_I son interpolée sur $H_N = Vect\{\phi_i,\ i = 1, \dots, N\}$, où ϕ_i désigne la i-ème fonction de base élément fini $P1$ associée au noeud x_i d'un maillage élément fini de $]0,1[$. Alors il existe $c \in \mathbb{R}$ ne dépendant que de u, tel que*

$$\|u - u_I\|_{H^1} \leq Ch. \tag{4.6.21}$$

Démonstration : On veut estimer

$$\|u - u_I\|_{H^1}^2 = |u - u_I|_0^2 + |u - u_I|_1^2$$

où $|v|_0 = \|v\|_{L^2}$ et $|v|_1 = \|Dv\|_{L^2}$. Calculons $|u - u_I|_1^2$:

$$|u - u_I|_1^2 = \int_0^1 |u' - u_I'|^2 dx = \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} |u'(x) - u_I'(x)|^2 dx.$$

Or pour $x \in]x_i, x_{i+1}[$ on a

$$u_I' = \frac{u(x_{i+1}) - u(x_i)}{h_i} = u'(\xi_i),$$

pour un certain $\xi_i \in]x_i, x_{i+1}[$. On a donc :

$$\int_{x_i}^{x_{i+1}} |u'(x) - u_I'(x)|^2 dx = \int_{x_i}^{x_{i+1}} |u'(x) - u'(\xi_i)|^2 dx.$$

On en déduit que :

$$\int_{x_i}^{x_{i+1}} |u'(x) - u'_I(x)|^2 dx = \int_{x_i}^{x_{i+1}} |\int_{\xi_i}^{x} u''(t)dt|^2 dx,$$

et donc, par l'inégalité de Cauchy-Schwarz,

$$\begin{aligned}\int_{x_i}^{x_{i+1}} |u'(x) - u'_I(x)|^2 dx &\leq \int_{x_i}^{x_{i+1}} \int_{\xi_i}^{x} |u''(t)|^2 dt |x - \xi_i| dx \\ &\leq h_i \int_{x_i}^{x_{i+1}} \left(\int_{\xi_i}^{x} |u''(t)|^2 dt \right) dx,\end{aligned}$$

car $|x - \xi_i| \leq h_i$. En réappliquant l'inégalité de Cauchy-Schwarz, on obtient :

$$\int_{x_i}^{x_{i+1}} |u'(x) - u'_I(x)|^2 dx \leq h_i^2 \int_{x_i}^{x_{i+1}} |u''(t)|^2 dt$$

En sommant sur i, ceci entraine :

$$|u - u_I|_1^2 \leq h^2 \int_0^1 |u''(t)|^2 dt. \tag{4.6.22}$$

Il reste maintenant à majorer $|u - u_I|_0^2 = \int_0^1 |u - u_I|^2 dx$. Pour $x \in [x_i, x_{i+1}]$

$$|u(x) - u_I(x)|^2 = \left(\int_{x_i}^{x} (u'(t) - u'_I(t)) dt \right)^2.$$

Par l'inégalité de Cauchy-Schwarz, on a donc :

$$|u(x) - u_I(x)|^2 \leq \int_{x_i}^{x} (u'(t) - u'_I(t))^2 dt \underbrace{|x - x_i|}_{\leq h_i}$$

Par des calculs similaires aux précédents, on obtient donc :

$$\begin{aligned}|u(x) - u_I(x)|^2 &\leq \int_{x_i}^{x} h_i \left(\int_{x_i}^{x_{i+1}} |u''(t)|^2 dt \right) dx h_i \\ &\leq h_i^3 \int_{x_i}^{x_{i+1}} |u''(t)|^2 dt.\end{aligned}$$

En intégrant sur $[x_i, x_{i+1}]$, il vient :

$$\int_{x_i}^{x_{i+1}} |u(x) - u_I(x)|^2 dx \leq h_i^4 \int_{x_i}^{x_{i+1}} (u''(t))^2 dt,$$

et en sommant sur $i = 1, \ldots, N$:

$$\int_0^1 (u(x) - u_I(x))^2 dx \leq h^4 \int_0^1 (u''(t))^2 dt.$$

On a donc :

$$|u - u_I|_0 \leq h^2 |u|_2$$

ce qui entraine, avec (4.6.22) :

$$\begin{aligned}\|u - u_I\|^2 &\leq h^4|u|_2^2 + h^2|u|_2^2 \\ &\leq (1+h^2)h^2|u|_2^2\end{aligned}$$

On en déduit le résultat annoncé. ∎

On en déduit le résultat d'estimation d'erreur suivant:

Corollaire 4.19 (Estimation d'erreur, P1, dimension 1) *Soit Ω un ouvert polygonal convexe de $\mathbb{R}^d$, $d \geq 1$; soit $f \in L^2(\Omega)$ et $u \in H_0^1(\Omega)$ l'unique solution du problème*

$$\left\{ \begin{array}{l} u \in H_0^1(\Omega) \\ a(u,v) = \displaystyle\int_\Omega \nabla u(x)\nabla v(x)dx = \int f(x)v(x)dx, \end{array} \right. ,$$

et $u_\mathcal{T}$ L'approximation éléments finis P1 obtenue sur un maillage admissible $\mathcal{T}$ de pas $h_\mathcal{T} = \max_{i=1,\ldots,N}\{|h_i|\}$. Alors il existe $C \in \mathbb{R}$ ne dépendant que de Ω et f tel que $\|u - u_\mathcal{T}\| < Ch$.

Ces résultats se généralisent au cas de plusieurs dimensions d'espace (voir Ciarlet), sous des conditions géométriques sur le maillage, nécessaires pour obtenir le résultat d'interpolation. Par exemple pour un maillage triangulaire en deux dimensions d'espace, intervient une condition d'angle : on demande que la famille de maillages considérée soit telle qu'il existe $\beta > 0$ tel que $\beta \leq \theta \leq \pi - \beta$ pour tout angle θ du maillage.

Remarque 4.20 (Sur les techniques d'estimations d'erreur pour les différentes méthodes de discrétisation
Lorsqu'on a voulu montrer des estimations d'erreur pour la méthode des différences finies, on a utilisé le principe de possitivité, la consistance et la stabilité en norme L^∞. En volumes finis et éléments finis, on n'utilise pas le principe de positivité. En volumes finis, la stabilité en norme L^2 est obtenue grâce à l'inégalité de Poincaré discrète, et la consistance est en fait la consistance des flux. Notons qu'en volumes finis on se sert aussi de la conservativité des flux numériques pour la preuve de convergence. Enfin, en éléments finis, la stabilité est obtenue grâce à la coercivité de la forme bilinéaire, et la consistance provient du contrôle de l'erreur d'interpolation.
Même si le principe de positivité n'est pas explicitement utilisé pour les preuves de convergence des éléments finis et volumes finis, il est toutefois intéressant de voir à quelles conditions ce principe est respecté, car il est parfois très important en pratique.
Reprenons d'abord le cas du schéma volumes finis sur un maillage $\mathcal{T}$ admissible pour la discrétisation de l'équation (3.1.1).

$$\left\{ \begin{array}{ll} -\Delta u = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega. \end{array} \right.$$

Rappelons que le schéma volumes finis s'écrit :

$$\sum_{K\in\mathcal{C}} \left(\sum_{\sigma\in\xi_K\cap\xi_{int}} \tau_{K,L}(u_K - u_L) + \sum_{\sigma\in\xi_K\cap\xi_{ext}} \tau_{K,\sigma}u_K \right) = |K|f_K, \tag{4.6.23}$$

avec

$$\tau_{K,L} = \frac{|K|L|}{d(x_K,x_L)} \text{ et } \tau_{K,\sigma} = \frac{|\sigma|}{d(x_K,\partial\Omega)},$$

où $|K|$, (resp. $|\sigma|$) désigne la mesure de Lebesque en dimension d (resp. $d-1$) de K (resp. σ).

Notons que les coefficients $\tau_{K,L}$ et $\tau_{K,\sigma}$ sont positifs, grâce au fait que le maillage est admissible (et donc $\vec{X_K X_L} = d(X_K,X_L)\vec{n_{KL}}$, où $\vec{X_K X_L}$ désigne le vecteur d'extrémités X_K et X_L et $\vec{n_{KL}}$ la normale unitaire à $K|L$ sortante de K.
Notons que le schéma (4.6.23) s'écrit comme une somme de termes d'échange entre les mailles K et L, avec des coefficients τ_{KL} positifs. C'est grâce à cette propriété que l'on montre facilement que le principe de positivité est vérifié. Considérons maintenant la méthode des éléments finis P1, pour la résolution du problème (3.1.1) sur maillage triangulaire. On sait (voir par exemple Ciarlet) que si le maillage satisfait la condition faible de Delaunay (qui stipule que la somme de deux angles opposés à une même arête doit être inférieure à π), alors le principe du maximum est vérifiée. Ce résultat peut se retrouver en écrivant le schéma éléments finis sous la forme d'un schéma volumes finis.

4.6.3 Super convergence

On considère ici un ouvert Ω polygonal convexe de $\mathbb{R}^d$, $d \geq 1$, et on suppose que $f \in L^2(\Omega)$. On s'intéresse à l'approximation par éléments finis $P1$ de la solution $u \in H_0^1(\Omega)$ du problème (3.1.5). On a vu dans le paragraphe précédent (corollaire 4.19) qu'on peut estimer l'erreur en norme L^2 entre la solution exacte u et la solution approchée par éléments finis $P1$; en effet, comme l'erreur d'interpolation est d'ordre h, on déduit une estimation sur l'erreur de discrétisation, également d'ordre h. En fait, si la solution u de (3.1.1) est dans H^2, il se produit un "petit miracle", car on peut montrer grâce à une technique astucieuse, dite"truc d'Aubin-Nitsche", que l'erreur de discrétisation en norme L^2 est en fait d'ordre 2.

Théorème 4.21 (Super convergence des éléments finis $P1$) *Soit Ω un ouvert polygonal convexe de $\mathbb{R}^d$,$d \geq 1$; soit $f \in L^2(\Omega)$, u solution de (3.1.5), $u_\mathcal{T}$ la solution apporchée obtenue par éléments finis P1, sur un maillage éléments finis $\mathcal{T}$. Soit*

$$h_\mathcal{T} = \max_{K \in \mathcal{T}} diam K.$$

Alors il existe $C \in \mathbb{R}$ ne dépendant que de Ω et f tel que :

$$\|u - u_\mathcal{T}\|_{H^1(\Omega)} \leq Ch \text{ et } \|u - u_\mathcal{T}\|_{L^2(\Omega)} \leq Ch^2.$$

Démonstration : Par le théorème de régularité 3.9 page 118, il existe $C_1 \in \mathbb{R}_+$ ne dépendant que de Ω tel que

$$\|u\|_{H^2(\Omega)} \leq C_1 \|f\|_{L^2(\Omega)}.$$

Grâce à ce résultat, on a obtenu (voir le théorème 4.19) qu'il existe C_2 ne dépendant que de Ω, β et tel que

$$\|u - u_\mathcal{T}\|_{H^1(\Omega)} \leq C_2 \|f\|_{L^2} h$$

Soit maintenant $e_\mathcal{T} = u - u_\mathcal{T}$ et $\varphi \in H_0^1(\Omega)$ vérifiant

$$\int_\Omega \nabla\varphi(x).\nabla\psi(x)dx = \int_\Omega e_\mathcal{T}(x)\psi(x)dx, \forall \psi \in H_0^1(\Omega). \tag{4.6.24}$$

On peut aussi dire que φ est la solution faible du problème

$$\begin{cases} -\Delta\varphi = e_\mathcal{T} \text{ dans } \Omega \\ \varphi = 0 \text{ sur } \partial\Omega. \end{cases}$$

Comme $e \in L^2(\Omega)$, par le théorème 3.9, il existe $C_3 \in \mathbb{R}_+$ ne dépendant que Ω tel que

$$\|\varphi\|_{H^2(\Omega)} \leq C_3 \|e_{\mathcal{T}}\|_{L^2(\Omega)}.$$

Or $\|e_{\mathcal{T}}\|^2_{L^2(\Omega)} = \displaystyle\int_\Omega e_{\mathcal{T}}(x) e_{\mathcal{T}}(x) dx = \int_\Omega \nabla\varphi(x).\nabla e(x) dx$, en prenant $\psi = e_{\mathcal{T}}$ dans (4.6.24).
Soit $\varphi_{\mathcal{T}}$ la solution approchée par éléments finis P1 du problème (4.6.24), c.à.d solution de :

$$\left\{ \begin{array}{l} \varphi_{\mathcal{T}} \in V_{\mathcal{T},0} = \{v \in C(\bar{\Omega}); v|_K \in P_1, \forall K \in \mathcal{T}, v|_{\partial\Omega} = 0\} \\ \displaystyle\int_\Omega \nabla\varphi_{\mathcal{T}}(x) \nabla v(x) dx = \int_\Omega e(x) v(x) dx, \forall v \in V_{\mathcal{T},0} \end{array} \right. \tag{4.6.25}$$

On sait que $u_{\mathcal{T}}$ vérifie :

$$\int_\Omega \nabla\varphi_{\mathcal{T}}(x).\nabla(u - u_{\mathcal{T}})(x) dx = 0;$$

on peut donc écrire que :

$$\|e_{\mathcal{T}}\|^2_{L^2(\Omega} = \int_\Omega \nabla(\varphi - \varphi_{\mathcal{T}})(x).\nabla(u - u_{\mathcal{T}})(x) dx \geq \|\varphi - \varphi_{\mathcal{T}}\|_{H^1(\Omega)} \|u - u_{\mathcal{T}}\|_{H^1(\Omega)}$$

D'après le théorème 4.19, on a :

$$\|\varphi - \varphi_{\mathcal{T}}\|_{H^1(\Omega)} \leq C_2 \|e\|_{L^2(\Omega)} h_{\mathcal{T}} \text{ et } \|u - u_{\mathcal{T}}\|_{H^1(\Omega)} \leq C_2 \|f\|_{L^2(\Omega)} h_{\mathcal{T}}.$$

On en déduit que :

$$\|e_{\mathcal{T}}\|_{L^2(\Omega)} \leq C_2^2 \|f\|_{L^2(\Omega)} h^2_{\mathcal{T}}.$$

Ce qui démontre le théorème.

4.6.4 Traitement des singularités

Les estimations d'erreur qu'on a obtenues au paragraphe précédent reposent sur la régularité H^2 de u. Que se passe-t-il si cette régularité n'est plus vérifiée ? Par exemple, si le domaine Ω possède un coin rentrant, on sait que dans ce cas, la solution u du problème (3.1.5) n'est plus dans $H^2(\Omega)$, mais dans un espace $H^{1+s}(\Omega)$, où s dépend de l'angle du coin rentrant. Considérons donc pour fixer les idées le problème

$$\left\{ \begin{array}{l} -\Delta u = f \text{ dans } \Omega, \text{ où } \Omega \text{ est un ouvert} \\ u = 0 \text{ sur } \partial\Omega \text{ polygônal avec un coin rentrant.} \end{array} \right.$$

Pour approcher correctement la singularité, une première technique consiste à raffiner le maillage dans le voisinage du coin. Cependant, une approche plus efficace, lorsque cela est possible, consiste à modifier l'espace d'approximation pour tenir compte de la singularité. En effet, dans le cas d'un polygône avec un coin rentrant par exemple, on sait trouver $\psi \in H^1_0(\Omega)$ (et $\psi \notin H^2(\Omega)$) telle que si u est solution de (3.1.5) avec $f \in L^2(\Omega)$, alors il existe un unique $\alpha \in \mathbb{R}$ tel que $u - \alpha\psi \in H^2(\Omega)$.
Plaçons nous dans le cas d'une approximation par éléments finis de Lagrange. Dans le cas où u est régulière, l'espace d'approximation est donc

$$V_{\mathcal{T}} = Vect\{\phi_i, i = 1, N_{\mathcal{T}}\},$$

où $N_{\mathcal{T}}$ est le nombre de noeuds internes du maillage $\mathcal{T}$ de Ω considéré et $(\phi_i)_{i=1,N_{\mathcal{T}}}$ la famille des fonctions de forme associées aux noeuds.

Dans le cas d'une singularité portée par la fonction ψ introduite ci-dessus, on modifie l'espace V et on prend maintenant :

$$V_{\mathcal{T}} = Vect\{\phi_i, i = 1, N_{\mathcal{T}}\} \oplus \mathbb{R}\psi$$

Notons que $V_{\mathcal{T}} \subset H_0^1(\Omega)$, car $\psi \in H_0^1(\Omega)$.
Examinons maintenant l'estimation d'erreur. Grâce au lemme de Céa, on a toujours

$$\|u - u_{\mathcal{T}}\|_{H^1} \leq \frac{M}{\alpha}\|u - w\|_{H^1(\Omega)}, \forall w \in V_{\mathcal{T}}.$$

On a donc également :

$$\|u - u_{\mathcal{T}}\|_{H^1} \leq \frac{M}{\alpha}\|u - \alpha\psi - w\|_{H^1(\Omega)}, \forall w \in V_{\mathcal{T}}.$$

puisque $\alpha\psi + w \in V_{\mathcal{T}}$. Or, $u - \alpha\psi = \widetilde{u} \in H^2(\Omega)$.
Donc $\|u - u_{\mathcal{T}}\|_{H^1} \leq \frac{M}{\alpha}\|\widetilde{u} - w\|_{H^1(\Omega)}, \forall w \in V_{\mathcal{T}}$.
Et grâce aux résultats d'interpolation qu'on a admis, si on note $\widetilde{u}_I$ l'interpolée de $\widetilde{u}$ dans $V_{\mathcal{T}}$, on a :

$$\|u - u_{\mathcal{T}}\|_{H^1(\Omega)} \leq \frac{M}{\alpha}\|\widetilde{u} - \widetilde{u}_I\|_{H^1(\Omega)} \leq \frac{M}{\alpha}C_2\|\widetilde{u}\|_{H^2(\Omega)}h.$$

On obtient donc encore une estimation d'erreur en h.
Examinons maintenant le système linéaire obtenu avec cette nouvelle approximation. On effectue un développement de Galerkin sur la base de $V_{\mathcal{T}}$. On pose

$$u_{\mathcal{T}} = \sum_{i=1,N_{\mathcal{T}}} u_i\phi_i + \gamma\psi.$$

Le problème discrétisé revient donc à chercher

$$\begin{cases} (u_s)_{i=1,N_{\mathcal{T}}} \subset \mathbb{R}^N \text{ et } \gamma \in \mathbb{R} \quad t.q. \\ \sum_{j=1,N_{\mathcal{T}}} u_j \int_\Omega \nabla\phi_j(x) \cdot \nabla\phi_i(x)dx + \gamma \int_\Omega \nabla\psi(x) \cdot \nabla\phi_i(x)dx = \int_\Omega f(x)\phi_i(x)dx, \forall i = 1, N_{\mathcal{T}} \\ \sum_{j=1,N_{\mathcal{T}}} u_j \int_\Omega \nabla\phi_i(x) \cdot \nabla\psi(x)dx + \gamma \int_\Omega \nabla\psi(x) \cdot \nabla\psi(x)dx = \int f(x)\psi(x)dx. \end{cases}$$

On obtient donc un système linéaire de $N_{\mathcal{T}} + 1$ équations à $N_{\mathcal{T}} + 1$ inconnues.

4.7 Exercices

Exercice 38 (Eléments finis P1 pour le problème de Dirichlet) *Corrigé en page 192*

Soit $f \in L^2(]0,1[)$. On s'intéresse au problème (3.4.38) qu'on rappelle ci–dessous, et dont on a étudié une formulation faible à l'exercice 3.6.

$$\begin{aligned} -u_{xx}(x) &= f(x),\ x \in]0,1[, \\ u(0) &= 0,\ u(1) = 0. \end{aligned}$$

Soient $N \in \mathbb{N}$, $h = 1/(N+1)$ et $x_i = ih$, pour $i = 0, \ldots, N+1$, et $K_i = [x_i, x_{i+1}]$, pour $i = 0, \ldots, N$. Soit $H_N = \{v \in C([0,1],\mathbb{R}) \text{ t.q. } v|_{K_i} \in P_1, i = 0, \ldots, N, \text{ et } v(0) = v(1) = 0\}$, où P_1 désigne l'ensemble des polynômes de degré inférieur ou égal à 1.

1. Montrer que $H_N \subset H_0^1$.

2. Pour $i = 1, \ldots, N$, on pose:

$$\begin{array}{l} \phi_i(x) = 1 - \dfrac{|x - x_i|}{h} \text{ si } x \in K_i \cup K_{i-1}, \\ \phi_i(x) = 0 \text{ sinon.} \end{array} \tag{4.7.26}$$

Montrer que $\phi_i \in H_N$ pour tout $i = 1, \ldots, N$ et que $H_N = Vect\{\phi_1, \ldots, \phi_N\}$.

3. Donner le système linéaire obtenu en remplaçant H par H_N dans la formulation faible de (3.4.38). Comparer avec le schéma obtenu par différences finies.

Exercice 39 (Conditions aux limites de Fourier et Neumann) *Corrigé en page 193*

Soit $f \in L^2(]0,1[)$. On s'intéresse au problème (3.4.40) qu'on rappelle:

$$\begin{array}{l} -u_{xx}(x) + u(x) = f(x),\ x \in]0,1[, \\ u'(0) - u(0) = 0,\ u'(1) = -1. \end{array} \tag{4.7.27}$$

On a étudié à l'exercice 39 l'existence et unicité d'une solution faible. On s'intéresse maintenant à la discrétisation.

1. Ecrire une discrétisation de (3.4.40) par différences finies pour un maillage non uniforme. Ecrire le système linéaire obtenu.

2. Ecrire une discrétisation de (3.4.40) par volumes finis de (3.4.40) pour un maillage non uniforme. Ecrire les système linéaire obtenu.

3. Ecrire une discrétisation par éléments finis conformes de type Lagrange P_1 de (3.4.40) pour un maillage non uniforme. Ecrire le système linéaire obtenu.

Exercice 40 (Conditions aux limites de Fourier et Neumann,bis) *Corrigé en page 197*

Soit $f \in L^2(]0,1[)$. On s'intéresse au problème (3.4.41) qu'on rappelle:

$$\left\{ \begin{array}{l} -u''(x) - u'(x) + u(x) = f(x),\ x \in]0,1[, \\ u(0) + u'(0) = 0,\ u(1) = 1 \end{array} \right.$$

1. Ecrire une discrétisation par éléments finis conformes de type Lagrange P_1 de (3.4.41) pour un maillage uniforme. Ecrire le système linéaire obtenu.

2. Ecrire une discrétisation par volumes finis centrés de (3.4.41) pour un maillage uniforme. Ecrire le système linéaire obtenu.

3. Ecrire une discrétisation par différences finies centrés de (3.4.41) pour un maillage uniforme. Ecrire le système linéaire obtenu.

4. Quel est l'ordre de convergence de chacune des méthodes étudiées aux questions précédentes?

Exercice 41 (Eléments finis pour un problème avec conditions mixtes) *Corrigé en page 200*

Soit $f \in L^2(]0,1[$. On s'intéresse ici au problème

$$\begin{array}{l} -\mathrm{div}(p(x)\nabla u(x)) + q(x)u(x) = f(x), x \in =]0,1[, \\ u(0) = 0, \\ u'(1) = 0, \end{array} \tag{4.7.28}$$

où: Notons que ce problème est un cas particulier du problème 3.4.44 page 139 étudié à l'exercice 34 page 139, en prenant: $\Omega =]0,1[$, $p \equiv 1$ et $q \equiv 1$, $\Gamma_0 = \{0\}$, $\Gamma_1 = \{1\}$, $g_0 \equiv 0$, $g_1 \equiv 0$ et $\sigma = 0$. On s'intéresse ici à la discrétisation du problème (4.7.28). Soient $N \in \mathbb{N}$, $h = 1\ (N+1)$ et $x_i = ih$, pour $i = 0, \ldots, N+1$, et $K_i = [x_i, x_{i+1}]$, pour $i = 0, \ldots, N$. On cherche une solution approchée de (3.4.44), notée u_h, en utilisant les éléments finis $(K_i, \{x_i, x_{i+1}\}, P_1)_{i=0}^N$. Déterminer l'espace d'approximation V_h.

3.1 Montrer que les fonctions base globales sont les fonctions Φ_i de [0,1] dans $\mathbb{R}$ définies par $\Phi_i(x) = (1 - \frac{|x-x_i|}{h})^+$, pour $i = 1, \ldots, N+1$.

3.2. Construire le système linéaire à résoudre et comparer avec les systèmes obtenus par différences finies et volumes finis.

3.3 A-t-on $u_h'(1) = 0$?

Exercice 42 (Eléments finis $Q1$) *Corrigé en page 203*

On considère le rectangle Ω de sommets $(-1,0),(2,0),(-1,1),(2,1)$. On s'intéresse à la discrétisation par éléments finis de l'espace fonctionnel $H^1(\Omega)$.
I. On choisit de découper Ω en deux éléments e_1 et e_2 définis par les quadrilatères de sommets respectifs $M_1(-1,1)$, $M_2(0,1)$, $M_5(1,0)$, $M_4(-1,0)$ et $M_2(0,1)$, $M_3(2,1)$, $M_6(2,0)$, $M_5(1,0)$.
On prend comme noeuds les points $M_1, \ldots, M_6$ et comme espace par élément l'ensemble des polynômes Q_1. On note $\Sigma_1 = \{M_4, M_5, M_2, M_1\}$ et $\Sigma_2 = \{M_5, M_6, M_3, M_2\}$.
On a donc construit la discrétisation $\{(e_1, \Sigma_1, Q_1), (e_2, \Sigma_2, Q_1)\}$.

I.1 Montrer que les éléments (e_1, Σ_1, Q_1) et (e_2, Σ_2, Q_1) sont des éléments finis de Lagrange.

I.2. Montrer que l'espace de dimension finie correspondant à cette discrétisation n'est pas inclus dans $H^1(\Omega)$ (construire une fonction de cet espace dont la dérivée distribution n'est pas dans L^2). Quelle est dans les hypothèses appelées en cours "cohérence globale" celle qui n'est pas vérifiée?

II. On fait le même choix des éléments et des noeuds que dans I. On introduit comme élément de référence e le carré de sommets $(\pm 1, \pm 1)$, Σ est l'ensemble des sommets de e et $P = Q_1$.

II.1. Quelles sont les fonctions de base locales de (e, Σ, P). On note ces fonctions $\Phi_1, \ldots, \Phi_4$.

II.2 A partir des fonctions $\Phi_1, \ldots, \Phi_4$, construire des bijections F_1 et F_2 de e dans e_1 et e_2. Les fonctions F_1 et F_2 sont elles affines?

II.3 On note $P_{e_i} = \{f : e_i \to \mathbb{R}, f \circ F_i|_e \in Q_1\}$, pour $i = 1,2$, où les F_i sont définies à la question précédente. Montrer que les éléments (e_1, Σ_1, P_{e_1}) et (e_2, Σ_2, P_{e_2}) sont des éléments finis de Lagrange et que l'espace vectoriel construit avec la discrétisation $\{(e_1, \Sigma_1, P_{e_1}), (e_2, \Sigma_2, P_{e_2})\}$ est inclus dans $H^1(\Omega)$ (i.e. vérifier la "cohérence globale" définie en cours). On pourra pour cela montrer que si $S = e_1 \cap e_2 = \{(x,y); x+y = 1\}$, alors $\{f|_S, f \in P_{e_i}\} = \{f : S \to \mathbb{R}; f(x,y) = a + by, a, b \in \mathbb{R}\}$.

Exercice 43 (Eléments affine–équivalents) *Corrigé en page 206*

Soit Ω un ouvert polygonal de $\mathbb{R}^2$, et $\mathcal{T}$ un maillage de Ω.
Soient $(\bar{K}, \bar{\Sigma}, \bar{P})$ et (K, Σ, P) deux éléments finis de Lagrange affine - équivalents. On suppose que les fonctions de base locales de $\bar{K}$ sont affines.

Montrer que toute fonction de P est affine.
En déduire que les fonctions de base locales de (K,Σ,P) affines.

Exercice 44 (Eléments finis $P2$ en une dimension d'espace) *Corrigé en page 206*

On veut résoudre numériquement le problème aux limites suivant

$$\begin{cases} u''(x) + u(x) = x^2, \quad 0 < x < 1 \\ u(0) = 0, \\ u'(1) = 1. \end{cases} \tag{4.7.29}$$

1. Donner une formulation faible du problème (4.7.29)
2. Démontrer que le problème (4.7.29) admet une unique solution.
3. On partage l'intervalle $]0,1[$ en N intervalles égaux et on approche la solution par une méthode d'éléments finis de degré 2. Ecrire le système qu'il faut résoudre.

Exercice 45 (Comparaison de schémas)

Soit $f \in L^2(]0,1[)$. On s'intéresse au problème suivant:

$$\begin{array}{l} -u''(x) - u'(x) + u(x) = f(x),\ x \in]0,1[, \\ u(0) + u'(0) = 0,\ u(1) = 1 \end{array} \tag{4.7.30}$$

1. Donner une formulation faible du problème de la forme

$$\begin{cases} \text{Trouver } u \in H^1(]0,1[); u(1) = 1, \\ a(u,v) = T(v), \forall v \in H. \end{cases} \tag{4.7.31}$$

où $H = \{v \in H^1(]0,1[); v(1) = 0\}$, a et T sont respectivement une forme bilinéaire sur $H^1(]0,1[)$ et une forme linéaire sur $H^1(]0,1[)$, à déterminer.

2. Y-a-t-il existence et unicité de solutions de cette formulation faible?

3. Ecrire une discrétisation par éléments finis conformes de type Lagrange P_1 de (4.7.30) pour un maillage uniforme. Ecrire le système linéaire obtenu.

4. Ecrire une discrétisation par volumes finis centrés de (4.7.30) pour un maillage uniforme. Ecrire le système linéaire obtenu.

5. Ecrire une discrétisation par différences finies centrés de (4.7.30) pour un maillage uniforme. Ecrire le système linéaire obtenu.

6. Ecrire une discrétisation par éléments finis $P2$ du problème pour un maillage uniforme.

7. Quel est l'ordre de convergence de chacune des méthodes étudiées aux questions précédentes?

Exercice 46 (Intégration numérique) *Corrigé en page 209*

1. Vérifier que l'intégration numérique à un point de Gauss, donné par le centre de gravité du triangle, sur l'élément fini P_1 sur triangle, est exacte pour les polynômes d'ordre 1.
2. Vérifier que l'intégration numérique à trois points de Gauss définis sur le trangle de rérérence par $p_1 = \left(\frac{1}{2},0\right)$, $p_2 = \left(\frac{1}{2},\frac{1}{2}\right)$, $p_3 = \left(0,\frac{1}{2}\right)$. avec les poids d'intégration $\omega_1 = \omega_2 = \omega_3 = \frac{1}{6}$, est exacte pour les polynômes d'ordre 2.

Exercice 47 (Eléments finis $Q2$) *Corrigé en page 209*

On note C le carré [0,1] × [0,1] de sommets

$$a_1 = (0,0), \quad a_2 = (1,0), \quad a_3 = (1,1), \quad a_4 = (0,1).$$

On note

$$a_5 = (1/2,0), \quad a_6 = (1,1/2), \quad a_7 = (1/2,1), \quad a_8 = (0,1/2), \quad a_9 = (1/2,1/2)$$

et

$$\sum = \{a_i, 1 \leq i \leq 8\}.$$

1. Montrer que pour tout $p \in P_2$

$$\sum_{i=1}^{i-4} p(a_i) - 2\sum_{i=5}^{i-8} p(a_i) + 4p(a_9) = 0.$$

2. En déduire une forme linéaire ϕ telle que, si $P = \{p \in Q_2, \phi(p) - 0\}$, alors

$$p \in P \text{ et } \forall i, 1 \leq i \leq 8 p(a_i) = 0 \text{ entraine } p = 0.$$

3. Calculer les fonctions de base de l'élément fini $(C, P \sum)$.

Exercice 48 (Eléments finis Q_2^*) *Corrigé en page 210*

Soit $C = [-1,1] \times [-1,1]$. On note $a_1, \ldots, a_8$ les noeuds de C, définis par

$$a_1 = (-1,-1), \quad a_2 = (1,-1), \quad a_3 = (1,1), \quad a_4 = (-1,1),$$

$$a_5 = (0,-1), \quad a_6 = (1,0), \quad a_7 = (0,1), \quad a_8 = (-1,0).$$

On rappelle que $Q_2 = Vect\{1, x, y, xy, x^2, y^2, x^2y, xy^2, x^2y^2\}$ et que dim $Q_2 = 9$. On note Q_2^* l'espace de polynôme engendré par les fonctions $\{1, x, y, xy, x^2, y^2, x^2y, xy^2\}$.

a) Construire $(\varphi_i^*)_{i=1,\ldots,8} \subset Q_2^*$ tel que

$$\varphi_j^*(a_i) = \delta_{ij} \qquad \forall i,j = 1, \ldots, 8.$$

b) Montrer que Σ est Q_2^*-unisolvant, avec $\Sigma = \{a_1, \ldots, a_8\}$.

c) Soit $S = [-1,1] \times \{1\}$, $\Sigma_S = \Sigma \cap S$, et soit P l'ensemble des restrictions à S des fonctions de Q_2^*, i.e. $P = \{f|_S; f \in Q_2^*\}$. Montrer que Σ_S est P-unisolvant. La propriété est elle vraie pour les autres arêtes de C?

Exercice 49 (Eléments finis $P1$ sur maillage triangulaire) *Corrigé en page 210*

On veut résoudre numériquement le problème

$$\begin{aligned} -\Delta u(x,y) &= f(x,y), \quad (x,y) \in D = (0,a) \times (0,b), \\ u(x,y) &= 0, \qquad\qquad (x,y) \in \partial D, \end{aligned}$$

où f est une fonction donnée, appartenant à $L^2(D)$. Soient M, N deux entiers. On définit

$$\Delta x = \frac{a}{M+1}, \Delta y = \frac{b}{N+1}$$

et on pose

$$x_k = k\Delta x, 0 \leq k \leq M+1, y_l - l\Delta y, 0 \leq l \leq N+1$$

On note

$T^0_{k+1/2,l+1/2}$ le triangle de sommets $(x_k,y_l),(x_{k+1},y_l),(x_{k+1},y_{l+1})$,

$T^1_{k+1/2,l+1/2}$ le triangle de sommets $(x_k,y_l),(x_k,y_{l+1}),(x_{k+1},y_{l+1})$.

Ecrire la matrice obtenue en discrétisant le problème avec les éléments finis triangulaires linéaires (utilisant le maillage précédent).

4.8 Corrigés des exercices

Corrigé de l'exercice 38 page 187

1.Soit $v \in H_N$, comme $H_N \subset C([0,1])$, on a $v \in L^2(]0,1[)$. D'autre part, comme $v|_{K_i} \in P_1$, on a $v|_{K_i}(x) = \alpha_i x + \beta_i$, avec $\alpha_i, \beta_i \in \mathbb{R}$. Donc v admet une dérivée faible dans $L^2(]0,1[)$, et $Dv|_{K_i} = \alpha_i$ on a donc :

$$\|Dv\|_{L^2} \leq \max_{i=1,\ldots,N} |\alpha_i| < +\infty.$$

De plus $v(0) = v(1) = 0$, donc $v \in H^1_0(]0,1[)$.
On en déduit que $H_N \subset H^1_0(]0,1[)$.
2. On a :

$$\phi_i(x) = \begin{cases} 1 - \dfrac{x - x_i}{h} \text{ si } x \in K_i \\ 1 + \dfrac{x - x_i}{h} \text{ si } x \in K_{i-1} \\ 0 \text{ si } x \in]0,1[\ K_i \cup K_{i-1} \end{cases}$$

On en déduit que $\phi_i|_{K_j} \subset P_1$ pour tout $j = 0,\ldots,N$.
De plus, les fonctions ϕ_i sont clairement continues. Pour montrer que $\phi_i \in H_N$, il reste à montrer que $\phi_i(0) = \phi_i(1) = 0$. Ceci est immédiat pour $i = 2,\ldots,N-1$, car dans ce cas $\phi_i|_{K_0} = \phi_i|_{K_{N+1}} = 0$. On vérifie alors facilement que $\phi_1(0) = 1 - \frac{h}{h} = 0$ et $\phi_N(1) = 0$.
Pour montrer que $H_N = Vect\{\phi_1,\ldots,\phi_N\}$, il suffit de montrer que $\{\phi_1,\ldots,\phi_N\}$ est une famille libre de H_N.
En effet, si $\sum_{i=1}^{N} a_i\phi_i = 0$, alors en particulier $\sum_{i=1}^{N} a_i\phi_i(x_k) = 0$, pour $k = 1,\ldots,N$, et donc $a_k = 0$ pour $k = 1,\ldots,N$.
3. Soit $u = \sum_{j=1}^{N} u_j\phi_j$ solution de

$$a(u,\phi_i) = T(\phi_i) \qquad \forall i = 1,\ldots,N.$$

La famille $(u_j)_{j=1,\ldots,N}$ est donc solution du système linéaire

$$\sum_{j=1}^{N} \mathcal{K}_{i,j} u_j = \mathcal{G}_i \qquad i = 1,\ldots,N$$

où $\mathcal{K}_{i,j} = a(\phi_j,\phi_i)$ et $\mathcal{G}_i = T(\phi_i)$. Calculons $\mathcal{K}_{i,j}$ et $\mathcal{G}_i$; on a :

$$K_{i,j} = \int_0^1 \phi_j'(x)\phi_i'(x)dx\,;\ \text{or } \phi_i'(x) = \begin{cases} \dfrac{1}{h} \text{ si } x \in]x_{i-1}x_i[\\ -\dfrac{1}{h} \text{ si } x \in]x_i,x_{i+1}[, \\ 0 \text{ ailleurs} \end{cases}$$

On en déduit que :

$$\mathcal{K}_{i,i} = \int_0^1 (\phi_i'(x))^2 dx = 2h\frac{1}{h^2} = \frac{2}{h} \text{ pour } i = 1,\ldots,N,$$

$$\mathcal{K}_{i,i+1} - \int_0^1 \phi_i'(x)\phi_{i+1}'(x)dx = -h \times \frac{1}{h^2} = -\frac{1}{h}, \text{ pour } i = 1,\ldots,N-1,$$

$$\mathcal{K}_{i,i-1} = \int_0^1 \phi_i'(x)\phi_{i-1}'(x)dx = -\frac{1}{h} \text{ pour } i = 2,\ldots,N,$$

$$\mathcal{K}_{i,j} = 0 \text{ pour } |i-j| > 1.$$

Calculons maintenant $\mathcal{G}_i$:

$$\mathcal{G}_i = \int_{x_i-1}^{x_{i+1}} f(x)\phi_i(x)dx$$

Si f est constante, on a alors $\mathcal{G}_i = f\displaystyle\int_{x_{i-1}}^{x_{i+1}} \phi_i(x)dx = hf$. Si f n'est pas constante, on procède à une intégration numérique. On peut, par exemple, utiliser la formule des trapèzes pour le calcul des intégrales $\displaystyle\int_{x_{i-1}}^{x_i} f(x)\phi_i(x)dx$ et $\displaystyle\int_{x_i}^{x_{i+1}} f(x)\phi_i(x)dx$. On obtient alors :

$$\mathcal{G}_i = hf(x_i).$$

Le schéma obtenu est donc :

$$\begin{cases} 2u_i - u_{i-1} - u_{i+1} = h^2 f(x_i) \qquad i = 1,\ldots,N \\ u_0 = u_{N+1} = 0 \end{cases}$$

C'est exactement le schéma différences finis avec un pas constant h.

Corrigé de l'exercice 39 page 188

1. Soit $(x_i)_{i=1,\ldots,N+1}$ une discrétisation de l'intervalle [0,1], avec $0 = x_0 < x_1 < \cdots x_i < x_{i+1} < x_N < x_{N+1} = 1$. Pour $i = 1,\ldots,N$, on pose $h_{i+\frac{1}{2}} = x_{i+1} - x_i$. L'équation (3.4.40) au point x_i s'écrit :

$$-u_{xx}(x_i) + u(x_i) = f(x)$$

On écrit les développements de Taylor de $u(x_{i+1})$ et $u(x_{i-1})$:

$$u(x_{i+1}) = u(x_i) + h_{i+\frac{1}{2}}u'(x_i) + \frac{1}{2}h^2_{i+\frac{1}{2}}u''(x_i) + \frac{1}{6}h^3_{i+\frac{1}{2}}u'''(\zeta_i), \text{ avec } \zeta_i \in [x_i,x_{i+1}],$$

$$u(x_{i-1}) = u(x_i) - h_{i-\frac{1}{2}}u'(x_i) + \frac{1}{2}h^2_{i-\frac{1}{2}}u''(x_i) - \frac{1}{6}h^3_{i-\frac{1}{2}}u'''(\theta_i), \text{ avec } \theta_i \in [x_{i-1},x_i],$$

En multipliant la première égalité par $h_{i-\frac{1}{2}}$, la deuxième par $h_{i+\frac{1}{2}}$ et en additionnant :

$$\begin{aligned} u''(x_i) &= \frac{2}{h_{i+\frac{1}{2}}h_{i-\frac{1}{2}}(h_{i+\frac{1}{2}}+h_{i-\frac{1}{2}})}\Big[h_{i-\frac{1}{2}}u(x_{i+1})+h_{i+\frac{1}{2}}u(x_{i-1})+(h_{i+\frac{1}{2}}+h_{i-\frac{1}{2}})u(x_i)\Big] \\ &\quad -\frac{1}{6}\frac{h_{i+\frac{1}{2}}^2}{h_{i+\frac{1}{2}}+h_{i-\frac{1}{2}}}u'''(\zeta_i)+\frac{1}{6}\frac{h_{i-\frac{1}{2}}^2}{h_{i+\frac{1}{2}}+h_{i-\frac{1}{2}}}u'''(\theta_i). \end{aligned}$$

En posant $\gamma_i = \frac{2}{h_{i+\frac{1}{2}}h_{i-\frac{1}{2}}(h_{i+\frac{1}{2}}+h_{i-\frac{1}{2}})}$, on déduit donc l'approximation aux différences finies suivante pour tous les noeuds internes :

$$\gamma_i\Big[h_{i-\frac{1}{2}}u_{i+1}+h_{i+\frac{1}{2}}u_{i-1}+(h_{i+\frac{1}{2}}+h_{i-\frac{1}{2}})u_i\Big]+u_i = f(x_i),\ i=1,\ldots,N.$$

La condition de Fourier en 0 se discrétise par

$$\frac{u_1-u_0}{h_{\frac{1}{2}}}-u_0=0,$$

et la condition de Neumann en 1 par :

$$\frac{u_{N+1}-u_N}{h_{N+\frac{1}{2}}}=-1.$$

On obtient ainsi un système linéaire carré d'ordre $N+1$.

2. On prend maintenant une discrétisation volumes finis non uniforme, c'est–à–dire qu'on se donne $N \in \mathbb{N}^\star$ et $h_1,\ldots,h_N > 0$ t.q. $\sum_{i=1}^N h_i = 1$. On pose $x_{\frac{1}{2}} = 0$, $x_{i+\frac{1}{2}} = x_{i-\frac{1}{2}} + h_i$, pour $i = 1,\ldots,N$ (de sorte que $x_{N+\frac{1}{2}} = 1$), $h_{i+\frac{1}{2}} = \frac{h_{i+1}+h_i}{2}$, pour $i = 1,\ldots,N-1$, et $f_i = \frac{1}{h_i}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)dx$, pour $i = 1,\ldots,N$.
En intégrant la première équation de (3.4.40), et en approchant les flux $u'(x_{i+\frac{1}{2}})$ par le flux numérique $F_{i+\frac{1}{2}}$, on obtient le schéma suivant :

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} + h_i u_i = h_i f_i,\ i \in \{1,\ldots,N\}, \tag{4.8.32}$$

où $(F_{i+\frac{1}{2}})_{i\in\{0,\ldots,N\}}$ donné en fonction des inconnues discrètes $(u_1,\ldots,u_N)$ par les expressions suivantes, tenant compte des conditions aux limites :

$$F_{i+\frac{1}{2}} = -\frac{u_{i+1}-u_i}{h_{i+\frac{1}{2}}},\ = i \in \{1,\ldots,N-1\}, \tag{4.8.33}$$

$$F_{\frac{1}{2}} = -\frac{u_1-u_0}{\frac{x_1}{2}}, \tag{4.8.34}$$

$$F_{\frac{1}{2}} - u_0 = 0 \tag{4.8.35}$$

$$F_{N+\frac{1}{2}} = -1. \tag{4.8.36}$$

Notons que u_0 peut être éliminé des équations (4.8.52) et(4.8.53). On obtient ainsi un système linéaire de N équations à N inconnues :

$$
\begin{aligned}
-\frac{u_2-u_1}{h_{\frac{3}{2}}}+\frac{u_1}{1-\frac{x_1}{2}}+h_1u_1 &= h_1f_1, && (4.8.37)\\
-\frac{u_{i+1}-u_i}{h_{i+\frac{1}{2}}}+\frac{u_i-u_{i-1}}{h_{i-\frac{1}{2}}}+h_iu_i &= h_if_i,\ i\in\{2,\ldots,N-1\}, && (4.8.38)\\
-1+\frac{u_N-u_{N-1}}{h_{N-\frac{1}{2}}}+h_Nu_N &= h_Nf_N, && (4.8.39)
\end{aligned}
$$

3. Comme pour les différences finies, on se donne $(x_i)_{i=1,\ldots,N+1}$ une discrétisation de l'intervalle [0,1], avec $0=x_0<x_1<\cdots x_i<x_{i+1}<x_N<x_{N+1}=1$. Pour $i=1,\ldots,N$, on pose $h_{i+\frac{1}{2}}=x_{i+1}-x_i$ et $K_{i+\frac{1}{2}}=[x_i,x_{i+1}]$, pour $i=0,\ldots,N$. On définit l'espace d'approximation $H_N=\{v\in C([0,1],\mathbb{R})$ t.q. $v|_{K_{i+\frac{1}{2}}}\in P_1, i=0,\ldots,N\}$, où P_1 désigne l'ensemble des polynômes de degré inférieur ou égal à 1. Remarquons que l'on a bien $H_N\subset H$.
Pour $i=1,\ldots,N$, on pose:

$$
\begin{aligned}
&\phi_i(x)=\frac{1}{h_{i-\frac{1}{2}}}(x-x_{i-1}) \text{ si } x\in K_{i-\frac{1}{2}},\\
&\phi_i(x)=\frac{1}{h_{i+\frac{1}{2}}}(x_{i+1}-x) \text{ si } x\in K_{i+\frac{1}{2}},\\
&\phi_i(x)=0 \text{ sinon},
\end{aligned}
\qquad (4.8.40)
$$

et on pose également

$$
\begin{aligned}
&\phi_{N+1}(x)=\frac{1}{h_{i-\frac{1}{2}}}(x-x_{i-1}) \text{ si } x\in K_{N+\frac{1}{2}},\\
&\phi_{N+1}(x)=0 \text{ sinon},
\end{aligned}
\qquad (4.8.41)
$$

$$
\begin{aligned}
&\phi_0(x)=\frac{1}{h_{\frac{1}{2}}}(x_1-x) \text{ si } x\in K_{\frac{1}{2}},\\
&\phi_0(x)=0 \text{ sinon},
\end{aligned}
\qquad (4.8.42)
$$

On vérifie facilement que $\phi_i\in H_N$ pour tout $0=1,\ldots,N+1$ et que $H_N=Vect\{\phi_0,\ldots,\phi_{N+1}\}$.
La formulation éléments finis s'écrit alors:

$$
\begin{aligned}
&u^{(N)}\in H_N,\\
&a(u^{(N)},v)=T(v),\forall v\in H_N,
\end{aligned}
\qquad (4.8.43)
$$

Pour construire le système linéaire à résoudre, on prend successivement $v=\phi_i$, $i=0,\ldots,N+1$ dans (4.8.43). Soit $u^{(N)}=\sum_{j=0}^{N+1}u_j\phi_j$ solution de

$$
a(u^{(N)},\phi_i)=T(\phi_i)\qquad \forall i=0,\ldots,N+1.
$$

La famille $(u_j)_{j=0,\ldots,N+1}$ est donc solution du système linéaire

$$
\sum_{j=0}^{N}\mathcal{K}_{i,j}u_j=\mathcal{G}_i\qquad i=0,\ldots,N+1,
$$

où $\mathcal{K}_{i,j} = a(\phi_j,\phi_i)$ et $\mathcal{G}_i = T(\phi_i)$. Calculons $\mathcal{K}_{i,j}$ et $\mathcal{G}_i$; on a : $\mathcal{K}_{ij} = \int_0^1 \phi_j'(x)\phi_i'(x)dx + \int_0^1 \phi_j(x)\phi_i(x)dx$.
Or

$$\phi_i'(x) = \begin{cases} \dfrac{1}{h_{i-\frac{1}{2}}} \text{ si } x \in]x_{i-1}x_i[\\ -\dfrac{1}{h_{i+\frac{1}{2}}} \text{ si } x \in]x_i,x_{i+1}[\\ 0 \text{ ailleurs.} \end{cases}$$

Donc si $1 \leq i = j \leq N$, on a

$$\mathcal{K}_{i,i} = \int_0^1 (\phi_i'(x))^2 dx + \int_0^1 (\phi_i(x))^2 dx + \int_0^1 \phi_i(x)\phi_i'(x)dx = \frac{1}{h_{i-\frac{1}{2}}} + \frac{1}{h_{i+\frac{1}{2}}} + \frac{h_{i-\frac{1}{2}}}{3} + \frac{h_{i+\frac{1}{2}}}{3}.$$

Si $i = j = N+1$, alors

$$\mathcal{K}_{N+1,N+1} = \int_0^1 (\phi_{N+1}'(x))^2 dx + \int_0^1 (\phi_{N+1}(x))^2 dx = \frac{1}{h_{N+\frac{1}{2}}} + \frac{h_{N+\frac{1}{2}}}{3}.$$

Si $i = j = 0$, alors

$$\mathcal{K}_{0,0} = \int_0^1 (\phi_0'(x))^2 dx + \int_0^1 (\phi_0(x))^2 dx + \phi_0^2 = \frac{1}{h_{\frac{1}{2}}} + \frac{h_{\frac{1}{2}}}{3} + 1.$$

Si $0 \leq i \leq N$ et $j = i+1$, on a :

$$\mathcal{K}_{i,i+1} = \int_0^1 \phi_i'(x)\phi_{i+1}'(x)dx + \int_0^1 \phi_i(x)\phi_{i+1}(x)dx = -h_{i+\frac{1}{2}} \times \frac{1}{h_{i+\frac{1}{2}}^2} + \frac{h_{i+\frac{1}{2}}}{2} - \frac{h_{i+\frac{1}{2}}}{3} = -\frac{1}{h_{i+\frac{1}{2}}} + \frac{h_{i+\frac{1}{2}}}{6}.$$

La matrice étant symétrique, si $1 \leq i \leq N+1$ et $j = i-1$, on a :

$$\mathcal{K}_{i,i-1} = \mathcal{K}_{i-1,i} = -\frac{1}{h_{i-\frac{1}{2}}} + \frac{h_{i-\frac{1}{2}}}{6}.$$

Calculons maintenant $\mathcal{G}_i$.

$$\mathcal{G}_i = \int_{x_i-1}^{x_{i+1}} f(x)\phi_i(x)dx + \phi_i(1).$$

Si f est constante, on a alors $\mathcal{G}_i = f\int_{x_{i-1}}^{x_{i+1}} \phi_i(x)dx + \phi_i(1) = \frac{1}{2}(h_{i-\frac{1}{2}} + h_{i+\frac{1}{2}})f + \phi_i(1)$.
Si f n'est pas constante, on procède à une intégration numérique. On peut, par exemple, utiliser la formule des trapèzes pour le calcul des intégrales $\int_{x_{i-1}}^{x_i} f(x)\phi_i(x)dx$ et $\int_{x_i}^{x_{i+1}} f(x)\phi_i(x)dx$. On obtient alors :

$$\mathcal{G}_i = \frac{1}{2}(h_{i-\frac{1}{2}} + h_{i+\frac{1}{2}})f(x_i) + \phi_i(1).$$

Le schéma obtenu est donc :

$$\left\{\begin{array}{l} (\frac{1}{h_{i-\frac{1}{2}}}+\frac{1}{h_{i+\frac{1}{2}}}+\frac{h_{i-\frac{1}{2}}}{3}+\frac{h_{i+\frac{1}{2}}}{3})u_i+(\frac{h_{i-\frac{1}{2}}}{6}-\frac{1}{h_{i-\frac{1}{2}}})u_{i-1}+(\frac{h_{i+\frac{1}{2}}}{6}-\frac{1}{h_{i+\frac{1}{2}}})u_{i+1} \\ \qquad\qquad = \frac{1}{2}(h_{i-\frac{1}{2}}+h_{i+\frac{1}{2}})f(x_i) \qquad i=1,\ldots,N \\ (\frac{1}{h_{\frac{1}{2}}}+\frac{h_{\frac{1}{2}}}{3}+1)u_0+(\frac{1}{h_{i+\frac{1}{2}}}+\frac{h_{i+\frac{1}{2}}}{6})u_1=\frac{1}{2}h_{\frac{1}{2}}f(x_0) \\ \\ (\frac{1}{h_{N+\frac{1}{2}}}+\frac{h_{N+\frac{1}{2}}}{3})u_{N+1}(\frac{h_{N+\frac{1}{2}}}{6}-\frac{1}{h_{N+\frac{1}{2}}}=\frac{1}{2}h_{N+\frac{1}{2}}f(x_{N+1})+1. \end{array}\right.$$

Corrigé de l'exercice 40 page 188

1. On se donne $(x_i)_{i=1,\ldots,N+1}$, discrétisation de l'intervalle [0,1], avec $h=1/N$, et $x_i=(i-1)h, i=1,\ldots,N+1$. Pour $i=1,\ldots,N$, on pose $K_{i+\frac{1}{2}}=[x_i,x_{i+1}]$, pour $i=0,\ldots,N$. On définit l'espace d'approximation $H_N=\{v\in C([0,1],\mathbb{R}) \text{ t.q. } v(1)=0 \text{ et } v|_{K_{i+\frac{1}{2}}}\in P_1, i=0,\ldots,N\}$, où P_1 désigne l'ensemble des polynômes de degré inférieur ou égal à 1. Remarquons que l'on a bien $H_N\subset H$.
Pour $i=1,\ldots,N$, on pose:

$$\begin{array}{l} \phi_i(x)=\dfrac{1}{h}(x-x_{i-1}) \text{ si } x\in K_{i-\frac{1}{2}}, \\ \phi_i(x)=\dfrac{1}{h}(x_{i+1}-x) \text{ si } x\in K_{i+\frac{1}{2}}, \\ \phi_i(x)=0 \text{ sinon}, \end{array} \tag{4.8.44}$$

$$\begin{array}{l} \phi_0(x)=\dfrac{2}{h}(x_1-x) \text{ si } x\in K_{\frac{1}{2}}, \\ \phi_0(x)=0 \text{ sinon}, \end{array} \tag{4.8.45}$$

On vérifie facilement que $\phi_i\in H_N$ pour tout $0=1,\ldots,N+1$ et que $H_N=Vect\{\phi_0,\ldots,\phi_N\}$.
On pose enfin

$$\begin{array}{l} \phi_{N+1}(x)=\dfrac{1}{h}(x-x_{i-1}) \text{ si } x\in K_{N+\frac{1}{2}}, \\ \phi_{N+1}(x)=0 \text{ sinon}, \end{array} \tag{4.8.46}$$

La formulation éléments finis s'écrit alors :

$$\begin{array}{l} u^{(N)}=\tilde{u}^{(N)}+\phi_{N+1}, \text{ avec } \tilde{u}^{(N)}\in H_N, \\ a(\tilde{u}^{(N)}+\phi_{N+1},v)=T(v), \forall v\in H_N, \end{array} \tag{4.8.47}$$

Pour construire le système linéaire à résoudre, on prend successivement $v=\phi_i$, $i=0,\ldots,N$ dans (4.8.47).
Soit $\tilde{u}^{(N)}=\sum_{j=0}^{N}u_j\phi_j$ solution de

$$a(\tilde{u}^{(N)}+\phi_{N+1},\phi_i)=T(\phi_i) \qquad \forall i=0,\ldots,N+1.$$

La famille $(u_j)_{j=0,\ldots,N+1}$ est donc solution du système linéaire

$$\sum_{j=0}^{N}\mathcal{K}_{i,j}u_j=\mathcal{G}_i \qquad i=0,\ldots,N+1,$$

où $\mathcal{K}_{i,j} = a(\phi_j,\phi_i)$ et $\mathcal{G}_i = T(\phi_i) - a(\phi_{N+1},\phi_j)$. Calculons $\mathcal{K}_{i,j}$ et $\mathcal{G}_i$; on a :

$$\mathcal{K}_{i,j} = \int_0^1 \phi_j'(x)\phi_i'(x)dx + \int_0^1 \phi_j'(x)\phi_i(x)dx + \int_0^1 \phi_j(x)\phi_i(x)dx.$$

Or

$$\phi_i'(x) = \begin{cases} \dfrac{1}{h} \text{ si } x \in]x_{i-1}x_i[\\ -\dfrac{1}{h} \text{ si } x \in]x_i,x_{i+1}[\\ 0 \text{ ailleurs.} \end{cases}$$

Donc si $1 \leq i = j \leq N$, on a

$$\mathcal{K}_{i,i} = \int_0^1 (\phi_i'(x))^2dx + \int_0^1 (\phi_i(x))^2dx + \int_0^1 \phi_i(x)\phi_i'(x)dx = \frac{1}{h} + \frac{1}{h} + \frac{h}{3} + \frac{h}{3}.$$

Si $i = j = 0$, alors

$$\mathcal{K}_{0,0} = \int_0^1 (\phi_0'(x))^2dx + \int_0^1 (\phi_0(x))^2dx + \int_0^1 \phi_0'(x)\phi_0(x)dx = \frac{2}{h} + \frac{h}{6} + 1.$$

Si $0 \leq i \leq N$ et $j = i+1$, on a :

$$\mathcal{K}_{i,i+1} = \int_0^1 \phi_i'(x)\phi_{i+1}'(x)dx + \int_0^1 \phi_i(x)\phi_{i+1}(x)dx + \int_0^1 \phi_i'(x)\phi_{i+1}(x)dx \tag{4.8.48}$$

$$= -h \times \frac{1}{h^2} + \frac{h}{2} - \frac{h}{3} = -\frac{1}{h} + \frac{h}{6} - \frac{1}{2}. \tag{4.8.49}$$

La matrice tant symétrique, si $1 \leq i \leq N+1$ et $j = i-1$, on a :

$$\mathcal{K}_{i,i-1} = \mathcal{K}_{i-1,i} = -\frac{1}{h_{i-\frac{1}{2}}} + \frac{h_{i-\frac{1}{2}}}{6} - \frac{1}{2}.$$

Calculons maintenant $\mathcal{G}_i$.

$$\mathcal{G}_i = \int_{x_i-1}^{x_{i+1}} f(x)(\phi_i(x) - \phi_{N+1}(x))dx.$$

2. On se donne $N \in \mathbb{N}^\star$ et $h > 0$ t.q. $Nh = 1$. On pose $x_{\frac{1}{2}} = 0$, $x_{i+\frac{1}{2}} = x_{i-\frac{1}{2}} + h$, pour $i = 1,\ldots,N$ (de sorte que $x_{N+\frac{1}{2}} = 1$), $f_i = \frac{1}{h}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)dx$, pour $i = 1,\ldots,N$.
En intégrant la première équation de (3.4.41), et en approchant les flux $u'(x_{i+\frac{1}{2}})$ par le flux numérique $F_{i+\frac{1}{2}}$, et $u(x_{i+\frac{1}{2}})$ par $u_{i+\frac{1}{2}}$, on obtient le schéma suivant :

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} + u_{i+\frac{1}{2}} - u_{i-\frac{1}{2}} + hu_i = hf_i,\ i \in \{1,\ldots,N\}, \tag{4.8.50}$$

où les flux numériques de diffusion $(F_{i+\frac{1}{2}})_{i\in\{0,\ldots,N\}}$ sont donnés en fonction des inconnues discrètes $(u_1,\ldots,u_N)$ par les expressions suivantes, tenant compte des conditions aux limites :

$$
\begin{aligned}
F_{i+\frac{1}{2}} &= -\frac{u_{i+1}-u_i}{h}, \ = i \in \{1,\ldots,N-1\}, && (4.8.51)\\
F_{\frac{1}{2}} &= -\frac{u_1-u_0}{\frac{h}{2}}, && (4.8.52)\\
-F_{\frac{1}{2}}+u_0 &= 0 && (4.8.53)\\
F_{N+\frac{1}{2}} &= -\frac{1-u_N}{\frac{h}{2}}, && (4.8.54)
\end{aligned}
$$

Avec un choix centré, on écrit : $u_{i+\frac{1}{2}} = \frac{1}{2}(u_{i+1}+u_i)$ pour $i = 1,\ldots,N$.
Notons que u_0 peut être éliminé des équations (4.8.52) et(4.8.53) ; on obtient $u_0 = u_1(1-\frac{h}{2})$. On obtient ainsi un système linéaire de N équations à N inconnues :

$$
\begin{aligned}
-\frac{u_2-u_1}{h}+u_1(1-\frac{h}{2})+\frac{1}{2}(u_2-u_1(1-\frac{h}{2}))+hu_1 &= hf_1, && (4.8.55)\\
-\frac{u_{i+1}-u_i}{h}+\frac{u_i-u_{i-1}}{h}+\frac{1}{2}(u_{i+1}-u_{i-1})+hu_i &= hf_i,\ i \in \{2,\ldots,N-1\}, && (4.8.56)\\
-1+\frac{u_N-u_{N-1}}{h}+1-\frac{1}{2}(u_N+u_{N-1})+hu_N &= h_N f_N, && (4.8.57)
\end{aligned}
$$

3. Soit $(x_i)_{i=1,\ldots,N+1}$ une discrétisation de l'intervalle [0,1], avec $x_i = ih$, pour $0 = 1,\ldots,N+1$, avec $h = \frac{1}{N+1}$. L'équation (3.4.40) au point x_i s'écrit :

$$-u_{xx}(x_i)+u(x_i) = f(x)$$

On écrit les développements de Taylor de $u(x_{i+1})$ et $u(x_{i-1})$:

$$u(x_{i+1}) = u(x_i)+hu'(x_i)+\frac{1}{2}h^2u''(x_i)+\frac{1}{6}h^3u'''(\zeta_i), \text{ avec } \zeta_i \in [x_i,x_{i+1}],$$

$$u(x_{i-1}) = u(x_i)-hu'(x_i)+\frac{1}{2}h^2u''(x_i)-\frac{1}{6}h^3u'''(\theta_i), \text{ avec } \theta_i \in [x_{i-1},x_i],$$

En multipliant les deux égalités par h et en additionnant, on obtient :

$$
\begin{aligned}
u''(x_i) &= \frac{1}{h^2}\Big[u(x_{i+1})+u(x_{i-1})+2u(x_i)\Big]\\
&\quad +\frac{h}{12}(-u'''(\zeta_i)+u'''(\theta_i)).
\end{aligned}
$$

On déduit donc l'approximation aux différences finies suivante pour tous les noeuds internes :

$$\frac{1}{h^2}\Big[u_{i+1}+u_{i-1}+2u_i\Big]+u_i = f(x_i),\ i = 1,\ldots,N.$$

La condition de Fourier en 0 se discrétise par

$$2\frac{u_1-u_0}{h}-u_0 = 0,$$

et la condition de Neumann en 1 par :

$$2\frac{u_{N+1}-u_N}{h}=-1.$$

On obtient ainsi un système linéaire carré d'ordre $N+1$.

4.. Quel est l'ordre de convergence de chacune des méthodes étudiées aux questions précédentes ?
Les trois premières méthodes (EF P1, volumes finis et différences finies sur maillage uniforme) sont d'ordre 2. La méthode d'éléments finis $P2$ est d'ordre 3.

Corrigé de l'exercice 41 page 188

Le problème (4.7.28) s'écrit :

$$\begin{cases} -u''(x)+u(x)=f(x), \quad x\in]0,1[\\ u(0)=0 \\ u'(1)=0. \end{cases} \tag{4.8.58}$$

L'espace $H^1_{\Gamma_0,g_0}$ défini à l'exercice 34 page 139 est donc dans ce cas égal à $H=\{u\in H^1(]0,1[,u(0)=0\}$. L'espace d'approximation V_h est donc l'ensemble des fonctions continues, affines sur chaque maille $K_i=[x_i,x_{i+1}]$, et nulles en 0.
1.1. L'ensemble V_h est engendré par les fonctions de base éléments finis $P1$ aux noeuds $x_i, i=1,\ldots,N+1$. Ces fonctions s'écrivent justement

$$\phi_i(x)=\left(1-\frac{|x-x_i|}{h}\right)^+ \quad i=1,\ldots,N+1.$$

1.2. Le système linéaire à résoudre s'écrit :

$$\mathcal{K}u=\mathcal{G},$$

où $\mathcal{K}$ est une matrice d'ordre $N+1, \mathcal{G}\in\mathbb{R}^{N+1}$, et

$$\mathcal{K}_{ij}=\int_0^1[\phi_i'(x)\phi_j'(x)+\phi_i(x)\phi_j(x)]dx,$$

$$\mathcal{G}_i=\int_0^1 f(x)\phi_i(x)dx, \qquad i=1,\ldots,N+1.$$

Les intégrales $\displaystyle\int_0^1\phi_i'(x)\phi_j'(x)dx$ et $\displaystyle\int_0^1 f(x)\phi_i(x)dx$ sont calculées par exemple à l'exercice 38 page 187, pour $i,j=1,\ldots,N$. Il reste à calculer $b_{ij}=\displaystyle\int_0^1\phi_i(x)\phi_j(x)dx$, pour $i,j=1,\ldots,N$, et les intégrales faisant intervenir le noeud $N+1$. En ce qui concerne le calcul de b_{ij} pour $i,j=1,\ldots,N$, trois cas se présentent.

1. Si $j=i$, $b_{ii}=\int_{x_{i-1}}^{x_i}\left(1+\frac{x-x_i}{h}\right)^2dx+\int_{x_i}^{x_{i+1}}\left(1-\frac{x-x_i}{h}\right)^2dx$, par changement de variable, $\xi=1+\dfrac{x-x_i}{h}$ dans la première intégrale et $\xi=1-\dfrac{x-x_i}{h}$ dans la seconde, on a donc :

$$b_{ii}=2h\int_0^1\xi^2d\xi=\frac{2h}{3}$$

2. Si $j = i+1$, $\quad b_{ii+1} = \int_{x_i}^{x_{i+1}} \left(1 - \frac{x-x_i}{h}\right)\left(1 + \frac{x-x_{i+1}}{h}\right) dx$. Posons $\xi = \dfrac{x-x_i}{h}$, on a donc $\dfrac{x-x_{i+1}}{h} = \dfrac{x-x_i+x_i-x_{i+1}}{h} = \xi - 1$. Donc $b_{ii+1} = \displaystyle\int_0^1 (1-\xi)\xi h d\xi = h\left[\frac{1}{2} - \frac{1}{3}\right] = \frac{h}{6}$.

3. De même, par symétrie, si $j = i-1$, $\quad b_{i,i+1} = \dfrac{h}{6}$.

On a donc finalement :

$$\mathcal{K}_{ii} = \frac{2}{h} + \frac{2h}{3} \quad \text{pour } i = 1, \ldots, N$$

$$\mathcal{K}_{ii+1} = -\frac{1}{h} + \frac{h}{6} \quad \text{pour } i = 1, \ldots, N$$

$$\mathcal{K}_{i-1,i} = -\frac{1}{h} + \frac{h}{6} \quad \text{pour } i = 2, \ldots, N+1$$

En ce qui concerne le noeud $N+1$, on a :

$$\int_0^1 \phi'_{N+1}(x)\phi'_{N+1}(x)dx = h$$

et

$$b_{N+1,N+1} = \int_{x_N}^{x_{N+1}} \left(1 + \frac{x - x_{N+1}}{h}\right)^2 dx = h\int_0^1 \xi^2 d\xi = \frac{h}{3}$$

Donc $\mathcal{K}_{N+1,N+1} = \dfrac{1}{h} + \dfrac{h}{3}$.

D'autre part, avec une intégration numérique par la méthode des trapèzes pour le calcul de $\mathcal{G}_i$, on obtient

$$\mathcal{G}_i = hf(x_i) \qquad i = 1, \ldots, N$$

$$\mathcal{G}_{N+1} = \frac{h}{2} f(x_{N+1}).$$

Le schéma éléments finis s'écrit donc finalement :

$$\begin{cases} \left(\dfrac{2}{h} + \dfrac{2h}{3}\right) u_i + \left(-\dfrac{1}{h} + \dfrac{h}{6}\right) u_{i-1} + \left(-\dfrac{1}{h} + \dfrac{h}{6}\right) u_{i+1} = hf(x_i), \quad i = 2, \ldots, N \\ \left(\dfrac{2}{h} + \dfrac{2h}{3}\right) u_1 + \left(-\dfrac{1}{h} + \dfrac{h}{6}\right) u_2 = hf(x_i) \\ \left(\dfrac{1}{h} + \dfrac{h}{3}\right) u_{N+1} + \left(-\dfrac{1}{h} + \dfrac{h}{6}\right) u_N = \dfrac{h}{2} f(x_{N+1}). \end{cases}$$

Le schéma différences finies pour le problème (4.8.58) s'écrit :

$$\begin{cases} \dfrac{1}{h^2}[2u_i - u_{i-1} - u_{i+1}] + u_i = f(x_i) i = 1, \ldots, N \\ u_0 = 0 \\ u_{N+1} = u_N. \end{cases}$$

Ce qui s'écrit encore :

$$\begin{cases} \left(\frac{2}{h}+h\right)u_i - \frac{1}{h}u_{i-1} - \frac{1}{h}u_{i+1} = hf(x_i), \quad i=2,\ldots,N \\ \left(\frac{2}{h}+h\right)u_1 - \frac{1}{h}u_2 = hf(x_1) \\ \left(\frac{1}{h}+h\right)u_N - \frac{1}{h}u_{N-1} = hf(x_N). \end{cases}$$

Donc le schéma EF P1 et le schéma différences finies ne sont pas équivalents.
Pour discrétiser le problème (4.8.58) par un schéma volumes finis, on commence par intégrer l'équation (4.8.58) sur chaque maille $K_i = [x_{i-\frac{1}{2}} x_{i+\frac{1}{2}}]$

$$-u'(x_{i+\frac{1}{2}}) + u'(x_{i\frac{1}{2}}) + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x)dx = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)dx.$$

La discrétisation de cette équation donne alors, en fonction des inconnues discrètes $u_1,\ldots,u_N$ (en tenant compte des conditions aux limites).

$$\begin{cases} -\frac{u_{i+1}-u_i}{h} + \frac{u_i - u_{i-1}}{h} + hu_i = hf_i \qquad i=2,\ldots,N-1 \\ -\frac{u_2-u_1}{h} + \frac{2u_1}{h} + hu_1 = hf_1 \\ \frac{u_N - u_{N-1}}{h} + hu_N = hf_N, \end{cases}$$

avec $f_i = \frac{1}{h}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x)dx$. Ceci s'écrit encore :

$$\begin{cases} \left(\frac{2}{h}+h\right)u_i - \frac{1}{h}u_{i-1} - \frac{1}{h}u_{i+1} = hf(x_i), \quad i=2,\ldots,N \\ \left(\frac{3}{h}+h\right)u_1 - \frac{1}{h}u_2 = hf(x_1) \\ \left(\frac{1}{h}+h\right)u_N - \frac{1}{h}u_{N-1} = hf(x_N). \end{cases}$$

Les schémas éléments finis, différences finis et volumes finis ne sont pas exactement les mêmes pour le problème (4.8.58).
3. Dans le cas du schéma élément fini, on a

$$u_h|_{[x_N,x_{N+1}]} = u_N\phi_{n-1} + u_{N+1}\phi_N.$$

Donc $u_h'(1) = \frac{1}{h}(u_{N+1} - u_N)$. Cette quantité n'est pas forcément égale à 0 (prendre par exemple le cas de trois mailles).

Correction de l'exercice 42 page 189

I.1. On note x,y les deux variables de $\mathbb{R}^2$. L'espace Q_1 est l'ensemble des polynômes de la forme $a+bx+cy+dxy$ avec a,b,c,$d \in \mathbb{R}$. On a donc dim $Q_1 = 4 =$ Card $\Sigma_1 =$ Card Σ_2 pour montrer que (e_1,Σ_1,Q_1) est un élément fini de Lagrange, il suffit de montrer que $f \in Q_1$ et $f|_{\Sigma_1} = 0$ implique $f = 0$. Soient donc a,b,c,d, $\in \mathbb{R}$. On pose $f(x,y) = a+bx+cy+dxy$ pour $(x,y)^t \in e_1$ et on suppose que $f|_{\Sigma_1} = 0$, c'est à dire : $f(-1,1) = 0$,$f(0,1) = 0$,$f(1,0) = 0$ et $f(-1,0) = 0$. On a donc :

$$\begin{cases} a-b+c-d = 0 \\ a+c = 0 \\ a+b = 0 \\ a-b = 0 \end{cases}$$

Les deux dernières équations donnent que $a = b = 0$, la troisième donne alors que $c = 0$, et la première donne enfin que $d = 0$. On a donc montré que $f = 0$. On en déduit que (e_1,Σ_1,Q_1) est un élément fini de Lagrange. Pour montrer que (e_2,Σ_2,Q_1) est un élément fini de Lagrange, on procède de la même façon : soient a,b,c,$d \in \mathbb{R}$ et $f(x,y) = a+bx+cy+dxy$ pour $(x,y)^t \in e_2$. On suppose que $f|_{\Sigma_2} = 0$, c'est à dire : $f(0,1) = 0$,$f(2,1) = 0$,$f(2,0) = 0$ et $f(1,0) = 0$. On a donc :

$$\begin{cases} a+c = 0 \\ a+2b+c+2d = 0 \\ a+2b = 0 \\ a+b = 0 \end{cases}$$

Les deux dernières équations donnent $a = b = 0$, la première donne alors $c = 0$ et, finalement, la quatrième donne $d = 0$. On a donc montré que $f = 0$. On en déduit que (e_2,Σ_2,Q_1) est un élément fini de Lagrange.

I.2. L'espace (de dimension finie) associé à cette discrétisation est engendré par les six fonctions de base globales. On va montrer que la fonction de base associée à M_1 (par exemple) n'est pas dans $H^1(\Omega)$. On note ϕ_1 cette fonction de base. On doit avoir $\phi_{1|e_1} \in Q_1$, $\phi_{1|e_2} \in Q_1$ et $\phi_1(M_1) = 1$, $\phi_1(M_i) = 0$ si $i \neq 1$. On en déduit que $\phi_1 = 0$ sur e_2 et $\phi_1(x,y) = -xy$ si $(xy)^t \in e_1$. On a bien $\phi_1 \in L^2(\Omega)$ mais on va montrer maintenant que ϕ_1 n'a pas de dérivée faible dans $L^2(\Omega)$ (et donc que $\phi_1 \notin H^1(\Omega)$). On va s'intéresser à la dérivée faible par rapport à x (mais on pourrait faire un raisonnement similaire pour la dérivée faible par rapport à y). On suppose que ϕ_1 a une dérivée faible par rapport à x dans $L^2(\Omega)$ (et on va montrer que ceci mène à une contradiction). Supposons donc qu'il existe une fonction $\psi \in L^2(\Omega)$ telle que

$$I = \int_{-1}^{2}\int_{0}^{1} \phi_1(x,y)\frac{\partial\varphi}{\partial x}(x,y)dxdy = \int_{-1}^{2}\int_{0}^{1} \psi(x,y)\varphi(x,y)dxdy, \text{ pour tout } \varphi \in C_c^\infty(\Omega). \tag{4.8.59}$$

Soit $\varphi \in C_c^\infty(\Omega)$, comme ϕ_1 est nulle sur e_2, on a $I = \int\int_{e_1} \phi_1(x,y)\frac{\partial\varphi}{\partial x}(x,y)dxdy$ et donc :

$$I = \int_0^1 \left(\int_{-1}^{1-y} (-xy)\frac{\partial\varphi}{\partial x}(x,y)dx \right) dy.$$

Par intégration par parties, en tenant compte du fait que φ est à support compact sur Ω, on obtient :

$$\begin{aligned} I &= \int_0^1 \left[\int_{-1}^{1-y} y\ \varphi(x,y)dx - (1-y)y\varphi(1-y,y) \right] dy \\ &= \int_0^1 \int_{-1}^{2} y1_{e_1}(x,y)\varphi(x,y)dx - \int_0^1 (1-y)y\varphi(1-y,y)dy. \end{aligned}$$

En posant $\widetilde{\psi}(x,y) = -\psi(x,y) + y1_{e_1}(x,y)$, on a $\widetilde{\psi} \in L^2(\Omega)$ et :

$$\int_0^1 (1-y)y\varphi(1-y,y)dy = \int_{-1}^{2}\int_0^1 \widetilde{\psi}(x,y)\varphi(x,y)dxdy. \tag{4.8.60}$$

Pour aboutir à une contradiction, on va montrer que (4.8.60) est fausse pour certains $\varphi \in C_c^\infty(\Omega)$. On remarque tout d'abord qu'il existe $\varphi \in C_c^\infty(\Omega)$ t.q.

$$\int_0^1 (1-y)(y)\varphi(1-y,y)dy > 0.$$

(Il suffit de choisir $\varphi \in C_c^\infty(\Omega)$ t.q. $\varphi \leq 0$ et $\varphi(1-y,y) > 0$ pour $y = \dfrac{1}{2}$, par exemple.) On se donne maintenant une fonction $\varphi \in C_c^\infty(\mathbb{R})$ t.q. $\varphi(0) = 1$ et $\rho = 0$ sur $[-1,1]^c$ et on écrit (4.8.60) avec φ_n au lieu de φ, où φ_n est définie par :

$$\varphi_n(x,y) = \varphi(x,y)\rho(n(x+y-1))$$

(noter que l'on a bien $\varphi_n \in C_c^\infty(\Omega)$ car $\rho \in C^\infty(\mathbb{R})$ et $\varphi \in C_c^\infty(\Omega)$) On a donc

$$\int_0^1 (1-y)y\varphi_n(1-y,y)dy = \int_{-1}^{2}\int_0^1 \widetilde{\psi}(x,y)\varphi_n(x,y)\ dxdy.$$

Le terme de gauche de cette égalité est indépendant de n et non nul car $\varphi_n(1-y,y) = \varphi(1-y,y)$ pour tout n et tout $y \in [0,1]$. Le terme de droite tend vers 0 quand $n \to \infty$ par convergence dominée car

$$\widetilde{\psi}\varphi_n \to 0 \text{ p.p., et } |\widetilde{\psi}\varphi_n| \leq \|\rho\|_\infty |\widetilde{\psi}|\ |\varphi| \in L^1(\Omega).$$

Ceci donne la contradiction désirée et donc que $\phi_1 \notin H^1(\Omega)$. L'hypothèse non vérifiée (pour avoir la cohérence globale) est l'hypothèse (4.2.7). En posant $S = \bar{e}_1 \cap \bar{e}_2$, on a

$$\Sigma_1 \cap S = \Sigma_2 \cap S = \{M_2, M_5\},$$

et on a, bien sûr, $\varphi_1|_S = \varphi_1|_S$ mais on remarque que $(\{M_2,M_5\},Q_1|_S)$ n'est pas unisolvant car $\text{Card}(\{M_2,M_5\}) = 2$ et $\dim(Q_1|_S) = 3$.

II.1. Les quatre fonctions de base de (e,Σ,P) sont :

$$\phi_1(x,y) = \frac{1}{4}(x+1)(y+1)$$

$$\phi_2(x,y) = -\frac{1}{4}(x+1)(y-1)$$

$$\phi_3(x,y) = -\frac{1}{4}(x-1)(y+1)$$

$$\phi_4(x,y) = \frac{1}{4}(x-1)(y-1).$$

II.2.

Construction de F_1 Pour $(x,y)^t \in e$, on pose

$$F_1(x,y) = M_1\phi_3(x,y) + M_2\phi_1(x,y) + M_5\phi_2(x,y) + M_4\phi_4(x,y),$$

ce qui donne

$$4F_1(x,y) = \begin{pmatrix} -1 \\ 1 \end{pmatrix}(1-x)(1+y) + \begin{pmatrix} 0 \\ 1 \end{pmatrix}(1+x)(1+y) + \begin{pmatrix} 1 \\ 0 \end{pmatrix}(1+x)(1-y) + \begin{pmatrix} -1 \\ 0 \end{pmatrix}(1-x)(1-y)$$

et donc

$$4F_1(x,y) = \begin{pmatrix} -1+3x-y-xy \\ 2(1+y) \end{pmatrix}.$$

Pour $y \in [-1,1]$ fixé, la première composante de $F_1(x,y)$ est linéaire par rapport à x et $F_1(\cdot,y)$ est une bijection de $[-1,1] \times \{y\}$ dans $[-1,\frac{1-y}{2}] \times \{\frac{1+y}{2}\}$. On en déduit que F_1 est une bijection de e dans e_1.

Construction de F_2 Pour $(x,y)^t \in e$, on pose

$$F_2(x,y) = M_2\phi_3(x,y) + M_3\phi_1(x,y) + M_6\phi_2(x,y) + M_5\phi_4(x,y),$$

ce qui donne

$$4F_2(x,y) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}(1-x)(1+y) + \begin{pmatrix} 2 \\ 1 \end{pmatrix}(1+x)(1+y) + \begin{pmatrix} 2 \\ 0 \end{pmatrix}(1+x)(1-y) + \begin{pmatrix} 1 \\ 0 \end{pmatrix}(1-x)(1-y)$$

et donc $4F_2(x,y) = \begin{pmatrix} 5+3x-y+xy \\ 2+2y \end{pmatrix}$ Pour $y \in [-1,1]$ fixé, la première composante de $F_2(x,y)$ est linéaire par rapport à x et $F_2(.,y)$ est une bijection de $[-1,1] \times \{y\}$ dans $\left[\frac{1-y}{2},2\right] \times \left\{\frac{1+y}{2}\right\}$ On en déduit que F_2 est une bijection de e dans e_2. Les fonctions F_1 et F_2 ne sont pas affines.

II.3. Les éléments (e_1,Σ_1,P_{e_1}) et (e_2,Σ_2,P_{e_2}) sont les éléments finis de Lagrange construits à partir de l'élément fini de Lagrange (e,Σ,P) et des bijections F_1 et F_2 (de e dans e_1 et de e dans e_2), voir la proposition 4.10 page 162. Pour montrer que l'espace vectoriel construit avec (e_1,Σ_1,P_{e_1}) et (e_2,Σ_2,P_{e_2}) est inclus dans $H^1(\Omega)$, il suffit de vérifier la propriété de "cohérence globale" donnée dans la proposition 4.11 page 163. On pose

$$\begin{aligned} S = \bar{e}_1 \cap \bar{e}_2 &= \{(x,y) \in \bar{\Omega}, x+y=1\} \\ &= \{(1-y,y), y \in [0,1]\} \end{aligned}$$

On remarque tout d'abord que $\Sigma_1 \cap S = \Sigma_2 \cap S = \{M_2,M_5\}$. On détermine maintenant $P_{e_1|_S}$ et $P_{e_2|S}$. Soit $f \in P_{e_1}$. Soit $(x,y) \in S$ (c'est à dire $y \in [0,1]$ et $x+y=1$) on a $f(x,y) = f \circ F_1(1,2y-1)$. (On a utilisé ici le fait que $F_1(\{1\} \times [-1,1]) = 5$). Donc $P_{e_1|_S}$ est l'ensemble des fonctions de S dans $\mathbb{R}$ de la forme : $(x,y) \mapsto g(1,2y-1)$, où $g \in Q_1$, c'est à dire l'ensemble des fonctions de S dans $\mathbb{R}$ de la forme :

$$(x,y) \mapsto \alpha + \beta + \gamma(2y-1) + \delta(2y-1),$$

avec α, β, γ, et $\delta \in \mathbb{R}$. On en déduit que $P_{e_1}|_S$ est l'ensemble des fonctions de S dans $\mathbb{R}$ de la forme $(x,y) \mapsto a+by$ avec $a,b \in \mathbb{R}$. On a donc $P_{e_1}|_S = P_{e_2}|_S$. Ceci donne la condition (4.2.6) page 163. Enfin, la condition (4.2.7) est bien vérifiée, c'est à dire $(\Sigma_1,P_{e_1}|_S)$ est unisolvant, car un élément de $P_{e_1}|_S$ est bien déterminé de manière unique par ses valeurs en (0,1) et (1,0).

Correction de l'exercice 43 page 189(Eléments affine–équivalents)

Si les fonctions de base de $(\bar{K},\bar{\Sigma},\bar{P})$ sont affines, alors l'espace $\bar{P}$ est constitué des fonctions affines, on peut donc écrire.

$$\bar{P} = \{\bar{f} : \bar{K} \to \mathbb{R},\bar{x} = (\bar{x}_1,\bar{x}_2)^t \mapsto f(\bar{x}) = a_1\bar{x}_1 + a_2\bar{x}_2 + b\}.$$

Comme $(\bar{K},\bar{\Sigma},\bar{P})$ et $((K,\Sigma,P)$ sont affines équivalents, on a par définition :

$$P = \{f : K \to \mathbb{R}; f = \bar{f} \circ F^{-1},\bar{f} \in \bar{P}\},$$

où F est une fonction affine de $\bar{K}$ dans K la fonction F^{-1} est donc aussi affine et s'écrit donc sous la forme :

$$F^{-1}(x) = F^{-1}((x_1,x_2)^t) = (\alpha_1x_1 + \alpha_1x_2 + +\gamma,\beta_1x_1 + \beta_2x_2 + \delta)^t$$

Donc si $f = \bar{f} \circ F^{-1} \in P$, on a

$$\begin{aligned} f(x) &= \bar{f} \circ F^{-1}((x_1,x_2)^t) \\ &= \bar{f}[(\alpha_1x_1 + \alpha_2x_2 + \gamma,\beta_1x_1 + \beta_2x_2 + \delta)^t] \\ &= A_1x_1 + A_2x_2 + B \end{aligned}$$

où A_1,A_2 et $B \in \mathbb{R}^2$. On en déduit que f est bien affine. L'espace P est donc constitué de fonctions affines. Pour montrer que les fonctions de base locales sont affines, il suffit de montrer que l'espace P est constitué de toutes les fonctions affines. En effet, si f est affine, i.e. $f(x_1,x_2) = A_1x_1 + A_2x_2 + B$, avec $A_1,A_2,B \in \mathbb{R}^2$, on montre facilement que $\bar{f} : f \circ F \in \bar{P}$, ce qui montre que $f \in P$.

Corrigé de l'exercice 44 page 190

1) Pour obtenir une formulation faible, on considère une fonction test $\varphi \in C^2([0,1],\mathbb{R})$; on multiplie la première équation de (2.5.23) par φ et on intègre par partie, on obtient alors :

$$u'(1)\varphi(1) - u'(0)\varphi(0) + \int_0^1 (u'(x)\varphi'(x) + u(x)\varphi(x))dx = \int_0^1 x^2\varphi(x)dx.$$

Comme $u'(1) = 0$, si on choisit $\varphi \in H = \{v \in H^1(]0,1[; v(0) = 0\}$, cette dernière égalité s'écrit :

$$\int_0^1 [u'(x)\varphi'(x) + u(x)\varphi(x)]\,dx = \int_0^1 x^2\varphi(x)\ dx - \varphi(1).$$

En posant

$$a(u,v) = \int_0^1 (u'(x)v'(x) + u(x)v(x))dx \text{ et } L(v) = \int_0^1 x^2v(x)dx,$$

on obtient la formulation variationnelle suivante :

$$\left\{ \begin{array}{l} u \in H \\ a(u,v) = L(v) \quad \forall v \in H \end{array} \right. \tag{4.8.61}$$

On a donc montré que si u est solution de (3.4.39), alors u est solution de (4.8.61).
Montrons maintenant la réciproque : Soit $u \in C^2([0,1])$ solution de (4.8.61) ; en intégrant par partie, il vient :

$$-u'(1)\varphi(1) + u'(0)\varphi(0) + \int_0^1 -u''(x)\varphi(x) + u(x)\varphi(x)dx = \int_0^1 x^2(x)\varphi(x)dx - \varphi(1).$$

Comme $\varphi(0) = 0$, on obtient donc :

$$\varphi(1)(-u'(1)+1) + \int_0^1 (-u''(x)\varphi(x) + u(x)\varphi(x))dx = \int_0^1 x^2\varphi(x)dx \tag{4.8.62}$$

Si on choisit φ à support compact, on obtient :

$$\int_0^1 (-u''(x) + u(x) + x^2)\varphi(x)dx = 0,$$

et comme cette égalité est vraie pour toute fonction φ à support compact sur [0,1], on en déduit que

$$-u'' + u = x^2 \quad \text{p.p.}$$

En tenant compte de cette relation dans (4.8.62), on a donc

$$(u'(1)+1)\varphi(1) = 0,$$

et comme ceci est vrai pour toute fonction $\varphi \in C^2([0,1],\mathbb{R})$, on en déduit que $u'(1) = 1$. Donc u est solution de (4.8.61).

2) Pour montrer que le problème (4.8.61) admet une unique solution, on va montrer que les hypothèses du théorème de Lax-Milgram sont satisfaites.

3) La forme bilinéaire a est en fait le produit scalaire sur H^1. Elle est donc évidemment continue et coercive sur H^1. Donc le théorème de Lax-Milgram s'applique (d'ailleurs, le théorème de Riesz suffit).

4) On cherche des fonctions de base dans l'espace $P_2 = \{P : x \mapsto ax^2 + bx + c, a,b,c \in \mathbb{R}\}$ des polynômes de degré 2 sur $\mathbb{R}$. Sur l'élément de référence [0,1], on considère les noeuds $a_1 = 0$,$a_2 = \dfrac{1}{2}$ et $a_3 = 1$, et les degrés de liberté des trois fonctions de base associées à ces noeuds sont les valeurs aux noeuds. Soient ϕ_1,ϕ_2,ϕ_3 les fonctions de base locales associées aux noeuds a_1,a_2,a_3, on a donc

$$\begin{aligned}
\phi_1(x) &= 2(x - \frac{1}{2})(x-1),\\
\phi_2(x) &= 4x(1-x),\\
\phi_3(x) &= 2x(x - \frac{1}{2}).
\end{aligned}$$

Donnons nous maintenant un maillage de [0,1], défini par

$$\begin{aligned}
&x_0 = 0,\\
&x_{i/2} = \frac{ih}{2}, \quad i = 1,\ldots,N,\\
&x_i = ih, \quad i = 1,\ldots,N.
\end{aligned}$$

On a donc un noeud lié ($x_0 = 0$) et $2N$ noeuds libres. Par "recollement" des fonctions de base locales, on obtient l'expression des fonctions de base globales : Pour $i = 1$ à N, on a :

$$\phi_i(x) = \begin{cases}
\dfrac{2}{h^2}(x - x_{i-\frac{1}{2}})(x - x_{i-1}) \text{ si } x \in [x_{i-1},x_i]\\
\dfrac{2}{h^2}(x - x_{i-\frac{1}{2}})(x - x_{i+1}) \text{ si } x \in [x_i,x_{i+1}]\\
0 \text{ sinon .}
\end{cases}$$

et

$$\phi_{i+\frac{1}{2}}(x) = \begin{cases} -\dfrac{4}{h^2}(x-x_i)(x-x_{i+1}) \text{ si } x \in [x_{i-1},x_i] \\ \\ 0 \text{ sinon .} \end{cases}$$

Notons que ces fonctions de forme ne sont pas de classe C^1. Remarquons que $Supp\ \phi_i = [x_{i-1},x_{i+1}]$ et $Supp\ \phi_{i+\frac{1}{2}} = [x_i,x_{i+1}]$, pour$i = 1,\ldots,N-1$. On en déduit que

$$a(\phi_i,\phi_{j+\frac{1}{2}}) = 0 \text{ si } j \neq i \text{ ou } j \neq i-1$$

$$a(\phi_i,\phi_j) = 0 \text{ si } j > i+1 \text{ ou } j < i-1$$

$$a(\phi_{i+\frac{1}{2}},\phi_{j+\frac{1}{2}}) = 0 \text{ dès que } j \neq i$$

Pour obtenir le système linéaire à résoudre, on approche H par H_N où $H_N = Vect\left\{\phi_i,\phi_{i+\frac{1}{2}}, i = 1,\ldots,N,\right\}$ dans la formulation faible (4.8.61), et on développe u sur la base des fonctions $(\phi_i, \quad \phi_{i+\frac{1}{2}})_{i=1,\ldots,N}$. On obtient donc :

$$\begin{cases} u = \displaystyle\sum_{i=1}^{N} u_i\phi_i + \sum_{i=1}^{N} u_{i+\frac{1}{2}}\phi_{i+\frac{1}{2}} \in H_N \\ \\ \displaystyle\sum_{j=1}^{N} u_j a(\phi_j,\phi_i) + \sum_{j=1}^{N} u_{j+\frac{1}{2}} a(\phi_{j+\frac{1}{2}},\phi_i) = L(\phi_i) i = 1,\ldots,N \\ \\ \displaystyle\sum_{j=1}^{N} u_j a(\phi_j,\phi_{i+\frac{1}{2}}) + \sum_{j=1}^{N} u_{j+\frac{1}{2}} a(\phi_{j+\frac{1}{2}},\phi_{i+\frac{1}{2}}) = L(\phi_{i+\frac{1}{2}}) i = 1,\ldots,N \end{cases}$$

Si on "range" les inconnues discrètes dans l'ordre naturel $\frac{1}{2},1,\frac{3}{2}\ldots i,i+\frac{1}{2},i+1\ldots,N$, on obtient un système linéaire pentadiagonal, de la forme : $AU = b$ avec

$$U = (u_{\frac{1}{2}},u_1,u_{\frac{3}{2}},\ldots,u_i,u_{i+\frac{1}{2}},u_{i+1},\ldots,u_N)^t,$$

$$b = (b_{\frac{1}{2}},b_1,b_{\frac{3}{2}},\ldots,b_i,b_{i+\frac{1}{2}},b_{i+1},\ldots,b_N)^t,$$

où

$$b_\alpha = \int_0^1 x^2\phi_\alpha(x)dx \qquad \text{si } \alpha < N$$

$$b_N = \int_0^1 x^2\phi_N(x)dx - 1,$$

et A est une matrice pentadiagonale (5 diagonales non nulles) de coefficients

$$A_{\alpha\beta} = a(\phi_\alpha,\phi_\beta) = \int_0^1 (\phi_\alpha(x)\phi_\beta(x) + \phi'_\alpha(x)\phi'_\beta(x))dx, \text{ avec } \alpha,\beta = \frac{1}{2},1,\ldots,N-\frac{1}{2},N.$$

Corrigé de l'exercice 46 page 190

1. Soit K le triangle de référence, de sommets (0,0), (1,0) et (0,1). On veut montrer que si p est un polynôme de degré 1, alors

$$\int\int_K p(x,y)\ dxdy = \int\int_K dxdy\, p(x_G,y_G) \tag{4.8.63}$$

où (x_G,y_G) est le centre de gravité de K. Comme K est le triangle de sommets (0,0), (1,0) et (0,1), on a $x_G = y_G = \frac{1}{3}$. Pour montrer (4.8.63), on va le montrer pour $p \equiv 1$, pour $p(x,y) = x$ et pour $p(x,y) = y$. On a

$$\int\int_K dxdy = \int_0^1 \int_0^{1-x} dydx = \frac{1}{2}.$$

On a donc bien (4.8.63) si $p \equiv 1$. Et

$$\int\int_K x\ dxdy = \int_0^1 x \int_0^{1-x} dydx = \int_0^1 (x - x^2)\ dx = \frac{1}{6}$$

Or si $p(x,y) = x$, on a $p(x_G,y_G) = \frac{1}{3}$, et donc on a encore bien (4.8.63). Le calcul de $\int\int_K y\ dxdy$ est identique ; on a donc bien montré que l'intégration numérique à un point de Gauss est exacte pour les polynômes d'ordre 1.

2. On veut montrer que pour tout polynôme p de degré 2, on a :

$$\int\int_K p(x,y)dxdy = L(p), \text{ où on a posé } L(p) = \frac{1}{6}\left(p\left(\frac{1}{2},0\right) + p\left(\frac{1}{2},\frac{1}{2}\right) + p\left(0,\frac{1}{2}\right)\right) \tag{4.8.64}$$

On va démontrer que (4.8.64) est vérifié pour tous les monômes de P_2. Si $p \equiv 1$, on a $L(p) = \frac{1}{2}$, et (4.8.64) est bien vérifiée. Si $p(x,y) = x$, on a $L(p) = \frac{1}{6}$, et on a vu à la question 1 que $\int\int_K xdxdy = \frac{1}{6}$. On a donc bien (4.8.64). Par symétrie, si $p(x,y) = y$ vérifie aussi (4.8.64). Calculons maintenant $I = \int\int_K xydxdy = \int_0^1 \times \int_0^{1-x} ydydx$. On a donc

$$I = \int_0^1 \times \frac{(1-x)^2}{2}dx = \frac{1}{2}\int_0^1 (x - 2x^2 + x^3)dx = \frac{1}{24},$$

et si $p(x,y) = xy$, on a bien : $L(p) = \frac{1}{6} \times \frac{1}{4}$. Donc (4.8.64) est bien vérifiée. Il reste à vérifier que (4.8.64) est vérifiée pour $p(x,y) = x^2$ (ou $p(x,y) = y^2$, par symétrie). Or, $J = \int\int_K x^2 dxdy = \int_0^1 x^2 \int_0^{1-x} dydx = \int_0^1 (x^2 - x^3)dx$. Donc $J = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$. Et pour $p(x,y) = x^2$, on a bien : $L(p) = \frac{1}{6}\left(\frac{1}{4} + \frac{1}{4}\right) = \frac{1}{12}$.

Corrigé de l'exercice 47 page 190

I. Comme $p \in P_2$, p est de la forme : $p(x,y) = a + bx + cy + dxy + \alpha x^2 + \beta y^2$, on a par développement de Taylor (exact car $p''' = 0$) :

$$2p(a_9) - p(a_6) - p(a_8) = p_{xx}(a_9) = \alpha$$

$$2p(a_9) - p(a_5) - p(a_7) = p_{yy}(a_9) = \beta$$

d'où on déduit que

$$4p(a_9) - \sum_{i=5}^{8} p(a_i) = \alpha + \beta. \tag{4.8.65}$$

De même, on a :

$$2p(a_5) - p(a_1) - p(a_2) = \alpha$$
$$2p(a_7) - p(a_3) - p(a_4) = \alpha$$
$$2p(a_6) - p(a_2) - p(a_3) = \beta$$
$$2p(a_8) - p(a_1) - p(a_4) = \beta.$$

Ces quatre dernières égalités entraînent :

$$\sum_{i=5}^{8} p(a_i) - \sum_{i=1}^{4} p(a_i) = \alpha + \beta \tag{4.8.66}$$

De (4.8.65) et (4.8.66), on déduit que :

$$\sum_{i=1}^{4} p(a_i) - 2\sum_{i=5}^{8} p(a_i) + 4p(a_9) = 0.$$

2. La question précédente nous suggère de choisir $\phi : Q_2 \to \mathbb{R}$ définie par

$$\phi(p) = \sum_{i=1}^{4} p(a_i) - 2\sum_{i=5}^{8} p(a_i) + 4p(a_9).$$

Soit $p \in \mathcal{P}$ tel que $p(a_i) = 0$, $i = 1, \ldots, 8$. Comme $p \in Q_2$, p est une combinaison linéaire des fonctions de base $\varphi_1, \ldots, \varphi_9$, associées aux noeuds $a_1, \ldots, a_9$, et comme $p(a_i) = 0$, $i = 1, \ldots, 8$, on en déduit que $p = \alpha\varphi_9$, $\alpha \in \mathbb{R}$. On a donc $\phi(p) = \alpha\phi(\varphi_9) = 4\alpha = 0$, ce qui entraîne $\alpha = 0$. On a donc $p = 0$.
3. Calculons les fonctions de base $\varphi_1^*, \ldots, \varphi_8^*$ associées aux noeuds $a_1, \ldots, a_8$ qui définissent Σ. On veut que $\varphi_i^* \in \mathcal{P}$ et $\varphi_i^*(a_j) = \delta_{ij}$ pour $i,j = 1, \ldots, 8$. Or $\varphi_9(a_j) = 0 \quad \forall i = 1, \ldots, 8$, et $\phi(\varphi_9) = 4$. Remarquons alors que pour $i = 1$, à 4 on a

$$p(\varphi_i) = 1, \text{ et donc si } \varphi_i^* = \varphi_i - \frac{1}{4}\,\varphi_9,$$

on a $p(\varphi_i^*) = 0$ et $\varphi_i^*(a_j) = \delta_{ij}$ pour $j = 1, \ldots, 8$. De même, pour $i = 5$ à 8, on a $p(\varphi_i) = -2$, et donc si $\varphi_i^* = \varphi_i + \frac{1}{2}\varphi_9$, on a $p(\varphi_i^* = 0$ et $\varphi_i^*(a_j) = \delta_{ij}$, pour $j = 1, \ldots, 8$. On a ainsi trouvé les fonctions de base de l'élément fini $(C,\mathcal{P},\Sigma)$. Notons que cet élément fini n'est autre que l'élément fini (C,Q_2^*,Σ) vu en cours (voir paragraphe 4.3.3 page 169 et que Ker $\phi = \mathcal{P} = Q_2^*$.

Corrigé de l'exercice 48 page 191

Corrigé en cours d'élaboration.

Corrigé de l'exercice 49 page 191

La formulation faible du problème s'écrit :

$$\begin{cases} \displaystyle\int_D \nabla u(x)\nabla v(x)dx = \int_D f(x)v(x)dx, \forall v \in H_0^1(\Omega) \\ u \in H_0^1(\Omega) \end{cases}$$

On note $I = \{(k,\ell),1 \leq k \leq M,1 \leq \ell \leq N\}$ noter que Card $I = MN$). L'espace vectoriel de dimension finie dans lequel on cherche la solution approchée (en utilisant les éléments finis suggérés par l'énoncé) est donc $H = Vect\ \{\phi_i,i \in I\}$, où ϕ_i est la fonction de base globale associée au noeud i. Cette solution approchée s'écrit $u = \sum_{j\in I} u_j\phi_j$ où la famille $\{u_j,j \in I\}$ est solution du système linéaire :

$$\sum_{j\in I} a_{ij}u_j = b_i,\forall i \in I \tag{4.8.67}$$

avec $b_i = \int_D f(x,y)\phi_i(x,y)dxdy$, pour tout $i \in I$ et $a_{ij} = \int_D \nabla\phi_i(x,y)\nabla\phi_j(x,y)dxdy$, pour tout $i.j \in I$. La matrice de ce système linéaire est donc donnée par le calcul de a_{ij} pour $i,j \in I$ et un ordre de numérotation des inconnues, plus précisément, soit $\varphi : I \to \{1,\ldots,MN\}$ bijective. On note ψ la fonction réciproque de φ. Le système (4.8.67) peut alors s'écrire :

$$\sum_{n=1}^{MN} a_{i,\psi(n)}u_{\psi(n)} = b_i, \forall i \in I$$

ou encore :

$$\sum_{n=1}^{MN} a_{\psi(m),\psi(n)}u_{\psi(n)} = b_{\psi(m)},\forall m \in \{1,\ldots,MN\},$$

$\{u_j,j \in I\}$ est donc solution de (4.8.67) si et seulement si $u_{\psi(n)} = \lambda_n$ pour tout $n \in \{1,\ldots,MN\}$ où $\lambda = (\lambda_1,\ldots,\lambda_{MN})^t \in \mathbb{R}^{MN}$ est solution du système linéaire :

$$A\lambda = C$$

avec $C = (C_1,\ldots,C_{MN})^t$, $C_m = b_{\psi(m)}$ pour tout $m \in \{1,\ldots,MN\}$ et $A = (A_{m,n})_{m,n=1}^{MN} \in \mathbb{R}^{MN}$ avec $A_{m,n} = a_{\psi(m),\psi(n)}$ pour tout $m,n \in \{1,\ldots,MN\}$. Il reste donc à calculer a_{ij} pour $i,j \in I$. Un examen de support des fonctions ϕ_i et ϕ_j et le fait que le maillage soit à pas constant nous montrent que seuls 4 nombres différents peuvent apparaitrent dans la matrice :

1. $i = j$. On pose alors $a_{ii} = \alpha$.
2. $i = (k,\ell),j = (k \pm 1,\ell)$. On pose alors $a_{ij} = \beta$.
3. $i = (k,\ell),j = (k,\ell \pm 1)$. On pose alors $a_{ij} = \gamma$.
4. $i = (k,\ell),j = (k+1,\ell+1)$ ou $(k-1,\ell-1)$. On pose alors $a_{ij} = \delta$.

En dehors des quatre cas décrits ci-dessus, on a nécessairement $a_{ij} = 0$ (car les supports de ϕ_i et ϕ_j sont disjoints). Calculons maintenant α,β,γ et δ.

Calcul de β On prend ici $i = (k,\ell)$ et $j = (k+1,\ell)$ On calcule tout d'abord $\int_{T^0} \nabla\phi_i - \nabla\phi_j dx$ avec $T^0 = T^0_{k+\frac{1}{2},j+\frac{1}{2}}$. Un argument d'invariance par translation permet de supposer que $x_k = y_\ell = 0$. On a alors

$$\phi_i(x,y) = \frac{\Delta x - x}{\Delta x} \text{ et } \phi_j(x,y)\frac{x\ \Delta y - y\ \Delta x}{\Delta x\ \Delta y},$$

de sorte que

$$\nabla\phi_i(x,y) - \nabla\phi_j(x,y) = -\left(\frac{1}{\Delta x}\right)^2.$$

On a donc

$$\int_{T^0} \nabla\phi_i - \nabla\phi_j dx = -\left(\frac{1}{\Delta x}\right)^2 \frac{\Delta x\ \Delta y}{2} = -\frac{\Delta y}{2\Delta x}$$

Un calcul similaire donne l'intégrale de $\nabla\phi_i \cdot \nabla\phi_j$ sur le deuxième triangle commun aux supports de ϕ_i et ϕ_j. Sur ce deuxième triangle, formé par les points (k,ℓ),$k+1,\ell)$ et $(k,\ell-1)$, noté T^2, on a

$$\phi_i(x,y) = 1 - \frac{x\ \Delta y - y\ \Delta x}{\Delta x\ \Delta y} \text{ et } \phi_j(x,y) = \frac{x}{\Delta x},$$

de sorte que

$$\nabla\phi_i(x,y) \cdot \nabla\phi_j(x,y) = -\left(\frac{1}{\Delta x}\right)^2 \text{ et } \int_{T^2} \nabla\phi_i(x,y) \cdot \nabla\phi_j(x,y)\ dxdy = -\left(\frac{1}{\Delta x}\right)^2 \frac{\Delta x\ \Delta y}{2} = -\frac{\Delta y}{2\Delta x}.$$

On a donc, finalement,

$$\beta = \int_D \nabla\phi_i(x,y) \cdot \nabla\phi_j(x,y)\ dxdy = -\frac{\Delta y}{\Delta x}.$$

Calcul de γ Le calcul de γ est le même que celui de β en changeant les rôles de Δx et Δy, on obtient donc

$$\gamma = -\frac{\Delta x}{\Delta y}$$

Calcul de δ On prend ici $i = (k,\ell)$ et $j = (k+1,\ell+1)$. On a donc, en notant $T^0 = T^0_{k+\frac{1}{2},\ell+\frac{1}{2}}$ et $T^1 = T^1_{k+\frac{1}{2},\ell+\frac{1}{2}}$,

$$\delta = \int_{T^0} \nabla\phi_i(x,y) \cdot \nabla\phi_j(x,y)\ dxdy + \int_{T^1} \nabla\phi_i(x,y) \cdot \nabla\phi_j(x,y)\ dxdy.$$

On peut supposer (par translation) que $x_k = 0 = y_\ell$. Sur T_1 , on a alors $\phi_i(x,y) = \dfrac{\Delta y - y}{\Delta y}$ et $\phi_j(x,y) = \dfrac{x}{\Delta x}$ de sorte que $\displaystyle\int_{T^1} \nabla\phi_i(x,y) \cdot \nabla\phi_j(x,y)\ dxdy = 0$ (car $\nabla\phi_i \cdot \nabla\phi_j = 0$). En changeant les rôles de x et y, on a aussi $\displaystyle\int_{T^0} \nabla\phi_i(x,y) \cdot \nabla\phi_j(x,y)\ dxdy = 0$. On a donc $\delta = 0$.

Calcul de α On prend ici $i = j = (k,\ell)$. On peut toujours supposer que $x_k = y_\ell = 0$ En reprenant les notations précédentes, on a, par raison de symétrie :

$$\alpha = \int_D \nabla\phi_i - \nabla\phi_i dx = 2\int_{T^0} |\nabla\phi_i|^2(x,y)\ dxdy + 2\int_{T^1} |\nabla\phi_i|^2(x,y)\ dxdy + 2\int_{T^2} |\nabla\phi_i|^2(x,y)\ dxdy.$$

Sur T^0, on a $\phi_i(x,y) = \dfrac{\Delta x - x}{\Delta x}$ et donc

$$\int_{T^0} |\nabla\phi_i|^2(x,y)\ dxdy = \left(\frac{1}{\Delta x}\right)^2 \frac{\Delta x \Delta y}{2} = \frac{1}{2}\ \frac{\Delta y}{\Delta x}$$

Sur T_1, on a $\phi_1(x,y) = \dfrac{\Delta y - y}{\Delta y}$ et donc

$$\int_{T^2} |\nabla \phi_i|^2(x,y)\ dxdy = \left[\left(\frac{1}{\Delta x} \right)^2 + \left(\frac{1}{\Delta y} \right)^2 \right] \frac{\Delta x\ \Delta y}{2} = \frac{1}{2}\ \frac{\Delta y}{\Delta x} + \frac{1}{2}\ \frac{\Delta x}{\Delta y}$$

On en déduit

$$\alpha = 2\frac{\Delta x}{\Delta y} + 2\frac{\Delta y}{\Delta x}.$$

Chapitre 5

Méthodes de volumes finis pour les problèmes hyperboliques

5.1 Exemple

L'exemple type d'équation hyperbolique est obtenu lorsqu'on modélise un phénomène de transport. Supposons par exemple, qu'on connaisse l'emplacement d'une nappe de pétrole due au dégazement intempestif d'un supertanker au large des côtes, et qu'on cherche à prévoir son déplacement dans les heures à venir, par exemple pour la mise en oeuvre efficace de barrages. On suppose connu $v : \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}^2$, le champ des vecteurs vitesse des courants marins, qui dépend de la variable d'espace x et du temps t; ce champ de vecteurs est donné par exemple par la table des marées (des exemples de telles cartes de courants sont données en Figure 5.1). A $t = 0$, on connaît $\rho_0(x)$: la densité d'hydrocarbure initiale, et on

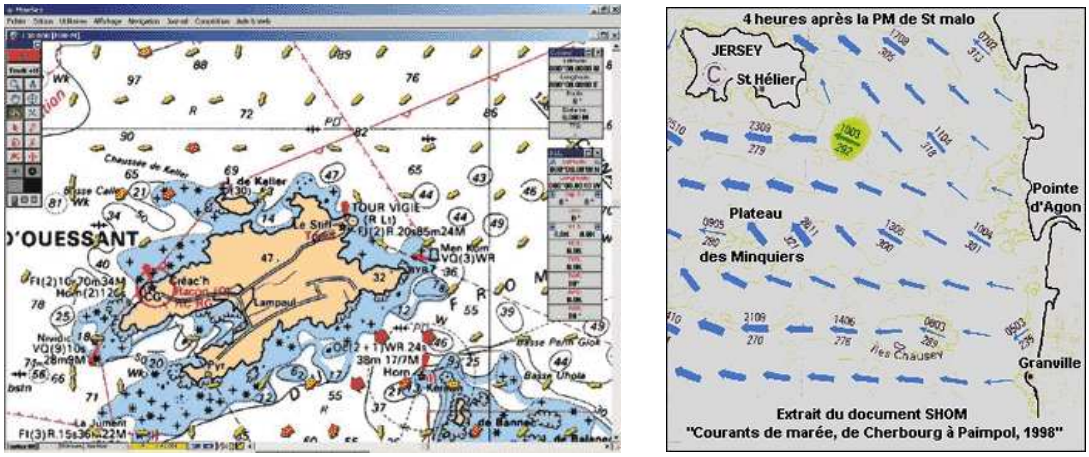

FIG. *5.1* – Exemples de cartes de courants marins au large de côtes de Bretagne (source: SHOM)

cherche à calculer $\rho(x,t)$ = densité de d'hydrocarbure au point x et au temps t. On écrit alors l'équation

de conservation de la masse :

$$\rho_t + div(\rho v) = 0, \tag{5.1.1}$$

$$\rho_0(x) = \begin{cases} 1 & x \in A, \\ 0 & x \in A^c, \end{cases} \tag{5.1.2}$$

où A représente le lieu initial de la nappe de pétrole. Dans le cas d'un déplacement maritime, le vecteur $v : \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}^2$, n'est évidemment pas constant (la marée n'est pas la même à Brest qu'à Saint Malo). De plus le déplacement de la nappe dépend également du vent, qui affecte donc le vecteur v. On supposera pourtant ici, pour simplifier l'exposé, que v soit constant en espace et en temps. Alors le problème (5.1.1) - (5.1.2) admet comme solution :

$$\rho(x,t) = \rho_0(x - vt), \tag{5.1.3}$$

qui exprime le transport de la nappe à la distance vt du point de départ dans la direction de V, au temps t. En fait, il est clair que (5.1.3) n'est pas une solution "classique" de puisque ρ n'est pas continue, et que ces dérivées en temps ne sont donc pas définies au sens habituel. Nous verrons par la suite comment on peut donner une formulation correcte des solutions de (5.1.1) - (5.1.2). Notons que les systèmes d'équations hyperboliques sont très importants en mécanique des fluides ; les équations d'Euler, par exemple sont utilisées pour modéliser pour modéliser l'écoulement de l'air autour d'une aile d'avion. Dans le cadre de ce cours, nous n'étudierons que le cas des équations scalaires. Par souci de simplicité, nous n'aborderons dans le cadre de ce cours que les problèmes posés en une dimension d'espace, tout d'abord dans le cas relativement simple d'une équation linéaire (section 5.2 page 215, puis dans le cas d'une équation non linéaire (section 5.4 page 223. Par souci de clarté, les schémas numériques seront introduits pour l'équation linéaire $u_t + u_x = 0$ (section 5.3 page 219. On donnera ensuite quelques schémas pour les équations hyperboliques non linéaires (section 5.5 page 233.

5.2 Equation hyperbolique linéaire en une dimension d'espace

Commençons par étudier le cas d'une équation hyperbolique linéaire :

$$\begin{cases} u_t + cu_x = 0, x \in \mathbb{R}, t > 0, \\ u(x,0) = u_0(x), \quad x \in \mathbb{R}. \end{cases} \tag{5.2.4}$$

où la vitesse de transport $c \in \mathbb{R}$ et la condition initiale $u_0 : \mathbb{R} \to \mathbb{R}$ sont données. Le problème (5.2.4) s'appelle "problème de Cauchy". On cherche $u : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$, solution de ce problème. Nous commençons par une étude succinte du problème continu, pour lequel on peut trouver une solution exacte explicite.

Solution classique et solution faible

Définition 5.1 (Solution classique) *On dit qu'une fonction $u : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ est solution classique du problème (5.2.4) si $u \in C^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R})$ et u vérifie (5.2.4).*

Une condition nécessaire pour avoir une solution classique est que $u_0 \in C^1(\mathbb{R})$.

Théorème 5.2 *Si $u_0 \in C^1(\mathbb{R})$, alors il existe une unique solution classique du problème (5.2.4), qui s'écrit $u(x,t) = u_0(x - ct)$.*

Démonstration : Pour montrer l'existence de la solution, il suffit de remarquer que u définie par (5.1) est de classe C^1, et que $u_t + cu_x = 0$ en tout point. Pour montrer l'unicité de la solution, on va introduire la notion de caractéristique, qui est d'ailleurs aussi fort utile dans le cadre de la résolution numérique. Soit u solution classique de (5.2.4). On appelle droite caractéristique de (5.2.4) issue de x_0 la droite d'équation $x(t) = ct + x_0$, qui est illustrée sur la figure 5.2. Montrons que si u est solution de (5.2.4), alors u est constante sur la droite $\mathcal{D}_{x_0}$, pour tout $x_0 \in \mathbb{R}$. Soit $x_0 \in \mathbb{R}$, et φ_{x_0} la fonction de $\mathbb{R}_+$ dans $\mathbb{R}$ définie

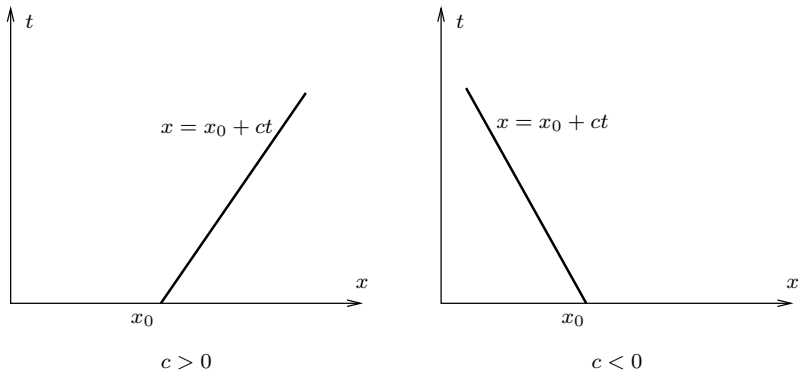

FIG. 5.2 – *Droites caractéristiques, cas linéaire*

par $\varphi_{x_0}(t) = u(x_0 + ct,t)$. Dérivons φ_{x_0} par rapport au temps :

$$\begin{aligned} \varphi'_{x_0}(t) &= cu_x(x_0 + ct,t) + u_t(x_0 + ct,t) \\ &= (u_t + cu_x)(x_0 + ct,t) = 0, \end{aligned}$$

car u est solution de (5.2.4). On en déduit que

$$\varphi_{x_0}(t) = \varphi_{x_0}(0) = u_0(x_0), \forall t \in \mathbb{R}_+.$$

On a donc $u(x_0 + ct,t) = u_0(x_0), \forall x_0 \in \mathbb{R}$, donc u est constante sur la droite caractéristique $\mathcal{D}_{x_0}$, et en posant $x = x_0 + ct$:

$$u(x,t) = u_0(x - ct),$$

ce qui prouve l'existence et l'unicité de (5.2.4). ∎

Remarque 5.3 (Terme source) *Le modèle physique peut amener à une équation avec terme source au second membre $f \in C(\mathbb{R} \times \mathbb{R}_+,\mathbb{R})$:*

$$\begin{cases} u_t + cu_x = f(x,t), \\ u(x,0) = u_0(x), \end{cases} \tag{5.2.5}$$

et $u_0 \in C(\mathbb{R})$. Ceci peut modéliser un dégazage sur un temps plus long, comme dans le cas du Prestige sur les côtes de Galice en 2003 par exemple. Pour montrer l'unicité de la solution de (5.2.5), on suppose que u est solution classique et on pose : $\varphi_{x_0}(t) = u(x_0 + ct,t)$. Par un calcul identique au précédent, on a

$$\varphi'_{x_0}(t) = f(x_0 + ct,t).$$

Donc $\varphi_{x_0}(t) = \varphi_{x_0}(0) + \int_0^t f(x_0 + cs,s)ds$ *On en déduit que :*

$$u(x_0 + ct,t) = \varphi_{x_0}(0) + \int_0^t f(x_0 + cs,s)ds.$$

on pose alors : $x = x_0 + ct$*, et on obtient :*

$$u(x,t) = u_0(x - ct) + \int_0^t f(x - c(t - s),s)ds,$$

ce qui prouve l'unicité. On obtient alors l'existence en remarquant que la fonction $u(x,t)$ *ainsi définie est effectivement solution de (5.2.5), car elle est de classe* C^1 *et elle vérifie* $u_t + cu_x = f$*.*

Dans ce qui précède, on a fortement utilisé le fait que u_0 est C^1. Ce n'est largement pas toujours le cas dans la réalité. Que faire si, par exemple, $u_0 \in L^\infty(\mathbb{R})$?

Définition 5.4 (Solution faible) *On dit que* u *est solution faible de (5.2.4) si* $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+,\mathbb{R})$ *et* u *vérifie :*

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} [u(x,t)\varphi_t(x,t) + cu(x,t)\varphi_x(x,t)]\, dtdx + \int_{\mathbb{R}} u_0(x)\varphi(x,0)dx = 0, \forall \varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+,\mathbb{R}). \quad (5.2.6)$$

Notons que dans la définition ci-dessus, on note $\mathbb{R}_+ = [0,+\infty[$, et $C_c^1(\mathbb{R} \times \mathbb{R}_+)$ l'ensemble des restrictions à $\mathbb{R} \times \mathbb{R}_+$ des fonctions $C_c^1(\mathbb{R} \times \mathbb{R})$. On insiste sur le fait qu'on peut donc avoir $\varphi(x,0) \neq 0$. Voyons maintenant les liens entre solution classique et solution faible.

Proposition 5.5 *Si* u *est solution classique de (5.2.4) alors* u *est solution faible. Réciproquement, si* $u \in C^1(\mathbb{R}\times]0, + \infty[) \cap C(\mathbb{R} \times [0, + \infty[)$ *est solution classique de (5.4.17) alors* u *est solution forte de (5.2.4).*

La démonstration de cette proposition est effectuée dans le cadre plus géneŕal des équations hyperboliques non linéaires (voir Proposition 5.20 Par contre, notons que si on prend $\varphi \in C_c^1(\mathbb{R}\times]0, + \infty[,\mathbb{R})$ au lieu de $\varphi \in C_c^1(\mathbb{R} \times [0, + \infty[,\mathbb{R})$ dans (5.2.6), on obtient :

$$u_t + cu_x = 0,$$

mais on ne récupère pas la condition initiale. Il est donc essentiel de prendre des fonctions test dans $C_c^1(\mathbb{R} \times [0, + \infty[,\mathbb{R})$.

Théorème 5.6 (Existence et unicité de la solution faible) *Si* $u_0 \in L^\infty_{loc}(\mathbb{R})$*, il existe une unique fonction* u *solution faible de (5.2.4).*

Démonstration : On va montrer que $u(x,t) = u_0(x - ct)$ est solution faible. Comme $u_0 \in L^\infty_{loc}(\mathbb{R})$, on a $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$. Soit $\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+,\mathbb{R})$, on veut montrer que :

$$\int\int_{\mathbb{R}\times\mathbb{R}_+} u(x,t)\varphi_t(x,t)dxdt + \int\int_{\mathbb{R}\times\mathbb{R}_+} cu(x,t)\varphi_x(x,t)dxdt + \int_{\mathbb{R}} u_0(x)\varphi(x,0)dx = 0.$$

Posons

$$A = \int\int_{\mathbb{R}\times\mathbb{R}_+} u(x,t)\varphi_t(x,t)dxdt + \int\int_{\mathbb{R}\times\mathbb{R}_+} cu(x,t)\varphi_x(x,t)dxdt.$$

Si $u(x,t) = u_0(x - ct)$, on a donc :

$$A = \int\int_{\mathbb{R}\times\mathbb{R}_+} [u_0(x - ct)\varphi_t(x,t) + cu_0(x - ct)\varphi_x(x,t)]\, dxdt.$$

En appliquant le changement de variable $y = x - ct$ et en utilisant le théorème de Fubini, on obtient :

$$A = \int_{\mathbb{R}} u_0(y) \int_{\mathbb{R}_+} [\varphi_t(y + ct,t) + c\varphi_x(y + ct,t)]\, dtdy.$$

Posons alors

$$\psi_y(t) = \varphi(y + ct,t).$$

On a donc :

$$A = \int_{\mathbb{R}} \left(u_0(y) \int_0^{+\infty} \psi'_y(t)dt \right) dy,$$

et comme ψ est à support compact sur $[0, +\infty[$,

$$A = -\int_{\mathbb{R}} u_0(y)\psi_y(0)dy,$$

donc finalement :

$$A = -\int_{\mathbb{R}} u_0(y)\varphi(y,0)dy.$$

On a ainsi démontré que la fonction u définie par $u(x,t) = u_0(x - ct)$ est solution faible de l'équation (5.2.4). On a donc existence d'une solution faible. Montrons maintenant que celle-ci est unique. Soient u et v deux solutions faibles de (5.2.4). On pose $w = u - v$ et on va montrer que $w = 0$. Par définition, w satisfait :

$$\int\int_{\mathbb{R}\times\mathbb{R}_+} w(x,t)(\varphi_t(x,t) + c\varphi_x(x,t))dxdt = 0, \quad \forall \varphi \in C_c^1(\mathbb{R} \times \mathbb{R}^+,\mathbb{R}) \tag{5.2.7}$$

Par le lemme (5.7) donné ci-dessous, pour toute fonction $f \in C_c^\infty(\mathbb{R} \times \mathbb{R}_+^*)$ il existe $\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+,\mathbb{R})$, telle que $\varphi_t + c\varphi_x = f$, et on a donc par (5.2.7) :

$$\int\int_{\mathbb{R}\times\mathbb{R}_+} w(x,t)f(x,t)dxdt = 0, \forall f \in C_c(\mathbb{R} \times \mathbb{R}_+^*,\mathbb{R})$$

Ceci entraîne que $w = 0$ p.p.. ∎

Lemme 5.7 (Résultat d'existence) *Soit $f \in C_c(\mathbb{R} \times \mathbb{R}_+^*,\mathbb{R})$, alors il existe*

$$\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+,\mathbb{R}) \text{ telle que } \varphi_t + c\varphi_x = f$$

Démonstration : Soit $f \in C_c(\mathbb{R} \times \mathbb{R}_+^*,\mathbb{R})$, et $T > 0$ tel que $f(x,t) = 0$ si $t \geq T$. On considère le problème :

$$\begin{cases} \varphi_t + c\varphi_x = f \\ \varphi(x,T) = 0 \end{cases} \tag{5.2.8}$$

On vérifie facilement que le problème (5.2.8) admet une solution classique

$$\varphi(x,t) = -\int_t^T f(x - c(s - t),s)ds$$

En effet, avec ce choix de φ, on a effectivement

$$\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+,\mathbb{R}) \text{ et } \varphi_t + c\varphi_x = f.$$

De plus, comme f est à support compact, φ est à support compact. ∎

Remarque 5.8 (Sur les propriétés de la solution) *Remarquons que la solution faible de (5.2.4) possède les propriétés suivantes :*

1. *Si $u_0 \geq 0$ pp alors $u \geq 0$ p.p.,*
2. $\|u(.,t)\|_{L^p(\mathbb{R})} = \|u_0(x)\|_{L^p(\mathbb{R})} \qquad \forall p \quad p \in [1, +\infty]$.

Lors de l'élaboration de schémas numériques pour la recherche d'une approximation, on s'attachera à vérifier que ces propriétés sont encore satisfaites par la solution approchée.

5.3 Schémas numériques pour $u_t + u_x = 0$

On considère ici le problème de transport linéaire :

$$\begin{cases} u_t + u_x = 0, \\ u(x,0) = u_0(x), u_0 \in L^\infty(\mathbb{R}). \end{cases} \tag{5.3.9}$$

On sait que la solution de ce problème s'écrit :

$$u(x,t) = u_0(x - ct).$$

On rappelle que u est une solution classique si $u \in C^1(\mathbb{R})$, et que u est une solution faible si $u_0 \in L^\infty(\mathbb{R})$. On va chercher à retrouver cette solution par une approximation numérique. Notons que dans le cas linéaire, l'utilisation d'un schéma numérique, n'est évidemment pas utile, mais nous commençons par ce cas par souci pédagogique.

5.3.1 Schéma explicite différences finies centrées

On effectue une discrétisation espace temps en se donnant un pas de discrétisation en espace h, et en posant : $x_i = ih$, $\forall i \in \mathbb{Z}$; de même on se donne un pas de discrétisation en temps k, et on pose $t_n = nk$, $\forall n \in \mathbb{N}$. Ecrivons le schéma d'Euler explicite pour l'approximation de u_t et un schéma centré pour l'approximation de u_x . On approche donc

$$u_t(x_i,t_n) \text{ par } \frac{u(x_i,t_{n+1}) - u(x_i,t_n)}{k}$$

et .

$$u_x(x_i,t_n) \text{ par } \frac{u(x_{i+1},t_n) - u(x_{i-1},t_n)}{2h}.$$

Le schéma centré s'écrit donc, en fonction des inconnues discrètes :

$$\begin{cases} \dfrac{u_i^{n+1} - u_i^n}{k} + \dfrac{u_{i+1}^n - u_{i-1}^n}{2h} = 0, \\ u_i^0 = u_0(x_i). \end{cases} \tag{5.3.10}$$

(où on a supposé $u_0 \in C(\mathbb{R})$) Ce schéma est inconditionnellement instable, et il faut donc éviter de l'utiliser. On peut en effet, montrer que :

1. Le schéma (5.3.10) ne respecte pas la positivité, car $u_0(x) \geq 0 \forall x$ n'entraine pas forcément $u_i^n \geq 0$. En effet si u_0 est telle que
$$\begin{cases} u_i^0 = 0, \forall i \leq 0, \\ u_i^0 = 1, \forall i > 0. \end{cases}$$

Alors :

$$u_i^{n+1} = u_i^n - \frac{k}{2h}(u_{i+1}^n - u_{i-1}^n)$$

donne, pour $n = 0$

$$u_0^1 = -\frac{k}{2h} < 0$$

2. Le schéma (5.3.10) n'est pas L^∞ stable :

$$\|u^n\|_\infty \leq C \text{ n'entraine pas } \|u^{n+1}\|_\infty \leq C.$$

3. Le schéma (5.3.10) n'est pas L^2 stable :

$$\|u^n\|_2 \leq C \text{ n'entraine pas que } \|u^{n+1}\|_2 \leq C.$$

4. Le schéma n'est pas stable au sens de Von Neumann. En effet, si

$$u_0(x) = e^{ipx}, \text{ où } i^2 = -1 \text{ et } p \in \mathbb{Z},$$

la solution exacte est $u(x,t) = e^{ip(x-t)}$. Une discrétisation de u_0 s'écrit :

$$u_j^0 = e^{ipjh} \quad j \in \mathbb{Z}.$$

On a donc :

$$\begin{aligned} u_i^1 &= u_i^0 - \frac{k}{2h}(u_{i+1}^0 - u_{i-1}^0) \\ &= e^{ipjh} - \frac{k}{2h}(e^{ip(j+1)h} - e^{ip(j-1)h}) \\ &= \mathcal{J}_{k,h} u_j^0, \end{aligned}$$

avec $\mathcal{J}_{kh} = 1 - \frac{ik}{h}\sin ph$. On a donc $|\mathcal{J}_{kh}| > 1$ si $\sin ph \neq 0$, ce qui montre que le schéma n'est pas stable au sens de Von Neuman.

5. Le schéma (5.3.10) n'est pas convergent. En effet, on peut montrer qu'il existe u_0,k et h telle que la solution approchée $u_{h,k} : (u_i^n)_{i\in\mathbb{Z}}^{n\in\mathbb{N}}$ ne converge pas vers u lorsque h et k tendent vers 0.

5.3.2 Schéma différences finies décentré amont

On utilise toujours le schéma d'Euler explicite pour la discrétisation en temps, mais on approche maintenant

$$u_x(x_i,t_n) \text{ par } \frac{u(x_i,t_n) - u(x_{i-1},t_n)}{h_{i-1/2}}.$$

On considère de plus un pas de discrétisation variable, défini par $h_{i-1/2} = x_i - x_{i-1}$. Le schéma par

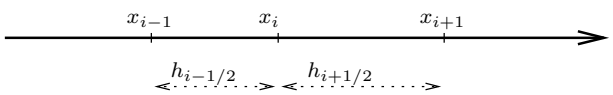

FIG. 5.3 – *Maillage volumes finis*

différences finies avec décentrement amont s'écrit :

$$\begin{cases} \frac{u_i^{n+1}-u_i^n}{k} + \frac{u_i^n - u_{i-1}^n}{h_{i-1/2}} = 0, \\ u(x,0) = u_0(x). \end{cases} . \tag{5.3.11}$$

Proposition 5.9 *Le schéma (5.3.11) est stable sous condition de Courant-Friedrichs-Levy (CFL)*

$$k \leq h = \inf_{i \in \mathbb{Z}} h_{i-1/2} > 0. \tag{5.3.12}$$

c'est à dire que si $A \leq u_i^n \leq B$, *alors* $A \leq u_i^{n+1} \leq B$.

Démonstration : On a : $u_i^{n+1} = u_i^n(1-\alpha_i) + \alpha_i u_{i-1}^n$ avec $\alpha_i = \dfrac{k}{h_{i-1/2}}$. Donc, si la condition (5.3.12) est vérifiée, u_i^{n+1} est une combinaison convexe de u_i^n et u_i^{n+1}, et donc $u_i^{n+1} \in [u_{i-1}^n, u_i^n]$. ∎

Théorème 5.10 (Convergence du schéma (5.3.11)) *On suppose que* $u_0 \in C^2(\mathbb{R},\mathbb{R})$ *et que* u_0, u_0', u_0'' *sont bornées. Soit* $A = \inf_{x \in \mathbb{R}} u_0(x)$ *et* $B = \sup_{x \in \mathbb{R}} u_0(x)$. *Alors :*

1. $A \leq u_i^n \leq B, \forall i \in \mathbb{Z}, \forall n \in \mathbb{N}$.
2. *Soit* $\bar{u}_i^n = u(x_i, t_n)$, *à* u *est la solution exacte de (5.3.9), alors :*

$$\sup_{\substack{i \in \mathbb{Z} \\ nk \leq T}} |u_i^n - \bar{u}_i^n| \leq T C_{u_0}(k+h),$$

où $TC_{u_0} \geq 0$ *ne dépend que de* u_0.

Démonstration : le point 1 se démontre par récurrence sur n à partir de la propriété précédente. Le point 2 (estimation d'erreur) se démontre en remarquant d'abord que l'erreur de consistance

$$\frac{\bar{u}_i^{n+1} - \bar{u}_i^n}{k} + \frac{\bar{u}_i^n - \bar{u}_{i-1}^n}{h_{i-1/2}} = R_i^n$$

vérifie :

$$|R_i^n| \leq C_{u_0}(\bar{h} + k)$$

où $\bar{h} = \max i \in \mathbb{Z}\, h_{i-1/2}$. On a donc :

$$\bar{u}_i^{n+1} = \bar{u}_i^n \left(1 - \frac{k}{h_{i-1/2}}\right) + \frac{k}{h_{i-1/2}} \bar{u}_{i-1}^n + kR_i^n,$$

or le schéma numérique s'écrit :

$$u_i^{n+1} = u_i^n \left(1 - \frac{k}{h_{i-1/2}}\right) + \frac{k}{h_{i-1/2}} u_{i-1}^n$$

Par différence, on obtient :

$$u_i^{n+1} - \bar{u}_i^{n+1} = (u_i^n - \bar{u}_i^n)\left(1 - \frac{k}{k_{i-1/2}}\right) + (u_{i-1}^n - \bar{u}^n i - 1)\frac{k}{h_{i-1/2}} - kR_i^n$$

et donc :

$$|u_i^{n+1} - \bar{u}_i^{n+1}| \leq |u_i^n - \bar{u}_i^n| \left(1 - \frac{k}{h_{i-1/2}}\right) + |u_{i-1}^n - \bar{u}_{i-1}^n| \frac{k}{h_{i-1/2}} + kC_{u_0}(\bar{h} + k) \tag{5.3.13}$$

On effectue alors l'hypothèse de récurrence :

$$\sup |u_i^n - \bar{u}_i^n| \leq (n-1)kC_{u_0}(k + \bar{h}) \tag{5.3.14}$$

grâce à (5.3.13) et (5.3.14), on obtient :

$$|u_i^{n+1} - \bar{u}_i^{n+1}| \leq (n-1)kC_{u_0}(k + \bar{h}) + k(C_{u_0}(k + \bar{h}).$$

Donc finalement :

$$|u_i^{n+1} - \bar{u}_i^{n+1}| \leq TC_{u_0}(k + \bar{h}).$$

Remarque 5.11 (Décentrement) . *Pour une équation de transport telle que (5.3.9), le choix du décentrement est crucial. Ici, on a approché $u_x(x_i)$ par $\frac{u_i - u_{i-1}}{h_{i-1/2}}$. Dans le cas où on étudie une équation de transport de type, $u_t + cu_x = 0$, avec $c \in \mathbb{R}$, le choix décentré amont sera toujours*

$$\frac{u_i - u_{i-1}}{h} \text{ si } c > 0,$$

par contre, si $c < 0$, le choix amont donnera

$$\frac{u_i - u_{i+1}}{h}.$$

Regardons ce qui se passe si l'on effectue un "mauvais" décentrement. Considérons toujours l'équation $u_t + u_x = 0$. Effectuer le "mauvais décentrement" amène au schéma :

$$\frac{u_i^{n+1} - u_i^n}{k} + \frac{u_{i+1}^n - u_i^n}{h} = 0,$$

c'est à dire :

$$u_i^{n+1} = u_i^n \left(1 + \frac{k}{h}\right) - \frac{k}{h} u_{i+1}^n.$$

Examinons le comportement de la solution approchée donnée par le schéma si on prend une condition initiale u_0 telle que $u_0(x) = 0, \forall x \geq 0$. Dans ce cas, on sait que $u(x,t) \neq 0$ pour t assez grand, or après calculs on obtient $u_{-1}^{n+1} = u_{-1}^n \left(1 + \frac{k}{h}\right) + 0 = u_{-1}^0 \left(1 + \frac{k}{h}\right)^n$, alors que $u_i^{n+1} = 0 \quad \forall i \geq 0$. On en déduit que la solution approchée est très mauvaise.

Remarque 5.12

1. *Dans le cas non linéaire, la démonstration précédente de convergence ne s'adapte pas car les solutions ne sont pas régulières.*
2. *On a défini (5.3.11) pour $u_0 \in C(\mathbb{R})$ Si $u_0 \notin C(\mathbb{R})$, on peut prendre comme donnée initiale $u_i^0 = \int_{x_{i-1/2}}^{x_{i+1/2}} u_0(x)dx$.*

5.3.3 Schéma volumes finis décentrés amont

On considère toujours le problème (5.3.9), avec condition initiale $u_0 \in L^\infty(\mathbb{R})$. On se donne une discrétisation en espace, c'est à dire un ensemble de points $(x_{i+1/2})_{i \in \mathbb{Z}}$, tels que $x_{i+1/2} > x_{i-1/2}$, et on

note $h_i = x_{i+1/2} - x_{i-1/2}$. On approche toujours la dérivée en temps par un schéma d'Euler explicite, on intègre (5.3.9) sur la maille $]x_{i-1/2},x_{i+1/2}[$, et on obtient :

$$\int_{x_{i-1/2}}^{x_{i+1/2}} (u_t + u_x)dx = 0.$$

En approchant $u(x_{i+1/2})$ (resp. $u(x_{i-1/2})$) par u_i^n (resp. u_{i-1}^n) et en approchant u_t par un schéma d'Euler explicite, on obtient :

$$\left\{ \begin{array}{l} h_i \dfrac{u_i^{n+1} - u_i^n}{k} + u_i^n - u_{i-1}^n = 0, \\ u_i^0 = \dfrac{1}{h_i} \displaystyle\int_{x_{i-1/2}}^{x_{i+1/2}} u_0(x)dx. \end{array} \right. \tag{5.3.15}$$

Proposition 5.13 *Soit $(u_i^n)_{\substack{n\in\mathbb{N} \\ i\in\mathbb{Z}}}$ la solution de (5.3.15). Si $k \leq h = \min h_i$ et si $A \leq u_0(x) \leq B$, alors $A \leq u_i^n \leq B \qquad \forall i \in \mathbb{Z}, \forall n \in \mathbb{N}$.*

La démonstration est similaire à celle de la proposition (5.9), et laissée à titre d'exercice.

Définition 5.14 (Solution approchée) *Soit $\mathcal{T}$ un maillage volumes finis de $\mathbb{R}$ défini par $\mathcal{T} = (K_i)_{i\in\mathbb{Z}}$ avec $K_i =]x_{i-1/2},x_{i+1/2}[$. On appelle solution approchée de (5.3.9) par le schéma (5.3.15) la fonction $u_{\mathcal{T},k} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$, définie par*

$$u_{\mathcal{T},k}(x,t) = u_i^n \text{ si } x \in K_i \text{ et } t \in [nk,nk+1[\tag{5.3.16}$$

On admettra le théorème de convergence suivant (voir aussi exercice 5.3.11) :

Théorème 5.15 (Convergence du schéma 5.3.15) *Soit $u_0 \in L^\infty(\mathbb{R})$, on suppose que $k \leq h = \inf(h_i)$, alors $u_{\mathcal{T},k}$ converge vers u dans $L^1_{loc}(\mathbb{R} \times \mathbb{R}_+)$ lorsque h (et k) tend vers 0, c'est à dire qu'on a : $\displaystyle\int_C |u_{\mathcal{T},k} - u|dxdt \to 0$ pour tout compact C de $\mathbb{R} \times \mathbb{R}_+$, lorsque h (et k) tend vers 0.*

5.4 Equations hyperboliques non linéaires

On se donne $f \in C^1(\mathbb{R},\mathbb{R})$ et $u_0 \in C(\mathbb{R})$ et on considère maintenant l'équation hyperbolique non linéaire :

$$\left\{ \begin{array}{l} u_t + (f(u))_x = 0, \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+, \\ u(x,0) = u_0(x). \end{array} \right. \tag{5.4.17}$$

Commençons par donner la définition de solution classique de ce problème même si, comme nous le verrons après, celle-ci n'a pas grand intérêt puisque le problème (5.4.17) n'a pas, en général de solution classique.

Définition 5.16 (Solution classique) *On suppose que $u_0 \in C^1(\mathbb{R})$ et $f \in C^2(\mathbb{R},\mathbb{R})$. Alors u est solution classique de (5.4.17) si $u \in C^1(\mathbb{R} \times \mathbb{R}_+,\mathbb{R})$ et u vérifie*

$$\left\{ \begin{array}{l} (u_t + (f(u))_x)(x,t) = 0, \quad \forall (x,t) \in \mathbb{R} \times \mathbb{R}_+, \\ u(x,0) = u_0(x), \qquad \forall x \in \mathbb{R}. \end{array} \right.$$

Avant d'énoncer le théorème de non existence, rappelons que dans le cas d'une équation différentielle du type non linéaire,

$$\left\{ \begin{array}{l} x'(t) = f(x(t)), \quad t \in \mathbb{R}_+, \\ x(0) = x_0, \end{array} \right.$$

si on note $T_{\max}$ le temps d'existence de la solution, et si $T_{\max} < +\infty$ alors $\|x(t)\| \to +\infty$ lorsque $t \to T_{\max}$. Donnons maintenant la définition des courbes caractéristiques de l'équation (5.4.17), qui permet le lien entre les équations hyperboliques non linéaires et les équations différentielles ordinaires.

Définition 5.17 (Courbe caractéristique) *On appelle courbe caractéristique du problème (5.4.17) issue de $x_0 \in \mathbb{R}$, la courbe définie par le problème de Cauchy suivant :*

$$\begin{cases} x'(t) = f'(u(x(t),t)) \\ x(0) = x_0 \end{cases} \tag{5.4.18}$$

Théorème 5.18 (Non existence) *Soit $f \in C^1(\mathbb{R},\mathbb{R})$, on suppose que f' n'est pas constante, alors il existe $u_0 \in C_c^\infty(\mathbb{R})$ telle que (5.4.17) n'admette pas de solution classique.*

Démonstration : Comme $f \in C^2(\mathbb{R},\mathbb{R})$, on a $f' \in C^1(\mathbb{R},\mathbb{R})$, et donc le théorème de Cauchy-Lipschitz s'applique. Il existe donc une solution maximale $x(t)$ définie sur $[0,T_{\max}[$, et $x(t)$ tend vers l'infini lorsque t tend vers $T_{\max}$ si $T_{\max} < +\infty$. Les quatre étapes de la démonstration sont les suivantes :

1. $u(x(t),t) = u_0(x_0)$, $\forall t \in [0,T_{\max}[$, et donc que toute solution de (5.4.17) est constante sur les caractéristiques.
2. Les courbes caractéristiques sont des droites.
3. $T_{\max} = +\infty$ et donc $u(x,t) = u_0(x_0) \quad \forall t \in [0,+\infty[$.
4. On en déduit alors qu'on n'a pas de solution classique de (5.4.17).

Détaillons maintenant ces étapes.

1. Soit φ définie par $\varphi(t) = u(x(t),t)$; en dérivant φ, on obtient : $\varphi'(t) = u_t(x(t),t) + u_x(x(t),t)x'(t)$. Comme x vérifie (5.4.18), ceci entraîne : $\varphi'(t) = u_t(x(t),t) + f'(u(x(t),t))u_x(x(t),t)$, et donc

$$\varphi'(t) = (u_t + (f(u))_x)(x(t),t) = 0.$$

   La fonction φ est donc constante, et on a :

$$u(x(t),t) = \varphi(t) = \varphi(0) = u(x(0),0) = u(x_0,0) = u_0(x_0), \forall t \in [0,T_{\max}[.$$

2. Comme $u(x(t),t) = u_0(x_0), \forall t \in [0,T_{\max}[$, on a donc $x'(t) = f'(u_0(x_0))$. Donc en intégrant, on obtient que le système (5.4.18) décrit la droite d'équation :

$$x(t) = f'(u_0(x_0))t + x_0. \tag{5.4.19}$$

3. Puisque x vérifie (5.4.19), on a donc

   $\lim_{t \to T_{\max}} |x(t)| < +\infty$. On en déduit que $T = T_{\max}$.

4. Comme f' est non constante, il existe v_0, v_1 tel que $f'(v_0) > f'(v_1)$, et on peut construire $u_0 \in C_c^\infty(\mathbb{R},\mathbb{R})$ telle que $u_0(x_0) = v_0$ et $u_0(x_1) = v_1$, où x_0 et x_1 sont donnés et $x_0 < x_1$, voir figure 5.4. Supposons que u soit solution classique avec cette donnée initiale. Alors :

$$u(x_0 + f'(u_0(x_0))t,t) = u_0(x_0) = v_0 \text{ et } u(x_1 + f'(u_0(x_1))t,t) = u_0(x_1) = v_1.$$

   Soit T tel que $x_0 + f'(v_0)T = x_1 + f'(v_1)T = \bar{x}$, c'est à dire

$$T = \frac{x_1 - x_0}{f'(v_0) - f'(v_1)}.$$

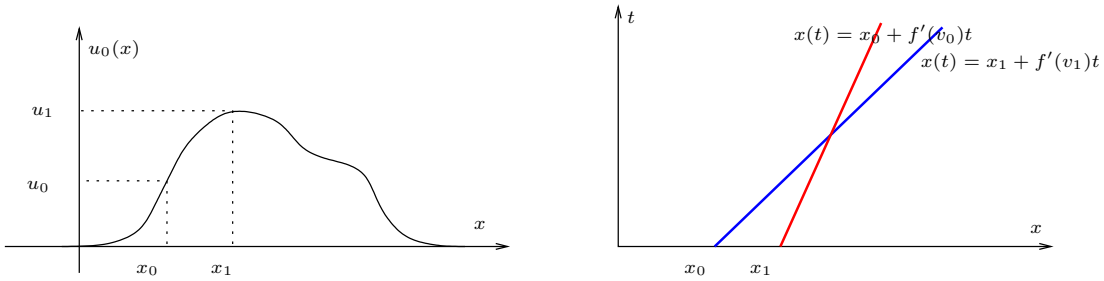

FIG. 5.4 – *Droites caractéristiques, cas non linéaire*

On a alors :

$$u(\bar{x},T) = u_0(x_0) = v_0 = u_0(x_1) = v_1,$$

ce qui est impossible. On en conclut que (5.4.17) n'admet pas de solution classique pour cette donnée initiale.

■

Définition 5.19 (Solution faible) *Soit $u_0 \in L^\infty(\mathbb{R})$ et $f \in C^1(\mathbb{R},\mathbb{R})$, On appelle solution faible de (5.4.17) une fonction $u \in L^\infty(\mathbb{R}\times\mathbb{R}_+)$ telle que*

$$\int\int_{\mathbb{R}\times\mathbb{R}_+} [u(x,t)\varphi_t(x,t) + f(u(x,t))\varphi_x(x,t)]dxdt + \int_{\mathbb{R}} u_0(x)\varphi(x,0)dx = 0, \forall \varphi \in C_c^1(\mathbb{R}\times\mathbb{R}_+,\mathbb{R}). \quad (5.4.20)$$

Donnons maintenant les liens entre solution classique et solution faible.

Proposition 5.20 *Soient $f \in C^1(\mathbb{R},\mathbb{R})$ et $u_0 \in C(\mathbb{R},\mathbb{R})$ des fonctions données.*

1. *Si u est solution classique de (5.4.17) alors u est solution faible de (5.4.17).*
2. *Si $u \in C^1(\mathbb{R}\times]0,+\infty[)\cap C(\mathbb{R}\times[0,+\infty[)$ est solution faible de (5.4.17) alors u est solution classique de (5.4.17).*
3. *Soit $\sigma \in \mathbb{R}$, $D_1 = \{(x,t) \in \mathbb{R}\times\mathbb{R}_+; x < \sigma t\}$ et $D_2 = \{(x,t) \in \mathbb{R}\times\mathbb{R}_+; x > \sigma t\}$. Alors si $u \in C(\mathbb{R}\times\mathbb{R}_+)$ est telle que $u_{|D_i} \in C^1(D_i,\mathbb{R})$, $i = 1,2$ et que (5.4.17) est vérifié pour tout $(x,t) \in D_i$, $i = 1,2$, alors u est solution faible de (5.4.17).*

Démonstration :

1. Supposons que u est solution classique de (5.4.17), c.à.d. de :

$$\begin{cases} u_t + (f(u))_x = 0, & (x,t) \in \mathbb{R}\times\mathbb{R}_+, \\ u(x,0) = u_0(x). \end{cases}$$

Soit $\varphi \in C_c^1(\mathbb{R}\times\mathbb{R}_+,\mathbb{R})$. Multiplions (5.4.17) par φ et intégrons sur $\mathbb{R}\times\mathbb{R}_+$. On obtient :

$$\int_{\mathbb{R}}\int_{\mathbb{R}_+} u_t(x,t)\varphi(x,t)dtdx + \int_{\mathbb{R}}\int_{\mathbb{R}_+} (f(u))_x(x,t)\varphi(x,t)dtdx = 0.$$

L'application du théorème de Fubini et une intégration par parties donnentalors :

$$\int_{\mathbb{R}} u(x,0)\varphi(x,0)dx - \int_{\mathbb{R}}\int_{\mathbb{R}_+} u(x,t)\varphi_t(x,t)dtdx - \int_{\mathbb{R}_+}\int_{\mathbb{R}} f(u)(x,t)\varphi_x(x,t)dxdt = 0,$$

(car supp(φ) est compact). Et on obtient donc bien la relation (5.4.20), grâce à lacondition initiale $u(x,0) = u_0(x)$.

2. Soit donc u une solution faible de (5.4.17), qui vérifie de plus $u \in C^1(\mathbb{R}\times]0,+\infty[)\cap C(\mathbb{R}\times[0,+\infty[)$. On a donc suffisamment de régularité pour intégrer par parties dans (5.4.20).
Commençons par prendre φ à support compact dans $\mathbb{R}\times]0,+\infty[$. On a donc $\varphi(x,0) = 0$, et une intégration par parties dans (5.4.20) donne :

$$-\int_{\mathbb{R}}\int_{\mathbb{R}_+} u_t(x,t)\varphi(x,t)dtdx - \int_{\mathbb{R}_+}\int_{\mathbb{R}} (f(u))_x(x,t)\varphi(x,t)dxdt = 0.$$

On a donc :

$$\int_{\mathbb{R}}\int_{\mathbb{R}_+} (u_t(x,t) + (f(u))_x(x,t))\,\varphi(x,t)dtdx = 0, \forall\varphi \in C_c^1(\mathbb{R}\times]0,+\infty[).$$

Comme $u_t + (f(u))_x$ est continue, on en déduit que $u_t + (f(u))_x = 0$. En effet, on on rappelle que si $\int_R f(x)\varphi(x)dx = 0$ pour toute fonction φ continue de $\mathbb{R}$ dans $\mathbb{R}$, alors $f = 0$ p.p. ; si de plus f est continue, alors $f = 0$ partout.
On prend alors $\varphi \in C_c^1(\mathbb{R}\times\mathbb{R}_+)$. Dans ce cas, une intégration par parties dans (5.4.20) donne

$$\int_{\mathbb{R}} u(x,0)\varphi(x,0)dx - \int_{\mathbb{R}}\int_{\mathbb{R}_+} (u_t(x,t) + (f(u))_x(x,t))\,\varphi(x,t)dtdx - \int_{\mathbb{R}} u_0(x)\varphi(x,0)dx = 0.$$

Mais on vient de montrer que $u_t + (f(u))_x = 0$. On en déduit que

$$\int_{\mathbb{R}} (u_0(x) - u(x,0))\varphi(x,0)dx = 0, \forall\varphi \in C_c^1(\mathbb{R}).$$

Comme u est continue, ceci entraîne $u(x,0) = u_0(x)$. Donc u est solution classique de (5.4.17).

3. Soit $u \in C(\mathbb{R}\times\mathbb{R}_+)$ telle que $u|_{D_i}$ vérifie (5.4.17), pour tout $(x,t) \in D_i$. Montrons que u est solution faible. Pour cela, calculons :

$$X = \int_{\mathbb{R}}\int_{\mathbb{R}_+} u(x,t)\varphi_t(x,t)dtdx + \int_{\mathbb{R}_+}\int_{\mathbb{R}} f(u)(x,t)\varphi_x(x,t)dxdt.$$

On a donc $X = X_1 + X_2$, avec

$$X_1 = \int_{\mathbb{R}}\int_{\mathbb{R}_+} u(x,t)\varphi_t(x,t)dtdx \text{ et } X_2 = \int_{\mathbb{R}_+}\int_{\mathbb{R}} (f(u))(x,t)\varphi_x(x,t)dxdt.$$

Calculons X_1. Comme u n'est de classe C^1 que sur chacun des domaines D_i, on n'a pas le droit d'intégrer par parties sur $\mathbb{R}\times\mathbb{R}_+$ entier. On va donc décomposer l'intégrale sur D_1 et D_2 ; supposons par exemple $\sigma < 0$, voir figure 5.5. (Le cas $\sigma > 0$ se traite de façon similaire). On a alors $D_2 = \{(x,t); x \in \mathbb{R}_- \text{ et } 0 < t < \frac{x}{\sigma}\}$ et $D_1 = \mathbb{R}_+\times\mathbb{R}_+ \cup \{(x,t); x \in \mathbb{R}_- \text{ et } \frac{x}{\sigma} < t < +\infty\}$.
On a donc :

$$X_1 = \int_{\mathbb{R}_-}\int_0^{x/\sigma} u(x,t)\varphi_t(x,t)dtdx + \int_{\mathbb{R}_-}\int_{\frac{x}{\sigma}}^{+\infty} u(x,t)\varphi_t(x,t)dtdx + \int_{\mathbb{R}_+}\int_{\mathbb{R}_+} u(x,t)\varphi_t(x,t)dtdx.$$

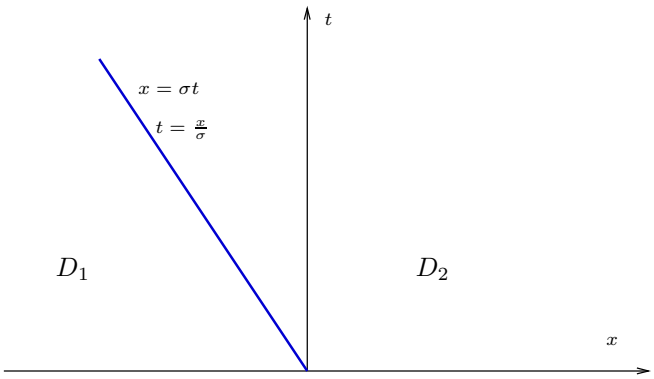

FIG. 5.5 – *Les domaines D1 et D2*

Comme u est de classe C^1 sur chacun des domaines, on peut intégrer par parties, ce qui donne :

$$X_1 = \int_{\mathbb{R}_-} u(x,\frac{x}{\sigma})\varphi(x,\frac{x}{\sigma})dx - \int_{\mathbb{R}_-} u(x,0)\varphi(x,0)dx - \int_{\mathbb{R}_-}\int_0^{\frac{x}{\sigma}} u_t(x,t)\varphi(x,t)dtdx,$$

$$+ \int_{\mathbb{R}_-} u(x,\frac{x}{\sigma})\varphi(x,\frac{x}{\sigma})dx - \int_{\mathbb{R}_-}\int_{\frac{x}{\sigma}}^{+\infty} u_t(x,t)\varphi(x,t)dtdx.$$

$$+ \int_{\mathbb{R}_+} (-u(x,0)\varphi(x,0)dx - \int_{\mathbb{R}_+}\int_{\mathbb{R}_+} u_t(x,t)\varphi(x,t)dtdx.$$

En simplifiant il vient :

$$X_1 = -\int_{\mathbb{R}} u(x,0)\varphi(x,0)dx - \int\int_{D_1} u_t(x,t)\varphi(x,t)dtdx - \int\int_{D_2} u_t(x,t)\varphi(x,t)dtdx.$$

On décompose de même X_2 sur $D_1 \cup D_2$, en remarquant maintenant que $D_1 = \{(x,t) \in \mathbb{R}\times\mathbb{R}_+; x < \sigma t\}$ et $D_2 = \{(x,t) \in \mathbb{R} \times \mathbb{R}_+; x > \sigma t\}$:

$$X_2 = \int_{\mathbb{R}_+}\int_{-\infty}^{\sigma t} f(u)(x,t)\varphi_x(x,t)dxdt + \int_{\mathbb{R}_+}\int_{\sigma t}^{+\infty} f(u)(x,t)\varphi_x(x,t)dxdt.$$

La fonction u est de classe C^1 sur chacun des domaines, on peut là encore intégrer par parties. Comme φ est à support compact sur $\mathbb{R} \times \mathbb{R}_+$, on obtient après simplification :

$$X_2 = -\int\int_{D_1} (f(u))_x(x,t)\varphi(x,t)dxdt - \int\int_{D_2} (f(u))_x(x,t)\varphi(x,t)dxdt.$$

Comme $u_t + (f(u))_x = 0$ sur D_1 et D_2, on a donc :

$$X = X_1 + X_2 = -\int_{\mathbb{R}} u(x,0)\varphi(x,0)dx,$$

ce qui prouve que u est solution faible de (5.4.17).

■

Notons qu'il existe souvent plusieurs solutions faibles. On a donc besoin d'une notion supplémentaire pour les distinguer. C'est la notion de solution entropique, qui nous permettra d'obtenir l'unicité. Donnons tout d'abord un exemple de non-unicité de la solution faible. Pour cela on va considérer une équation modèle, appelée équation de Burgers, qui s'écrit

$$u_t + (u^2)_x = 0. \tag{5.4.21}$$

Pour calculer les solutions du problème de Cauchy associé à cette équation de manière analytique, on considère une donnée initiale particulière, qui sécrit

$$u_0(x) = \begin{cases} u_g \text{ si } x < 0, \\ u_d \text{ si } x > 0, \end{cases}$$

Ces données initiales définissent un problème de Cauchy particulier, qu'on appelle problème de Riemann, que nous étudierons plus en détails par la suite.
Considérons alors le problème suivant (dit problème de Riemann, voir définition 5.28) pour l'équation de Burgers:

$$\begin{cases} u_t + (u^2)_x = 0, \\ u_0(x) = \begin{cases} u_g = -1 \text{ si } x < 0, \\ u_d = 1 \text{ si } x > 0. \end{cases} \end{cases} \tag{5.4.22}$$

On cherche une solution faible de la forme :

$$u(x,t) = \begin{cases} u_g \text{ si } x < \sigma t, \\ \\ u_d \text{ si } x > \sigma t. \end{cases} \tag{5.4.23}$$

Notons que cette éventuelle solution est discontinue au travers de la droite d'équation $x = \sigma t$ dans le plan (x,t). On remplace $u(x,t)$ par ces valeurs dans (5.4.20). Après calculs (voir exercice 57 page 238, ou aussi la proposition 5.29 plus loin), on s'aperçoit que u est solution faible si la condition suivante, dite condition de Rankine et Hugoniot, est vérifiée:

$$\sigma(u_d - u_g) = (f(u_d) - f(u_g)), \tag{5.4.24}$$

ce qui avec la condition initiale particulière choisie ici, donne $2\sigma = \frac{1^2-(-1)^2}{=}0$.
Mais on peut trouver d'autres solutions faibles : en effet, on sait que sur les caractéristiques, qui ont pour équation $x = x_0 + f'(u_0(x_0))t$, la fonction u est constante. Comme $f'(u) = 2u$, les caractéristiques sont donc des droites de pente -2 si $x_0 < 0$, et de pente 2 si $x_0 > 0$. Construisons ces caractéristiques sur la figure 5.6 : Dans la zone du milieu, où l'on a représenté un point d'interrogation, on cherche u sous la forme $u(x,t) = \varphi\left(\frac{x}{t}\right)$, et telle que u soit continue sur $\mathbb{R} \times \mathbb{R}_+$. La fonction u suivante convient :

$$u(x,t) = \begin{cases} -1 \text{ si } x < -2t, \\ \\ \dfrac{x}{2t} \text{ si } -2t < x < 2t, \\ \\ 1 \text{ si } x > 2t. \end{cases} \tag{5.4.25}$$

Comment choisir la "bonne" solution faible, entre (5.4.23) et(5.4.25) ? Comme les problèmes hyperboliques sont souvent obtenus en négligeant les termes de diffusion dans des équations paraboliques, une technique pour choisir la solution est de chercher la limite du problème de diffusion associé qui s'écrit :

$$u_t + (f(u))_x - \varepsilon u_{xx} = 0, \tag{5.4.26}$$

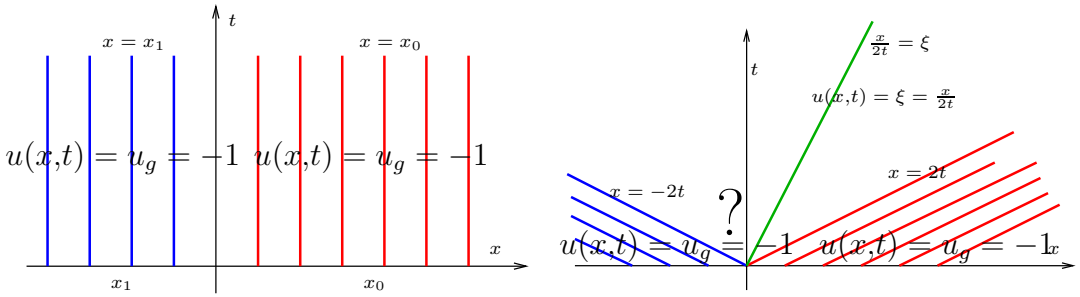

FIG. 5.6 – *Problème de Riemann pour léquation de Burgers*

lorsque le terme de diffusion devient négligeable, c.à.d. lorsque ε tend vers 0. Soit u_ε la solution de (5.4.26) (on admettra l'existence et l'unicité de u_ε). On peut montrer que u_ε tend vers u lorsque ε tend vers 0, où u est la "solution faible entropique" de (5.4.26), définie comme suit.

Définition 5.21 (Solution entropique) *Soit $u_0 \in L^\infty(\mathbb{R})$ et $f \in C^1(\mathbb{R})$, on dit que $u \in L^\infty(\mathbb{R}\times\mathbb{R}_+)$ est solution entropique de (5.4.26) si pour toute fonction $\eta \in C^1(\mathbb{R})$ convexe, appelée "entropie", et pour toute fonction $\phi \in C^1$ telle que $\phi' = f'\eta'$, appelé "flux d' entropie", on a :*

$$\int_{\mathbb{R}}\int_{\mathbb{R}_+} (\eta(u)\varphi_t + \phi(u)\varphi_x)dxdt + \int_{\mathbb{R}} \eta(u_0(x))\varphi(x,0)dx \geq 0, \forall \varphi \in C_c^1(\mathbb{R}\times\mathbb{R}_+,\mathbb{R}_+). \tag{5.4.27}$$

Remarque 5.22 (Condition initiale) *Noter que dans la définition 5.21, on prend une fois de plus $\varphi \in C_c^1(\mathbb{R}\times\mathbb{R}_+,\mathbb{R}_+)$ de manière à bien prendre en compte la condition initiale ; ceci n'est pas toujours fait de cette manière dans les travaux plus anciens sur le sujet, mais entraîne des difficultés lorqu'on s'intéresse à la convergence des schémas numériques.*

On admettra le théorème suivant (dû à Kruskov, 1955)

Théorème 5.23 (Kruskov) *Soient $u_0 \in L^\infty(\mathbb{R})$ et $f \in C^1(\mathbb{R})$ alors il existe une unique solution entropique de (5.4.17) au sens de la définition 5.21.*

Proposition 5.24 *Si u est solution classique de (5.4.17), alors u est solution entropique.*

Démonstration : Soit $u \in C^1(\mathbb{R}\times\mathbb{R}_+,\mathbb{R})$, Soit $\eta \in C^1(\mathbb{R})$, convexe, une entropie et ϕ tel que $\phi' = f'\eta'$, le flux associé. Multiplions (5.4.17) par $\eta'(u)$:

$$\eta'(u)u_t + f'(u)u_x\eta'(u) = 0$$

Soit encore, puisque $\phi' = f'\eta'$,

$$(\eta(u))_t + \phi'(u)u_x - 0$$

On a donc finalement :

$$(\eta(u))_t + (\phi(u))_x - 0 \tag{5.4.28}$$

De plus, comme

$$u(x,0) = u_0(x), \text{ on a aussi : } \eta(u(x,0)) = \eta(u_0(x)) \tag{5.4.29}$$

Soit $\varphi \in C_c^1(\mathbb{R}\times\mathbb{R}_+,\mathbb{R}_+)$, on multiplie (5.4.28) par φ, on intègre sur $\mathbb{R}\times\mathbb{R}_+$ et on obtient (5.4.27) (avec égalité) en intégrant par parties. Dans le cas d'une solution classique, l'inégalité d'entropie est une égalité. ∎

On a de même le résultat suivant :

Proposition 5.25 *Si u est solution faible entropique de (5.4.17), alors u est solution faible.*

Démonstration : Il suffit de prendre $\eta(u) = u$ et $\eta(u) = -u$ dans (5.4.27) pour se convaincre du résultat. ∎

On déduit de la proposition 5.24, et du théorème 5.23 de Kruskov, que si on a plusieurs solutions faibles au problème 5.4.17 page 223 et que l'une d'entre elles est régulière, alors cette dernière est forcément la solution entropique. Enfin, la caractérisation suivante, que l'on admettra, est souvent utilisée en pratique :

Proposition 5.26 (Entropies de Kruskov) *Soit $u_0 \in L^\infty(\mathbb{R})$ et $f \in C^1(\mathbb{R})$, alors $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$ est solution entropique de (5.4.26) au sens de la définition 5.21 si et seulement si pour tout $k \in \mathbb{R}$, alors (5.4.27) est vérifiée avec η définie par $\eta(s) = |s - k|$, et ϕ, flux d'entropie associée, défini par :*

$$\phi(u) = \max(f(u),k) - \min(f(u),k).$$

Notons que η n'est pas de classe C^1.

Notons que les solutions d'une équation hyperbolique non linéaire respectent les bornes de la solution initiale. Plus précisément, on a le résultat suivant, qu'on admettra :

Proposition 5.27 *Si $u_0 \in L^\infty(\mathbb{R})$ et soit A et $B \in \mathbb{R}$ tels que $A \leq u_0 \leq B$ p.p.. Soit $f \in C^1(\mathbb{R})$, alors la solution entropique $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$ de (5.4.17) vérifie : $A \leq u(x) \leq B$ p.p. dans $\mathbb{R} \times \mathbb{R}_+$.*

Cette propriété est essentielle dans les phénomènes de transport, et il est souhaitable qu'elle soit préservée pour la solution approchée donnée par un schéma numérique.

Avant d'aborder l'étude des schémas numériques pour les équations hyperboliques, nous terminons par un résultat sur les solutions du problème de Riemann, dont nous nous sommes d'ailleurs servis pour montrer la non unicité des solutions faibles de (5.4.22).

Définition 5.28 (Problème de Riemann) *Soient $f \in C^1(\mathbb{R},\mathbb{R})$, on appelle problème de Riemann avec données u_g, $u_d \in \mathbb{R}$, le problème suivant :*

$$\left\{ \begin{array}{l} u_t + (f(u))_x = 0,\ x \in \mathbb{R},\ t > 0 \\ u(0,x) = \left\{ \begin{array}{l} u_g \ si\ x < 0 \\ u_d \ si\ x > 0 \end{array} \right. \end{array} \right. \tag{5.4.30}$$

Lorsque la fonction f est convexe ou concave, les solutions du problème de Riemann se calculent facilement ; en effet, a le rsultat suivant (voir aussi exercice 58 page 239) :

Proposition 5.29 *Soit $f \in C^1(\mathbb{R},\mathbb{R})$ strictement convexe, et soient u_g et $u_d \in \mathbb{R}$.*

1. *Si $u_g > u_d$, on pose*

$$\sigma = \frac{[f(u)]}{[u]} \ \text{avec}\ [f(u)] = f(u_d) - f(u_g)\ \text{et}\ [u] = u_d - u_g. \tag{5.4.31}$$

alors la fonction u définie par

$$\left\{ \begin{array}{l} u(x,t) = u_g \ si\ x < \sigma t \\ u(x,t) = u_d \ si\ x > \sigma t \end{array} \right. \tag{5.4.32}$$

est l'unique solution entropique de (5.4.30). Une solution de la forme (5.4.32) est appellée une onde de "choc".

2. Si $u_g < u_d$, alors la fonction u définie par

$$\begin{cases} u(x,t) = u_g & si\ x < f'(u_g)t \\ u(x,t) = u_d & si\ x > f'(u_d)t \\ u(x,t) = \xi & si\ x = f'(\xi)t\ avec\ u_g < \xi < u_d \end{cases} \tag{5.4.33}$$

est l' unique solution entropique de (5.4.30). Notons que dans ce cas, la solution entropique est continue. Une solution de la forme (5.4.33) est appelée une onde de "détente".

Démonstration : 1. Cherchons u sous la forme (5.4.32). Commençons par déterminer σ pour que u soit solution faible. On suppose, pour fixer les idées, que $\sigma > 0$ (mais le même raisonnement marche pour $\sigma < 0$). Soit $\varphi \in C_c^\infty(\mathbb{R} \times \mathbb{R}_+, \mathbb{R})$. On veut montrer que

$$X = X_1 + X_2 = -\int_{\mathbb{R}} u(x,0)\varphi(x,0)dx,$$

où $X_1 = \int_{\mathbb{R}}\int_{\mathbb{R}_+} u(x,t)\varphi_t(x,t)dtdx$ et $X_2 = \int_{\mathbb{R}}\int_{\mathbb{R}_+} f((u(x,t))\varphi_x(x,t)dtdx$.
Calculons donc X_1 et X_2 :

$$\begin{aligned} X_1 &= \int_{-\infty}^0\int_0^{+\infty} u(x,t)\varphi_t(x,t)dtdx + \int_0^{+\infty}\int_0^{\frac{x}{\sigma}} u(x,t)\varphi_t(x,t)dtdx + \int_0^{+\infty}\int_{\frac{x}{\sigma}}^{+\infty} u(x,t)\varphi_t(x,t)dtdx \\ &= -\int_{\infty}^0 u_g\varphi(x,0)dx + \int_0^{+\infty} u_d\left(\varphi(x,\frac{x}{\sigma}) - \varphi(x,0)\right)dx + \int_0^{+\infty} u_g\left(-\varphi(x,\frac{x}{\sigma}\right)dx \\ &= -\int_{\mathbb{R}} u(x,0)\varphi(x,0)dx + \int_0^{+\infty}(u_d - u_g)\varphi(x,\frac{x}{\sigma})dx. \end{aligned}$$

De même

$$\begin{aligned} X_2 &= \int_0^{+\infty}\int_{-\infty}^{\sigma t} f(u)\varphi_x(x,t)dxdt + \int_0^{+\infty}\left(\int_{\sigma t}^{+\infty} f(u)(x,t)\varphi_x(x,t)\right)dxdt \\ &= \int_0^{+\infty} f(u_g)\varphi(\sigma t,t)dt - \int_0^{+\infty} f(u_d)\varphi(\sigma t,t)dt. \end{aligned}$$

En posant $[u] = u_d - u_g$ et $[f(u)] = f(u_d) - f(u_g)$, on obtient :

$$\begin{aligned} X + \int_{\mathbb{R}} u(x,0)\varphi(x,0)dx &= \int_0^{+\infty}[u]\varphi(x,\frac{x}{\sigma})dx - \int_0^{+\infty}[f(u)]\varphi(\sigma t,t)dt \\ &= \int_0^{+\infty}[u]\varphi(\sigma t,t)\sigma dt - \int_0^{+\infty}[f(u)]\varphi(\sigma t,t)dt. \end{aligned}$$

On en déduit que

$$X + \int_{\mathbb{R}} u(x,0)\varphi(x,0)dx = 0 \text{ si } \sigma[u] - [f(u)] = 0,$$

ce qui est vrai si la condition suivante, dite de Rankine et Hugoniot :

$$\sigma[u] = [f(u)] \tag{5.4.34}$$

est vérifiée.
Voyons maintenant si u est bien solution entropique. Pour cela, on considère $\eta \in C^1$ une "entropie", et $\phi \in C^1$ le flux d'entropie associé, t.q. $\phi' = \eta' f'$. Le même calcul que le précédent, en remplaçant u par $\eta(u)$ et $f(u)$ par $\phi(u)$ donne que :

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} \eta(u)(x,t)\varphi_t(x,t)dtdx + \int_{\mathbb{R}_+} \int_{\mathbb{R}} \phi(u)(x,t)\varphi_x(x,t)dxdt$$

$$+ \int_{\mathbb{R}} \eta(u_0(x))dx = \int_0^{+\infty} (\sigma[\eta(u)] - [\phi(u)])\varphi(\sigma t,t)dt.$$

Pour que u soit solution entropique, il faut (et il suffit) donc que

$$\sigma[\eta(u)] \geq [\phi(u)] \tag{5.4.35}$$

Il reste à vérifier que cette inégalité est vérifiée pour σ donné par (5.4.34), c.à.d.

$$\frac{f(u_d) - f(u_g)}{u_d - u_g}(\eta(u_d) - \eta(u_g)) \geq \phi(u_d) - \phi(u_g)$$

Ceci s'écrit encore :

$$(f(u_d) - f(u_g))(\eta(u_d) - \eta(u_g)) \leq (\phi(u_d) - \phi(u_g))(u_d - u_g).$$

Cette inégalité est vérifiée en appliquant le lemme suivant avec $b = u_g > u_d = a$.

Lemme 5.30 *Soient a,$b \in \mathbb{R}$ tels que $a < b$, soient f et $\eta \in C^1(\mathbb{R})$ des fonctions convexes et $\phi \in C^1(\mathbb{R})$ telle que $\phi' = \eta' f'$, alors :*

$$\int_a^b \phi'(s)ds(b-a)ds \geq \int_a^b f'(s)ds \int_a^b \eta'(s)ds$$

Démonstration : On a

$$\begin{aligned} \int_a^b \phi'(x)dx &= \int_a^b f'(x)\eta'(x)dx \\ &= \int_a^b f'(x)(\eta'(x) - \eta'(y))dx + \int_a^b f'(x)\eta'(y)dx, \quad \forall y \in \mathbb{R}. \end{aligned}$$

On a donc, en intégrant par rapport à y entre a et b :

$$(b-a)\int_a^b \phi'(x)dx = \int_a^b \int_a^b f'(x)(\eta'(x) - \eta'(y))dxdy + \int_a^b f'(x)dx \int_a^b \eta'(y)dy$$

Or

$$\int_a^b \int_a^b f'(x)[\eta'(x) - \eta'(y)]dxdy = \int_a^b \int_a^b f'(y)(\eta'(y) - \eta'(x))dxdy$$

et donc

$$(b-a)\int_a^b \int_a^b \phi'(x)dx = \int_a^b \int_a^b (f'(x) - f'(y)(\eta'(x) - \eta'(y)dxdy + \left(\int_a^b f'(x)dx\right)\left(\int_a^b \eta'(y(y)dy\right).$$

Comme f' et η' sont croissantes, la première intégrale du second membre est nulle, et on a donc bien le résultat annoncé. ∎

2. On vérifie facilement que la fonction u définie par (5.4.33) est continue sur $\mathbb{R} \times \mathbb{R}_+^*$, et qu'elle vérifie $u_t + (f(u))_x = 0$ dans chacun des domaines D_1,D_2,D_3 définis par

$$D_1 = \{t > 0, x < f'(u_g)t\}, D_2 = \{t > 0, f'(u_g)t < x < f'(u_d)t\} \text{ et } D_3 = \{t > 0, x > f'(u_d)t\}.$$

Donc par le point 3 de la proposition 5.20 page 225, on sait que u est solution faible (mais attention, ce n'est pas une solution classique car u n'est pas forcément C^1 sur $\mathbb{R} \times \mathbb{R}_+$ tout entier).
Soit $\eta \in C^1(\mathbb{R},\mathbb{R})$ une entropie (convexe) et ϕ le flux d'entropie associé, comme $u_t + (f(u))_x = 0$ dans D_i pour $i = 1$ à 3, en multipliant par $\eta'(u)$, on a également que $(\eta(u))_t + (\phi(u))_x = 0$ dans D_i pour $i = 1$ à 3. Soit maintenant $\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+,\mathbb{R}_+)$, on va montrer que

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} (\eta(u))(x,t)\varphi_t(x,t)dtdx + \int_{\mathbb{R}} \int_{\mathbb{R}} (\phi(u))(x,t)\varphi_x(x,t)dxdt + \int_{\mathbb{R}} \eta(u_0(x))\varphi(x,0)dx = 0$$

(dans le cas d'une solution continue, l'inégalité d'entropie est une égalité). En effet, en intégrant par parties les trois termes précédents sur D_1,D_2,D_3, comme on l'a fait dans les questions 1 et 2, comme la fonction u est continue, les traces des fonctions sur le bord des domaines s'annulent deux à deux, et il ne reste donc que la condition initiale. On montre ainsi (faire le calcul pour s'en convaincre...) que

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} (\eta(u))(x,t)\varphi_x(x,t)dtdx + \int_{\mathbb{R}} \int_{\mathbb{R}_+} \phi(u)(x,t)(\varphi_x(x,t))dxdt = -\int_{\mathbb{R}} \eta(u_0(x))\varphi(x,0),$$

ce qui prouve que u est la solution entropique.

∎

5.5 Schémas pour les équations non linéaires

On se donne $u_0 \in L^\infty(\mathbb{R}$ et $f \in C^1(\mathbb{R})$, et on cherche à trouver une approximation de la solution entropique du problème (5.4.17). On utilise les mêmes notations que pour le schéma (5.3.15). En intégrant l'équation $u_t + (f(u))_x = 0$ sur une maille K_i, on obtient, au temps $t = t_n$:

$$\int_{K_i} u_t(x,t_n)dxdt + f(u(x_{i+1/2},t)) - f(u(x_{i-1/2},t_n)) = 0.$$

En utilisant le schéma d'Euler explicite pour la discrétisation de la dérivée temporelle, et en notant $f_{i+1/2}^n$ le flux numérique, c'est à dire l'approximation de $f(u(x_{i+1/2},t_n))$ on obtient le schéma numérique suivant :

$$\begin{cases} h_i \frac{u_i^{n+1} - u_i^n}{k} + f_{i+1/2}^n - f_{i-1/2}^n = 0 \\ u_i^0 = \dfrac{1}{h_i} \displaystyle\int_{K_i} u_0(x)dx. \end{cases} \tag{5.5.36}$$

Pour que ce schéma soit complètement défini, il reste à préciser $f_{i+1/2}^n$ en fonction des inconnues discrètes u_i^n. Un premier choix possible est le schéma centré,

$$f_{i+1/2}^n = \frac{f(u_{i+1}^n) + f(u_i^n)}{2}$$

dont on a vu qu'il est à proscrire, puisque, dans le cas linéaire, il est instable. Rappelons que dans le cas linéaire, le choix décentré amont donne

$$\text{si } f(u) = u, f^n_{i+1/2} = f(u^n_i), \text{ et}$$
$$\text{si } f(u) = -u, f^n_{i+1/2} = f(u^n_{i+1}).$$

Dans le cadre de ce cours , on va s'intéresser aux schémas les plus simples à trois points, c.à.d. que léquation associée à l'inconnue u^n_i fait intervenir les trois inconnues discrètes u^n_i, u^n_{i-1} et u^n_{i+1}. Le flux numérique g s'écrit sous la forme

$$f^n_{i+1/2} = g(u^n_i,u^n_{i+1}).$$

Pour obtenir un "bon" schéma, on va choisir un flux "monotone", au sens suivant :

Définition 5.31 *On dit que qu'une fonction g définie de $\mathbb{R}^2$ dans $\mathbb{R}$ est un flux monotone pour la discrétisation de (5.4.17), si*

1. *g est consistante par rapport à f, c.à.d. $g(u,u) = f(u)$,*
2. *g est croissante par rapport à la première variable et décroissante par rapport à la deuxième variable,*
3. *g est lipschitzienne sur $[A,B]$, où $A = \inf_{\mathbb{R}} u_0$ et $B = \sup_{\mathbb{R}} u_0$.*

Remarque 5.32 (Flux monotones et schémas monotones) *Si le schéma 5.3.15 est à flux monotone, et s'il vérifie la condition de CFL, on peut alors montrer que le schéma est monotone, c.à.d. qu'il s'écrit sous la forme :*

$$u^{n+1}_i = H(u^n_{i-1},u^n_i,u^n_{i+1}),$$

où H est une fonction croissante de ses trois arguments.

Cas où f est monotone Pour illustrer le choix de g, supposons par exemple que f soit croissante. Un choix très simple consiste alors à prendre $g(u^n_i,u^n_{i+1}) = f(u^n_i)$. On vérifie (exercice) que dans ce cas, les trois conditions ci-dessus sont vérifiées, ce schéma est dit décentré amont. On vérifiera qu'on retrouve le schéma décentré amont exposé dans le cas linéaire. De même si f est décroissante on peut facilement vérifier que le choix $g(u^n_i,u^n_{i+1}) = f(u^n_{i+1})$ convient.

Schéma à décomposition de flux Le schéma à décomposition de flux, appelé aussi "flux splitting" en anglais, consiste comme le nom l'indique à décomposer $f = f_1 + f_2$, où f_1 est croissante et f_2 décroissante, et à prendre pour g :

$$g(u^n_i,u^n_{i+1}) = f_1(u^n_i) + f_2(u^n_{i+1})$$

Schéma de Lax Friedrich Le schéma de Lax Friedrich consiste à modifier le schéma centré de manière à le rendre stable. On écrit donc :

$$g(u^n_i,u^n_{i+1}) = \frac{1}{2}(f(u^n_i) + f(u^n_{i+1})) + D(u^n_i - u^n_{i+1})$$

où $D \geq 0$ est il faut avoir D suffisamment grand pour que g soit croissante par rapport à la première variable et décroissante par rapport à la seconde variable.

Schéma de Godunov Le schéma de Godunov est un des plus connus pour les équations hyperboliques non linéaires. De nombreux schémas pour les systèmes ont été inspirés par ce schéma. Le flux numérique du schéma de Godunov s'écrit :

$$g(u_i^n,u_{i+1}^n) = f(w_R(u_i^n,u_{i+1}^n)) \tag{5.5.37}$$

où $w_R(u_i^n,u_{i+1}^n)$ est la solution en 0 du problème de Riemann avec conditions u_i^n,u_{i+1}^n, qui s'écrit :

$$\begin{cases} u_t + (f(u))_x = 0 \\ u_0(x) = \quad \begin{cases} u_g = u_i^n \quad w < 0 \\ u_d = u_{i+1}^n \quad w > 0 \end{cases} \end{cases}$$

On peut montrer que le flux de Godunov (5.5.37) vérifie les conditions de la définition 5.31.

Schéma de Murman Une manière de simplifier le schéma de Godunov est de remplacer la résolution du problème de Riemann linéaire. On prend alors $g(u_i^n,u_{i+1}^n) = f(\widetilde{w}_R(u_i^n,u_{i+1}^n)$ où $\widetilde{w}_R(u_i^n,u_{i+1}^n)$ est solution de

$$\begin{cases} u_t + \alpha u_x = 0 \\ u_0(x) = \quad \begin{cases} u_i^n \quad x < 0 \\ u_{i+1}^n \quad x > 0 \end{cases} \end{cases}$$

Comme le problème est linéaire, la solution de ce problème est connue : $u(x,t) = u_0(x-\alpha t)$. Le schéma est donc très simple, malheureusement, le schéma de Murman n'est pas un schéma monotone (voir exercice (60), car le flux n'est pas monotone par rapport aux deux variables. De fait on peut montrer que les solutions approchées peuvent converger vers des solutions non entropiques. On peut alors envisager une procédure "correction d'entropie"...

Théorème 5.33 (Stabilité et convergence) *Soit* $(u_i^n)_{\substack{i\in\mathbb{Z} \\ n\in\mathbb{N}}}$ *donnée par le schéma*

$$\begin{cases} h_i \frac{u_i^{n+1}-u_i^n}{k} + g(u_i^n,u_{i+1}^n) - g(u_{i-1}^n,u_i^n) = 0 \\ u_i^0 = \dfrac{1}{h_i}\displaystyle\int_{K_i} u_0(x)dx \end{cases}$$

On suppose que g est un flux monotone au sens de la définition 5.31. On suppose de plus que :

$$k \leq \frac{\alpha h}{2M}, \text{ et } \alpha h \leq h_i \leq h, \forall i,$$

où M est la constante de Lipschitz de g sur [A,B], et A et B sont tels que $A \leq u_0(x) \leq B$ *p.p.. On a alors* $A \leq u_i^n \leq B$ *p.p., et* $\|u_{\tau,k}\| \leq \|u_0\|_\infty$. *Sous les mêmes hypothèses, si on note* $u_{\tau,k}$ *la solution approchée définie par (5.3.16), alors*
$u_{\tau,k}$ *tend vers u, solution entropique de (5.4.17) dans* $L^1_{loc}(\mathbb{R}\times\mathbb{R}_+)$ *lorsque h (et k) tend vers 0.*

5.6 Exercices

Exercice 50 (Problème linéaire en dimension 1) *Corrigé en page 245*

Calculer la solution faible du problème :

$$\left\{ \begin{array}{l} u_t - 2u_x = 0,\ x \in \mathbb{R}, t \in \mathbb{R}_+ \\ u(x,0) = \left\{ \begin{array}{l} 0 \text{ si } x < 0, \\ 1 \text{ sinon.} \end{array} \right. \end{array} \right. \tag{5.6.38}$$

1.Tracer sur un graphique la solution a $t = 0$ et à $t = 1$, en fonction de x. Cette solution faible est-elle solution classique de (5.6.38) ?

2. Même question en remplaçant la condition initiale par $u(x,0) = \sin x$.

Exercice 51 (Problème linéaire en dimension 2) *Suggestions en page 245, Corrigé en page 245*

Soit $\mathbf{v} \in \mathbb{R}^2$ et soit $u_0 \in C^1(\mathbb{R}^2,\mathbb{R})$. On considère le problème de Cauchy suivant:

$$\left\{ \begin{array}{l} u_t + div(\mathbf{v}u) = 0, \\ u(x,0) = u_0(x), \end{array} \right. \tag{5.6.39}$$

Calculer la solution du problème (5.6.39) en tout point $(x,t) \in \mathbb{R}^2 \times \mathbb{R}$.

Exercice 52 (Stabilité du schéma amont dans le cas linéaire) *Corrigé en page 246*

On considère le problème hyperbolique linéaire (5.3.9), avec $u_0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$, dont on calcule une solution approchée par le schéma volumes finis amont (5.3.15). Montrer que ce schéma est stable pour les normes L^1, L^2 et L^∞, c.à.d. que la solution approchée satisfait les propriétés suivantes:

1. $\|u_{\mathcal{T},k}(.,n)\|_{L^1(\mathbb{R})} \leq \|u_0\|_{L^1(\mathbb{R})}, \forall n \in \mathbb{N}$,
2. $\|u_{\mathcal{T},k}(.,n)\|_{L^2(\mathbb{R})} \leq \|u_0\|_{L^2(\mathbb{R})}, \forall n \in \mathbb{N}$,
3. $\|u_{\mathcal{T},k}(.,n)\|_{L^\infty(\mathbb{R})} \leq \|u_0\|_{L^\infty(\mathbb{R})}, \forall n \in \mathbb{N}$,

où $u_{\mathcal{T},k}$ désigne la solution approchée calculée par le schéma (voir (5.3.16)).

Exercice 53 (Convergence des schémas DFDA et VFDA dans le cas linéaire)
Corrigé en page 246

Soit $u_0 \in C^2(\mathbb{R},\mathbb{R})$ et $T \in \mathbb{R}_+^\star$. On suppose que u_0, u_0' et u_0'' sont bornées (sur $\mathbb{R}$). On considère le problème suivant :

$$\begin{aligned} u_t(x,t) + u_x(x,t) &= 0, x \in \mathbb{R}, t \in [0,T], & (5.6.40) \\ u(x,0) &= u_0(x). & (5.6.41) \end{aligned}$$

Ce problème admet une et une seule solution classique, notée u. On se donne un pas de temps, k, avec $k = \frac{T}{N+1}$ ($N \in \mathbb{N}$), et des points de discrétisation en espace, $(x_i)_{i \in \mathbb{Z}}$. On pose $t_n = nk$, pour $n \in \{0,\ldots,N+1\}$, et $h_{i+\frac{1}{2}} = x_{i+1} - x_i$, pour $i \in \mathbb{Z}$. On note $\overline{u}_i^n = u(t_n,x_i)$ (pour $n \in \{0,\ldots,N+1\}$ et $i \in \mathbb{Z}$), et on cherche une approximation de $\overline{u}_i^n$.

1. Soient $\alpha,\beta \in \mathbb{R}$. On suppose que, pour un certain $h \in \mathbb{R}$, $\alpha h \leq h_{i+\frac{1}{2}} \leq \beta h$, pour tout $i \in \mathbb{Z}$. On considère, dans cette question le schéma suivant, appelé DFDA (pour Différences Finies Décentré Amont) :

$$\frac{u_i^{n+1}-u_i^n}{k}+\frac{1}{h_{i-\frac{1}{2}}}(u_i^n-u_{i-1}^n) = 0, n\in\{0,\ldots,N\}, i\in\mathbb{Z}, \quad (5.6.42)$$

$$u_i^0 = u_0(x_i), i\in\mathbb{Z}. \quad (5.6.43)$$

(a) (Stabilité) Montrer que $k\leq \alpha h \Rightarrow \inf(u_0)\leq u_i^n\leq \sup(u_0), \forall n\in\{0,\ldots,N+1\}, \forall i\in\mathbb{Z}$.

(b) (Convergence) Montrer que, si $k\leq \alpha h$, on a :

$$\sup_{i\in\mathbb{Z}}|u_i^n-\overline{u}_i^n|\leq CT(k+h), \forall n\in\{0,\ldots,N+1\},$$

où C ne dépend que de u_0 et β.

2. On suppose maintenant que x_i est le centre de la maille $M_i=[x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}]$, pour $i\in\mathbb{Z}$. On pose $h_i=x_{i+\frac{1}{2}}-x_{i-\frac{1}{2}}$. Soient $\alpha,\beta\in\mathbb{R}$. On suppose que, pour un certain $h\in\mathbb{R}$, $\alpha h\leq h_i\leq \beta h$, pour tout $i\in\mathbb{Z}$. On considère, dans cette question le schéma suivant, appelé VFDA (pour Volumes Finis Décentré Amont) :

$$h_i\frac{u_i^{n+1}-u_i^n}{k}+(u_i^n-u_{i-1}^n) = 0, n\in\{0,\ldots,N\}, i\in\mathbb{Z}, \quad (5.6.44)$$

$$u_i^0 = \frac{1}{h_i}\int_{M_i}u_0(x)dx, i\in\mathbb{Z}. \quad (5.6.45)$$

(a) (Stabilité) Montrer que $k\leq \alpha h \Rightarrow \inf(u_0)\leq u_i^n\leq \sup(u_0), \forall n\in\{0,\ldots,N+1\}, \forall i\in\mathbb{Z}$.

(b) Etudier la consistance du schéma au sens DF.

(c) (Convergence) On pose $\overline{\overline{u}}_i^n=u(t_n,x_{i+\frac{1}{2}})$. Montrer que, si $k\leq \alpha h$, on a :

$$\sup_{i\in\mathbb{Z}}|u_i^n-\overline{\overline{u}}_i^n|\leq C_1(k+h), \forall n\in\{0,\ldots,N+1\},$$

où C_1 ne dépend que de u_0, β et T. En déduire que :

$$\sup_{i\in\mathbb{Z}}|u_i^n-\overline{u}_i^n|\leq C_2(k+h), \forall n\in\{0,\ldots,N+1\},$$

où C_2 ne dépend que de u_0, β et T.

Exercice 54 (Eq. lin., sol. faible, conv. des schémas VFDA et DFDA, méthode VF)
Corrigé en page 248

Soit $u_0\in L^\infty(\mathbb{R})\cap L^2(\mathbb{R})$ et $T\in\mathbb{R}_+^\star$. On considère le problème suivant :

$$u_t(x,t)+u_x(x,t) = 0, x\in\mathbb{R}, t\in[0,T], \quad (5.6.46)$$

$$u(x,0) = u_0(x). \quad (5.6.47)$$

Ce problème admet une et une seule solution faible, notée u. On se donne un pas de temps, k, avec $k=\frac{T}{N+1}$ ($N\in\mathbb{N}$), et on pose $t_n=nk$, pour $n\in\{0,\ldots,N+1\}$; On se donne des points de discrétisation en espace, $(x_i)_{i\in\mathbb{Z}}$, et on suppose que x_i est le centre de la maille $M_i=[x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}]$, pour $i\in\mathbb{Z}$. On pose $h_i=x_{i+\frac{1}{2}}-x_{i-\frac{1}{2}}$ et $h_{i+\frac{1}{2}}=x_{i+1}-x_i$. Soient $\alpha,\beta\in\mathbb{R}$. On suppose que, pour un certain $h\in\mathbb{R}$, $\alpha h\leq h_i\leq \beta h$, pour tout $i\in\mathbb{Z}$. On considère le schéma (5.6.44),(5.6.45) (schéma "VFDA").

1. (Stabilité L^∞) Montrer que $k\leq \alpha h\Rightarrow |u_i^n|\leq \|u_0\|_\infty, \forall n\in\{0,\ldots,N+1\}, \forall i\in\mathbb{Z}$.

2. Montrer que, pour tout $n = 0,\dots,N$, on a $u_i^n \to 0$ lorsque $i \to +\infty$ ou $i \to -\infty$.
3. (Estimation "BV faible") Soient $\zeta > 0$. Montrer que :

$$k \leq (1-\zeta)\alpha h \Rightarrow \sum_{n=0,\dots,N} \sum_{i\in\mathbb{Z}} k(u_i^n - u_{i-1}^n)^2 \leq C(\zeta,u_0),$$

où $C(\zeta,u_0)$ ne dépend que de ζ et u_0 (multiplier (5.6.44) par ku_i^n et sommer sur i et n.)
4. (convergence) On pose $\mathcal{T} = (M_i)_{i\in\mathbb{Z}}$ et on définit la solution approchée sur $[0,T]\times\mathbb{R}$, notée $u_{\mathcal{T},k}$, donnée par (5.6.44),(5.6.45), par $u_{\mathcal{T},k}(t,x) = u_i^n$, si $x \in \mathcal{M}_i$ et $t \in [t_n,t_{n+1}[$.
On admet que $u_{\mathcal{T},k} \to u$, pour la topologie faible-$\star$ de $L^\infty(]0,T[\times\mathbb{R})$, quand $h \to 0$, avec $k \leq (1-\zeta)\alpha h$ (ζ fixé). Montrer que u est la solution faible de (5.6.46)-(5.6.47).
Remarques : On peut montrer le même résultat avec (5.6.42) au lieu de (5.6.44). On peut aussi montrer (cf. la suite du cours...) que la convergence est forte dans $L^p_{loc}(]0,T[\times\mathbb{R})$, pour tout $p < \infty$.

Exercice 55 (Construction d'une solution faible) *Corrigé en page 251*

1/ Construire une solution faible du problème

$$\begin{cases} u_t + (u^2)_x = 0 \\ u(x,0) = u_0(x) = \begin{cases} 1 & si\ x < 0 \\ 1-x & si\ x \in [0,1] \\ 0 & six > 1 \end{cases} \end{cases}$$

2/ Même question (mais nettement plus difficile...) pour le problème

$$\begin{cases} u_t + (u^2)_x = 0 \\ u(x,0) = u_0(x) = \begin{cases} 0 & si\ x < 0 \\ 1-x & si\ x \in [0,1] \\ 1 & six > 1 \end{cases} \end{cases}$$

Exercice 56 (Problème de Riemann)

Soit f la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par $f(s) = s^4$. Soit u_d et u_g des réels. Calculer la solution entropique du problème de Riemann (5.4.30) avec données u_d et u_g en fonction de u_d et u_g.

Exercice 57 (Non unicité des solutions faibles) *Corrigé en page 252*

On considère l'équation

$$\begin{cases} u_t + (u^2)_x = 0 \\ u(0,x) = \begin{cases} u_g & si\ x < 0 \\ u_d & si\ x > 0 \end{cases} \end{cases} \tag{5.6.48}$$

avec $u_g < u_d$.

1. Montrer qu'il existe $\sigma \in \mathbb{R}$ tel que si $\begin{cases} u(t,x) = u_g & si\ x < \sigma t \\ u(t,x) = u_d & si\ x > \sigma t \end{cases}$ alors u est solution faible de (5.6.48). Vérifier que u n'est pas solution entropique de (5.6.48).

2. Montrer que u définie par :

$$\begin{cases} u(t,x) = u_g & si\ x < 2u_g t \\ u(t,x) = \frac{x}{2t} & si\ 2u_g t \leq x \leq 2u_d t \\ u(t,x) = u_d & si\ x > 2u_d t \end{cases} \tag{5.6.49}$$

alors u est solution faible entropique de (5.6.48).

Exercice 58 (Problème de Riemann)

1. Déterminer la solution entropique de (5.4.30) dans le cas où f est strictement concave.
2. On se place dans le cas où f est convexe puis concave : plus précisément, on considère $f \in C^2(\mathbb{R},\mathbb{R})$ avec
   (i) $f(0) = 0$, $f'(0) = f'(1) = 0$
   (ii) $\exists a \in]0,1[$, tel que f est strictement convexe sur$]0,a[$, f est strictement concave sur $]a,1[$.

   On supposera de plus $u_g = 1$, $u_d = 0$.
   (a) Soit b l'unique élément $b \in]a,1[$ tel que $\frac{f(b)}{b} = f'(b)$; montrer que u définie par :

$$\begin{cases} u(t,x) = 1 & si\ x \leq 0 \\ u(t,x) = \xi & si\ x = f'(\xi)t,\ b < \xi < 1 \\ u(t,x) = 0 & si\ x > f'(b)t \end{cases}$$

   est la solution faible entropique de (5.4.30) (sous les hypothèses précédentes).
   (b) Construire la solution entropique du problème de Riemann dans le cas $f(u) = \frac{u^2}{u^2 + \frac{(1-u)^2}{4}}$ et u_g, $u_d \in [0,1]$. [Compliqué. On distinguera plusieurs cas.)

Exercice 59 (Stabilité de schémas numériques) *Corrigé en page 252*

Soient $f \in C^1(\mathbb{R},\mathbb{R})$ et $u_0 \in L^\infty(\mathbb{R})$. On considère le problème suivant :

$$u_t(x,t) + (f(u))_x(x,t) = 0, x \in \mathbb{R}, t \in [0,T], \tag{5.6.50}$$

$$u(x,0) = u_0(x). \tag{5.6.51}$$

On utilise ci dessous les notations du cours. On discrétise le problème (5.6.50),(5.6.51) par l'un des schémas vu en cours ("Flux-splitting", "Godunov", "Lax-Friedrichs modifié" et "Murman"). Montrer qu'il existe M (dépendant de la fonction "flux numérique" et de u_0) tel que $k \leq Mh_i$, pour tout $i \in \mathbb{Z}$, implique :

1. $\|u^{n+1}\|_\infty \leq \|u^n\|_\infty$ pour tout $n \in \mathbb{N}$.
2. (Plus difficile) $\sum_{i\in\mathbb{Z}} |u_{i+1}^{n+1} - u_i^{n+1}| \leq \sum_{i\in\mathbb{Z}} |u_{i+1}^n - u_i^n|$ pour tout $n \in \mathbb{N}$. (Cette estimation n'est intéressante que si $\sum_{i\in\mathbb{Z}} |u_{i+1}^0 - u_i^0| < \infty$, ce qui n'est pas toujours vrai pour $u_0 \in L^\infty(\mathbb{R})$. Cela est vrai si u_0 est une fonction à "variation bornée".)

Exercice 60 (Schéma de Murman) *Corrigé en page 254*

Soient $f \in C^1(\mathbb{R},\mathbb{R})$ et $u_0 \in L^\infty(\mathbb{R})$. On suppose que $A \leq u_0 \leq B$, p.p. sur $\mathbb{R}$. On s'intéresse au problème suivant :

$$u_t(x,t) + (f(u))_x(x,t) = 0, x \in \mathbb{R}, t \in \mathbb{R}_+, \tag{5.6.52}$$

$$u(x,0) = u_0(x),\ x \in \mathbb{R}. \tag{5.6.53}$$

Pour discrétiser le problème (5.6.52)-(5.6.53), on se donne un pas d'espace $h > 0$ et un pas de temps $k > 0$. On pose $M_i =]ih,ih + h[$ et on note u_i^n l'approximation recherchée de la solution exacte dans la maille M_i à l'instant nk. On considère le schéma de Murmann :

$$h\frac{u_i^{n+1} - u_i^n}{k} + (f_{i+\frac{1}{2}}^n - f_{i-\frac{1}{2}}^n) = 0, n \in \mathbb{N}, i \in \mathbb{Z}, \tag{5.6.54}$$

$$u_i^0 = \frac{1}{h}\int_{M_i} u_0(x)dx, i \in \mathbb{Z}, \tag{5.6.55}$$

avec $f_{i+\frac{1}{2}}^n = g(u_i^n,u_{i+1}^n)$ et $g \in C(\mathbb{R} \times \mathbb{R},\mathbb{R})$ définie par :
$g(a,a) = f(a)$ et, pour $a \neq b$,

$$g(a,b) = \begin{array}{l} f(a) \text{ si } \frac{f(b)-f(a)}{b-a} \geq 0, \\ f(b) \text{ si } \frac{f(b)-f(a)}{b-a} < 0. \end{array}$$

1. (Stabilité) Montrer qu'il existe M, ne dépendant que de f, A et B (on donnera la valeur de M en fonction de f, A et B) t.q. pour $k \leq Mh$ on ait :
   (a) (Stabilité L^∞) $A \leq u_i^n \leq B$, pour tout $n \in \mathbb{N}$ et tout $i \in \mathbb{Z}$,
   (b) (Stabilité BV) $\sum_{i\in\mathbb{Z}} |u_{i+1}^{n+1} - u_i^{n+1}| \leq \sum_{i\in\mathbb{Z}} |u_{i+1}^n - u_i^n|$ pour tout $n \in \mathbb{N}$. (Cette estimation n'est intéressante que si $\sum_{i\in\mathbb{Z}} |u_{i+1}^0 - u_i^0| < \infty$, ce qui n'est pas toujours vrai pour $u_0 \in L^\infty(\mathbb{R})$. Cela est vrai si u_0 est une fonction à "variation bornée".)

2. On prend, dans cette question, $f(s) = s^2$.
   (a) (Non monotonie) Montrer que si $A < 0$ et $B > 0$, la fonction g n'est pas "croissante par rapport à son premier argument et décroissante par rapport à son deuxième argument" sur $[A,B]^2$.
   (b) (Exemple de non convergence) Donner un exemple de non convergence du schéma. Plus précisément, donner u_0 t.q., pour tout $h > 0$ et tout $k > 0$, on ait $u_i^n = u_i^0$ pour tout $i \in \mathbb{Z}$ et pour tout $n \in \mathbb{N}$ (la solution discrète est donc "stationnaire") et pourtant $u(.,T)$ (u est la solution exacte de (5.6.52)-(5.6.53)) est différent de u_0 pour tout $T > 0$ (la solution exacte n'est donc pas stationnaire).
3. (Schéma "ordre 2", question plus difficile) Pour avoir un schéma "plus précis", on pose maintenant $p_i^n = \text{minmod}(\frac{u_{i+1}^n - u_{i-1}^n}{2h}, 2\frac{u_{i+1}^n - u_i^n}{h}, 2\frac{u_i^n - u_{i-1}^n}{h})$ et on remplaçe, dans le schéma précédent, $f_{i+\frac{1}{2}}^n = g(u_i^n,u_{i+1}^n)$ par $f_{i+\frac{1}{2}}^n = g(u_i^n + (h/2)p_i^n, u_{i+1}^n - (h/2)p_{i+1}^n)$. Reprendre les 2 questions précédentes (c'est à dire : "Stabilité L^∞", "Stabilité BV", "non monotonie" et "Exemple de non convergence").

Exercice 61 (Flux monotones)

Soient $f \in C^1(\mathbb{R},\mathbb{R})$ et $u_0 \in L^\infty(\mathbb{R})$; on considère l'équation hyperbolique non linéaire (5.4.17) qu'on rappelle :

$$\begin{cases} u_t + (f(u))_x = 0, & (x,t) \in \mathbb{R} \times \mathbb{R}_+, \\ u(x,0) = u_0(x). \end{cases} \tag{5.6.56}$$

On se donne un maillage $(]x_{i-1/2},x_{i+1/2}[)_{i\in\mathbb{Z}}$ de $\mathbb{R}$ et $k > 0$ et, pour $i \in \mathbb{Z}$, on définit une condition initiale approchée : $u_i^0 = \frac{1}{h_i}\int u_0(x)dx$, avec $h_i = x_{i+1/2} - x_{i-1/2}$.

Pour calculer la solution entropique de l'équation (5.4.17), on considère un schéma de type volumes finis explicite à trois points, défini par un flux numérique g, fonction de deux variables.

1. Ecrire le schéma numérique (i.e. donner l'expression de u_i^{n+1} en fonction des $(u_i^n)_{i\in\mathbb{Z}}$).

2. On suppose dans cette question que le flux g est monotone et lipschitzien en ses deux variables, c.à.d. qu'il existe $M \geq 0$ tel que pour tout $(x,y,z) \in \mathbb{R}^3$, $|g(x,z) - g(y,z)| \leq M|x-y|$ et $|g(x,y) - g(x,z)| \leq M|y-z|$. Montrer que le schéma numérique de la question précédente peut s'écrire sous la forme

$$u_i^{n+1} = H(u_{i-1}^n, u_i^n, u_{i+1}^n)$$

où H est une fonction croissante de ses trois arguments si k satisfait une condition de type $k \leq Ch_i$ pout tout i, où C est une constante à déterminer.

3. Montrer que si la fonction g est croissante par rapport à son premier argument et décroissant par rapport au second, et si $a,b \in \mathbb{R}$ sont tels que $a \leq b$, alors $g(a,b) \leq g(\xi,\xi)$ pour tout $\xi \in [a,b]$.

4. En déduire que si le flux g est monotone, alors il vérifie la propriété suivante:

$$\forall (a,b) \in \mathbb{R}^2, \begin{cases} g(a,b) \leq \min_{s\in[a,b]} f(s) \text{ si } a \leq b \\ g(a,b) \geq \max_{s\in[a,b]} f(s) \text{ si } a \geq b. \end{cases}$$

5. Soit g un flux monotone qui vérifie de plus $g(a,b) = f(u_{a,b})$. Montrer que

$$\forall (a,b) \in \mathbb{R}^2, \begin{cases} g(a,b) = \min_{s\in[a,b]} f(s) \text{ si } a \leq b \\ g(a,b) = \max_{s\in[b,a]} f(s) \text{ si } a \geq b. \end{cases}$$

Exercice 62 (Schémas pour les problèmes hyperboliques)

Soient $f \in C^2(\mathbb{R},\mathbb{R})$, $T > 0$ et $u_0 \in L^\infty(\mathbb{R}) \cap BV(\mathbb{R})$; on cherche une approximation de la solution de l'equation hyperbolique avec condition initiale :

$$\begin{aligned} u_t(x,t) + (f(u))_x(x,t) &= 0, x \in \mathbb{R}, t \in [0,T], && (5.6.57) \\ u(x,0) &= u_0(x). && (5.6.58) \end{aligned}$$

On note h (resp. $k = \frac{1}{N+1}$) le pas (constant, pour simplifier) de la discrétisation en espace (resp. en temps), et u_i^n la valeur approchée recherchée de u au temps nk dans la maille $M_i = [(i-\frac{1}{2})h,(i+\frac{1}{2})h]$, pour $n \in \{0,\ldots,N+1\}$ et $i \in \mathbb{Z}$. On considère le schéma obtenu par une discrétisation par volumes finis explicite à trois points :

$$\begin{aligned} \frac{u_i^{n+1} - u_i^n}{k} + \frac{1}{h}(f_{i+\frac{1}{2}}^n - f_{i-\frac{1}{2}}^n) &= 0, n \in \{0,\ldots,N+1\}, i \in \mathbb{Z}, && (5.6.59) \\ u_i^0 &= \frac{1}{h}\int_{M_i} u_0(x)dx, && (5.6.60) \end{aligned}$$

avec $f_{i+\frac{1}{2}}^n = g(u_i^n, u_{i+1}^n)$, où $g \in C^1(\mathbb{R},\mathbb{R})$.

1. Montrer que le schéma (5.6.59),(5.6.60) possède la propriété de "consistance des flux" ssi g est telle que :

$$g(s,s) = f(s), \forall s \in \mathbb{R}. \quad (5.6.61)$$

2. Montrer que le schéma, vu comme un schéma de différences finies, est, avec la condition (5.6.61), d'ordre 1 (c.à.d. que l'erreur de consistance est majorée par $C(h+k)$, où C ne dépend que de f et de la solution exacte, que l'on suppose régulière). Montrer que si le pas d'espace est non constant, la condition (5.6.61) est (en général) insuffisante pour assurer que le schéma (5.6.59),(5.6.60) (convenablement modifié) est consistant au sens des différences finies, et que le schéma est alors d'ordre 0.
3. On étudie, dans cette question, le schéma de Godunov, c.à.d. qu'on prend :

$$g(u_g,u_d) = f\big(u_{u_g,u_d}(0,t)\big),$$

où u_{u_g,u_d} est la solution du problème de Riemann :

$$\begin{aligned} u_t(x,t) + (f(u))_x(x,t) &= 0, \ (x,t) \in \mathbb{R} \times \mathbb{R}_+ && (5.6.62)\\ u(x,0) &= u_g \text{ si } x < 0, && (5.6.63)\\ u(x,0) &= u_d \text{ si } x > 0. && (5.6.64) \end{aligned}$$

   (a) Montrer que le schéma (5.6.59), (5.6.60) peut s'écrire :

$$u_i^{n+1} = u_i^n + C_i(u_{i+1}^n - u_i^n) + D_i(u_{i-1}^n - u_i^n),$$

   avec : $C_i = \dfrac{k}{h}\dfrac{f(u_i^n) - g(u_i^n,u_{i+1}^n)}{u_{i+1}^n - u_i^n} \geq 0$ et $D_i = \dfrac{k}{h}\dfrac{g(u_{i-1}^n,u_i^n) - f(u_i^n)}{u_{i-1}^n - u_i^n} \geq 0.$

   (b) On pose $A = \|u_0\|_\infty$, $M = \sup_{s\in[-A,A]=} |f'(s)|$, et h le pas (constant) d'espace. On suppose que k et h vérifient la condition de CFL :

$$k \leq \frac{h}{2M}.$$

   On note u^n la fonction définie par : $u^n(x) = (u_i^n)$ si $x \in M_i$; montrer que :

   – Stabilité L^∞ : $\|u^{n+1}\|_\infty \leq \|u^n\|_\infty (\leq \ldots \leq \|u^0\|_\infty), \forall\, n \in \{0,\ldots,N+1\}$. $(E1)$
   – Stabilité BV : $\|u^{n+1}\|_{BV} \leq \|u^n\|_{BV} (\leq \ldots \leq \|u^0\|_{BV}), \forall\, n \in \{0,\ldots,N+1\}$. $(E2)$

   On rappelle que, comme u_n est une fonction constante par morceaux, on a :

$$\|u^n\|_{BV} = \sum_{i\in\mathbb{Z}} |u_{i+1}^n - u_i^n|.$$

   (c) Remarque: on peut montrer (ce n'est pas facile) que si on a la condition "CFL" le schéma de Godunov converge.

4. On suppose maintenant $f(u) = au$, $a \in \mathbb{R}$, et on prend $g(\lambda,\mu) = \dfrac{\lambda+\mu}{2}$ (schéma centré).
   Montrer que pour tous $k,h > 0$, les conditions $(E1)$ et $(E2)$ sont fausses, c.à.d. qu'il existe $u_0 (\in L^\infty \cap BV)$ t.q. $\|u^1\|_\infty \not\leq \|u_0\|_\infty$, et $\|u_1\|_{BV} \not\leq \|u_0\|_{BV}$.
5. On étudie maintenant un schéma de type "MUSCL", i.e. On prend dans le schéma (5.6.59) $f_{i+\frac{1}{2}}^n = f(u_i^n + \frac{h}{2}p_i^n)$, où :

$$p_i^n = \begin{cases} \dfrac{\varepsilon_i^n}{2h}\min\big(|u_{i+1}^n - u_{i-1}^n|, \quad 4|u_{i+1}^n - u_i^n|, 4|u_i^n - u_{i-1}^n|\big), \text{ où } \varepsilon_i^n = \text{sign}(u_{i+1}^n - u_{i-1}^n) \\ \qquad\qquad\qquad\qquad \text{si } \text{sign}(u_{i+1}^n - u_{i-1}^n) = \text{sign}(u_{i+1}^n - u_i^n) = \text{sign}(u_i^n - u_{i-1}^n) \\ 0 \text{ sinon.} \end{cases}$$

(a) Montrer que $\frac{1}{h}(f^n_{i+\frac{1}{2}} - f^n_{i-\frac{1}{2}})$ est une approximation d'ordre 2 de $\big(f(u)\big)_x(x_i,t_n)$ aux points où $u \in C^2$ et $u_x \neq 0$.

(b) Montrer que sous une condition de type $k \leq Ch$, où C ne dépend que de u_0 et f, les conditions de stabilité $(E1)$ et $(E2)$ sont vérifiées.

Exercice 63 (Eléments finis pour une équation hyperbolique)

Soit $f \in C^1(\mathbb{R},\mathbb{R})$, $u_0 \in C(\mathbb{R})$ t.q. u_0 bornée; on considère la loi de conservation scalaire suivante:

$$\frac{\partial u}{\partial t}(x,t) + \frac{\partial}{\partial x}(f(u))(x,t) = 0, x \in \mathbb{R}, t \in \mathbb{R}_+, \tag{5.6.65}$$

avec la condition initiale:

$$u(x,0) = u_0(x). \tag{5.6.66}$$

On se donne un pas de discrétisation en temps constant k, on note $t_n = nk$ pour $n \in \mathbb{N}$, et on cherche à approcher $u(.,t_n)$. On note $u^{(n)}$ la solution approchée recherchée.

1. Montrer qu'une discrétisation par le schéma d'Euler explicite en temps amène au schéma en temps suivant:

$$\frac{1}{k}(u^{(n+1)} - u^{(n)}) + \frac{\partial}{\partial x}\Big(f(u^{(n)})\Big)(x) = 0, x \in \mathbb{R}, n \in \mathbb{N}^*, \tag{5.6.67}$$

$$u^0(x) = u_0(x). \tag{5.6.68}$$

On cherche à discrétiser (5.6.67) par une méthode d'éléments finis. On se donne pour cela une famille de points $(x_i)_{i\in\mathbb{Z}} \subset \mathbb{R}$, avec $x_i < x_{i+1}$.

2. On introduit les fonctions de forme P_1, notées $\Phi_i, i \in \mathbb{Z}$, des éléments finis associés au maillage donné par la famille de points $(x_i)_{i\in\mathbb{Z}}$; on effectue un développement de Galerkin de $u^{(n)}$ sur ces fonctions de forme dans (5.6.67) et (5.6.68); on multiplie l'équation ainsi obtenue par chaque fonction de forme, et on approche le terme $f(\sum_{j\in\mathbb{Z}} u^{(n)}_j \Phi_j)$ par $\sum_{j\in\mathbb{Z}} f(u^{(n)}_j)\Phi_j$, et on intègre sur $\mathbb{R}$. Montrer qu'on obtient ainsi un système d'équations de la forme:

$$\sum_{j\in\mathbb{Z}} a_{i,j}\frac{u^{(n+1)}_j - u^{(n)}_j}{k} + \sum_{j\in\mathbb{Z}} b_{i,j} f(u^{(n)}_j) = 0,\ i \in \mathbb{Z},\ n \in \mathbb{N}^*. \tag{5.6.69}$$

$$u^0_i = u_0(x_i)\ i \in \mathbb{Z}. \tag{5.6.70}$$

(les $a_{i,j}$ et $b_{i,j}$ sont à déterminer).

3. On effectue une "condensation de la matrice de masse", c.à.d. qu'on remplace les $a_{i,j}$ dans (5.6.69) par $\tilde{a}_{i,j}$ avec $\tilde{a}_{i,j} = 0$ si $i \neq j$ et $\tilde{a}_{i,i} = \sum_{j\in\mathbb{Z}} a_{i,j}$. Montrer que le schéma ainsi obtenu est identique à un schéma volumes finis sur le maillage $(K_i)_{i\in\mathbb{Z}}$ où $K_i =]x_{i-1/2},x_{i+1/2}[$, $x_{i+1/2} = (x_i + x_{i+1})/2$, avec approximation centrée du flux.

4. Montrer que ce schéma est instable, dans un (ou plusieurs) sens à préciser.

5. On remplace le flux numérique centré $F_{i+1/2}$ du schéma volumes finis obtenu à la question 3 par $G_{i+1/2} = F_{i+1/2} + D_{i+1/2}(u^{(n)}_i - u^{(n)}_{i+1})$. Montrer que l'approximation du flux reste consistante et que si les $D_{i+1/2}$ sont bien choisis, le nouveau schéma est stable sous une condition de CFL à préciser.

On considère maintenant la même équation de conservation, mais sur $\mathbb{R}^2$ (avec $f \in C^1(\mathbb{R},\mathbb{R}^2)$, $u_0 \in C(\mathbb{R}^2)$, bornée.

$$u_t(x,t) + \mathrm{div}(f(u))(x,t) = 0, x \in \mathbb{R}^2, t \in \mathbb{R}_+, \tag{5.6.71}$$

$$u(x,0) = u_0(x). \tag{5.6.72}$$

Soit $\mathcal{T}$ un maillage en triangles de $\mathbb{R}^2$, admissible pour une discrétisation par éléments finis P_1. Soit $\mathcal{S}$ l'ensemble des noeuds de ce maillage et $(\Phi_j)_{j\in\mathcal{S}}$ la famille des fonctions de forme éléments finis bilinéaires P_1. En conservant la même discrétisation en temps, on cherche une approximation de $u(.,t_n)$ dans l'espace engendré par les fonctions Φ_j.

6. Montrer qu'en suivant la même démarche qu'aux questions 2 et 3, on aboutit au schéma:

$$\frac{u_i^{(n+1)} - u_i^{(n)}}{k} \int_{\mathbb{R}^2} \Phi_i(x)dx - \sum_{j\in\mathcal{S}} f(u_j^{(n)}) \cdot \int_{\mathbb{R}^2} \Phi_j(x)\nabla\Phi_i(x)dx = 0, n \in \mathbb{N}^* \tag{5.6.73}$$

7. Montrer que ce schéma peut encore s'écrire:

$$\frac{u_i^{(n+1)} - u_i^{(n)}}{k} \int_{\mathbb{R}^2} \Phi_i(x)dx + \sum_{j\in\mathcal{S}} E_{i,j} = 0, \tag{5.6.74}$$

avec

$$E_{i,j} = \frac{1}{2}(f(u_i^{(n)}) + f(u_j^{(n)})) \cdot \int_{\mathbb{R}^2} (\Phi_i(x)\nabla\Phi_j(x) - \Phi_j(x)\nabla\Phi_i(x))dx.$$

Montrer que ce schéma est instable.

8. Dans le schéma (5.6.74), on remplace $E_{i,j}$ par

$$\tilde{E}_{i,j}^n = E_{i,j}^n + D_{i,j}(u_i^n - u_j^n),$$

où $D_{i,j} = D_{j,i}$ (pour que le schéma reste conservatif). Montrer que pour un choix judicieux de $D_{i,j}$, le schéma ainsi obtenu est à flux monotone et stable sous condition de CFL.

5.7 Suggestions pour les exercices

Exercice 51 page 236

Chercher les solutions sous la forme $u(\mathbf{x},t) = u_0(\mathbf{x} - \mathbf{v}t)$.

5.8 Corrigés des exercices

Corrigé de l'exercice 50 page 236

1. En appliquant les ré'sultats de la section 5.2 page 215, la solution faible du problème s'écrit $u(x,t) = u_0(x + 2t)$, pour $x \in \mathbb{R}$, et $t \in \mathbb{R}_+$, c.à.d.

$$\left\{ \begin{array}{l} u(x,t) = \left\{ \begin{array}{l} 0, \text{ si } x < -2t, \\ 1 \text{ si } x > -2t. \end{array} \right. \end{array} \right. \tag{5.8.75}$$

La représentation graphique de la solution a $t = 0$ et à $t = 1$, en fonction de x est donnée en Figure 5.7. Cette solution faible n'est pas solution classique de (5.8.75) car elle n'est pas continue, donc ses dérivées

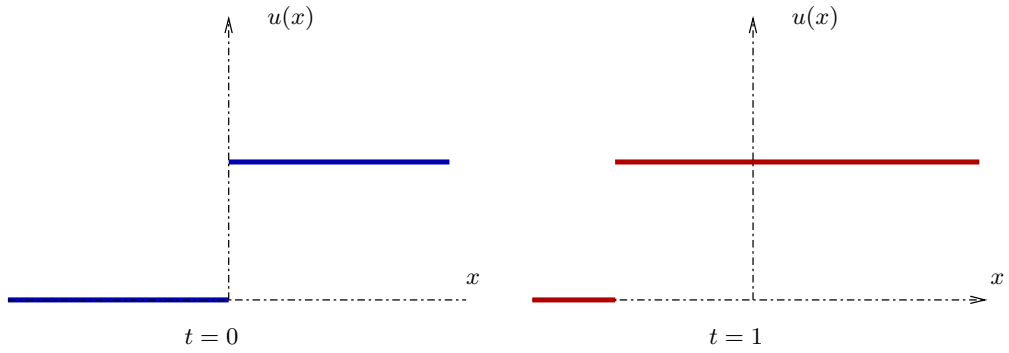

FIG. 5.7 – *Représentation graphique de la solution*

en temps et espace ne sont pas définies partout.
2. Dans le cas où $u_0(x) = \sin x$, la solution faible du problème s'écrit $u(x,t) = \sin(x + 2t)$, pour $x \in \mathbb{R}$, et $t \in \mathbb{R}_+$, et cette solution est régulière, donc solution classique.

Corrigé de l'exercice 51 page 236

Pour $(x,t) \in \mathbb{R}^2 \times \mathbb{R}$, on pose $u(x,t) = u_0(x - \mathbf{v}t)$. Comme $u_0 \in C^1(\mathbb{R},\mathbb{R})$, on a $u \in C^1(\mathbb{R}^2 \times \mathbb{R}_+,\mathbb{R})$; on peut donc calculer les dérivées partielles de u par rapport à au temps t, qu'on notera $\partial_t u$ et par rapport aux deux variables d'espace x_1 et x_2, qu'on notera $\partial_1 u$ et $\partial_2 u$. On a : $\partial_t u(x,t) = \nabla u_0(x - \mathbf{v}t) \cdot \mathbf{v}$. Or $\mathrm{div}(\mathbf{v}u) = \mathbf{v} \cdot \nabla u$ car $\mathbf{v}$ est constant, et $\nabla u = \nabla u_0$. On en déduit que $u_t(x,t) + \mathrm{div}(\mathbf{v}u)(x,t) = 0$, et donc u est solution (classique) de (5.6.39).

Corrigé de l'exercice 52 page 236 (Stabilité du schéma amont dans le cas linéaire)

On considère le problème hyperbolique linéaire (5.3.9), avec $u_0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$, dont on calcule une solution approchée par le schéma volumes finis amont (5.3.15). Montrer que ce schéma est stable pour les normes L^2 et L^∞, c.à.d. que la solution approchée satisfait les propriétés suivantes:

1. $\|u_{\mathcal{T},k}(.,n)\|_{L^2(\mathbb{R})} \leq \|u_0\|_{L^2(\mathbb{R})}, \forall n \in \mathbb{N}$,
2. $\|u_{\mathcal{T},k}(.,n)\|_{L^\infty(\mathbb{R})} \leq \|u_0\|_{L^\infty(\mathbb{R})}, \forall n \in \mathbb{N}$,

où $u_{\mathcal{T},k}$ désigne la solution approchée calculée par le schéma (voir (5.3.16)).
Le schéma (5.3.15) s'écrit encore :

$$h_i(u_i^{n+1} - u_i^n) = k(u_i^n - u_{i-1}^n).$$

Multiplions par u_i^{n+1}. On obtient:

$$\frac{1}{2}h_i(u_i^{n+1} - u_i^n)^2 + \frac{1}{2}h_i(u_i^{n+1})^2 - \frac{1}{2}h_i(u_i^n)^2 + k(u_i^{n+1} - u_i^n)(u_i^n - u_{i-1}^n) + ku_i^n(u_i^n - u_{i-1}^n) = 0.$$

Corrigé de l'exercice 53 page 236

1.a) Le schéma numérique s'écrit :

$$u_i^{n+1} = (1 - \frac{k}{h_{i-\frac{1}{2}}})u_i^n + \frac{k}{h_{i-\frac{1}{2}}}u_{i-1}^n \tag{5.8.76}$$

Comme $k \leq \alpha h \leq h_{i-\frac{1}{2}}$ on a $\dfrac{k}{h_{i-\frac{1}{2}}} \in$ [0,1] On a donc

$$\min(u_i^n, u_{i-1}^n) \leq u_i^{n+1} \leq \max(u_i^n, u_{i-1}^n)$$

d'où on déduit que

$$\min_j(u_j^n) \leq u_i^{n+1} \leq \max_j(u_j^n), \quad \forall i \in \mathbb{Z},$$

puis, par récurrence sur n, que

$$\inf u_0 \leq u_i^n \leq \sup u_0 \qquad \forall i \in \mathbb{Z}, \forall n \mathbf{n} \mathbb{N}.$$

1.b) Par définition de l'erreur de consistance, on a :

$$\begin{array}{l} \frac{\bar{u}_i^{n+1} - \bar{u}_i^n}{k} = u_t(x_i, tn) + R_i^n, \text{ où } |R_i^n| \leq \|u_{tt}\|_\infty k, \\ \frac{\bar{u}_i^n - \bar{u}_{i-1}^n}{h_{i-\frac{1}{2}}} = u_x(x_i, tn) + S_i^n, |S_i^n| \leq \|u_{xx}\|_\infty h.\beta \end{array}$$

En posant $e_i^n = \bar{u}_i^n - u_i^n$, on a donc

$$\frac{e_i^{n+1} - e_i^n}{k} + \frac{1}{h_{i-\frac{1}{2}}}(e_i^n - e_{i-1}^n) = R_i^n + S_i^n \leq C(u_0, \beta)(h + k),$$

avec $C(u_0\beta) = \|u_0''\|_\infty \max(\beta, 1)$, car $(u(t,x) = u_0(x,t))$ et donc $\|u_{tt}\|_\infty = \|u_{xx}\|_\infty = \|u_0''\|_\infty$. On pose $C(u_0, \beta) = C$, on obtient alors

$$e_i^{n+1} = \left(1 - \frac{k}{h_{i-\frac{1}{2}}}\right) e_i^n + \frac{k}{h_{i-\frac{1}{2}}} e_{i-1}^n + Ck(h + k)$$

donc $\sup_i |e_i^{n+1}| \leq \sup_j |e_j^n| + Ck(h+k)$. Par récurrence sur n, on en déduit

$$\sup_i |e_i^n| \leq Ckn(h+k) \text{ et donc } \sup_i |e_i^n| \leq CT(h+k) \text{ si } 0 \leq n \leq N+1, \text{ où } (N+1)k = T.$$

2.a) On a $\inf u_0 \leq u_i^0 \leq \sup u_0$ puis, pa récurrence :

$$u_i^{n+1} = (1 - \frac{k}{h_i})u_i^n + \frac{k}{h_i}u_{i-1}^n.$$

Comme $k \leq \alpha h \leq h_i$ on en déduit comme en 1) a) que :

$$\inf(u_0) \leq u_i^h \leq \sup(u_0).$$

2.b) Consistance

$$\frac{\bar{u}_i^{n+1} - \bar{u}_i^n}{k} = u_t(x_i,t_n) + R_i^n, |R_i^n| \leq \|u_{tt}\|_\infty k$$

mais

$$\frac{\bar{u}_i^n - \bar{u}_{i-1}^n}{h_i} = \frac{\bar{u}_i^n - \bar{u}_{i-1}^n}{h_{i-\frac{1}{2}}} \frac{h_{i-\frac{1}{2}}}{h_i} = [u_x(x_i,t_n) + S_i^n]\frac{h_{i-\frac{1}{2}}}{h}, \text{ avec } |S_i^n| \leq \|u_{xx}\|_\infty \beta h,$$

donc

$$\frac{\bar{u}_i^{n+1} - \bar{u}_i^n}{k} + \frac{\bar{u}_i^n - \bar{u}_{i-1}^n}{h_i} = (u_t + u_x)(x_i,t_n) + R_i^n + S_i^n \frac{h_{i-\frac{1}{2}}}{h_i} + T_i^n,$$

$$= R_i^n + S_i^n \frac{h_{i-\frac{1}{2}}}{h_i} + T_i^n,$$

avec :

$$|R_i^n| \leq \|u_{tt}\|_\infty k,$$
$$\left|\frac{h_{i-\frac{1}{2}}}{h_i}\right| |S_i^n| \leq \|u_{xx}\|_\infty \frac{\beta h}{\alpha h}\beta h = \frac{\beta^2}{\alpha}\|u_{xx}\|_\infty h,$$
$$T_i^n = u_x(x_i,t_n)\frac{h_{i-\frac{1}{2}} - h_i}{h_i} = u_x(x_i,t_n)\frac{h_{i-1} - h_i}{2h_i}.$$

En prenant par exemple un pas tel que $h_i = h$ si i est pair et $h_i = h/2$ si i est impair, on voit T_i^n ne tend pas vers 0 lorsque h tend vers 0; le schéma apparait donc comme non consistant au sens des différences finies.

c) Convergence. On a

$$\frac{\bar{\bar{u}}_i^{n+1} - \bar{\bar{u}}_i^n}{k} = u_t(x_{i+\frac{1}{2}},t_n) + R_i^n, |R_i^n| \leq \|u_{tt}\|_\infty k$$

$$\frac{\bar{\bar{u}}_i^n - \bar{\bar{u}}_{i-1}^n}{h_i} = u_x(x_{i+\frac{1}{2}},t_n) + S_i^n, |S_i^n| \leq \|u_{xx}\|_\infty \beta h$$

donc, avec $f_i^n = \bar{\bar{u}}_i^n - u_i^n$

$$f_i^{n+1} = (1 - \frac{k}{h_i})f_i^n + f_{i-1}^n(\frac{k}{h_i}) + k(S_i^n + R_i^n)$$

On a donc :

$$\begin{aligned} \sup_i |f_i^{n+1}| &\leq \sup_i |f_i^n| + k\|u_0''\|_\infty(k + \beta h) \\ &\leq \sup_i |f_i^n| + kC_1(k+h) \qquad C_1 = \|u_0''\|_\infty \max(\beta,1), \end{aligned}$$

et par récurrence sur n

$$\sup_i |f_i^n| \leq C_1 nk(k+h) + \|u_0'\|_\infty h\beta$$

car $\sup_i |f_i^0| \leq \|u_0'\|_\infty h\beta$. D'où on deéduit que

$$\begin{aligned}\sup_i |f_i^n| &\leq C_1 T(k+h) + \|u_0'\|_\infty \beta h \\ &\leq C_2(k+h), \qquad 0 \leq n \leq N+1.\end{aligned}$$

avec $C_2 = C_1 T + \beta\|u_0'\|_\infty$ Il reste à remarquer que $|\bar{u}_i^n - \bar{\bar{u}}_i^n| \leq \|u_0'\|_\infty \beta h$ pour avoir

$$\sup_i |\bar{u}_i^n - u_i^n| \leq C_3(h+k) \quad \text{avec}$$

$$C_3 = C_2 + \beta\|u_0'\|_\infty = \|u_0''\|_\infty \max(\beta,1)T + 2\beta\|u_0'\|_\infty.$$

Corrigé de l'exercice 54 page 237

1. On remarque d'abord que $|u_i^0| \in [-\|u_0\|_\infty, \|u_0\|_\infty]$. On a vu à la question 2) a) de l'exercice 53 que $u_i^{n+1} \in [u_i^n, u_{i-1}^n[$ ou $[u_{i-1}^n, u_i^n]$. On en déduit par une récurrence sur n que $u_i^n \in [-\|u_0\|_\infty, \|u_0\|_\infty] \quad \forall i, \forall n \geq 0$.

2. On va utiliser le fait que $u_0 \in L^2$ et montrer la propriété par récurrence sur n. Pour $n = 0$, on a :

$$|u_i^0|^2 \leq \int_{x_{i-\frac{1}{2}}}^{x_{i-\frac{1}{2}}} (u_0(x))^2 dx \frac{1}{h_i} \to 0 \text{ lorsque } i \to \pm\infty \tag{5.8.77}$$

En effet, comme $u_0 1_{[x,x+\eta[} \to 0 pp$, $u_0 1_{[x,x+\eta[} \leq u_0 \in L^2$ donc $\int_x^{x+\eta} |u_0|^2 dx \to 0$ lorsque $x \to +\infty$ par convergence dominée, pour tout $\eta > 0$. De plus, $h \geq \alpha h (\Rightarrow \frac{1}{h_i} \leq \frac{1}{\alpha h})$, d'où on déduit que (5.8.77) est vérifiée. On conclut ensuite par une récurrence immédiate sur n, que :

$$|u_i^{n+1}| \leq \max(|u_i^n||u_{i-1}^n|) \to 0 \text{ quand } i \to \pm\infty. \tag{5.8.78}$$

2. On veut montrer que $\sum_{n=0}^{N} \sum_{i\in\mathbb{Z}} k(u_i^n - u_{i-1}^n)^2 \leq C(\zeta, u_0)$. On multiplie le schéma par ku_i^n, on obtient :

$$h_i(u_i^{n+1} - u_i^n)u_i^n + (u_i^n - u_{i-1}^n)ku_i^n = 0,$$

ce qu'on peut réécrire :

$$h_i\left[-\frac{(u_i^{n+1} - u_i^n)^2}{2} + \frac{(u_i^{n+1})^2}{2} - \frac{(u_i^n)^2}{2}\right] + k\left[\frac{(u_i^n - u_{i-1}^n)^2}{2} + \frac{(u_i^n)^2}{2} - \frac{(u_{i-1}^n)^2}{2}\right] = 0.$$

Comme $|u_i^{n+1} - u_i^n| = \frac{k}{h_i}|u_i^n - u_{i-1}^n|$, ceci s'écrit aussi :

$$k(1 - \frac{k}{h_i})(u_i^n - u_{i-1}^n)^2 + h_i(u_i^{n+1})^2 - h_i(u_i^n)^2 + k(u_i^n)^2 - k(u_{i-1}^n)^2 = 0,$$

et comme $\frac{k}{h} \leq 1-\zeta$, on a donc $1-\frac{k}{h_i} \geq \zeta$ et $\zeta(u_i^n - u_{i-1}^n)^2 + h_i(u_i^{n+1})^2 - h_i(u_i^n)^2 + (u_i^n)^2 - (u_{i-1}^n)^2 \leq 0$ en sommant pour $i \in \{-M,\ldots,M\}$, et $h \in \{0,\ldots,N\}$, on obtient alors :

$$\zeta \sum_{i=-M}^{M} \sum_{n=0}^{M} (u_i^n - u_{i-1}^n)^2 + \alpha h \sum_{n=0}^{N} (u_M^n)^2 - \beta h \sum_{n=0}^{N} (u_{-M-1}^n)^2 \leq \sum_{i=-M}^{M} (u_i^0)^2.$$

En remarquant que

$$k \sum_{i=-M}^{M} (u_i^0)^2 \leq \sum_{i=-M}^{M} h_i (u_i^0)^2 \leq \|u_0\|_2^2$$

(voir (5.8.77)) et que $u_{-M}^n \to 0$ qd $M \to \infty$ (voir (5.8.78)), on en déduit

$$\zeta k \sum_{i=-\infty}^{\infty} \sum_{n=0}^{N} (u_i^n - u_{i-1}^n)^2 \leq \|u_0\|_2^2,$$

donc $C = \frac{\|u_0\|_2^2}{\zeta}$ convient.

3) (Convergence) Pour montrer la convergence, on va passer à la limite sur le schéma numérique. On aura pour cela besoin du lemme suivant :

Lemme 5.34 *Soit $(u_n)_{n\in\mathbb{N}}$ une suite bornée dans $L^\infty(\mathbb{R})$. Si $u_n \to u$ dans $L^\infty(\mathbb{R})$ pour la topologie faible $*$ lorsque $n \to +\infty$, (c.à.d*

$$\int_{\mathbb{R}} u_n(x)\varphi(x)dx \xrightarrow[n\to+\infty]{} \int_{\mathbb{R}} u(x)\varphi(x)dx,\ \forall \varphi \in L^1(\mathbb{R}),$$

et $v_n \to v$ dans L^1 lorsque $n \to +\infty$, alors

$$\int u_n(x)v_n(x)dx \xrightarrow[n\to+\infty]{} \int u(x)v(x)dx.$$

Démonstration :

$$\begin{aligned} &|\int u_n(x)v_n(x)dx - \int u(x)v(x)dx| \leq \|u_n\|_\infty \|v_n - v\|_1 + |\int u_n(x)v(x)dx - \int u(x)v(x)dx| \\ &\leq C\|u_n - v\|_1 + |\int u_n(x)v_n(x)dx - \int u(x)v(x)dx| \xrightarrow[n\to+\infty]{} 0, \end{aligned}$$

car $(u_n)_n$ est bornée dans L^∞. ∎

On multiplie le schéma numérique par $k\varphi_i^n$,$\varphi \in C_c^\infty(\mathbb{R}\times[0,T[)$ et $\varphi_i^n = \varphi(x,t_n)$, et en somme sur i et n (toutes les sommes sont finies, car φ est à support compact) ; on obtient :

$$\sum_{i\in\mathbb{Z}} \sum_{n\in\mathbb{N}} \frac{u_i^{n+1} - u_i^n}{k} k h_i \varphi_i^n + \sum_{i\in\mathbb{Z}} \sum_{n=1}^{N} (u_i^n - u_{i-1}^n) k \varphi_i^n = 0.$$

Comme $\varphi_i^n = 0$ si $n \geq N+1$, on a :

$$\sum_{i\in\mathbb{Z}} \sum_{n=1}^{N} h_i u_i^n (\varphi_i^{n-1} - \varphi_i^n) - \sum_i u_i^0 \varphi_i^0 h_i + \sum_{i\in\mathbb{Z}} \sum_n (\varphi_i^n - \varphi_{i+1}^n) u_i^n k = 0.$$

Or :

- $T_1 = \sum_{i\in\mathbb{Z}} u_i^0\varphi_i^0 h_i = \sum_i \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u_0(x)\varphi_0(x_i)dx \xrightarrow[h\to 0]{} \int u_0\varphi dx$.(avec $\varphi_0 = \varphi(.,0)$)
  car $\sum_i \varphi_0(x_i)1_{]x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}[} \to \varphi(.,0)$ dans L^1 quand $h \to 0$.
- $T_2 = \sum_{i\in\mathbb{Z}}\sum_{n=1} h_i u_i^n \frac{\varphi_i^{n-1}-\varphi_i^n}{k}k = -\int_{\mathbb{R}_+}\int_{\mathbb{R}} u_{\mathcal{T},k}\psi_{\mathcal{T},k}dxdt$. Soit

$$\psi_{\mathcal{T},k}(x,t) = \sum_{i\in\mathbb{Z}}\sum_{n=1}^{N} \frac{\varphi_i^{n-1}-\varphi_i^n}{k} 1_{]x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}[}1_{]nk,(n+1)k[}.$$

  En effet, pour $x \in \mathbb{R}$ et $t > 0$, $|\frac{\varphi_i^{n-1}-\varphi_i^n}{k} - \varphi_t(x,t)| \le k\|\varphi_{tt}\|_\infty$ si $(x,t) \in]x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}[\times[nk(n+1)k[$, pour $n \ge 1$. On a donc donc $\psi_{\mathcal{T},k} \to \varphi_t$ p.p. sur $\mathbb{R}\times]0,T[$, De plus, et $|\psi_{\mathcal{T},k}| \le \|\varphi_t\|_\infty 1_K$ si $\beta h \le 1$, où $K = [-a-1,a+1]\times[0,T]$, et a est tel que $\varphi = 0$ sur $([-a,a]\times[0,T])^c$ Donc, par convergence dominée, $\psi_{\mathcal{T},k} \to -\varphi_t$ dans $L^1(\mathbb{R}\times]0,T[)$ lorsque $k \to 0$. Comme $u_{\mathcal{T},k}$ converge vers u dans L^infty faible $*$, on en déduit par le lemme 5.34 que :

$$T_2 == -\int_{\mathbb{R}_+}\int_{\mathbb{R}} u_{\mathcal{T},k}(x,t)\psi_{\mathcal{T},k}(x,t)dxdt \xrightarrow[h,k\to 0]{} -\int_{\mathbb{R}_+}\int_{\mathbb{R}} u(x,t)\varphi_t(x,t)dxdt.$$

- $T_3 = \sum_{i\in\mathbb{Z}}\sum_{n\in\mathbb{N}} \frac{\varphi_i^n-\varphi_{i+1}^n}{h_i}u_i^n k h_i. = \sum_{i\in\mathbb{Z}}\sum_{n\in\mathbb{N}} k\varphi_i^n(u_i^n - u_{i-1}^n)$
  $= \sum_{i\in\mathbb{Z}}\sum_{n\in\mathbb{N}} k\varphi_{i-1}^n(u_i^n-u_{i-1}^n) + \sum_{i\in\mathbb{Z}}\sum_{n\in\mathbb{N}} k(\varphi_i^n - \varphi_{i-\frac{1}{2}}^n)(u_i^n-u_{i-1}^n) = T_4 + T_5$, avec :
  * $T_4 = \sum_i\sum_n kh_i\frac{\varphi_{i-\frac{1}{2}}^n - \varphi_{i+\frac{1}{2}}^n}{h}u_i^n = \int\int u_{\mathcal{T}k}(x)\chi_{\mathcal{T}k}(x)dx$ où

$$\chi_{\mathcal{T}k} = \frac{\varphi_{i-\frac{1}{2}}^n - \varphi_{i+\frac{1}{2}}^n}{h_i} \text{ sur }]x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}[\times]t_n,t_{n+1}[\ ;$$

    et donc $\chi_{\mathcal{T}k} \to -\varphi_x$ dans $L^1(\mathbb{R}\times]0,1[)$ et $T_4 \to -\int\int u(x)\varphi_x(x)dxdt$ lorsque $h \to 0$,
  * $T_5 \le \sum_{i=M}^{M_2}\sum_{n=0}^{N} k\beta h\|\varphi_x\|_\infty(u_i - u_{i-1}^h) \le \beta kh\|\varphi_x\|_\infty \sum_{n=0}^{N}\sum_{i=M}^{M_2}(u_i^n - u_{i-1}^n)$ si $\beta h \le 1$, où M_1 et M_2 sont tels que $i \notin \{M_1,\dots,M_2\} \Rightarrow]x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}[\subset [-a,a]^c$, et $\varphi = 0$ sur $([-a,a]\times[0,T])^c$. On a donc :

$$\begin{aligned}
&T_5 \le \beta kh\|\varphi_x\|(\sum_{i=M_1}^{M_2}\sum_{n=0}^{N}(u_i^n - u_{i-1}^n)^2)^{1/2}(\sum_{n=0}^{N}\sum_{i=M_1}^{M_2}1)^{1/2}\\
&\le \beta kh\|\varphi_x\|_\infty\tfrac{\sqrt{c}}{\sqrt{k}}(\textstyle\sum_{n=0}^{N}\sum_{i=M_1}^{M_2}1)^{1/2},\\
&\le \beta\sqrt{k}h\|\varphi_x\|_\infty\sqrt{c}\quad\sqrt{N+1}\sqrt{M_2-M_1}\quad (M_2-M_1)\alpha h \le 2a.\\
&\le \beta\sqrt{k}h\|\varphi_x\|_\infty\sqrt{c}\tfrac{\sqrt{T}}{\sqrt{k}}\quad\tfrac{\sqrt{2}a}{\sqrt{\alpha}h} = \beta\|\psi_x\|_\infty\sqrt{c}\tfrac{\sqrt{T}}{\sqrt{\alpha}}\sqrt{h}\\
&\to 0 \text{ quand } h \to 0.
\end{aligned}$$

  On en déduit que $T_3 \to -\int\int u(x)\varphi_x(x)dx$ quand $h \to 0$.

Comme $T_1 + T_2 + T_3 = 0$, on a donc

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} u(x,t)\varphi_t(x;t)dxdt + \int_{\mathbb{R}} \int_{\mathbb{R}_+} u(x,t)\varphi_x(x,t)dxdt + \int_{\mathbb{R}} u_0(x)\varphi(x,.)dx = 0$$

et donc u est solution faible de (5.6.46)-(5.6.47).

Corrigé de l'exercice 55 page 238

1/ Dans le premier cas, la solution est facile à construire par la méthode des caractéristiques, pour tout $t < 1/2$. En effet, les droites caractéristiques sont d'équation : $x(t) = 2u_0(x_0)t + x_0$, c'est-à-dire

$$x(t) = \begin{cases} 2t + x_0, \text{ si } x_0 < 0, \\ 2(1 - x_0)t + x_0, \text{ si } x_0 \in]0,1[, \\ 0 \text{ si } x_0 > 1. \end{cases}$$

Les droites caractéristiques se rencontrent à partir de $t = 1/2$, il y alors apparition d'un choc, dont la vitesse est donnée par la relation de Rankine–Hugoniot :

$$\sigma(u_g - u_d) = (u_g^2 - u_d^2), \text{ et donc } \sigma = u_g + u_d = 1.$$

La solution entropique est donc :

$$u(x,t) = \begin{cases} 1 \text{ si } t < \dfrac{1}{2} \text{ et } x < 2t \text{ ou si } t > \dfrac{1}{2} \text{ et } x < t + \dfrac{1}{2}, \\ \\ \frac{x-1}{2t-1} \\ 0 \text{ si } t < \frac{1}{2} \text{ et } x > 1 \text{ ou si } t > \frac{1}{2} \text{ et } x > t + \frac{1}{2}. \end{cases}$$

2/ On pourra montrer que la fonction définie par les formules suivantes est la solution pour $t < \frac{1}{2}$ (c'est-à-dire avant que les droites caractéristiques ne se rencontrent, la solution contient deux zones de détentes).

$$u(x,t) = 0, \text{ si } x < 0, t < \frac{1}{2},$$

$$u(x,t) = \frac{x}{2t}, \text{ si } 0 < x < 2t, t < \frac{1}{2},$$

$$u(x,t) = \frac{1-x}{1-2t}, \text{ si } 2t < x < 1, t < \frac{1}{2},$$

$$u(x,t) = \frac{x-1}{2t}, \text{ si } 1 < x < 1 + 2t, t < \frac{1}{2},$$

$$u(x,t) = 1, \text{ si } 1 + 2t < x, t < \frac{1}{2}.$$

En $t = \frac{1}{2}$, on pourra vérifier qu'un choc apparaît en $x = 1$ et se propage à la vitesse 1. On obtient alors pour $t > \frac{1}{2}$ la solution suivante :

$$u(x,t) = 0, \text{ si } x < 0, t > \frac{1}{2},$$

$$u(x,t) = \frac{x}{2t}, \text{ si } 0 < x < \frac{1}{2} + t, t > \frac{1}{2},$$

$$u(x,t) = \frac{x-1}{2t}, \text{ si } \frac{1}{2} + t < x < 1 + 2t,\ t > \frac{1}{2},$$

$$u(x,t) = 1, \text{ si } 1 + 2t < x,\ t > \frac{1}{2}.$$

Remarquons que, bien que la solution initiale soit discontinue, la solution entropique est continue pour $t \in]0,1/2[$.

Corrigé de l'exercice 57 page 238

1. La question 1 découle du point 1 de la proposition 5.29 page 230 (il faut que σ satisfasse la condition de Rankine–Hugoniot.
2. La question 2 découle du point 2 de la proposition 5.29 page 230.

Corrigé de l'exercice 59 page 239

Les quatre schémas s'écrivent sous la forme :

$$u_i^{n+1} = u_i^n - \frac{k}{h_i}(g(u_i^n,u_{i+1}^n) - g(u_i^n,u_i^n)) + \frac{k}{h_i}(g(u_{i-1}^n,u_i^n) - g(u_i^n,u_i^n))$$

soit encore

$$u_i^{n+1} = u_i^n + C_i^n(u_{i+1}^n - u_i^n) + D_i^n(u_{i-1}^n - u_i^n),$$

avec

$$C_i^n = \frac{k}{h_i}\frac{g(u_i^n,u_i^n) - g(u_i^n,u_{i+1}^n)}{u_{i+1}^n - u_i^n} \quad \text{si } u_i^n \neq u_{i+1}^n (0 \text{ sinon })$$

$$D_i^n = \frac{k}{h_i}\frac{g(u_{i-1}^n,u_i^n) - g(u_i^n,u_i^n)}{u_{i-1}^n - u_i^n} \quad \text{si } u_i^n \neq u_{i+1}^n (0 \text{ sinon })$$

On suppose que $A \leq u_0 \leq B$ p.p. et on remarque qu'il existe $L \in \mathbb{R}_+$ tel que :

$$\left.\begin{array}{l} |g(a,b) - g(a,c)| \leq L|b-c|, \\ |g(b,a) - g(c,a)| \leq L|b-c| \end{array}\right\} \qquad \forall a,b,c \in [A,B]$$

(On laisse le lecteur vérifier qu'un tel L existe pour les 4 schémas considérés).
1) Dans le cas des 3 premiers schémas (FS, Godunov et LFM), la fonction g est croissante par rapport au 1er argument et décroissante par rapport au 2ème argument. Donc si $u_i^n \in [A,B], \forall i$ (pour n fixé), on a $C_i^n \geq 0 \quad D_i^n \geq 0$. En prenant $2k \leq Lh_i \quad \forall i$ on a aussi : $C_i^n, D_i^n \leq \frac{1}{2}$ et donc u_i^{n+1} est une combinaison convexe de $u_{i-1}^n, u_i^n, u_{i+1}^n$ donc $u_i^{n+1} \in [A,B] \quad \forall i$ (et aussi $\|u^{n+1}\|_\infty \leq \|u^n\|_\infty$). Par récurrence sur n on en déduit :

$$u_i^n \in [A,B] \quad \forall i, \forall n \text{ si } k \leq uh_i \forall i \text{ avec } M = \frac{L}{2}$$

Dans le dernier cas (Murman), on a
$g(a,b) = f(a)$ si $\frac{f(b) - f(a)}{b-a} \geq 0 \quad (a \neq b)$, $g(a,b) = f(b)$ si $\frac{f(b) - f(a)}{b-a} < 0 \quad (a \neq b)$ et $g(a,a) = f(a)$.
Si $\frac{f(u_{i+1}^n) - f(u_i^n)}{u_{i+1}^n - u_i^n} \geq 0$, on a : $g(u_i^n,u_{i+1}^n) = f(u^n)$, donc $C_i^n = 0$.

Si $\dfrac{f(u_{i+1}^n)-f(u_i^n)}{u_{i+1}^n-u_i^n} < 0$, on a : $g(u_i^n,u_{i+1}^n) = f(u_{i+1}^n)$,$C_i^n = \frac{-f(u_{i+1}^n)+f(u^n)}{u_{i+1}^n-u_i^n} > 0$, et $C_i^n \leq \dfrac{1}{2}$ si $k \leq Mh_i$ avec $M = \dfrac{L}{2}$ (L est ici la constante de Lipschitz de f).
Le même calcul vaut pour D_i^n et on conclut comme précédemment car

$$u^{n+1} = (1 - C_i^n - D_i^n)u^n + C_i^n u_{i+1}^n + D_i^n u_{i-1}^n$$

2) On reprend la formule de 1) et la même limitation sur k (pour les 4 schémas). On a :

$$u_{i+1}^{n+1} = u_{i+1}^n + C_{i+1}^n(u_{i+2}^n - u_{i+1}^n) + D_{i+1}^n(u_i^n - u_{i+1}^n)$$

$$u_i^{n+1} = u_i^n + C_i^n(u_{i+1}^n - u_i^n) + D_i^n(u_{i-1}^n - u_i^n)$$

et donc, en soustrayant membre à membre :

$$u_{i+1}^{n+1} - u_i^{n+1} = (u_{i+1}^n - u_i^n)\underbrace{(1 - C_i^n - D_{i+1}^n)}_{\geq 0} + \underbrace{C_{i+1}^n}_{\geq 0}(u_{i+2}^n - u_{i+1}^n) + \underbrace{D_i^n}_{\geq 0}(u_i^n - u_{i-1}^n)$$

Par inégalité triangulaire, on a donc :

$$|u_{i+1}^{n+1} - u_i^{n+1}| \leq |u_{i+1}^n - u_i^n|(1 - C_i^n - D_{i+1}^n) + C_{i+1}^n|u_{i+2}^n - u_{i+1}^n| + D_i^n|u_i^n - u_{i-1}^n|$$

Sommons alors entre $i = -P$ à P :

$$\sum_{i=-P}^{P} \left|u_{i+1}^{n+1} - u_i^{n+1}\right| \leq \sum_{i=-P}^{P} \left|u_{i+1}^n - u_i^n\right| - \sum_{i=-P}^{P} C_i^n \left|u_{i+1}^n - u_i^n\right| + \sum_{i=-P}^{P} C_{i+1}^n \left|u_{i+2}^n - u_{i+1}^n\right|$$
$$- \sum_{i=-P}^{P} D_{i+1}^n \left|u_{i+1}^n - u_i^n\right| + \sum_{i=-P}^{P} D_i^n \left|u_i^n - u_{i-1}^n\right|.$$

En regroupant :

$$\sum_{i=-P}^{P} \left|u_{i+1}^{n+1} - u_i^{n+1}\right| \leq \sum_{i=-P}^{P} \left|u_{i+1}^n - u_i^n\right| + C_{P+1}^n \left|u_{P+2}^n - u_{P+1}^n\right| + D_{-P}^n \left|u_{-P}^n - u_{-P-1}^n\right|.$$

Or $C_{P+1}^n \in [0,1]$ et $D_{-P}^n \in [0,1]$ donc

$$\sum_{i=-P}^{P} \left|u_{i+1}^{n+1} - u_i^{n+1}\right| \leq \sum_{i=-P-1}^{P+1} \left|u_{i+1}^n - u_i^n\right| \leq \sum_{i=-\infty}^{+\infty} \left|u_{i+1}^n - u_i^n\right|.$$

Il ne reste plus qu'a faire tendre P vers $+\infty$ pour obtenir le résultat.

Corrigé de l'exercice 60 page 239

1) Cette question a été complètement traitée dans l'exercice 59.
Les estimations sont vérifiées avec $M = \dfrac{L}{2}$, où L est la constante de Lipschitz de f sur $[A,B]$.

2) Remarquons que si $f(s) = s^2$ alors $\dfrac{f(b) - f(a)}{b - a} = b + a$.
a) Soit $\bar{b} \in]0,B[, \bar{a} \in]A,0[$ tel que $\bar{b} + \bar{a} > 0$, (par exemple : $\bar{a} = -\frac{\epsilon}{2}, \bar{b} = \epsilon$ avec $0 < \epsilon < \min(-A,B)$). Soit $\alpha \in]0, \bar{a} + \bar{b}[$. Pour $a \in [\bar{a} - \alpha, \bar{a} + \alpha]$, on a $\bar{b} + a > 0$, et donc $g(a,\bar{b}) = f(a) = a^2$, ce qui prouve que sur l'ensemble $[\bar{a} - \alpha, \bar{a} + \alpha] \times \{\bar{b}\}$, la fonction g est décroissante par rapport à a.
b) Soit n u_0 définie par : $u_0 = \begin{cases} -1 \text{ sur } \mathbb{R}_- \\ +1 \text{ sur } \mathbb{R}_+ \end{cases}$ de sorte que $u_i^0 = \begin{cases} +1 \text{ si } i \geq 0 \\ -1 \text{ si } i < 0 \end{cases}$
Comme $f(u_i^0) = +1 \quad \forall i$ on a $u_i^1 = u_i^0 \quad \forall i$ et donc $u_i^n = u_i^0$ pour tout i et pour tout n. Par une récurrence facile, la solution approchée est donc stationnaire. La solution exacte n'est pas stationnaire (voir proposition 5.29, cas où f est strictement convexe et $u_g < u_d$).

Corrigé de l'exercice 61 page 240 (Flux monotones)

1. Le schéma s'écrit: $u_i^{n+1} = u_i^n - \frac{h_i}{k}(g(u_i^n, u_{i+1}^n) - g(u_i^n, u_{i-1}^n))$

2.

$$\begin{aligned} u_{i+1}^n) &= u_i^n + C_i^n(u_{i+1}^n - u_i^n) + D_i^n(u_{i-1}^n - u_i^n) \\ &= (1 - C_i^n - D_i^n)u_i^n + C_i^n u_{i+1}^n + D_i^n u_{i-1}^n \end{aligned}$$

$$\text{avec } C_i^n = \frac{h_i}{k}\frac{g(u_i^n, u_{i+1}^n) - g(u_i^n, u_i^n)}{u_{i+1}^n - u_i^n} \text{ si } u_{i+1}^n \neq u_i^n \text{ (et 0 sinon)}$$

$$\text{et } D_i^n = \frac{h_i}{k}\frac{g(u_i^n, u_{i+1}^n) - g(u_i^n, u_i^n)}{u_{i+1}^n - u_i^n} \text{ si } u_{i-1}^n \neq u_i^n \text{ (et 0 sinon)}$$

Remarquons que $C_i^n \geq 0$ et $D_i^n \geq 0$ car g est monotone. On en déduit que H définie par

$$H(u_{i-1}^n, u_i^n, u_{i+1}^n) = (1 - C_i^n - D_i^n)u_i^n + C_i^n u_{i+1}^n + D_i^n u_{i-1}^n$$

est une fonction croissante de ses arguments si $1 - C_i^n - D_i^n \geq 0$, ce qui est vérifié si $k \leq \frac{h_i}{2M}$ pour tout $i \in \mathbb{Z}$.

3. Comme $a \leq \xi$ $g(a,b) \leq g(\xi,b)$, et comme $\xi \leq b$, $g(\xi,b) \leq g(\xi,\xi)$.

4. D'après la question précédente, si $a \leq b$, on a bien $g(a,b) \leq \min\{g(\xi,\xi), \xi \in [a,b]\}$, et comme $g(\xi,\xi) = f(\xi)$, on a le résultat souhaité.
Si $a \geq b$, alors on vérifie facilement que : $g(a,b) \geq g(\xi,\xi)$ pour tout $\xi \in [b,a]$, ce qui prouve le résultat.

5. Comme $g(a,b) = f(u_{a,b})$, on a $\min_{s\in[a,b]} f(s) \leq g(a,b)$ si $a \leq b$ et $g(a,b) \leq \max_{s\in[b,a]} f(s)$ si $a \geq b$. On a donc égalité dans les inégalités de la question 3.
2. Montrer que sous une condition à préciser, le schéma peut s'écrire sour la forme

$$u_i^{n+1} = H(u_{i-1}^n, u_i^n, u_{i+1}^n)$$

où H est une fonction croissante de ses trois arguments.

Bibliographie

[1] Brezis, H. (1983), *Analyse Fonctionnelle: Théorie et Applications* (Masson, Paris).

[2] P.G. Ciarlet, Introduction a‘ l’analyse numérique et à l’optimisation, Masson 1982.

[3] Ciarlet, P.G. (1978), *The Finite Element Method for Elliptic Problems* (North-Holland, Amsterdam).

[4] Ciarlet, P.G. (1991), *Basic error estimates for elliptic problems* in: *Handbook of Numerical Analysis* **II** (North-Holland,Amsterdam) 17-352.

[5] R. Eymard, T. Gallouët and R. Herbin, Finite Volume Methods, *Handbook of Numerical Analysis*, Vol. VII, pp. 713-1020. Edited by P.G. Ciarlet and J.L. Lions (North Holland).

[6] T. Gallouët and R. Herbin, Théorie de l’intégration et de la mesure. http://www-gm3.univ-mrs.fr/ gallouet/licence.d/int-poly.pdf

[7] Godlewski E. and P. A. Raviart (1991), *Hyperbolic systems of conservation laws*, Ellipses.

[8] Godlewski E. and P.A. Raviart (1996), *Numerical approximation of hyperbolic systems of conservation laws*, *Applied Mathematical Sciences* **118** (Springer, New York).

[9] Godunov S. (1976), *Résolution numérique des problèmes multidimensionnels de la dynamique des gaz* (Editions de Moscou).

[10] R. Herbin, Analyse numérique. http://www.cmi.univ-mrs.fr/ herbin/PUBLI/polyananum.pdf

[11] Kröner D. (1997) *Numerical schemes for conservation laws in two dimensions*, Wiley-Teubner Series Advances in Numerical Mathematics. (John Wiley and Sons, Ltd., Chichester; B. G. Teubner, Stuttgart).

[12] LeVeque, R. J. (1990), *Numerical methods for conservation laws* (Birkhauser verlag).

[13] Quarteroni A., Sacco R., and Saleri F. Numerical mathematics. Springer, 2000.

[14] J. Rappaz and M. Picasso Introduction a‘ l’analyse numérique. Presses Polytechniques et Universitaires Romandes, Lausanne, 1998.

[15] P.A. Raviart and JM Thomas. Introduction a‘ l’analyse numérique des équations aux dérivées partielles.