Méthodes variationnelles pour les problèmes hyperbolique et paraboliques, application à la mécanique des fluides.

R. Abgrall
Mathématiques APpliquées de Bordeaux , Université Bordeaux I
Projet Scalapplix, INRIA FutURs

Avant propos

Dans ce cours, on s'interessera à l'approximation numérique des problèmes de type advection–diffusion, et plus généralement les équations de Navier Stokes, au moyen de formulation variationnelles. On pourra considérer des problèmes stationnaires ou instationnaires.

On partira de la formulation volume fini, vu comme une formulation variationnelle. Elle sera interprétée soit par une représentation des données aux noeuds d'un maillage, soit par une représentation constante par éléments. Cette méthode est d'ordre un, et on sait que quelque soit l'interpretation (aux termes visqueux pres), les schémas sont très proches. On s'interessera alors à la montée en précision, ce qui permettra alors de différencier les approches. On considèrera la méthode «Galerkin discontinu», généralisation naturelle de la méthode volume fini avec une représentation des inconues constante par élément. et aux méthodes d'éléments fini stabilisée (residual free bubbles, stream line diffusion, residual distribution).

On détaillera comment ces méthodes nouvelles interviennent dans les évolutions récentes des grands codes industriels que sont ElSa (ONERA), Tau (DLR=ONERA Allemand), CEDRE (ONERA) et EUGENI (DASSAULT).

Comment ne pas lire ce polycopié ...

Comme le titre l'indique, ce cours s'intéresse aux méthodes numériques permettant d'approcher les solutions d'équations hyperboliques non linéaires. Le sujet étant bien trop complexe, nous nous sommes contentés de décrire certaines méthodes d'approximation qui nous paraissent posséder les développements les plus prometteurs, en nous restreingant au cas scalaire. Le case des systèmes, en particulier celui des problèmes fluides n'est pas traité en détail dans ce cours. Chaque chapitre se clot néanmoins par des notes bibliographiques qui complètent les développements présentés dans le chapitre considéré en particulier du point de vue des systèmes.

Les méthodes numériques décrites dans ce cours partent toutes d'une formulation variationnelle qui reposent sur la formulation faible des problèmes hyperboliques non linéaires avec ou sans termes paraboliques. Ainsi, il faut définir un espace d'approximation et un espace de fonctions tests. De manière très grossière, deux classes de méthodes sont dérites ici, suivant qu'on considère des espaces d'approximations continues ou non.

Dans une première lecture, on suggère de parcourir les chapitres 1, 2, 4, 5 en omettant soigneusement les parties techniques, à savoir les sections 5.5.1 et 5.5.2. Ceci permettra d'avoir une idée de la méthode des volumes finis et de ses développements récents. Ici, on considère des espaces d'approximations discontinues. Pour les lecteurs intéressés par des formulations continues, on suggère la lecture du chapitre 6

Dans le chapitre 2, on pourra commencer par les sections 2.1 et 2.2 où sont décrits des exemples et des calculs exaustifs sont effectués. Les paragraphes 2.3, 2.4, 2.5 peuvent être abordés après avoir bien assimilé les paragraphes 2.1 et 2.2.

Le chapitre 3 d'ecrit quelques propriétés d'un modèle bien connu, celui des équations d'Euler. Il est relativement indépendant du reste du texte, mais permet –espérons–le– une meilleure compréhension du chapitre 4.

Dans le chapitre 4, il faut d'abord bien comprendre le paragraphe 4.1 puis 4.2. Le paragraphe 4.3 est plus «philosophique».

Dans le chapitre 5, il faut bien maîtriser les paragraphes 5.1, 5.3 et 5.4 ; les autres parties sont plus techniques et peuvent être abordées plus tard.

Notations

- $P_k(K)$ espace des polynômes de degré au plus k définis dans K.
- $x^t\ y$: produit scalaire des vecteurs x et y.
- Variables conservées scalaires : u, flux : f,
- Variables conservées vectorielles : W, flux : $\mathbf{f}$,
- $q = \frac{\partial u}{\partial x}$.
- $(v, w) = \displaystyle\int_\Omega vwdx$.
- $||v||_s = ||v||_{H^s(\Omega)}$.
- $\langle v, w\rangle = \displaystyle\int_\Gamma v\ w; \vec{\lambda}^t \vec{n} d\Gamma$.
- $\langle v, w\rangle_- = \displaystyle\int_{\Gamma^-} v\ w; \vec{\lambda}^t \vec{n} d\Gamma$.
- $\langle v, w\rangle_+ = \displaystyle\int_{\Gamma^+} v\ w; \vec{\lambda}^t \vec{n} d\Gamma$.
- $|v| = \left(\displaystyle\int_\Gamma v^2\ |\vec{\lambda}^t \vec{n}| d\Gamma\right)^{1/2}$.

Table des matières

Chapitre 1

Introduction

L'objet de ce cours est de donner quelques éléments sur l'approximation numérique des problèmes de la forme

$$\begin{aligned}&\frac{\partial u}{\partial t}+\text{ div }\mathbf{f}(u)=\text{div }\big(\mathbf{K}\cdot\nabla u\big)\quad x\in\Omega\subset\mathbb{R}^d, t>0\\&u(x,t=0)=u_0(x)\\&\text{ conditions aux limites}\end{aligned}\tag{1.1}$$

où $d=1,2,3$, $u\ :\Omega\times\mathbb{R}^+\rightarrow\mathbb{R}^p$, $p\in\mathbb{N}$, le flux $\mathbf{f}\ :\mathbb{R}^p\rightarrow\mathbb{R}^p\times\mathbb{R}^d$ est régulier, $\mathbf{K}$ est un champ de matrices définie positives (ou semblables à des matrices définies positives). On supposera l'opérateur div $\big(\mathbf{K}\nabla u\big)$ elliptique uniformément coercif. Enfin le sous–système

$$\frac{\partial u}{\partial t}+\text{ div }\mathbf{f}(u)=0\tag{1.2}$$

strictement hyperbolique.

Précisons ce dernier point. Le flux est $\mathbf{f}=(F_1,\ldots,F_d)^T$ ou F_i est de classe C^1 au moins et (1.2) se réécrit comme

$$\frac{\partial u}{\partial t}+\sum_{j=1}^d\frac{\partial F_j}{\partial x_j}(u)=0.\tag{1.3}$$

On considère alors les matrices jacobiennes $\mathbf{A}_j=\dfrac{\partial F_j}{\partial u}$ (matrices dim $u\times$ dim u. On dit que le système (1.3) est strictement hyperbolique si

1. quelque soit $\vec{n}=(n_1,\ldots,n_d)$ la matrices
$$\sum_{j=1}^d n_jA_j$$
sont diagonalisables dans $\mathbb{R}$ (i.e. possède dim u valeurs propres réelles, et un système de dim u vecteurs propres formant une base de $\mathbb{R}^d$,
2. les valeurs propres peuvent être numérotées $\lambda_1((\vec{n}),u),\ldots,\lambda_d((\vec{n}),u)$ indépendament de $\vec{n}$ et u telles que
$$\lambda_1((\vec{n}),u)<\lambda_2((\vec{n}),u)<\ldots<\lambda_{d-1}((\vec{n}),u)<\lambda_d((\vec{n}),u).$$

De manière classique, on note $\Lambda(\vec{(n)}, u)$ la matrice diagonale formée par les valeurs propres, $R(\vec{(n)}, u)$ la matrice formée des vecteurs propres (vecteurs propres à droite dans la littérature anglosaxone) et $L(\vec{(n)}, u)$ la matrice formée des formes linéaires propres (vecteurs propres à gauche dans la littérature anglosaxone), si bien que

$$\sum_{j=1}^{d} n_j A_j = R(\vec{(n)}, u)\Lambda(\vec{(n)}, u)L(\vec{(n)}, u).$$

Evidemment, posé tel quel, le problème est infiniment compliqué. Dans ce cours on se contentera d'étudier quelques cas particuliers de systèmes :

1. problèmes scalaires monodimensionnels

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x}(u) = \varepsilon \frac{\partial^2 u}{\partial x^2},$$

   où $u\ : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$ et le flux f est de classe C^1,

2. problèmes stationnaires de la forme

$$\alpha u + \sum_{j=1}^{n} \frac{\partial f_j}{\partial x_j}(u) = f + \varepsilon \Delta u$$

   où $\alpha > 0$ et f est donnée dans $L^2 \cap L^\infty$,

3. problèmes instationnaires scalaires.

Dans la pratique, le but ultime est l'approximation numérique de problèmes système tel celui de la mécanique des fluides. Ce point ne sera pas abordé ici, sauf dans les notes bibliographiques.

On pourra consulter les ouvrages suivants :

- Godlewski et Raviart, tome 1 et 2 [1, 2],
- R. LeVeque
- D. Kröner,
- D. Serre, tomes 1 et 2 [3, 4].

pour la théorie sur les systèmes hyperbolique avec diffusion.

Le plan du cours est le suivant. Sur un exemple «intuitif», on va introduire quelques propriétés et problèmes rencontrés dans ces équations. On s'intéressera ensuite à l'approximation numérique par méthode variationnelle. L'exemple le plus simple est la méthode des volumes finis dont il existe deux variantes : une méthode «centrée sommet» et une méthode «centrée cellules». Une différente essentielle entre ces deux classes de méthode réside dans la manière d'approcher les termes diffusifs.

Suivant l'interprétation dans laquelle on se place (centré cellule ou centré éléments), on aura deux grandes classes de méthodes : les méthodes utilisant des représentation discontinues (méthodes de type Galerkin discontinu), les méthodes utilisant des représentations continues des données (méthodes de type stream–line). On donnera un aperçu de ces deux classes de méthodes.

Chapitre 2

Etude de quelques exemples simples

Ici, on cherche à étudier quelques exemples modèles simples afin d'avoir une idées des caractéristiques des solutions. Une bonne compréhension des propriétés des solutions joue un rôle important dans l'approximation numérique. Ce qui concerne le modèle de traffic routier et son analyse est emprunté à [5].

2.1 Phénomène de couche limite.

On se donne le modèle le plus simple : un problème stationnaire scalaire en une dimension d'espace. La difficulté provient des conditions aux limites. Soit

$$\begin{array}{ll} u_\varepsilon' = \varepsilon {u_\varepsilon}'' + 1 & x \in]0,1[\\ u(0) = 0 & \\ u(1) = C & \end{array} \tag{2.1}$$

La solution exacte du problème est

$$u_\varepsilon(x) = x + (C-1)\frac{e^{\frac{x-1}{\varepsilon}} - e^{-\frac{1}{\varepsilon}}}{1 - e^{-\frac{1}{\varepsilon}}}.$$

On voit facilement que

$$\lim_{\varepsilon} u_\varepsilon(x) = x$$

uniformément sur tout compact de $[0,1[$, mais u_ε' devient aussi grand qu'on veut au voisinage de 1 (uniformément sur tout compact contenant 1).

On a un phénomène de couche limite. C'est un phénomène typique en mécanique des fluides pour des écoulements visqueux à grand nombre de Reynolds (le paramètre ε joue le rôle de l'inverse du nombre de Reynolds.)

Ceci implique un certain nombre de problèmes numériques si on souhaite capturer correctement une couche limite à cause du fait que u_ε' devient grand. Ces problèmes peuvent être du type perte de stabilité, ou manque de précision.

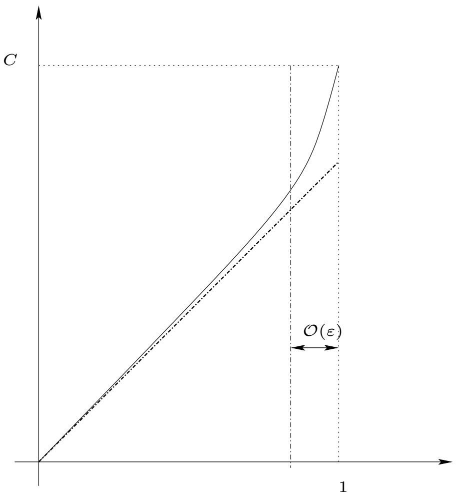


FIG. 2.1 – Comportement de la solution de (2.1) pour $C > 1$.

2.2 Modèle de traffic routier : un exemple de problème non linéaire

Ici, ρ représente la «densité» de voiture se déplaçant sur une autoroute (loin des branchements). Puisqu'on ne rajoute ni n'enlève de voiture (pas de source ni perte), si v désigne la vitesse locale des voiture, on a

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho v)}{\partial x} = 0 \tag{2.2}$$

Il s'agit d'une loi de conservation. Il faut relier la vitesse à la densité de voiture : on voit bien que c'est ce qui se passe en pratique .... Un modèle simple est de prendre

$$\rho v = f(\rho) - \nu \frac{\partial \rho}{\partial x}$$

avec $f(\rho) = \rho V(\rho)$. On suppose aussi que $\rho \mapsto V(\rho)$ est décroissante. Le terme $\nu \frac{\partial^2 \rho}{\partial x^2}$ dit qu'on ralentit quand la densité de voiture augmente. Le facteur ν indique en quelque sorte le degré de prudence d'un conducteur.

L'expression la plus simple pour la vitesse V est

$$V(\rho) = v_{\max}\left(1 - \frac{\rho}{\rho^*}\right).$$

On dit simplement qu'au delà d'une certaine densité de voiture, il n'est plus possible d'avancer.

Le problème est donc

$$\frac{\partial \rho}{\partial t} + \frac{\partial f}{\partial x}(\rho) = \nu \frac{\partial^2 \rho}{\partial x^2}, \qquad f(\rho) = v_{\max}\rho\left(1 - \frac{\rho}{\rho^*}\right).$$

Le flux est quadratique et convexe en ρ. Dans ce qui suit, on fait une étude succinte du modèle. Les résultats sont en fait générique pour des flux convexes. Pour le cas général, on pourra consulter [1].

Dans la pratique, on «voit» bien qu'il y a des ondes : c'est une expérience quotidienne sur un autoroute. On sait d'expérience qu'il y a deux type d'ondes : celles où la densité diminue, celles où elle augmente et qui, dans certains cas, conduisent à des bouchons. Pourquoi ?

On va chercher des solutions du problème de la forme $\rho(x,t) = \varphi(x - ct)$, ou plus précisément de la forme

$$\rho(x,t) = \varphi(\frac{x - ct}{\varepsilon})$$

pour des raisons de simplicité de calcul. On voit facilement que φ vérifie

$$-c\varphi' + \big(f(\varphi)\big)' = \varphi''.$$

On cherche de plus φ tel que

$$\lim_{x\to\pm\infty} \varphi(x) = \rho^\pm, \qquad \lim_{x\to\pm\infty} \varphi'(x) = 0 \tag{2.3}$$

où $\rho^\pm$ sont donnés. Ce type de solution correspond au cas où le flot de voiture nťest perturbé significativement quand dans une zone de «faible» ampleur. Expérimentalement, on sait qu'il existe des bouchons ou des zones où la densité décroit. On «sait» aussi d'expérience que les vraies solutions semblent être la superposition de telles solutions (bouchons, écoulements uniformes, zones d'accélération).

On intègre alors formellement, et φ doit vérifier

$$c\int_\infty^{+\infty} \varphi' dx + \int_\infty^{+\infty} \big(f(\varphi)\big)' dx = \int_\infty^{+\infty} \varphi'' dx$$

soit, en tenant compte des conditions précédentes

$$c = \frac{f(\rho^+) - f(\rho^-)}{\rho^+ - \rho^-}.$$

Dans ce cas, φ vérifie (en intégrant dans $[x, +\infty[$)

$$\varphi' = f(\varphi) - \big(f(\rho^+) - c(\rho^+ - \varphi)\big) \qquad \lim_{x\to-\infty} = \rho^-$$

De plus, φ vérifie la condition finale $\lim_{x\to+\infty} = \rho^+$. On résoud explicitement le problème. On voit que

$$f(\varphi) - \big(f(\rho^+) - c(\rho^+ - \varphi)\big) = -\frac{v_{\max}}{\rho^*}(\varphi - \rho^-)(\varphi - \rho^+)$$

donc l'équation à résoudre est

$$\frac{\varphi'}{(\varphi - \rho^-)(\varphi - \rho^+)} = -\frac{\rho^*}{v_{\max}}.$$

La décomposition en éléments simple de la fraction rationnelle est

$$\frac{1}{(\varphi - \rho^-)(\varphi - \rho^+)} = \frac{1}{\rho^- - \rho^+}\left(\frac{1}{\varphi - \rho^-} - \frac{1}{\varphi - \rho^+}\right),$$

si bien que la solution explicite du problème est

$$\varphi(s) = \frac{\rho^- - \rho^+ e^{-\rho^* \frac{\rho^- - \rho^+}{v_{\max}} s}}{1 - e^{-\rho^* \frac{\rho^- - \rho^+}{v_{\max}}(s - s_0)}}$$

où la constante s_0 est arbitraire[1] On peut voir facilement que φ possède une dérivée qui s'annule en $\pm\infty$. On examine les cas $\rho^- > \rho^+$ et $\rho^- \leq \rho^+$.

– Si $\rho^- > \rho^+$. On voit que
$$\begin{array}{l} \lim_{s\to-\infty} \varphi(s) = \rho^+ \\ \lim_{s\to+\infty} \varphi(s) = \rho^- \end{array}$$
Il faut donc que $\rho^- = \rho^*$.
– Si $\rho^- \leq \rho^+$, on voit que
$$\begin{array}{l} \lim_{s\to-\infty} \varphi(s) = \rho^- \\ \lim_{s\to+\infty} \varphi(s) = \rho^+ \end{array}$$
ce qui est correct. On remarque que le flux f est ici concave.

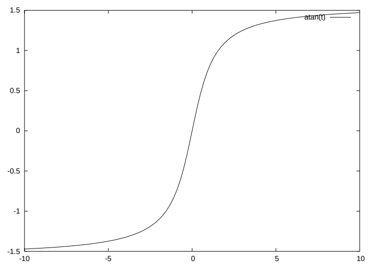


Fig. 2.2 – Profil typique d'onde.

Dans la figure 2.2, on a représenté un profil typique de traveling wave ($v_{\max} = 1$, $\rho^* = 1$, $\rho^\pm = \pm\frac{\pi}{2}$.

Il y a donc existence et unicité de la solution si et seulement si $\rho^+ \geq \rho^-$. Cela correspond au cas où il y a augmentation de la densité de voiture.

On regarde maintenant le comportement des termes visqueux. Le profil est de la forme $\rho(x,t) = \varphi\big(\frac{x-ct}{\varepsilon}\big)$ avec
$$\varphi(s) = \frac{a - be^{\alpha s}}{1 - be^{\alpha s}}$$
donc
$$\rho_{xx}(x,t) = \frac{1}{\varepsilon^2}\frac{d^2\varphi}{ds^2}\big(\frac{x-ct}{\varepsilon}\big)$$
et $\frac{d^2\varphi}{ds^2}$ est une fraction rationnelle du second degré en $e^{\alpha s}$ qui décroit très rapidement quand $\alpha s \to \pm\infty$. On voit donc que $\varepsilon\rho_{xx}$ est de l'ordre de $\frac{1}{\varepsilon}$ seulement pour des points tels que $|x - ct| \leq \mathcal{O}(\varepsilon)$. En d'autres termes,
$$\varepsilon\rho_{xx} \simeq \frac{1}{\varepsilon}e^{-|\alpha|(\frac{x-ct}{\varepsilon})}$$
et donc $\varepsilon\rho_{xx} \to 0$ quand $\varepsilon \to 0$ sauf pour $x = ct$. D'autre part,
$$\int_{+\infty}^{+\infty} \varepsilon\rho_{xx}(x,t)dx = 0$$

[1] Cela introduit une paramètre supplémentaire dans le problème. Ce paramètre peut être (par exemple) la position initiale du point où la dérivée de φ est maximale. Dans la discussion que l'on effectue, la valeur précise de ce paramètre n'a pas d'importance.

donc $\varepsilon\rho_{xx}$ ne peut pas converger vers un Dirac (au sens des distributions). On aurait donc tendance à dire que le terme $\varepsilon\rho_{xx}$ est négligeable, ce qui est loin d'être le cas. En effet, si on fait tendre formellement ε vers 0 dans (2.2), le problème devient mal posé.

On peut voir que le profil précédent tend, quand $\varepsilon \to 0$, le profil converge (uniformément sur tout compact ne contenant pas le point $(x = ct, t)$) vers

$$\rho(x,t) = \begin{cases} \rho^- & \text{si } x - ct < 0 \\ \rho^+ & \text{si } x - ct > 0. \end{cases}$$

Dans un premier temps, on va montrer la nécessite de considérer des solutions discontinues quand on pose $\varepsilon = 0$, et on va introduire la bonne notion, celle de solution faible. C'est l'objet du paragraphe (2.3). Dans un second temps, on va étudier le problème d'unicité, ce sera l'objet du paragraphe 2.4 et relier cela aux résultats que nous avons obtenus sur l'équation du traffic routier.

Introduisons maintenant la notion *de solution faible*. On peut facilement montrer que même si on part d'une condition initiale ρ_0 de classe C^∞, supposer que la solution de

$$\begin{aligned} &\frac{\partial \rho}{\partial t} + \frac{\partial f(\rho)}{\partial x} = 0 \qquad x \in \mathbb{R}, t > 0 \\ &\rho(x, t = 0) = \rho_0(x) \end{aligned} \tag{2.4}$$

peut conduire à une contradiction. Si la solution restait régulière, on aurait alors

$$\frac{\partial \rho}{\partial t} + v_{\max}(1 - \frac{\rho}{\rho^*})\frac{\partial \rho}{\partial x} = 0.$$

En employant la méthode des caractéristiques,

$$\begin{aligned} &\frac{dX}{dt} = f'(\rho(X(x,t),t) \\ &X(x, t = 0) = x \end{aligned}$$

on voit que $t \mapsto \rho(X(x,t),t)$ est constante suivant cette courbe caractéristique. On peut facilement construire des conditions initiales telles qu'on arrive à une contradiction. On peut consulter [1] pour plus de détails.

On introduit alors la notion de solution faible

2.3 Notion de solution faible

On fait cette étude dans un cadre plus large que celui du problème précédent. Au lieu d'être en dimension 1 d'espace, on suppoer être en dimension $d \geq 1$. On se restreint ici au cas des fonctions de la variable réelle, mais cette restriction est artificielle. Les calculs et définitions s'étendent naturellement (en ce qui concerne la notion d solution faible) au cas de systèmes.

On doit cette définition à P.D. Lax [6]. La remarque est la suivante. Supposons que u soit C^1. Si on multiplie

$$\frac{\partial u}{\partial t} + \operatorname{div} f(u) = 0 \tag{2.5}$$

par ϕ, fonction test C^1 à support compact dans $\mathbb{R}^d \times \mathbb{R}^+$, on a

$$\int_{\mathbb{R}^d\times\mathbb{R}^+} \varphi(\boldsymbol{x},t)\left(\frac{\partial u}{\partial t}(\boldsymbol{x},t) + \operatorname{div} f(u)(\boldsymbol{x},t)\right)\mathrm{d}\boldsymbol{x}\,\mathrm{d}t = 0.$$

En intégrant par partie, on voit que u vérifie aussi

$$\begin{aligned}\int_{\mathbb{R}^d\times\mathbb{R}^+} \left(u(\boldsymbol{x},t)\frac{\partial \varphi}{\partial t}(\boldsymbol{x},t) + f(u(\boldsymbol{x},t))\cdot\nabla\varphi(\boldsymbol{x},t)\right)\mathrm{d}\boldsymbol{x}\,\mathrm{d}t&\\ +\int_{\mathbb{R}^d} u_0(\boldsymbol{x})\,\varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x} = 0.&\end{aligned} \tag{2.6}$$

Réciproquement, si u, fonction de $\left(\mathrm{C}^1\left(\mathbb{R}^d\times\mathbb{R}^+\right)\right)^n$ vérifie (2.6), alors u est solution de (2.5).

Recherchons à quelles conditions les intégrales apparaissant dans (2.6) sont définies. Si f est de classe C^1, il suffit que u et u_0 soient bornées et mesurables pour que les intégrales de (2.6) soient définies. Il va de soit que les calculs précédents s'étendent aux systèmes (2.2).

On introduit alors la définition suivante, après avoir défini $\mathrm{C}_0^1(\mathbb{R}\times\mathbb{R}^+)$ comme l'ensemble des fonctions C^1 à support compact dans $\mathbb{R}^d\times\mathbb{R}^+$.

Définition 2.1. *On dit que $u \in \left(\mathrm{L}^\infty\left(\mathbb{R}^d\times\mathbb{R}^+\right)\right)^n$ est solution faible de (2.2) si et seulement si, quelque soit $\varphi \in \left(\mathrm{C}_0^1(\mathbb{R}\times\mathbb{R}^+)\right)^n$, on a*

$$\begin{aligned}\int_{\mathbb{R}^d\times\mathbb{R}^+} \left(u(\boldsymbol{x},t)\frac{\partial \varphi}{\partial t}(\boldsymbol{x},t) + f(u(\boldsymbol{x},t))\cdot\nabla\varphi(\boldsymbol{x},t)\right)\mathrm{d}\boldsymbol{x}\,\mathrm{d}t&\\ +\int_{\mathbb{R}^d} u_0(\boldsymbol{x})\,\varphi(\boldsymbol{x},0)\,\mathrm{d}\boldsymbol{x} &= 0.\end{aligned}$$

Cette définition autorise des solutions discontinues bornées, mais une solution classique (i.e. de classe C^1) est solution faible comme les calculs précédents le montrent [2].

On va maintenant regarder à quelle condition une fonction de classe C^1 par morceau peut être une solution faible. Une fonction est dite de classe C^1 par morceaux si elle est de classe C^1 sauf peut-être sur une surface Γ de $\mathbb{R}^d\times\mathbb{R}^+$ qui est de classe C^1 au moins. Notons par $\vec{n} = (\vec{n}_x, \vec{n}_t)$ une normale à Γ au point $(\boldsymbol{x},t)\in\Gamma$. Supposons qu'en chaque point $(\boldsymbol{x},t)\in\Gamma$, les limites u^+ et u^-

$$u^+(\boldsymbol{x},t) = \lim_{\xi\to 0^+} u((\boldsymbol{x},t)+\xi\vec{n}),$$

$$u^-(\boldsymbol{x},t) = \lim_{\xi\to 0^-} u((\boldsymbol{x},t)+\xi\vec{n})$$

existent, alors les sauts $[u] = u^+ - u^-$ et $[f(u)]\cdot\vec{n}_x = f(u^+)\cdot\vec{n}_x - f(u^-)\cdot\vec{n}_x$ vérifient les *relations de Rankine–Hugoniot*

$$n_t[u] + [f(u)]\cdot\vec{n}_x = 0. \tag{2.7}$$

Ce résultat se démontre en considérant φ une fonction à support compact à support dans un voisinage de $M\in\Gamma$ (voir figure 2.3). Notons Ω le support de φ, et Ω^+ (respectivement Ω^-) les

[2] Un résultat de calcul intégral montre que si g est bornée mesurable, définie dans un ouvert Ω telle que pour toute fonction Ψ de classe C^1, $\int_\Omega f\,\Psi\,\mathrm{d}\boldsymbol{x} = 0$, alors f est nulle presque partout, voir [7].

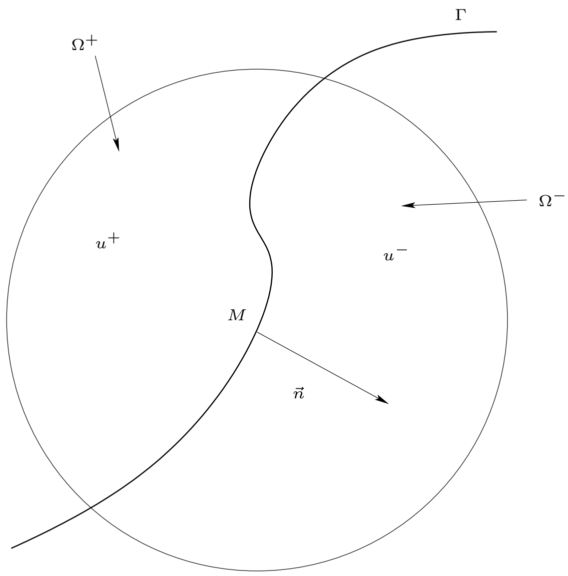


FIG. 2.3 –

ouverts situés « à droite » et « à gauche » de Γ. Leur existence ne pose pas de problème si Ω est un voisinage assez petit de Γ car c'est une surface régulière. Dans Ω^+, u est solution classique et donc

$$\int_{\Omega^+} \left(\varphi(\boldsymbol{x}, t) \frac{\partial u}{\partial t} + \varphi(\boldsymbol{x}, t) \operatorname{div} f(u) \right) \mathrm{d}\boldsymbol{x} \, \mathrm{d}t = 0. \tag{2.8}$$

Notons $\partial\Omega^+$ le bord de Ω^+. En appliquant le théorème de la divergence, et parce que φ est de support Ω, on transforme (2.8) en

$$\int_{\Omega^+} \left(\frac{\partial \varphi}{\partial t} u + \nabla\varphi \cdot f(u) \right) \mathrm{d}\boldsymbol{x} \, \mathrm{d}t + \int_{\partial\Omega\cap\Gamma} \varphi \left(u^+ n_t + \vec{n}_x \cdot f\left(u^+\right) \right) \mathrm{d}\sigma = 0. \tag{2.9}$$

Dans Ω^-, on a de même

$$\int_{\Omega} \left(\frac{\partial \varphi}{\partial t} u + \nabla\varphi \cdot f(u) \right) \mathrm{d}\boldsymbol{x} \, \mathrm{d}t - \int_{\partial\Omega\cap\Gamma} \varphi \left(u^- n_t + \vec{n}_x \cdot f\left(u^-\right) \right) \mathrm{d}\sigma = 0. \tag{2.10}$$

Le signe "$-$" dans (2.10) provient du fait que la normale extérieure à Ω^- en $(\boldsymbol{x}, t) \in \Gamma$ est l'opposée de la normale extérieure à Ω^+ en $(\boldsymbol{x}, t)$. Dans les intégrales (2.9) et (2.10) seules interviennent les contributions sur $\partial\Omega^- \cap \Gamma = \partial\Omega^+ \cap \Gamma$ car φ est nulle sur $\partial\Omega$.

En sommant (2.9–2.10) et tenant compte de (2.1), on obtient

$$\int_{\partial\Omega\cap\Gamma} \varphi \left([u] \cdot n_t + [f(u)] \cdot \vec{n}_x \right) \mathrm{d}\sigma = 0,$$

d'où (2.7) car φ est quelconque.

En dimension 1, la relation (2.7) prend une forme plus parlante. Considérons la figure 2.4. Supposons de plus que Γ admette un paramétrage de la forme $x = \Psi(t)$. Dans ce cas la normale à

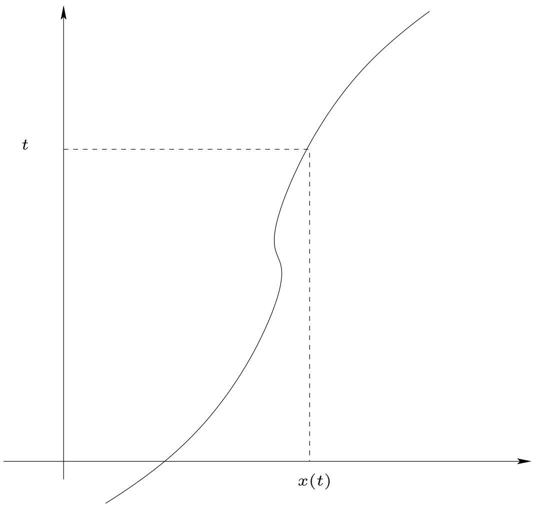


FIG. 2.4 – Cas d'une discontinuité pour une équation scalaire

un point courant de Γ est

$$\vec{n} = \begin{pmatrix} -\Psi'(t) \\ 1 \end{pmatrix}$$

et donc (2.7) prend la forme (si on note classiquement $s = \Psi'(t)$ la vitesse de la discontinuité)

$$f\left(u^{+}\right) - f\left(u^{-}\right) = s\left(u^{+} - u^{-}\right). \tag{2.11}$$

Il faut remarquer que la définition 2.1 ne garantit pas l'unicité des solutions. En effet, considérons l'équation (2.2) avec $\rho^{-} = \frac{1}{4}$, $\rho^{+} = \frac{3}{4}$, $v_{\max} = 1$ et $\rho^{*} = 1$ (donc $f(\rho) = \rho(1-\rho)$). On a $c = 0$. Si $\delta > 0$, on considère

$$\rho^{\delta}(x,t) = \begin{cases} \frac{3}{4} & \text{si } \delta < \frac{x}{t} \\ \frac{1}{4} - \delta & \text{si } 0 < \frac{x}{t} < \delta \\ \frac{3}{4} + \delta & \text{si } -\delta < \frac{x}{t} < 0 \\ \frac{1}{4} & \text{si } \frac{x}{t} < -\delta \end{cases}$$

On voit que quelque soit $\delta > 0$, ρ^{δ} est solution faible de (2.2)pour $\varepsilon = 0$. Il suffit de vérifier que les conditions de Rankine Hugoniot sont vérifiées.

En fait, l'équation (2.2) avec u_0 comme condition initiale admet une infinité de solutions. Il faut donc un critère supplémentaire pour "trier" les solutions excédentaires : on perd le caractère bien posé du problème.

2.4 Notion d'entropie

Revenons un instant sur le problème du traffic routier. Considérant des données initiales discontinues, on a construit pour le problème

$$\frac{\partial \rho_\varepsilon}{\partial t} + \frac{\partial f(\rho_\varepsilon)}{\partial x} = \varepsilon \frac{\partial^2 \rho_\varepsilon}{\partial x^2}$$
$$\rho_\varepsilon(x,0) = \begin{cases} \rho_- & \text{si } x < 0 \\ \rho_+ & textsix > 0 \end{cases}$$

telle que ρ_ε converge uniformément sur tout compact ne contenant pas $\{(x,t) \in \mathbb{R} \times \mathbb{R}^+, x = ct\}$ avec

$$c = \frac{f(\rho_+) - f(\rho_-)}{\rho_+ - \rho_-}$$

vers

$$\rho(x,t) = \begin{cases} \rho_- & \text{si } x - ct < 0 \\ \rho_+ & \text{sinon} \end{cases}$$

Cependant une telle solution ne peut exister que si $\rho_- \leq \rho_+$: on voit qu'il y a sélection. L'objet de ce paragraphe est de montrer que cette méthode de viscosité permet de définir la bonne classe de solution pour lesquelles (2.4) est bien posé. On se place en dimension $d \geq 1$, mais la discussion n'est valable que pour des fonctions à valeur réelles. Le cas système est beaucoup plus complexe et sort de loin le cadre de ces notes.

L'idée est de chercher la solution sous forme d'une limite de solutions de problèmes "visqueux" :

$$\begin{cases} \dfrac{\partial u_\varepsilon}{\partial t} + \operatorname{div} f(u_\varepsilon) = \varepsilon \Delta u_\varepsilon, & \boldsymbol{x} \in \mathbb{R}^d, \quad t > 0, \\ u_\varepsilon(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}), \end{cases} \tag{2.12}$$

où $u_0 \in \mathrm{L}^\infty(\mathbb{R}^d)$ et $\varepsilon \to 0$.

On essaie d'imiter la mécanique des fluides avec une viscosité cinématique qui tend vers 0 et un modèle de viscosité très simplifié : le terme de viscosité est modélisé par $\varepsilon \Delta u_\varepsilon$.[3] On peut montrer par une technique de point fixe, voir [1] pour plus de détails, que le problème (2.12) admet une unique solution.

Considérons une fonction de régularisation θ définie sur $\mathbb{R}^d$ et à valeurs dans $\mathbb{R}$ telle que

$$\theta(\boldsymbol{x}) \geq 0, \qquad \forall \boldsymbol{x} \in \mathbb{R}^d,$$

$$\int_{\mathbb{R}^d} \theta(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = 1.$$

Si $u_0 \in \mathrm{L}^\infty(\mathbb{R}^d) \cap \mathrm{L}^1(\mathbb{R}^d)$, on peut définir

$$u_{0\varepsilon} = \int_{\mathbb{R}^d} u_0(\boldsymbol{x}) \, \theta\left(\frac{\boldsymbol{x}}{\varepsilon}\right) \frac{\mathrm{d}\boldsymbol{x}}{\varepsilon^d},$$

et on sait (voir [7]) que

[3] Notons que c'est exactement ce qu'on a fait dans le cas du traffic routier.

1. $u_{0\varepsilon} \in \mathrm{L}^\infty\left(\mathbb{R}^d\right) \cap \mathrm{L}^1\left(\mathbb{R}^d\right)$;
2. La suite $(u_{0\varepsilon})_\varepsilon$ converge vers u_0 dans $\mathrm{L}^\infty\left(\mathbb{R}^d\right)$ pour la norme sup et dans $\mathrm{L}^1\left(\mathbb{R}^d\right)$ pour la norme L^1.

On a alors le

Théorème 2.1. *Si $u_0 \in \mathrm{L}^\infty\left(\mathbb{R}^d\right) \cap \mathrm{L}^1\left(\mathbb{R}^d\right) \cap \mathrm{BV}\left(\mathbb{R}^d\right)$, alors (2.12), avec la condition initiale $u_{0\varepsilon}$, admet une unique solution u_ε dans $\mathrm{C}^\infty\left(\mathbb{R}^d\right)$ qui vérifie pour tout $t \geq 0$*

1. $\|u_\varepsilon\|_\infty \leq \|u_0\|_\infty$,
2. $\displaystyle\int_{\mathbb{R}^d} u_\varepsilon(\boldsymbol{x}, t)\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^d} u_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$,
3. $\mathrm{TV}\left(u_\varepsilon(\,\cdot\,, t)\right) \leq \mathrm{TV}(u_0)$,
4. *quelque soit $t \in [0, T]$,*
$$\|u_\varepsilon(\,\cdot\,, t)\|_1 \leq \|u_0\|_1 + C\,T\,\mathrm{TV}(u_0),$$
où $C > 0$ est une constante indépendante de ε.

Voir encore [1] pour la preuve.

Dans ce résultat, on utilise la variation totale de la solution u, $\mathrm{TV}(u)$. Celle ci est définie de la manière suivante :

$$\begin{aligned}\mathrm{TV}(u) = \sup\Bigg\{ \int_{\mathbb{R}^d} u \cdot \nabla\varphi\,\mathrm{d}\boldsymbol{x}\ ; \quad \varphi \ \text{ fonction } \mathrm{C}^1\left(\mathbb{R}^d\right) \\ \text{à support compact telle que } \|\varphi\|_\infty \leq 1 \Bigg\}\end{aligned} \tag{2.13}$$

On dira qu'elle est à variation totale bornée si $\mathrm{TV}(u) < +\infty$. Dans le cas où u est C^1, on a

$$\int_{\mathbb{R}^d} u \cdot \nabla\varphi\,\mathrm{d}\boldsymbol{x} = -\int_{\mathbb{R}^d} \nabla u\,\varphi\,\mathrm{d}\boldsymbol{x},$$

ce qui permet de voir que $\mathrm{TV}(u) \leq \int_{\mathbb{R}^d} \|\nabla u\|\,\mathrm{d}\boldsymbol{x}$ [4]. Des arguments un peu plus sophistiqués [1] permettent de montrer que

$$\mathrm{TV}(u) = \int_{\mathbb{R}^d} \|\nabla u\|\,\mathrm{d}\boldsymbol{x}.$$

Un exemple très important de fonctions à variation totale décroissante est donné par les fonctions constantes par morceaux. En effet, les schémas numériques que nous verrons dans le chapitre 4 approchent la solution du problème de Cauchy comme une fonction constante par morceaux.

Considérons une subdivision croissante de $\mathbb{R}$, $(x_i)_{i\in\mathbb{Z}}$, et une suite $(u_i)_{i\in\mathbb{Z}}$. Si on définit la fonction u par

$$u(x) = u_i \quad \text{si} \quad x \in [x_i, x_{i+1}[\,,$$

alors

$$\mathrm{TV}(u) = \sum_{j\in\mathbb{Z}} |u_{i+1} - u_i|.$$

[4] ce qui justifie le choix de φ telle que $\|\varphi\|_\infty \leq 1$!

Dans la pratique, on souhaite s'affranchir de la condition « u est limite de u_ε ». Pour cela, on cherche une caractérisation *a priori* des solutions de (2.12) qui sera étendue formellement aux solutions de (2.2). On verra par la suite que cette définition, qui est compatible avec la classe des solutions classiques, garantit l'existence et l'unicité des solutions avec continuité de la solution par rapport à la condition initiale. D'un point de vue théorique et pratique, c'est donc la bonne notion.

Considérons U une fonction régulière définie sur $\mathbb{R}^d$ et à valeur réelle. On multiplie (2.12) par $U'(u_\varepsilon)$, pour obtenir :

$$\frac{\partial U(u_\varepsilon)}{\partial t} + U'(u_\varepsilon) \cdot \operatorname{div} f(u_\varepsilon) = \varepsilon U'(u_\varepsilon) \cdot \Delta u_\varepsilon.$$

Supposons de plus qu'il existe $G \in \mathrm{C}^1\left(\mathbb{R}^d\right)^n$ telle que

$$\text{pour tout } u \in \mathbb{R}^n,\ U'(u) \cdot \operatorname{div} f(u) = \operatorname{div} G(u). \tag{2.14}$$

alors

$$\frac{\partial U(u_\varepsilon)}{\partial t} + \operatorname{div} G(u_\varepsilon) = \varepsilon U'(u_\varepsilon) \cdot \Delta u_\varepsilon.$$

De plus, $U'(u_\varepsilon) \cdot \Delta u_\varepsilon = \Delta U(u_\varepsilon) - \nabla u_\varepsilon^{\mathrm{t}} U''(u_\varepsilon) \nabla u_\varepsilon$, donc

$$\frac{\partial U(u_\varepsilon)}{\partial t} + \operatorname{div} G(u_\varepsilon) = \varepsilon \Delta U(u_\varepsilon) - \varepsilon \nabla u_\varepsilon^{\mathrm{t}} U''(u_\varepsilon) \nabla u_\varepsilon \leq \varepsilon \Delta U(u_\varepsilon).$$

si U est *convexe*. Si maintenant, on recommence les calculs du paragraphe 2.3 avec une fonction test φ *positive ou nulle*, on voit que u_ε vérifie

$$\int_{\mathbb{R}^d \times \mathbb{R}^+} \left(U(u_\varepsilon)(\boldsymbol{x}, t) \frac{\partial \varphi}{\partial t} + \nabla \varphi \cdot G(u_\varepsilon) \right) \mathrm{d}\boldsymbol{x}\, \mathrm{d}t + \int_{\mathbb{R}^d} U(u_{0\varepsilon}) \varphi(\boldsymbol{x}, 0)\, \mathrm{d}\boldsymbol{x} \leq 0.$$

Si on suppose que $u_\varepsilon \to u$ dans L^1 par exemple, on en déduit que

$$\int_{\mathbb{R}^d \times \mathbb{R}^+} U(u)(\boldsymbol{x}, t) \frac{\partial \varphi}{\partial t} + \nabla \varphi \cdot G(u)\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}t + \int_{\mathbb{R}^d} U(u_0) \varphi(\boldsymbol{x}, 0)\, \mathrm{d}\boldsymbol{x} \leq 0. \tag{2.15}$$

On dira que le couple (U, G) est un couple entropie–flux.

Définition 2.2. *On dira que $u \in \left(\mathrm{L}^\infty\left(\mathbb{R}^d\right)\right)^n$ est solution entropique de (2.2) si elle vérifie, pour toute fonction test $\varphi \geq 0$ de classe $\mathrm{C}^1\left(R^d\right)$ à support compact et pour tout couple entropie–flux (U, G) vérifiant les relations (2.14), l'inégalité (2.15).*

Si u est solution classique en dehors d'une surface Γ de classe C^1, en utilisant la même méthode que pour établir les relations de Rankine–Hugoniot, on voit que

$$n_t [U] + \vec{n}_x \cdot [G(u)] \leq 0 \tag{2.16}$$

Dans (2.16), la définition de $[\cdot]$ et le choix de l'orientation de Γ sont liés. Si $\vec{n}$ est la normale qui pointe vers Ω^+ (voir figure 2.3), alors $[\cdot] = (\cdot)^+ - (\cdot)^-$.

On a le résultat suivant dans le cas *scalaire* (voir [1] pour une preuve)

Théorème 2.2 (Kruzkov). *Soit $u_0 \in \mathcal{E} = \mathrm{L}^\infty\left(\mathbb{R}^d\right) \cap \mathrm{L}^1\left(\mathbb{R}^d\right) \cap \mathrm{BV}\left(\mathbb{R}^d\right)$. On suppose le flux C^1. Alors (2.2) admet une unique solution entropique u définie sur $\mathbb{R}^d \times [0, +\infty[$ qui vérifie, quelque soit $t \geq 0$*

1. $||u(\cdot,t)||_\infty \leq \|u_0\|_\infty$

2. Quelque soit $t \geq 0$, $\int_{\mathbb{R}^d} u(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^d} u_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$

3. $\mathrm{TV}(u(\cdot,t)) \leq \mathrm{TV}(u_0)$,

4. Il existe C *ne dépendant que de* u_0 *tel que quels que soient* t_1, t_2,

$$\int_{\mathbb{R}^d} |u(\boldsymbol{x},t_2) - u(\boldsymbol{x},t_1)|\,\mathrm{d}\boldsymbol{x} \leq C\,\mathrm{TV}(u_0)\ |t_2 - t_1|$$

De plus, si $u_0, v_0 \in \mathcal{E}$, *les solutions entropique* u *et* v *associées à ces conditions initiales vérifient quelque soit* $R > 0$

$$\int_{\|\boldsymbol{x}\|\leq R} |u(\boldsymbol{x},t) - v(\boldsymbol{x},t)|\,\mathrm{d}\boldsymbol{x} \leq \int_{\|\boldsymbol{x}\|\leq R+Mt} |u_0(\boldsymbol{x}) - v_0(\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x}$$

où $M = \sup\{|f'(u)|,\ \textit{pour}\ |u| \leq \max(|u_0|_\infty, |v_0|_\infty)\}$. *Enfin, si* $u_0 \leq v_0$ *alors* $u \leq v$. *De plus*

$$\int_{\mathbb{R}^d} |u(\boldsymbol{x},t) - v(\boldsymbol{x},t)|\,\mathrm{d}\boldsymbol{x} \leq \int_{\mathbb{R}^d} |u_0(\boldsymbol{x}) - v_0(\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x}.$$

La propriété 1 exprime que la solution est toujours bornée par $\|u_0\|_\infty$. La seconde propriété est une propriété de conservation. La troisième propriété a pour conséquence, dans le cas scalaire, que u ne peut pas avoir plus d'extréma que u_0. En particulier, il ne peut pas y avoir création de nouveaux extremum. La cinquième propriété indique que le problème est hyperbolique : si u_0 et v_0 sont identiques dans $\{\boldsymbol{x}, \|\boldsymbol{x}\| \leq K\}$, alors u et v seront égales dans $\{\boldsymbol{x}, \|\boldsymbol{x}\| \leq K - M\,t\}$ où M est défini comme dans le théorème 2.2. Ceci montre que toute perturbation se propage à vitesse finie. L'avant dernière propriété est une propriété de monotonie.

2.5 Caractérisation des discontinuités admissibles dans le cas scalaire ($n = d = 1$)

Dans ce paragraphe, on va supposer que le flux f est convexe. Le cas général est traité dans [1] par exemple.

2.5.1 Entropies de Kruzkov

Dans la définition 2.2, on considère une entropie U convexe de classe C^1. Il est bien connu qu'une fonction scalaire de classe C^1 est limite uniforme sur tout compact de $\mathbb{R}$ de fonctions de la forme

$$\sum_{i=1,n} a_i |x - x_i|$$

où les a_i sont réels et $(x_i)_{i=1,n}$ est une subdivision croissante de $\mathbb{R}$. Les fonctions $x \mapsto |x - x_i|$ sont convexes mais non différentiables en $x = x_i$ et ne rentrent pas dans le cadre de la définition 2.2. Cependant, en considérant les fonctions U_k^ε, pour $\varepsilon \to 0$, définies par

$$U_k^\varepsilon(u) = \begin{cases} |u-k| & \text{si } |u-k| \geq \varepsilon \\ \\ \dfrac{(u-k)^2 + \varepsilon^2}{2\varepsilon} & \end{cases}$$

ces fonctions convergent simplement vers $U_k \ : \ u \mapsto |u-k|$, et sont bornées sur tout compact. En définissant $G_k^\varepsilon(u)$ comme une primitive de $(U_k^\varepsilon)' f'$ qui s'annule en $u=k$, G_k^ε converge simplement vers

$$G_k(u) = \mathrm{sign}(u-k)(f(u)-f(k))$$

en restant bornée. On voit donc, en appliquant le théorème de convergence dominée, que le couple (U_k, G_k) est un couple entropie–flux au sens de la définition 2.2. Les fonctions U_k sont appelées *entropies de Kruzkov* .

A l'aide de ces entropies particulières, on peut caractériser les solutions C^1 par morceaux qui vérifient à la fois les relations de Rankine–Hugoniot et les inégalités entropiques. Les relations de Rankine–Hugoniot pour les entropies de Kruzkov s'écrivent

$$[\mathrm{sign}(u-k)(f(u)-f(k))] \leq s\,[|u-k|]\,,$$

où $k \in \mathbb{R}$ est quelconque. On a alors trois cas à étudier suivant la valeur de k.

1^{er}cas : $k > u^+$ et $k > u^-$.

Alors $[|u-k|] = -[u]$ et $[\mathrm{sign}(u-k)(f(u)-f(k))] = -[f(u)]$. On a donc

$$s\ [u] \leq [f(u)]\,.$$

2^ecas : $k < u^+$ et $k < u^-$ on a donc

$$s\ [u] \geq [f(u)]\,.$$

En combinant les inégalités $s\ [u] \leq [f(u)]$ et $s\ [u] \geq [f(u)]$, ce qui est licite car k est quelconque, on retrouve les relations de Rankine–Hugoniot. Elles peuvent être vue, dans le cas scalaire, comme un cas particulier des inégalités d'entropie.

3^ecas : k entre u^+ et u^-. On écrit $k = au^+ + (1-a)u^-$ avec $a \in [0,1]$. On a

$$u^+ - k = (1-a)\left(u^+ - u^-\right), \quad u^- - k = -a\left(u^+ - u^-\right)$$

et

$$\begin{gathered}[\mathrm{sign}(u-k)(f(u)-f(k))] \\ = \\ \left(f(u^+) + f(u^-) - 2f(au^+ + (1-a)u^-)\right)\ \mathrm{sign}(u^+ - u^-)\end{gathered}$$

Utilisant l'expression de s en fonction de u^+ et u^- déduite des relations de Rankine–Hugoniot, on obtient

$$[\mathrm{sign}(u-k)(f(u)-f(k))] =$$

$$\begin{aligned}\mathrm{sign}(u^+ - u^-)(af(u^+) + \quad & (1-a)f(u^-) - \\ & f(au^+ + (1-a)u^-))\ \mathrm{sign}(u^+ - u^-) \leq 0.\end{aligned}$$

Si f est convexe, le cas $u^+ > u^-$ est à exclure. Une discontinuité sera admissible si et seulement si

$$u^+ \leq u^- \tag{2.17}$$

Ce résultat ne peut s'obtenir que si u vérifie *toutes* les inégalités d'entropies et non pas une sous–famille seulement. Si tel était le cas, on perdrait en général l'unicité. Si f est convexe (comme dans le cas du traffic routier dans le paragraphe 2.2), l'inégalité (2.17) devient

$$u^+ \geq u^-.$$

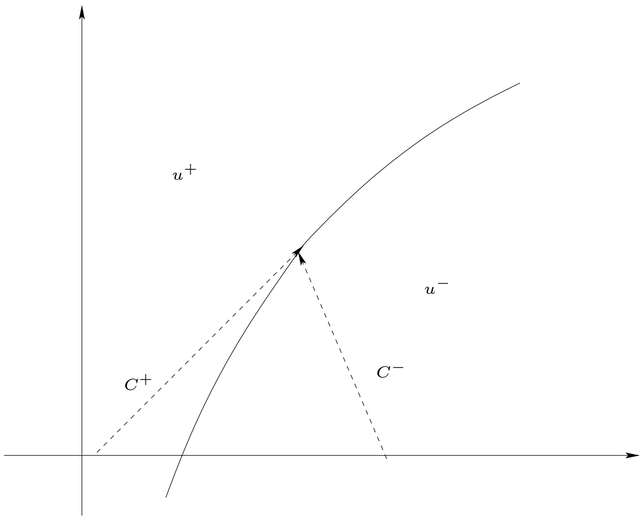


FIG. 2.5 – Condition de Lax

La relation (2.17) s'interprète géométriquement. Considérons deux états constants séparés par une discontinuité admissible. Les caractéristiques issues du domaine dans lequel $u = u^+$ sont des droites parallèles de pente $a(u^+)$. Les caratéristiques issues de points dans le domaine où $u \equiv u^-$ sont aussi des droites de pente $a(u^-)$. Puisque le flux f est convexe, $a = f'$ est croissant. Si (2.17) est réalisée, alors $a(u^+) \leq a(u^-)$, ce qui signifie géométriquement que les caractéristiques se coupent, comme pour l'équation de Burgers. De plus, la vitesse de la discontinuité vérifie la relation de Rankine–Hugoniot

$$s = \frac{f(u^+) - f(u^-)}{u^+ - u^-}$$

qui, d'après le théorème de Rolle, vaut $s = a(\xi)$, pour $\xi \in [u^-, u^+]$. la fonction $a = f'$ est croissante si f est convexe. On obtient donc *la condition de Lax*

$$a\left(u^+\right) < s < a\left(u^-\right). \tag{2.18}$$

Réciproquement, si deux domaines uniformes sont tels que les caractéristiques issues de l'un d'entre eux coupent celles issues du second, c'est à dire si la condition de Lax est vérifiée, la solution ne peut pas être régulière, une discontinuité se crée.

2.5.2 Application à la résolution du problème de Riemann dans le cas d'un flux convexe

On appelle *problème de Riemann* le problème suivant

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} &= 0 \quad \text{si } x \in \mathbb{R} \text{ et } t > 0 \\ u(x,0) = u_0(x) \quad &= \begin{cases} u_+ & \text{si } x > 0 \\ u_- & \text{si } x < 0 \end{cases} \end{aligned} \tag{2.19}$$

Comme $u_0 \in \mathrm{L}^\infty\mathbb{R} \cap \mathrm{L}^1(\mathbb{R}) \cap \mathrm{BV}(\mathbb{R})$, le théorème de Kruzkov assure l'existence et l'unicité d'une solution à (2.19).

Remarque 2.1. *Remarquons que si* $(x,t) \mapsto u(x,t)$ *est une solution, alors* $(x,t) \mapsto u(\lambda x, \lambda t)$ *(pour* $\lambda > 0$*) est aussi une solution car l'équation est invariante par transformation du type* $(x,t) \mapsto (\lambda x, \lambda t)$ *et la condition initiale est invariante par la transformation* $x \mapsto \lambda x$, $u_0(x) = u_0(\lambda x)$ *si* $\lambda > 0$. *On en déduit que* $u(x,t) = u(\lambda x, \lambda t)$ *dès que* $\lambda > 0$, *et donc* $u(x,t) = u\left(\frac{x}{t}, 1\right) \equiv v\left(\frac{x}{t}\right)$.

1^er^cas $u_+ < u_-$. Alors si $s = \frac{f(u_+)-f(u_-)}{u_+-u_-}$,

$$u(x,t) = \begin{cases} u_- & \text{si } \dfrac{x}{t} < s \\ \\ u_+ & \text{si } \dfrac{x}{t} > s \end{cases}$$

est une solution faible qui vérifie toutes les inégalités entropiques. C'est donc la solution entropique. Cette solution est bien sure autosimilaire, comme prévu par la remarque 2.1.

2^e^cas $u_+ \geq u_-$. La discontinuité n'est pas admissible. On cherche encore u sous la forme $u(x,t) = v\left(\frac{x}{t}\right)$. Si v était régulière, en injectant cette expression dans (2.19), on aurait

$$\frac{x}{t^2}\, v'\left(\frac{x}{t}\right) + \frac{1}{t} a\left(v\left(\frac{x}{t}\right)\right)\, v'\left(\frac{x}{t}\right) = 0$$

d'où, si $\xi = \dfrac{x}{t}$,

$$a(v(\xi)) = \xi.$$

Puisque f est convexe, $a = f'$ est croissante, l'équation précédente est inversible. Notons b la fonction réciproque de a. La solution est donc

$$u(x,t) = \begin{cases} u_- & \text{si } \dfrac{x}{t} \leq a(u_-) \\ b\left(\dfrac{x}{t}\right) & \text{si } u_- \leq \dfrac{x}{t} \leq u_+ \\ u_+ & \text{si } \dfrac{x}{t} \geq a(u_+) \end{cases}.$$

C'est effectivement une solution faible (même raisonnement que pour l'équation de Burgers) et elle vérifie les inégalités entropiques.

Chapitre 3

Les équations d'Euler de la mécanique des fluides

3.1 Introduction

Dans ce chapitre, on se propose d'examiner plus en détail les propriétés des solutions du système d'équations qui décrivent l'évolution d'un fluide compressible non-visqueux. Notre but est d'illustrer comment les propriétés que nous avons dégagées pour les problèmes scalaires sont en fait génériques, mais aussi de montrer les différences techniques.

On rappelle ce système d'équations ci-dessous :

$$\begin{aligned}
\frac{\partial \rho}{\partial t} &+ \sum_{i=1}^{d} \frac{\partial(\rho u_i)}{\partial x_i} &&= 0 \\
\frac{\partial(\rho u_i)}{\partial t} &+ \sum_{i=1}^{d} \frac{\partial(\rho u_i u_j)}{\partial x_j} + \frac{\partial p}{\partial x_i} &&= 0 \\
\frac{\partial(\rho e)}{\partial t} &+ \sum_{i=1}^{d} \frac{\partial[(\rho e + p)u_j]}{\partial x_j} &&= 0
\end{aligned} \tag{3.1}$$

où d est la dimension de l'espace[1], u_i désigne la i ième composante de la vitesse, p est la pression, e l'énergie totale spécifique. Le système (3.1) est un système d'équations aux dérivées partielles du premier ordre auquel on doit adjoindre une condition initiale :

$$\rho(0, \boldsymbol{x}) = \rho_0(\boldsymbol{x}), \boldsymbol{u}(0, \boldsymbol{x}) = \boldsymbol{u}_0(\boldsymbol{x}), e(0, \boldsymbol{x}) = e_0(\boldsymbol{x})$$

Il faut lui adjoindre une équation d'état, c'est à dire une relation liant la pression à deux variables thermodynamiques indépendantes. Ici, on choisit l'énergie interne $\varepsilon = E - \frac{1}{2}\rho \mathbf{u}^2$ et la densité ρ,

$$p = p(\rho, \varepsilon). \tag{3.2}$$

[1] Ici, $d = 1$, 2 ou 3.

Dans le cas particulier d'un gaz parfait caloriquement parfait, on a

$$p = (\gamma - 1)\Big(E - \tfrac{1}{2}\rho \mathbf{u}^2\Big)$$

où le rapport des chaleurs spécifiques γ est supposé constant.

Afin de pouvoir effectuer tous les calculs de manière explicite, nous avons *choisi* de nous concentrer sur le cas d'un fluide possédant une loi d'état de type gaz parfait. Cependant, les résultats que nous donnons s'étendent dans le cas général pourvu que la loi d'état du gaz satisfasse quelques hypothèses que nous ne détaillerons pas. On renvoit à [2, 8, 3] pour plus de détails.

Ce système requiert *à priori* que les fonctions $\rho(t,\boldsymbol{x}), \boldsymbol{u}(t,\boldsymbol{x}), e(t,\boldsymbol{x})$ soient au moins C^1. Une des caractéristiques majeures de ce système d'équation est l'explosion des dérivées au bout d'un temps fini T^* et ceci pour une large classe de conditions initiales même très régulières (C^∞ par exemple). Les solutions de (3.1) deviennent discontinues pour $t > T^*$ et il faut pouvoir donner un sens à (3.1) pour admettre ces discontinuités.

Reécrivons les équations (3.1) sous la forme :

$$\frac{\partial U}{\partial t} + \operatorname{div} \vec{F}(U) = \frac{\partial U}{\partial t} + \sum_{j=1}^{d} \frac{\partial F_j(U)}{\partial x_j} = 0 \tag{3.3}$$

en adoptant les notations suivantes :

$$U = \begin{pmatrix} \rho \\ \rho \boldsymbol{u} \\ \rho e \end{pmatrix}, \quad \vec{F} = \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix}, \quad \text{avec} \quad F_j(U) = \begin{pmatrix} \rho u_j \\ \rho u_j \boldsymbol{u} + p\boldsymbol{\delta}_j \\ (\rho e + p) u_j \end{pmatrix},$$

où $\boldsymbol{\delta}_j$ représente le j-ème vecteur colonne de la matrice identité $d \times d$.

Suivant la démarche du chapitre 2, on introduit la notion de solution faible en demandant que les équations (3.3) soient satisfaites au sens des distributions *i.e.* U sera une solution faible de (3.1) si pour toute fonction $\varphi(t,\boldsymbol{x})$ à support compact dans $\mathbb{R}^+ \times \mathbb{R}^d$ on a :

$$\int_0^{+\infty} \int_{\mathbb{R}^d} \left(\sum_{j=1}^{d} F_j(U) \frac{\partial \varphi}{\partial x_j} + U \frac{\partial \varphi}{\partial t} \right) \mathrm{d}\boldsymbol{x}\, \mathrm{d}t + \int_{\mathbb{R}^d} U_0(\boldsymbol{x}) \varphi(\boldsymbol{x}, 0)\, \mathrm{d}\boldsymbol{x} = 0 \tag{3.4}$$

La condition (3.4) est équivalente à demander que pour tout domaine borné $[t_1, t_2] \times \Omega \subset \mathbb{R}_+ \times \mathbb{R}^d$ on ait :

$$\int_\Omega U(\boldsymbol{x}, t_2)\, \mathrm{d}\boldsymbol{x} - \int_\Omega U(\boldsymbol{x}, t_1)\, \mathrm{d}\boldsymbol{x} + \int_{t_1}^{t_2} \int_{\partial\Omega} \vec{F}.\vec{n}\, \mathrm{d}\sigma\, \mathrm{d}t = 0 \tag{3.5}$$

et on vérifiera que cette relation s'obtient en intégrant sur un intervalle de temps $[t_1, t_2]$ les lois de conservation (3.3). La notion mathématique de solutions faibles correspond bien à la notion physique intuitive de loi de conservation, comme pour le cas scalaire vu au chapitre 2. De manière plus générale, si Ω est un ouvert régulier de $\mathbb{R}^d \times \mathbb{R}^+$ de bord $\mathcal{V}$, la formule de Green montre que

$$\int_{\partial\mathcal{V}} \left[U n_t + \sum_{j=1}^{d} F_j(U) n_j \right] d\partial\mathcal{V} = 0 \tag{3.6}$$

puisqu'on remarque que (3.3) peut (et doit) être comprise comme étant l'opérateur divergence espace–temps appliqué au champ $(U, \vec{F}(U))$. Ainsi, la relation (3.3) n'est rien d'autre qu'une relation de conservation en espace–temps, comme l'exprime (3.6).

La première propriété de ces équations sous leur forme (3.3) ou (3.5) est d'être invariante par changement de repère galiléen ainsi que par rotation.

3.2 Relations de Rankine–Hugoniot

On commence par s'intéresser aux solutions de (3.4) C^1 par morceaux. Dans les régions où la solution est C^1, il est clair que les solutions de (3.4) vérifient (3.3) presque partout. On veut maintenant caractériser le comportement des solutions de (3.4) au travers des frontières des régions de régularité.

On cherche donc des solutions C^1 sauf peut–être au travers d'une surface $\Sigma \subset \mathbb{R}^d \times \mathbb{R}^+$ de classe C^1. Considérons alors les sphères O centrées sur un point quelconque de Σ. Elles sont partagées par Σ en deux demi-sphères O_1 et O_2. On notera $\partial_1 = \partial O \cap \partial O_1$ et $\partial_2 = \partial O \cap \partial O_2$ et on supposera que $(\eta_t, \eta_1, \ldots, \eta_d)$ est la normale extérieure à Σ dirigée de O_1 vers O_2. En appliquant (3.6) à chacun des volumes O, O_1 et O_2 on obtient :

$$\int_{\partial_1 \cup \partial_2} \left[U n_t + \sum_{j=1}^{d} F_j(U) n_j \right] \mathrm{d}\partial\mathcal{V} = 0$$

$$\int_{\partial_1} \left[U n_t + \sum_{j=1}^{d} F_j(U) n_j \right] \mathrm{d}\partial\mathcal{V} + \int_{O\cap\Sigma} \left[U^1 \eta_t + \sum_{j=1}^{d} F_j^1(U) \eta_j \right] d\Gamma = 0$$

$$\int_{\partial_2} \left[U n_t + \sum_{j=1}^{d} F_j(U) n_j \right] \mathrm{d}\partial\mathcal{V} - \int_{O\cap\Sigma} \left[U^2 \eta_t - \sum_{j=1}^{d} F_j^2(U) \eta_j \right] d\partial\mathcal{V} = 0$$

où l'on a noté $(\cdot)^{1,2}$ les limites de part et d'autre de Σ. Alors, en additionnant les deux dernières équations et en retranchant à la somme obtenue la première équation, on obtient :

$$\int_{O\cap\Sigma} \left\{ U^1 \eta_t + \sum_{j=1}^{d} F_j^1(U) \eta_j - U^2 \eta_t - \sum_{j=1}^{d} F_j^2(U) \eta_j \right\} d\partial\mathcal{V} = 0$$

Cette relation étant vraie pour toutes les sphères O, on en déduit la relation de Rankine–Hugoniot :

$$[U]\, \eta_t + \sum_{j=1}^{d} [F_j(U)]\, \eta_j = 0 \tag{3.7}$$

où l'on a noté $[\cdot] = (\cdot)^1 - (\cdot)^2$.

Examinons le cas particulier des ondes discontinues vérifiant (3.7) et l'on va étudier des ondes planes constantes de chaque côté de la discontinuité. Considérons, pour $\sigma \in \mathbb{R}$ la surface de discontinuité donnée par :

$$\Omega = \{(\boldsymbol{x}, t) \in \mathbb{R}^d \times \mathbb{R}_+; \phi(\boldsymbol{x}, t) = \boldsymbol{x} \cdot \vec{n} - \sigma t = 0\}.$$

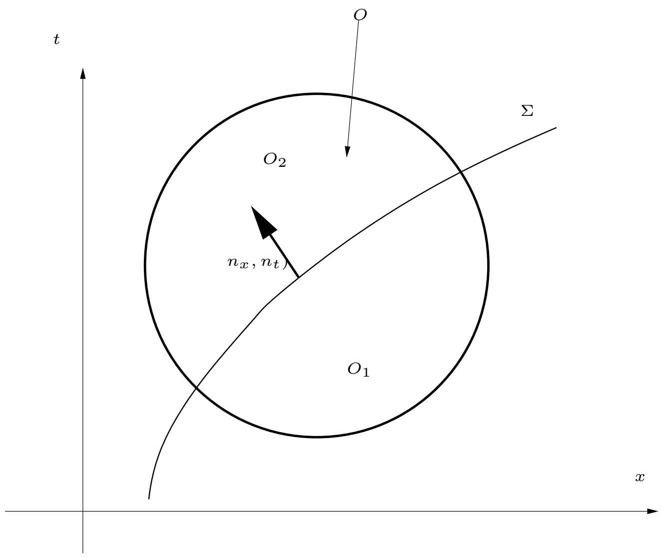


FIG. 3.1 – Discontinuité au travers d'une surface Σ. η_x désigne le vecteur de composantes $\eta_1, \ldots, \eta_d$.

On suppose que U_L (resp. U_R) est la valeur constante de U pour $\boldsymbol{x}\cdot\vec{n}-\sigma t<0$ (resp. $\boldsymbol{x}\cdot\vec{n}-\sigma t>0$). Le vecteur $\left(\dfrac{\partial\phi}{\partial t}, \dfrac{\partial\phi}{\partial x_1}, \ldots, \dfrac{\partial\phi}{\partial x_d}\right)$ est le vecteur normal à Ω et (3.7) devient :

$$-\sigma\,[U] + \sum_{j=1}^{d} [F_j(U)]\, n_j = 0 \tag{3.8}$$

En supposant par exemple[2] que le front se propage le long de l'axe x_d ($\vec{n} = (0,0,\ldots,0,1)$), la fonction discontinue :

$$U(\boldsymbol{x},t) = \begin{cases} U_G & \text{si} \quad x_d < \sigma t, \\ \\ U_D & \text{si} \quad x_d \geq \sigma t, \end{cases} \tag{3.9}$$

où U_G, U_D, σ satisfont (3.8) est une solution de (3.6). Plus précisément, on a

$$\begin{array}{l} \left[\rho u_d\right] = \sigma\left[\rho\right] \\ \left[\rho u_d^2 + p\right] = \sigma\left[\rho u_d\right] \\ \left[(\rho e + p)u_d\right] = \sigma\left[\rho e\right] \\ \left(\rho u_d u_j\right) = \sigma\left[\rho u_j\right], \qquad j = 1, \ldots, d-1. \end{array}$$

En compinant la première équation et les $d-1$ dernières équations, on voit que

$$\left[u_j\right] = 0, \qquad j = 1, \ldots, d-1 \tag{3.10}$$

Les j premières composantes de la vitesse sont invariantes au travers d'un choc. En utilisant ce résultat et en définisant $e' = \rho\varepsilon + \frac{1}{2}\rho u_d^2$, l'energie totale dans la direction x_d, on obtient

$$\begin{array}{l} \left[\rho u_d\right] = \sigma\left[\rho\right] \\ \left[\rho u_d^2 + p\right] = \sigma\left[\rho u_d\right] \\ \left[(\rho e' + p)u_d\right] = \sigma\left[\rho e'\right] \\ \left[u_j\right] = 0, \qquad j = 1, \ldots, d-1. \end{array}$$

[2]ceci sans restriction puisque les équations sont tivariantes par rotation

Ceci montre que le problème se découple en deux problèmes : des relations de choc pour la densité, la vitesse normale au choc et l'énergie totale normale, et $d-1$ relations sur les vitesses transverses au choc : on est donc ramené à étudier plus en détail de telles solutions dans le cas unidimensionnel $d=1$ pour les équations d'Euler.

Une discontinuité solution des équations d'Euler vérifie les relations de Rankine–Hugoniot qui s'écrivent dans ce cas :

$$[\rho u] = \sigma [\rho], \tag{3.11a}$$

$$[\rho u^2 + p] = \sigma [\rho u], \tag{3.11b}$$

$$[(\rho e + p)u] = \sigma [\rho e]. \tag{3.11c}$$

On notera (0) et (1) les deux états de part et d'autre du choc. Introduisons les vitesses v_i relatives au choc :

$$v_i = u_i - \sigma$$

ou, ce qui revient au même, plaçons-nous dans un repère se déplaçant à la vitesse (constante) de la discontinuité. Dans ce repère, la vitesse de la discontinuité est nulle et l'on a puisque les relations de Rankine Hugoniot sont invariantes par changement de repère galiléen :

$$\rho_0 v_0 = \rho_1 v_1, \tag{3.12a}$$

$$\rho_0 v_0^2 + p_0 = \rho_1 v_1^2 + p_1, \tag{3.12b}$$

$$(\rho_0 \varepsilon_0 + \rho_0 v_0^2/2 + p_0)v_0 = (\rho_1 \varepsilon_1 + \rho_1 v_1^2 + p_1)v_1. \tag{3.12c}$$

Posons $m = \rho_0 v_0 = \rho_1 v_1$, le flux de masse qui traverse la discontinuité. Deux cas peuvent se présenter.

1. Si $m = 0$, alors on a nécessairement :

$$\begin{aligned} u_0 &= u_1 = \sigma, \\ p_0 &= p_1 \end{aligned} \tag{3.13}$$

   et la seule possibilité pour avoir une discontinuité est d'avoir :

$$\rho_0 \neq \rho_1.$$

   On appelle ce type de discontinuité, une discontinuité de contact.

2. Lorsque $m \neq 0$, toutes les quantités varient à la traversée de la discontinuité de contact. On dit que l'on a affaire à un choc.

Dans le cas d'une onde de choc, on peut réécrire les relations (3.11a)–(3.11b)–(3.11c) sous la forme

$$m = (u_1 - u_0)/(\tau_1 - \tau_0), \tag{3.14a}$$

$$m^2 = -(p_1 - p_0)/(\tau_1 - \tau_0), \tag{3.14b}$$

$$\varepsilon_1 - \varepsilon_0 + (p_1 + p_0)\frac{(\tau_1 - \tau_0)}{2} = 0, \tag{3.14c}$$

où $\tau = 1/\rho$ est le volume spécifique.

Remarque 3.1. *1. relation (3.14b) montre que la pression et le volume spécifique varient en sens opposé.*

2. La relation (3.14c) est de nature purement thermodynamique. En fait puisque l'état thermodynamique d'un milieu est totalement défini par deux variables d'états. On peut écrire $\varepsilon = \varepsilon(p, \tau)$ et (3.14c) est une relation liant les variables p et τ. On appelle cette relation la relation d'Hugoniot. On notera que cette relation est valable sans qu'aucune hypothèse n'ait été faite sur la loi d'état. Cette relation est donc valable, non seulement pour les gaz parfaits, mais pour les mélanges de gaz parfaits, les gaz réels ou même pour des gaz chimiquement réactifs. Pour analyse plus complète de cette équation, on pourra consulter [2] ou, pour une analyse plus complète [8].

Dans le cas scalaire, on a vu que les relations de saut ne permettent pas de définir les solutions discontinues. Un problème identique existe avec les relations (3.14a)–(3.14b)–(3.14c). Il faut «sélectionner» les «bonnes» solutions. Pour cela, il faut revenir à l'origine des équations, limites formelle des équations de Navier Stokes.

3.3 Entropie

Les équations de Navier–Stokes s'écrivent comme

$$\frac{\partial U}{\partial t} + \sum_{j=1}^{d} \frac{\partial F_j(U)}{\partial x_j} = \sum_{j=1}^{d} \frac{\partial Q_j}{\partial x_j} \tag{3.15}$$

où $Q = Q^v + Q^c$ avec

$$Q_j^v := Q_j^v(u, \frac{\partial u}{\partial x_1}, \dots, \frac{\partial u}{\partial x_d}) = \begin{pmatrix} 0 \\ \tau_{j1} \vdots \\ \tau_{jd} \\ \sum_{l=1}^{d} \tau_{jl} u_l \end{pmatrix}, \qquad Q_j^c = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ q_j \end{pmatrix}$$

où $q = (q_1, \dots, q_d)$ est le flux de chaleur et $\tau = (\tau_{ij})$ est défini par

$$\tau = \lambda(\text{ div } u)\,\text{Id} + 2\mu\frac{\nabla u + \nabla u^T}{2}.$$

Le tenseur $D = \frac{\nabla u + \nabla u^T}{2}$ est le tenseur de déformation, les coefficients λ et μ sont les coefficients de Lamé. Le flux de chaleur est donné par la loi de Fourier

$$q = k\nabla T$$

où T est la température. Généralement λ, μ et k dépendent de T, aussi pour fermer le système, on a besoin de deux lois d'état $p = p(\rho, \varepsilon)$ et $T = T(\rho, \varepsilon)$. On observe que

$$Q_j^v = \sum_{l=1}^{d} A_{jl} \frac{\partial u}{\partial x_j}.$$

L'entropie spécifique d'un système est une loi d'état définie par

$$Tds = d\left(\frac{\varepsilon}{\rho}\right) + pd\left(\frac{1}{\rho}\right)$$

qui, d'après le second principe de la thermodynamique doit vérifier pour tout volume $\Omega(t)$ se déplaçant avec le fluide l'inégalité suivante

$$\frac{d}{dt}\int_{\Omega(t)} \rho s d\Omega \geq \int_{\Omega(t)} \frac{\mathcal{Q}}{T} d\Omega - \int_{\partial\Omega(t)} \frac{q}{T} d\partial\Omega \tag{3.16}$$

où $\mathcal{Q}$ est l'apport de chaleur par unité de volume. Il est dû à l'existence de terme sources ou de transformations chimique qu'on va négliger ici, donc pour nous $\mathcal{Q} = 0$.

Récrite au sens des distributions, (3.16) donne

$$\frac{D(\rho s)}{Dt} + \sum_{j=1}^{d} \frac{\partial (q_j/T)}{\partial x_j} \geq 0. \tag{3.17}$$

où on a utilisé la dérivée matérielle

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \sum_{j=1}^{d} u_j \frac{\partial}{\partial x_j}.$$

A partir de là, on obtient (en utilisant la conservation de l'énergie) la relation de Clausius Duhem

$$\rho\left(T\frac{Ds}{Dt} - \frac{D\varepsilon}{Dt}\right) - p\sum_{j=1}^{d} \frac{\partial u_j}{\partial x_j} + \sum_{i,j} \tau_{ij}\frac{\partial u_j}{\partial x_i} - \frac{1}{T}\sum_j q_j \frac{\partial T}{\partial x_j} \geq 0. \tag{3.18}$$

A partir de la relation (3.15), on peut montrer que le modèle est compatible avec le second principe si $k \geq 0$ et $2\lambda + 3\mu \geq 0$, ce que nous supposerons.[3]

Si maintenant, on multiplie (3.15) par $\nabla_U(\rho s)$ et en tenant compte de la relation de conservation de la masse, on obtient

$$\frac{\partial(\rho s)}{\partial t} + \sum_{j=1}^{d} \frac{\partial \rho u_j s}{\partial x_j} = \nabla_U(\rho s) \cdot \left[\sum_{j=1}^{d} \frac{\partial Q_j}{\partial x_j}\right]. \tag{3.19}$$

[3]Le terme de production de chaleur, $\sum_j q_j \frac{\partial T}{\partial x_j}$ s'écrit $k(\nabla T)^2$ d'où $k \geq 0$ car $T \geq 0$. Pour le terme de dissipation $\sum_{i,j} \tau_{ij}\frac{\partial u_j}{\partial x_i}$, on peut consulter [9] page 20, par exemple.

On peut estimer le second membre de (3.19) en utilisant l'inégalité de Clausius–Duhem (3.18), l'équation de conservation de la masse et la conservation de l'énergie totale,

$$\frac{\partial(\rho s)}{\partial t}+\sum_{j=1}^{d}\frac{\partial \rho u_j s}{\partial x_j}=\frac{1}{T}\dots\geq 0$$

La fonction $U\mapsto \rho s(U)$ est concave, donc $-\rho s$ est une entropie mathématique.

On remarque que

$$\begin{aligned}\nabla_U(\rho s)\cdot\left[\sum_{j=1}^{d}\frac{\partial Q_j}{\partial x_j}\right]&=\sum_{j=1}^{d}\frac{\partial[\nabla_U(\rho s)\cdot Q_j]}{\partial x_j}\\&\quad-\sum_{j,l}\frac{\partial u}{\partial x_j}^T\{\nabla^2(\rho s)A_{jl}\}\frac{\partial u}{\partial x_l}\end{aligned}$$

On peut montrer (voir [10] par exemple) que la matrice

$$\begin{pmatrix}\nabla^2(\rho s)A_{11} & \dots & \nabla^2(\rho s)A_{1d}\\ \vdots & & \vdots\\ \nabla^2(\rho s)A_{d1} & \dots & \nabla^2(\rho s)A_{dd}\end{pmatrix}$$

est symétrique positive (mais pas définie positive).

Dans [11], A. Harten construit, à la suite de [12], une famille d'entropie (c'est à dire de fonctionnelles strictement convexes) de la forme $\mathcal{S}=-\rho g(s)$ avec $g'>0$ et $g''/g'<\gamma^{-1}$, mais Hughes, Franca et Mallet [10] montre que la seule entropie telle que

$$-\begin{pmatrix}\nabla^2\mathcal{S}A_{11} & \dots & \nabla^2\mathcal{S}A_{1d}\\ \vdots & & \vdots\\ \nabla^2\mathcal{S}A_{d1} & \dots & \nabla^2\mathcal{S}A_{dd}\end{pmatrix}$$

soit symétrique et positive est obtenue pour $g(s)=c_0+c_1 s$ avec c_0 et $c_1>0$ constantes quelconques.

Dans le cas des équations d'Euler, on parlera de solution entropique quand pour toute entropie (mathématique) $\mathcal{S}$, il existe un flux (d'entropie) $\mathcal{G}=(\mathcal{G}_1,\dots,\mathcal{G}_d)$ tel que l'inégalité

$$\frac{\partial\mathcal{S}}{\partial t}+\sum_{j=1}^{d}\frac{\partial\mathcal{G}_j}{\partial x_j}\leq 0$$

est vraie au sens des distributions Cette inégalité signifie en particulier que $\nabla_U\mathcal{S}(\nabla F_1,\dots,\nabla F_d)=(\nabla\mathcal{G}_1,\dots,\nabla\mathcal{G}_d)$.

3.4 Hyperbolicité

Le fait que le système (3.3) admette des entropie convexes entrainent de nombreuses propriétés. En particulier, on a les résultats suivants dus à Mock [12]. On pose $\Omega=\{U=(\rho,\rho u,E)$ tels que $\rho>0$ et $E-\rho u^2/2>0\}$

Théorème 3.4.1 (Mock). *Soit $\mathcal{S}$: $\Omega \to \mathbb{R}$ une fonction strictement convexe. Une condition nécessaire et suffisante pour que $\mathcal{S}$ soit une entropie de (3.3) est que les matrices $\nabla_U^2 S \nabla_U F_j$, $1 \leq j \leq d$ soient symétriques. On dit alors que $\mathcal{S}$ symétrise (3.3).*

Démonstration. Voir [2], page 24. □

Dans la suite, on note par $A_j(U)$ les matrices jacobiennes définies par : $A_j(U) = \dfrac{\partial F_j(U)}{\partial U}$. Par exemple, pour les équations d'Euler, l'expression détaillée de ces matrices est (on se restreint ici au cas bi-dimensionnel pour simplifier les écritures) :

$$\frac{\partial F_1(U)}{\partial U} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -u^2 + (\gamma-1)E_c & (3-\gamma)u & -(\gamma-1)v & (\gamma-1) \\ -uv & v & u & 0 \\ u(\gamma-1)E_c - uH & H-(\gamma-1)u^2 & -(\gamma-1)uv & \gamma u \end{pmatrix} \tag{3.20}$$

et

$$\frac{\partial F_2(U)}{\partial U} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ -uv & v & u & 0 \\ -v^2 + (\gamma-1)E_c & -(\gamma-1)u & (3-\gamma)v & (\gamma-1) \\ v(\gamma-1)E_c - vH & -(\gamma-1)uv & H-(\gamma-1)v^2 & \gamma v \end{pmatrix} \tag{3.21}$$

où $E_c = (u^2+v^2)/2$ et H est l'enthapie totale spécifique $H = e + p/\rho$.

Pour tout $\omega \in \mathbb{R}^d$ on considère la matrice :

$$A(U,\omega) = \sum_j \omega_j \frac{\partial F_j}{\partial U} = \sum_j \omega_j A(U, \vec{e_i}) \tag{3.22}$$

Définition 3.1. *Le système d'équations :*

$$\frac{\partial U}{\partial t} + \sum_{j=1}^{d} A_j(U) \frac{\partial U}{\partial x_j} = 0$$

est dit hyperbolique si $\forall U \in G \subset \mathbb{R}^p$ et $\forall \omega \in \mathbb{R}^d, |\omega| = 1$, la matrice $A(U,\omega)$ définie en (3.22) est diagonalisable dans $\mathbb{R}$ i.e. cette matrice possède p valeurs propres réelles $\lambda_1(U,\omega) \leq \lambda_2(U,\omega) \leq \cdots \leq \lambda_p(U,\omega)$ et une famille complète de vecteurs propres à droite $r_1(U,\omega), r_2(U,\omega), \ldots, r_p(U,\omega)$ vérifiant :

$$A(U,\vec{n}) r_k(U,\vec{n}) = \lambda_k(U,\vec{n}) r_k(U,\vec{n}) \tag{3.23}$$

et constituant une base de $\mathbb{R}^p$.

Si toutes les valeurs propres réelles sont distinctes, on dit que le système est strictement hyperbolique.

Le fait que (3.3) soit symétrisable implique que le système est hyperbolique. En effet, A_1 est semblable à $[\nabla^2\mathcal{S}]^{1/2}A_1[\nabla^2\mathcal{S}]^{-1/2}$ et cette matrice est égale à la matrice symétrique

$$[\nabla^2\mathcal{S}]^{-1/2}\Big(\nabla^2\mathcal{S}A_1\Big)[\nabla^2\mathcal{S}]^{-1/2}$$

qui est donc diagonalisable. On procède de même avec A_2 puis $\omega_1 A_1+\omega_2 A_2$ d'où le résultat.

Dans le cas des équations d'Euler, on peut faire le calcul explicitement. Par un calcul direct, on montre que le système (3.1) des équations d'Euler est *hyperbolique*, (mais pas strictement hyperbolique pour $d=2,3$), les valeurs propres de la matrice $A(U,\vec{n})$ sont pour $d=2$:

$$\left\{\begin{array}{ll} \lambda_1(U,\vec{n}) = & \boldsymbol{u}\cdot\vec{n}-a\\ \lambda_2(U,\vec{n}) = & \boldsymbol{u}\cdot\vec{n}\\ \lambda_3(U,\vec{n}) = & \boldsymbol{u}\cdot\vec{n}\\ \lambda_4(U,\vec{n}) = & \boldsymbol{u}\cdot\vec{n}+a \end{array}\right. \tag{3.24}$$

avec $a=\sqrt{\dfrac{\gamma p}{\rho}}$, la vitesse du son. Cette matrice admet un système complet de vecteurs propres à droite $r_k(U,\vec{n})$ qui sont :

$$r_1(U,\vec{n}) = \begin{pmatrix} 1\\ u-a\omega_1\\ v-a\omega_2\\ H-\vec{u}.\vec{n}a \end{pmatrix}, \quad r_2(U,\vec{n}) = \begin{pmatrix} 1\\ u\\ v\\ (u^2+v^2)/2 \end{pmatrix} \tag{3.25}$$

$$r_3(U,\vec{n}) = \begin{pmatrix} 0\\ -\omega_2\\ \omega_1\\ \vec{u}.\vec{n}^\perp \end{pmatrix}, \quad r_4(U,\vec{n}) = \begin{pmatrix} 1\\ u+a\omega_1\\ v+a\omega_2\\ H+\vec{u}.\vec{n}a \end{pmatrix}, \tag{3.26}$$

Les formes linéaires propres sont données, pour un vecteur $U=(A,B,C,D)^T$ par

$$\ell_+(U) = \frac{1}{2}\left(\frac{\mathcal{A}}{a^2}+\frac{\mathcal{B}}{a}\right) \tag{3.27a}$$

$$\ell_0(U) = A-\frac{\mathcal{A}}{a^2} \tag{3.27b}$$

$$\ell_t(U) = -vB+uA-\vec{n}^\perp\cdot\vec{u} \tag{3.27c}$$

$$\ell_-(U) = \frac{1}{2}\left(\frac{\mathcal{A}}{a^2}-\frac{\mathcal{B}}{a}\right) \tag{3.27d}$$

où on a posé

$$\begin{array}{l} \mathcal{A} = D-uB-vC+\frac{u^2+v^2}{2}A\\ \mathcal{B} = Bu+Cv-\vec{n}\cdot\vec{u} \end{array} \tag{3.28}$$

On peut facilement voir que les vecteurs propres sont en fait orthogonaux pour la forme quadratique définie par l'entropie,

$$<U,U'>= -U^T\,\nabla^2(\rho s)\,U'.$$

3.5 Quelques solutions particulières

En vue d'étudier le problème de Riemann, on regarde deux classes de solutions, l'une formée de solutions régulières, l'autres de solutions discontinues. Pour faire cette étude, on se place en une dimension d'espace.

Le système des équations d'Euler s'écrit comme

$$\frac{\partial U}{\partial t} + \frac{\partial F(U)}{\partial x} = 0.$$

On remarque que si $(x,t) \mapsto U(x,t)$ est une solution faible du problème, si $\lambda \in \mathbb{R}$, la fonction $(x,t) \mapsto U(\lambda x, \lambda t)$ est encore une solution faible du problème. Par contre, pour que le second principe soit respecté, il faut se restreindre à $\lambda > 0$.

On va donc chercher des solutions de la forme

$$U(x,t) = V(x/t)$$

où $\xi \mapsto V(\xi)$ est une fonction à valeur dans $\mathbb{R}^3$. On appelle onde simple toute fonction[4] autosimilaire. Dans cette partie, on recherche donc des solutions de type onde simple.

3.5.1 Les ondes simples régulières

Si U est supposée régulière, V est aussi régulière et l'on a

$$\Big(A(V) - \frac{x}{t}\mathrm{Id}_3\Big)V'(\frac{x}{t}) = 0.$$

La fonction V vérifie donc, pour tout $\xi \in \mathbb{R}$,

$$\Big(A(V) - \xi\mathrm{Id}_3\Big)V'(\xi) = 0.$$

Il s'ensuit que ou bien V est constant, ou bien V' est vecteur propre de $A - \xi\mathrm{Id}_3$. On note encore $\lambda_1 = u - a$, $\lambda_2 = u$ et $\lambda_3 = u + a$. On note aussi ℓ_1, ℓ_2, ℓ_3 les formes linéaires associées aux vecteurs propres r_1, r_2, r_3 correspondant aux trois valeurs propres.. On a donc trois possibilités

1. $\lambda_2(\xi) = \xi$. Dans ce cas, $\ell_1(V') = \ell_3(V') = 0$ soit

$$\begin{array}{l} E' - u(\rho u)' + \frac{u^2}{2}\rho' = (\gamma - 1)p' = 0 \\ (\rho u)' - u\rho' = \rho u' = 0 \end{array}$$

   dont on déduit $u' = 0$ et $p' = 0$: u et p sont constants.

2. $\lambda_1(\xi) = \xi$. Dans ce cas, $\ell_2(V') = 0$ et $\ell_3(V') = 0$. On a donc

$$\begin{array}{l} \rho u' - (\gamma - 1)\frac{p'}{a} = 0 \\ \rho' - \frac{p'}{a^2} = 0 \end{array}$$

   En utilisant l'expression de a, $a^2 = \gamma\frac{p}{\rho}$, et le fait que γ est constant, on en déduit que

$$\begin{array}{l} u + 2\dfrac{a}{\gamma - 1} = \mathrm{Cte} \\ p\rho^{-\gamma} = \mathrm{Cte}. \end{array}$$

[4] régulière ou non

3. $\lambda_3(\xi) = \xi$. Des calculs similaires montrent que

$$\begin{aligned} u - 2\frac{a}{\gamma - 1} &= \text{Cte} \\ p\rho^{-\gamma} &= \text{Cte}. \end{aligned}$$

On voit que les deux premiers cas correspondent à une solution telle que $\frac{dU}{d\sigma}$ est colinéaire à r_1 (resp. r_3). Dans le dernier cas, $\frac{dU}{d\sigma}$ est colinéaire à r_2.

On appelle invariants de Riemann les fonctions $u \pm \frac{2}{\gamma-1}a$, $p\rho^{-\gamma}$ qui sont associées aux valeurs propres λ_1 et λ_3, ainsi que la vitesse et la pression qui sont associés à λ_2. On appelera onde simple centrée toute solution des équations d'Euler telles que l'une des trois situations suivante soit vraie

- ou bien $p\rho^{\gamma}$ et $u + \frac{2a}{\gamma-1}$ sont constants.
- ou bien $p\rho^{\gamma}$ et $u - \frac{2a}{\gamma-1}$ sont constants
- ou bien p et u sont constants.

On cherche maintenant une paramétrisation simple des ondes simples. Considérons pour commencer le cas d'une détente associée à la valeur propre $u - a$ issue d'un état(ρ_0, u_0, p_0) auquel correspond la vitesse du son a_0. On a, si

$$\begin{aligned} \frac{p}{\rho^{\gamma}} &= \frac{p_0}{\rho_0^{\gamma}} \\ u + \frac{2}{\gamma-1}a &= u_0 + \frac{2}{\gamma-1}a_0 \end{aligned}$$

De la première relation, on déduit

$$\frac{a}{a_0} = \left(\frac{p}{p_0}\right)^{\frac{\gamma-1}{2\gamma}}$$

d'où

$$u = u_0 - \frac{2}{\gamma-1}a_0\left(1 - \left(\frac{p}{p_0}\right)^{\frac{\gamma-1}{2\gamma}}\right).$$

On pose

$$V(p; p_0, \rho_0) = \frac{2}{\gamma-1}a_0\left(1 - \left(\frac{p}{p_0}\right)^{\frac{\gamma-1}{2\gamma}}\right). \tag{3.29}$$

La relation précédente s'écrit donc

$$u = u_0 - V(p; p_0, \rho_0). \tag{3.30}$$

De même, dans le cas d'une détente associée à l'onde $u + a$, on trouvera le paramètrage

$$u = u_0 + V(p; p_0, \rho_0). \tag{3.31}$$

On voit donc que les courbes se paramètrent dans les coordonnées (u, p).

Un calcul simple montre que la fonction V est décroissante en p car $\gamma > 1$. Ainsi, dans une détente associée à $u - a$, puisque

$$u - a = u_0 - a_0 - a_0\frac{\gamma-1}{\gamma+1}\left(\frac{p}{p_0}\right)^{\frac{\gamma-1}{2\gamma}},$$

la valeur propre λ_1 est croissante quand p croit. On voit aussi d'après (3.30) que u décroit quand p croit. De même, dans une détente associée à λ_3, la valeur propre λ_3 décroit quand p croit et u croit quand p croit.

3.5.2 Les ondes simples non régulières

On va chercher une solution régulière par morceau autosimilaire de la forme

$$U(x,t) = V(\frac{x}{t} - \sigma)$$

Les courbes où sont localisées les discontinuités sont nécessairement des droites. On va s'intéresser au cas où il n'y a qu'une discontinuité se propageant à la vitesse σ. Ces solutions vérifient donc les relations de Rankine Hugoniot (3.11a), (3.11b), (3.11c) qu'on réécrit dans le repère lié à la discontinuité. On considère donc plutôt (3.12a), (3.12b), (3.12c). On considère une onde de choc, donc $m = \rho_0 v_0 = \rho_1 v_1 \neq 0$. On pose

$$M_0 = \frac{v_0}{a_0}$$

le nombre de Mach dans le référentiel lié au choc.

Un calcul un peu long, mais qui n'emploie que l'étude des racines d'un polynôme du second degré donne

$$\begin{aligned} \frac{\rho_0}{\rho_1} &= \frac{\gamma-1}{\gamma+1} + \frac{2}{(\gamma+1)M_0^2}, \\ \frac{v_1}{v_0} &= \frac{\gamma-1}{\gamma+1} + \frac{2}{(\gamma+1)M_0^2}, \\ \frac{p_1}{p_0} &= -\frac{\gamma-1}{\gamma+1} + 2\frac{\gamma M_0^2}{\gamma+1} \end{aligned} \tag{3.32}$$

ce qui permet de paramétrer l'onde en fonction du nombre de choc amont ou aval suivant le choix effectué.

En inversant les indices 0 et 1, on obtient

$$\left(M_0^2 - \frac{\gamma-1}{2\gamma}\right)\left(M_1^2 - \frac{\gamma-1}{2\gamma}\right) = \left(\frac{\gamma+1}{2\gamma}\right)^2.$$

Cette relation montre que si $M_0^2 \geq 1$ alors $M_1^2 \leq 1$, et récriproquement.

La relation (3.32) permet d'étudier la variation d'entropie au travers d'une discontinuité. On a

$$s_1 - s_0 = c_v \log\left(\frac{p_1}{p_0}\left(\frac{\rho_0}{\rho_1}\right)^\gamma\right).$$

On voit facilement que $s_1 - s_0 \geq 0$ si et seulement $M_0^2 \geq 1$. Aussi, les discontinuités admissibles sont celles telles que $M_0^2 \geq 1$ et $M_1^2 \leq 1$. Si $M_0^2 \geq 1$, on voit que $p_1 \geq p_0$, $\rho_1 \geq \rho_0$: un choc est donc compressif. De plus, on a $v_1/v_0 \in [0,1]$. Si $M_0 \geq 0$ alors $M_1 \geq 0$. On a donc $u_0, u_1 \geq \sigma$ et de plus, $u_0 - \sigma \geq a_0$ et $u_1 - \sigma \leq a_1$ c'est à dire $u_0 - a_0 \geq \sigma$ et $u_1 - a_1 \leq \sigma$. De plus, le second principe de la thermodynamique donne

$$[\rho u s] \geq \sigma[\rho s]$$

c'est à dire

$$m[s] \geq 0 \tag{3.33}$$

avec $m = \rho_0(u_0 - \sigma) = \rho_1(u_1 - \sigma)$. Cette inégalité montre que l'entropie des particules augmente quand elles traversent les chocs. Ainsi, si $M_0 \geq 0$, $m \geq 0$ et donc $[s] > 0$: on a $p_1 > p_0$ si $M_0 > 0$ et $p_0 < p_1$ si $M_0 < 0$.

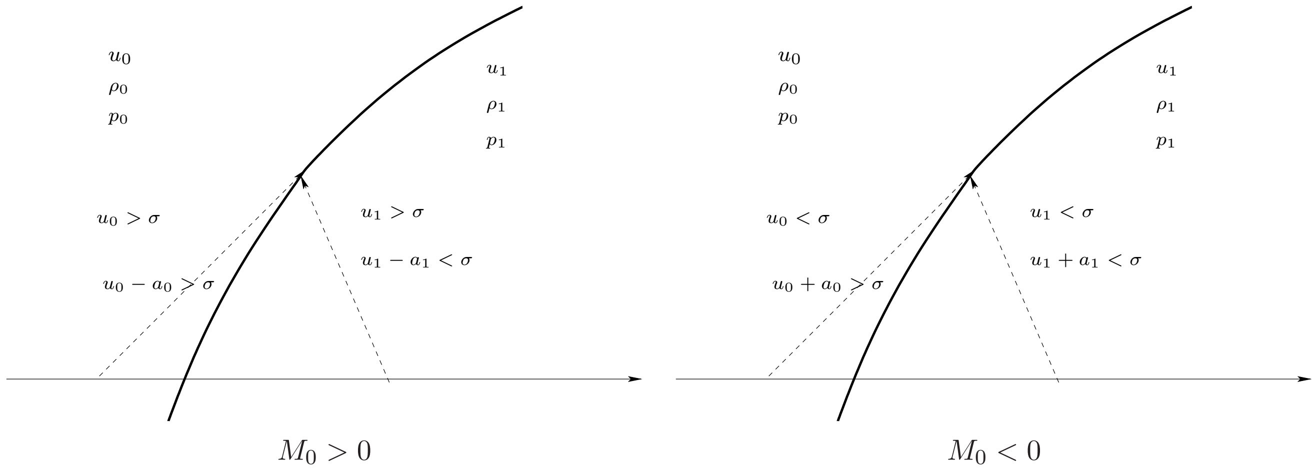


FIG. 3.2 – Illustration des conditions de Lax.

Si $M_0 \leq 0$ alors $M_1 \leq 0$ et donc $u_0, u_1 \leq \sigma$ et $u_0 + a_0 \geq \sigma$ et $u_1 + a_1 \leq \sigma$. Ce sont les conditions de Lax.

On a donc deux branches définissant le lieu des états admissible, l'une où $M_0 > 0$ et l'autre pour laquelle $M_0 < 0$. Quand $M_0 < 0$, l'onde de choc est associée à la valeur propre $u - a$, tandis que quand $M_0 > 0$ elle est associée à $u + a$, voir la figure 3.2.

Les relations (3.32) sont paramétrées par le nombre de Mach par rapport à l'un des états. On peut en fait paramétrer ces courbes dans les variables (u, p). En effet, reprenons les relations de Rankine Hugoniot. On a

$$\begin{aligned}
&\rho v = \rho_0 v_0 := m \\
&\rho v^2 + p = \rho_0 v_0^2 + p_0 \\
&\frac{e + p}{\rho} = \frac{e_0 + p_0}{\rho_0}
\end{aligned}$$

En combinant les deux premières équations, on a

$$p - p_0 = m(v_0 - v) = m(u_0 - u)$$

d'où

$$u = u_0 + \frac{p - p_0}{m}.$$

Il faut déterminer $m = \rho_0(u_0 - \sigma) = \rho_0 a_0 \frac{u_0 - \sigma}{a_0} = \rho_0 u_0 M_0$. Puisque

$$\frac{p}{p_0} = \frac{\gamma - 1}{\gamma + 1} + \frac{2}{(\gamma + 1) M_0^2}$$

on obtient

$$M_0 = \pm \sqrt{\frac{2}{(\gamma + 1)\frac{p}{p_0} - (\gamma - 1)}}.$$

Posons

$$W(p; p_0, \rho_0) = \sqrt{\frac{2}{(\gamma + 1)\frac{p}{p_0} - (\gamma - 1)}}. \tag{3.34}$$

Un calcul simple montre que W est croissante avec p.

La courbe de choc se paramètre comme

$$u = u_0 - W(p; p_0, \rho_0) \tag{3.35}$$

pour une onde de choc associée à $u - a$ issue de (ρ_0, u_0, p_0) et

$$u = u_0 + W(p; p_0, \rho_0) \tag{3.36}$$

pour une onde de choc associée à $u + a$. Notons que, à cause du second principe, (3.35) ne doit être considérée que pour $p > p_0$ et (3.36) pour $p \leq p_0$.

En comparant les relations (3.35) et (3.30), on constate que (3.36) ne doit être considérée que pour $p > p_0$, auquel cas u diminue. Pour (3.30), la situation est analogue : u diminue avec p. Pour les relations (3.31) et (3.36), on a un phénomène nalogue : u diminue quand p augmente.

Ceci conduit à définir les courbes d'onde issues de (p_0, u_0)

1. les 1–ondes définies par $u = u_0 - \Theta(p; p_0, \rho_0)$ avec

$$\Theta(p; p_0, \rho_0) = \begin{cases} V(p; p_0, \rho_0) & \text{pour } 0 \leq p \leq p_0 \\ W(p; p_0, \rho_0) & \text{pour } p \geq p_0 \end{cases} \tag{3.37}$$

2. les 3–ondes définies par $u = u_0 + \Theta(p; p_0, \rho_0)$.

Les courbes d'ondes issues de (p_0, u_0) sont de classe C^1, et on les a représentées sur la figure 3.3.

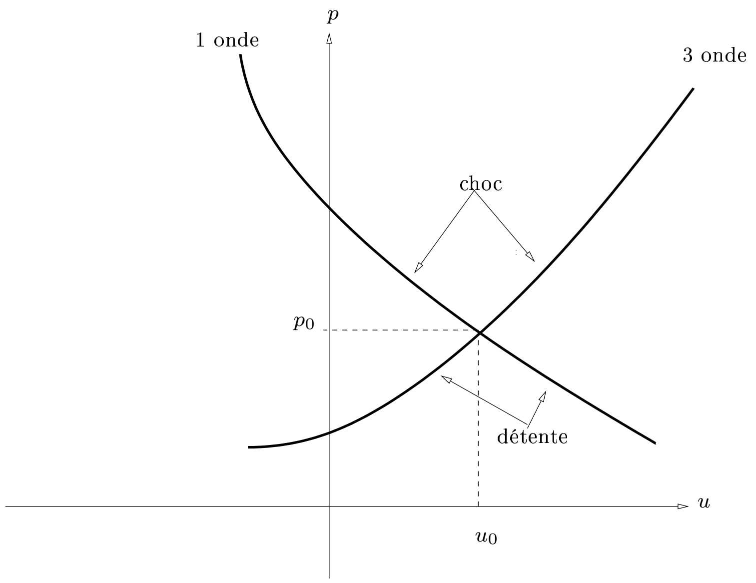


FIG. 3.3 – Représentation des courbes d'ondes.

3.6 Le problème de Riemann

On appelle problème de Riemann le problème de Cauchy suivant

$$\frac{\partial U}{\partial t} + \frac{\partial F(U)}{\partial x} = 0 U(x,0) = \begin{cases} U_G & \text{si } x < 0 \\ U_D & \text{si } x > 0 \end{cases}$$

Un calcul simple permet de voir que

$$\begin{aligned}
&\nabla_U \lambda_1(U(\sigma)) \cdot r_1(U(\sigma)) = -\frac{\gamma+1}{2\gamma} a < 0,\\
&\nabla_U \lambda_3(U(\sigma)) \cdot r_3(U(\sigma)) = \frac{\gamma+1}{2\gamma} a > 0,\\
&\nabla_U \lambda_2 \cdot r_2(U) = 0.
\end{aligned}$$

On dit que les champs caractéristiques associés à λ_1 et λ_3 sont vraiment non linéaires tandis que celui associé à λ_2 est linéairement dégénéré. Une propriété important des équations d'Euler est que la nature des champs ne change pas.

Une autre propriété de ce système est qu'il est strictement hyperbolique : les valeurs propres de la jacobienne vérifie toujours $\lambda_1(U) < \lambda_2(U) < \lambda_3(U)$.

On a alors le résultat général suivant du à P. Lax

Théorème 3.6.1 (Lax). *Soit*

$$\frac{\partial U}{\partial t} + \frac{\partial F}{\partial x}(U) = 0$$

un système de lois de conservations strictement hyperbolique défini dans un ouvert $\mathcal{U} \subset \mathbb{R}^n$ *contenant l'état* U_0. *On note* $\lambda_1 < \lambda_2 < \ldots < \lambda_{n-1} < \lambda_n$ *les valeurs propres de* $\nabla_U F$. *On suppose que chaque champ caractéristique est ou bien vraiment non linéaire ou linéairement dégénéré au voisinage de* U_0.

Pour tout voisinage $\mathcal{W} \subset \mathcal{U}$ *de* U_0, *il existe un voisinage* $\mathcal{V} \subset \mathcal{W}$ *de* U_0 *tel que pour tout* U_G, $U_D \in \mathcal{V}$, *le problème de Riemann entre* U_G *et* U_D *possède une unique solution qui se présente sous la forme d'une succession d'état constants*

$$U_0 = U_G, U_1, \cdots, U_n = U_D$$

où, quelque soit $0 \leq i \leq n-1$, *les états* U_i *et* U_{i+1} *séparés par onde simple.*

Démonstration. On peut trouver la preuve dans Godlewski-Raviart [2] pages 84–87 ou Serre [3], pages 141–145. Elle repose sur l'application du théorème d'inversion locale. La condition de champs uniformément vraiment non linéaire ou uniformément linéairement dégénéré joue un rôle fondamental. □

Dans le cas des équations d'Euler, on cherche donc la solution, voir la figure 3.6, que la solution du problème de Riemann comme une suite d'états formés d'une détente/onde de choc (la structure ℓ_1 de la figure 3.6) associée à l'onde $u-a$, suivi d'une discontinuité de contact, suivi d'une détente/onde de choc (la structure ℓ_3 de la figure 3.6). Puisque ℓ_1 et ℓ_3 se paramètrent au moyen de (u,p), et que la discontinuité de contacte laisse uniforme la vitesse et la pression, le problème se ramène à déterminer $u^\star$ et $p^\star$ la vitesse et la pression entre les états ℓ_1 et ℓ_3. On a donc

$$\begin{aligned}
u^\star &= u_G + \Theta(p^\star; p_G, u_G) \quad &\text{onde } \ell_1\\
u^\star &= u_D - \Theta(p^\star; p_D, u_D) \quad &\text{onde } \ell_2
\end{aligned}$$

d'où

$$u_G + \Theta(p^\star; p_G, u_G) = u_D - \Theta(p^\star; p_D, u_D). \tag{3.38}$$

On résoud cette équation par une méthode de point fixe ou une méthode de type Newton.

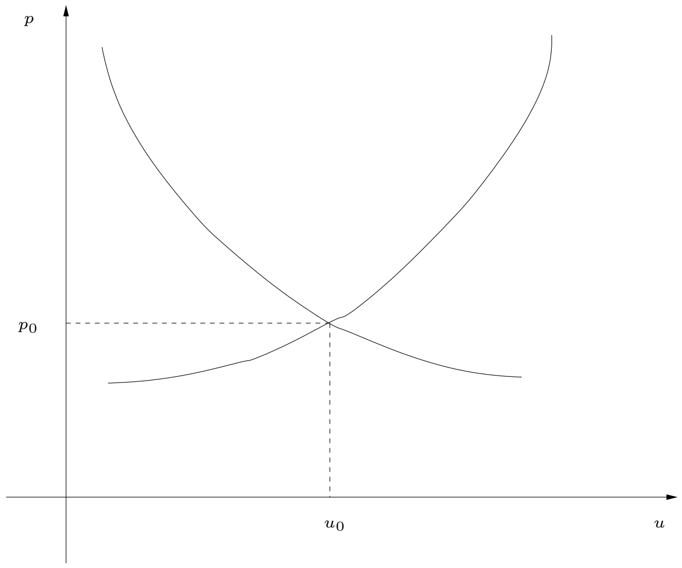


FIG. 3.4 – Etats admissibles.

Il existe des cas où cette équation n'a pas de solution : ce sont les cas où il y a apparition du vide. En effet, la fonction Θ (3.37) est définie pour $p \geq 0$. Les 1–ondes admettent donc une vitesse maximale

$$u_{\max} = u_0 + \Theta(0; p_0, \rho_0) = u_0 + \frac{2}{\gamma - 1} a_0$$

et les 3–ondes admettent une vitesse minimale

$$u_{\min} = u_0 - \frac{2}{\gamma - 1} a_0.$$

Ainsi, si

$$u_G + \frac{2}{\gamma - 1} a_G < u_D - \frac{2}{\gamma - 1} a_D,$$

voir figure 3.7. Dans cd cas, on complète la solution par une zone de vide, voir figure 3.8.

Remarque 3.2 (Cas du problème de Riemann en plusieurs dimension d'espace). *Quand on considère le problème de Cauchy pour l'équation* (3.1) *avec la condition initiale*

$$U(\boldsymbol{x}, t = 0) = \begin{cases} U_G & si\ \boldsymbol{x} \cdot \boldsymbol{n} < 0 \\ U_D & si\ \boldsymbol{x} \cdot \boldsymbol{n} > 0 \end{cases}$$

on se ramène au cas unidimensionnel en utilisant l'invariance par rotation des équations d'Euler et la même méthode que dans le paragraphe 3.2. On se ramène donc à étudier le problème de Riemann unidimensionnel où la donnée de Cauchy est

$$U'_G = (\rho_G, \rho_G (u_n)_G, e'_G)^T, \qquad U'_D = (\rho_D, \rho_D (u_n)_D, e'_D)^T$$

où $u_n = \boldsymbol{u} \cdot \boldsymbol{n}$*. On a donc une succession d'une 1–onde suivi d'une discontinuité de contact, suivi d'une 3–onde, les densités, pressions et vitesses normales sont définies comme en une dimension*

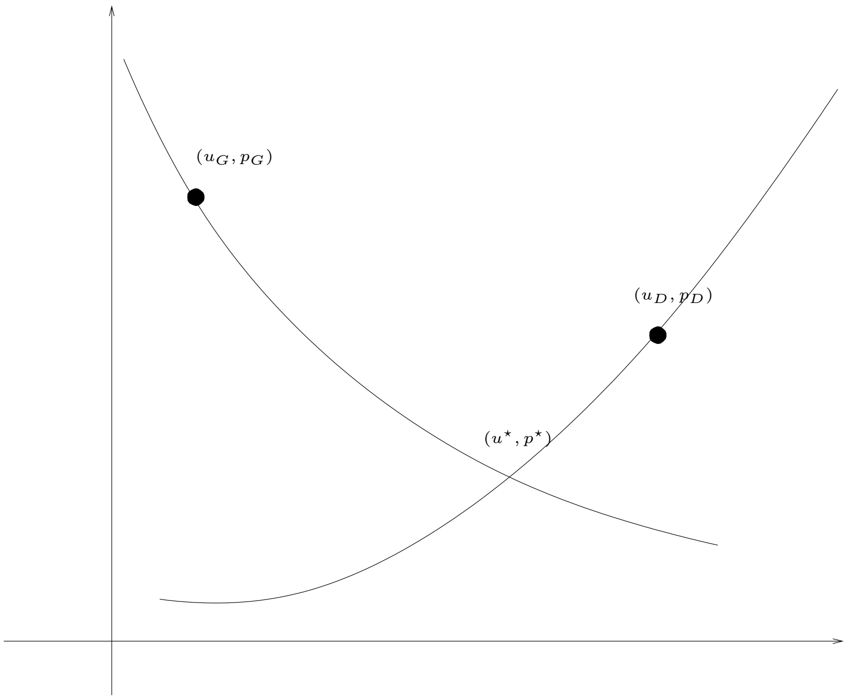


FIG. 3.5 – Solution du problème de Riemann.

d'espace. En particulier, on note u_n^* la vitesse du contact. De plus, les vitesses transverses $u_t = \boldsymbol{u} \cdot \boldsymbol{n}^\perp$ sont définies par

$$u_t = \begin{cases} (u_t)_G & si\ x_n < u_n^* t \\ (u_t)_D & si\ x_n > u_n^* t \end{cases}$$

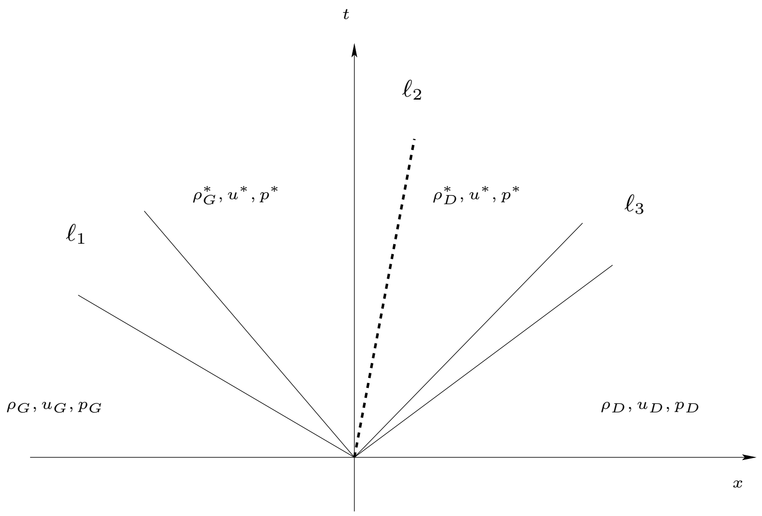


FIG. 3.6 –

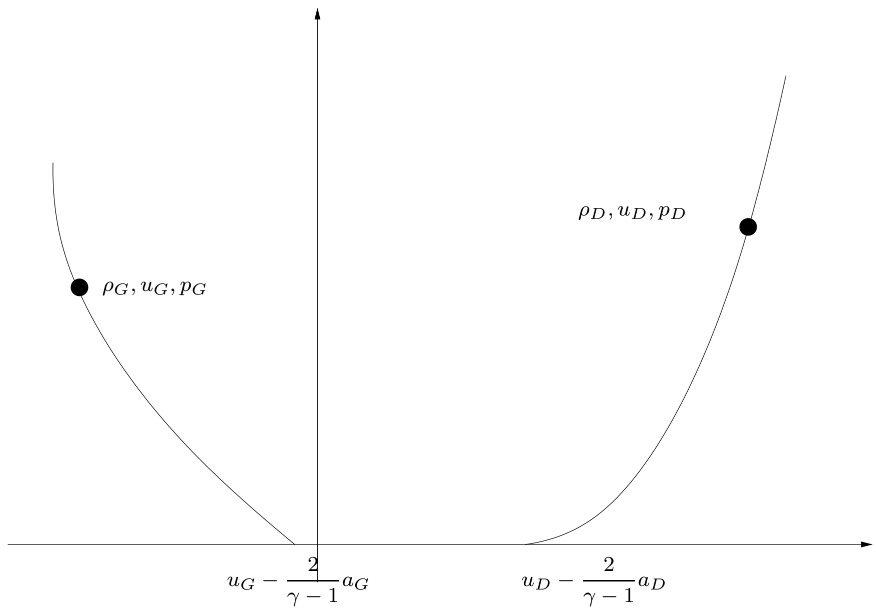


FIG. 3.7 – Cas d'apparition du vide.

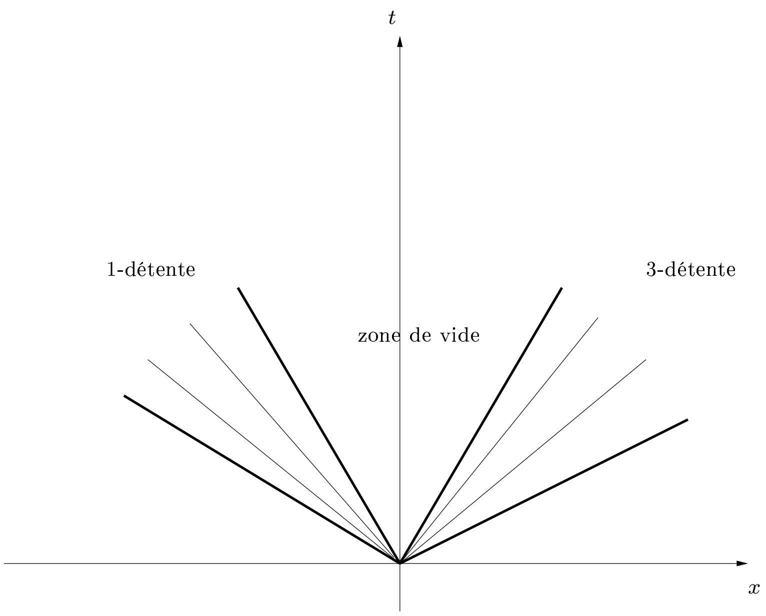


FIG. 3.8 – Solution du problème de Riemann dans le cas d'apparition du vide.

Chapitre 4

Approximation numérique par la méthode des volumes finis

4.1 Le principe des schémas conservatifs

On part du système

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial \mathbf{f}(u)}{\partial x} = 0 & x \in \mathbb{R}, t > 0 \\ u(x,0) = u_0(x) & x \in \mathbb{R}. \end{cases} \tag{4.1}$$

où la solution faible entropique u est définie sur $\mathbb{R} \times \mathbb{R}^+$ et à valeur dans un ouvert Ω de $\mathbb{R}^p$. Dans le cas des équations de la méanique des fluides, on a $p = 3$ et Ω est l'ensemble des états ayant une densité et une pression positives.

On considère une subdivision de $\mathbb{R}$, $(x_i)_{i\in\mathbb{Z}}$, et on note $x_{i+1/2} = \frac{x_i + x_{i+1}}{2}$. On définit ensuite les *cellules de contrôle* C_i par

$$C_i = [x_{i-1/2}, x_{i+1/2}]$$

et $\Delta x_i = x_{i+1/2} - x_{i-1/2}$ est sa largeur. La réunion des cellules C_i est $\mathbb{R}$. De même, si $\Delta t > 0$, on pose $t_n = n\Delta t$, $n \geq 0$ entier, et on subdivise $\mathbb{R}^+$ en intervalles $[t_n, t_{n+1}[$ dont la réunion constitue $\mathbb{R}^+$.

Formellement au moins[1], on peut intégrer (4.1) dans $C_i \times [t_n, t_{n+1}[$ afin d'obtenir (après division par Δx_i)

$$\begin{aligned} &\frac{1}{\Delta x_i}\int_{x_{i-1/2}}^{x_{i+1/2}} u(x,t_{n+1})\mathrm{d}x - \frac{1}{\Delta x_i}\int_{x_{i-1/2}}^{x_{i+1/2}} u(x,t_n)\mathrm{d}x \\ &+\frac{\Delta t}{\Delta x_i}\left(\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \mathbf{f}(u(x_{i+1/2},s))\mathrm{d}s - \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \mathbf{f}(u(x_{i-1/2},s))\mathrm{d}s\right) = 0 \end{aligned} \tag{4.2}$$

Cette écriture suggère :

1. qu'une description intéressante de u est donnée par

$$u_i^n \simeq \frac{1}{\Delta x_i}\int_{x_{i-1/2}}^{x_{i+1/2}} u(x,t_n)\mathrm{d}x \quad \text{pour} \quad i \in \mathbb{Z} \text{ et } n \geq 0.$$

[1] mais cela peut se justifier aisément, voir Godlewski–Raviart [1].

Ce sont les valeurs moyennes de u dans chaque cellule de contrôle. Un autre intérêt de cette description est que u peut ne pas être définie en un point, par exemple quand la solution présente une discontinuité. Par contre, les solutions recherchées sont intégrables donc leur valeur moyenne est toujours définie.

2. Le bilan d'accroissement de u_i^n au cours d'un pas de temps est donné par

$$\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \mathbf{f}(u(x_{i+1/2},s))\mathrm{d}s - \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \mathbf{f}(u(x_{i-1/2},s))\mathrm{d}s.$$

Le terme $\dfrac{1}{\Delta t}\displaystyle\int_{t_n}^{t_{n+1}} \mathbf{f}(u(x_{i+1/2},s))\mathrm{d}s$ s'interprète comme le flux moyen en temps entrant ou sortant de l'interface de C_i située en $x_{i+1/2}$. L'équation (4.2) se réécrit

$$u_i^{n+1} = u_i^n - \lambda_i \left(\mathbf{f}_{i+1/2}^n - \mathbf{f}_{i-1/2}^n\right) \tag{4.3}$$

où $\lambda_i = \frac{\Delta t}{\Delta x_i}$. On dit que ce « schéma » s'écrit sous forme *conservative.*

Les schémas sous forme conservative sont particulièrement étudiés à cause du théorème de Lax Wendroff (dans lequel on suppose que le pas d'espace est uniforme, $\Delta x_i = \Delta x$ quelque soit $i \in \mathbb{Z}$.). On se donne un flux numérique $\mathbf{f}$ qui dépend de $2k$ arguments $\mathbf{f}_{i+1/2} = \mathbf{f}(u_{i-k+1}, \cdots, u_{i+k})$ et le schéma numérique

$$u_i^{n+1} = u_i^n - \lambda \left(\mathbf{f}_{i+1/2}^n - \mathbf{f}_{i-1/2}^n\right). \tag{4.4}$$

où $\lambda = \frac{\Delta t}{\Delta x}$. On suppose que $\mathbf{f}$ est *consistant,*

$$\mathbf{f}(u, \cdots, u) = \mathbf{f}(u).$$

Théorème 4.1.1. *On considère un schéma écrit sous forme conservative, de flux numérique continu consistant. On suppose que λ reste constant. On considère une suite de pas de discrétisation en espace telle que $\Delta x_l \to 0$ quand $l \to +\infty$. Notons u_{Δ_l} la fonction constante par morceaux, associée à $(u_i^n)_{i\in\mathbb{Z}}$ suite définie par le schéma (4.4), associée à la condition initiale*

$$u_i^0 = \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} u_0(x)\mathrm{d}x$$

et définie par

$$u_{\Delta_k}(x,t) = u_i^n \text{ si } (x,t) \in C_i \times [t_n, t_{n+1}[.$$

S'il existe une constante C indépendante de l telle que $\max_{i,n} \|u_i^n\| \leq C$ et une fonction u_{Δ_k} dans $(\mathrm{L}^1_{loc}(\mathbb{R}\times\mathbb{R}^+))^p$ telle que $u_{\Delta_k} \to u$ dans $(\mathrm{L}^1_{loc}(\mathbb{R}\times\mathbb{R}^+))^p$, alors u est solution faible de (4.1).

La preuve est donnée, par exemple, dans [1].

Rien ne dit que u vérifie les conditions d'entropie . Pour obtenir ceci automatiquement, il suffit que le schéma vérifie des inégalités d'entropie discrètes. Plus précisément, si U est une entropie associée à un flux d'entropie G, on dira que (4.4) vérifie une inégalité d'entropie discrète s'il existe $\mathcal{G}$, un flux d'entropie

$$\mathcal{G}_{i+1/2} = \mathcal{G}(u_{i-k+1}, \cdots, u_{i+k})$$

consistant avec G au sens où

$$\mathcal{G}(u, \cdots, u) = G(u)$$

tel que $(u_i^n)_{i\in\mathbb{Z}}$ vérifie

$$U(u_i^{n+1}) \leq U(u_i^n) - \lambda_i \left(\mathcal{G}_{i+1/2}^n - \mathcal{G}_{i-1/2}^n\right). \tag{4.5}$$

Sous les hypothèses du Théorème de Lax Wendroff, si $\mathcal{G}$ est continue, on voit que u vérifie au sens des distributions

$$\frac{\partial U(u)}{\partial t} + \frac{\partial G(u)}{\partial x} \leq 0 \qquad \text{si } x \in \mathbb{R}, t > 0$$

$$U(u(x,0)) = U(u_0(x)) \quad x \in \mathbb{R}.$$

La preuve est identique à celle du théorème de Lax Wendroff, voir [1]. Cette inégalité permet de montrer que la limite des solutions discrètes vérifie une inégalité d'entropie, celle associée à U et G.

On va considérer trois schémas permettant l'intégration des équations d'Euler : le schéma de Godunov, le schéma de Roe et le schéma de Lax Friedrichs. Chacun illustre des aspects différents de l'approximation. Pour simplifier l'écriture, on supposera toujours que le pas d'espace est uniforme.

4.2 Quelques exemples de flux numériques

Dans chaque exemple, on commencera par la description du schéma dans le cas scalaire, puis on le fera dans le cas du système des équations d'Euler.

4.2.1 Le schéma de Godunov

Ce schéma repose sur la résolution exacte du problème de Riemann.

Commençons par le cas du problème scalaire

$$\begin{aligned} &\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0 \qquad x \in \mathbb{R},\ t > 0 \\ &u(x,0) = u_\Delta(x, t_n) \end{aligned} \tag{4.6}$$

où u_Δ est définie comme dans le théorème de Lax Wendroff.

Grâce au théorème de Kruzkhov, on sait que si u_Δ est bornée (ce que nous supposerons évidemment), si t est assez petit, l'influence des discontinuités ne se fait sentir que sur au plus une maille. Pour cela, il faut que

$$t \leq \frac{1}{2} \min_j \frac{\Delta x_j}{\max_{u\in[-||u^n||_\infty, +||u^n||_\infty]} |f'(u)|} = \Delta t_0.$$

On supposera donc que le pas de temps est plus petit que Δt_0 [2] D'après le théorème de Kruzkov, il y a existence et unicité de la solution notée u. Le schéma de Godunov est définit par

$$u_j^{n+1} = \frac{1}{\Delta x_j} \int_{x_{j-1/2}}^{x_{j+1/2}} u(x, \Delta t) dx^3 \tag{4.7}$$

Un calcul analogue au calcul effectué dans le paragraphe précédent montre que

$$u_j^{n+1} = u_j^n - \lambda_j \big[\hat{f}^G(u_{j-1}^n, u_j^n) - \hat{f}^G(u_j^n, u_{j+1}^n)\big] \tag{4.8}$$

[2]implicitement, en disant cela, on suppose que le maillage est régulier afin que $\Delta t_0 > 0$.

[3]Notons que la consition «initiale» est définie à $t = t_n$.

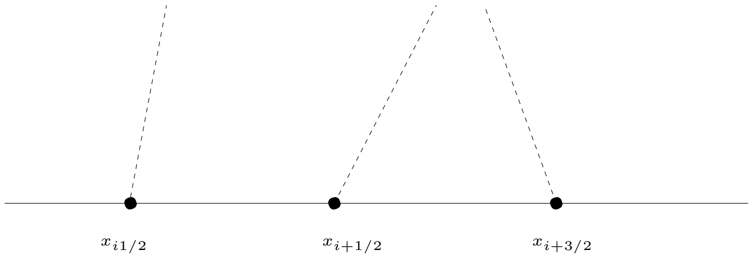


FIG. 4.1 – Illustration du schéma de Godunov.

où le flux de Godunov est définit par

$$\hat{f}^G(u^-, u^+) = f(u^*) \tag{4.9}$$

où u^* est la valeur (constante) de la solution du problème de Riemann en $x = 0$, lieu de la discontinuité initiale. Les calculs effectués dans le paragraphe 2.5.2 montrent que si u^* peut ne pas être définie (car $x = 0$ est le lieu d'une discontinuité stationanire), le flux $f(u^*)$ est lui parfaitement défini grâce aux relations de Rankine Hugoniot.

Passons au cas des équations d'Euler maintenant. Supposons que u_Δ soit défini comme dans le théorème de Lax Wendroff jusqu'à l'instant t_n, et considérons le problème de Cauchy suivant :

$$\begin{aligned} &\frac{\partial u}{\partial t} + \frac{\partial \mathbf{f}(u)}{\partial x} = 0 \qquad \text{si} \quad x \in \mathbb{R}, \quad t > 0, \\ &u(x,0) = u_\Delta(x, t_n) \quad x \in \mathbb{R}. \end{aligned} \tag{4.10}$$

Ce système étant hyperbolique, on peut montrer que si Δt est choisi de manière à ce que

$$\frac{\Delta t}{\Delta x} \max_i \left(|u_i^n| + a_i^n\right) \leq \frac{1}{2}, \tag{4.11}$$

la résolution de (4.10) dans n'importe quelle cellule C_i se ramène à la superposition de deux problèmes de Riemann indépendants. La condition CFL (4.11) fait intervenir la vitesse des ondes les plus rapides (les ondes acoustiques) définies par la condition « initiale » $u_\Delta(\cdot, t_n)$.

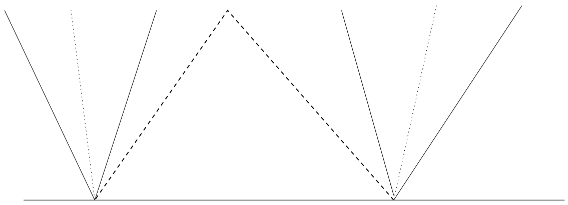


FIG. 4.2 – Résolution du problème de Cauchy (4.10). On a placé les 3 ondes issues des problèmes de Riemann aux interface des cellules et en pointillé les segments $x = x_{i\pm1/2} \pm \max_i \left(|u_i^n| + a_i^n\right) t$.

Grâce à cette remarque, la résolution du problème de Cauchy à l'interface $x_{i+1/2}$ entre les deux cellules adjacentes C_i et C_{i+1} ne dépend que de u_i^n et de u_{i+1}^n. Or, dans la relation (4.2), la variation de la valeur moyenne ne dépend que de la valeur moyenne à l'instant t_n et des contributions des flux aux noeuds $x_{i+1/2}$ et $x_{i-1/2}$. Si on *pose*

$$u_i^{n+1} = \frac{1}{\Delta x_i} \int_{x_{i-1/2}}^{x_{i+1/2}} u(x, t_{n+1}) \mathrm{d}x,$$

on obtient

$$u_i^{n+1} = u_i^n - \lambda_i \left(\mathbf{f}_{i+1/2} - \mathbf{f}_{i-1/2} \right) \tag{4.12}$$

avec

$$\mathbf{f}_{i+1/2} = \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \mathbf{f}(u(x_{i+1/2}, s)) \mathrm{d}s.$$

On a donc besoin de connaître $u(x_{i+1/2}, s)$ en $x = x_{i+1/2}$ et sur un pas de temps.

On va montrer que $\mathbf{f}(u(x_{i+1/2}, s))$ ne dépend pas de s. Pour cela, on peut se ramener au problème de Riemann séparant une interface en $x = 0$ séparant les états u_D et u_G. Notons $u(x, t; u_D, u_G)$ la solution du problème de Riemann

$$\begin{aligned} &\frac{\partial u}{\partial t} + \frac{\partial \mathbf{f}(u)}{\partial x} = 0 && \text{si } x \in \mathbb{R}, t > 0 \\ &u(x, 0) = \begin{cases} u_D & \text{si } x > 0 \\ u_G & \text{si } x < 0 \end{cases} \end{aligned} \tag{4.13}$$

On voit facilement que si u est solution, $u'(x, t) = u(\lambda x, \lambda t)$ est aussi solution car l'équation et la condition initiale sont autosimilaires. De manière à conserver le deuxième principe (l'inégalité d'entropie), il faut choisir $\lambda > 0$. Dans le chapitre 3, on a vu qu'il y avait unicité de la solution et donc

$$u(x, t; u_D, u_G) = u' \left(\frac{x}{t}; u_D, u_G \right) \text{ pour } t > 0.$$

u est définie en x si u n'a pas de discontinuité en x. Sinon, la solution doit vérifier les relation de Rankine–Hugoniot.

Supposons dans un premier temps que u n'a pas de discontinuité en $x = 0$. Alors, on a pour $t > 0$, $u(0, t; u_D, u_G) = u'(0; u_D, u_G)$ et donc l'intégrande en temps est constante.

Si u admet une discontinuité en $x = 0$, elle est nécessairement stationaire (puisque c'est $x = 0$!), et donc

$$\mathbf{f}(u(0^+, t; u_D, u_G)) = \mathbf{f}(u(0^-, t; u_D, u_G)) = \mathbf{f}(u'(0^{\pm}; u_D, u_G))$$

quelque soit t>0, et donc l'intégrande en temps est encore indépendante de t.

On a donc

$$\mathbf{f}_{i+1/2} = \mathbf{f}(u'(0^{\pm}; u_D, u_G). \tag{4.14}$$

Le schéma de Godunov est constitué des relations (4.12), (4.14), sous la condition CFL (4.11).

Une propreté intéressante est que le schéma de Godunov vérifie toutes les inégalités d'entropie possibles. Si (U, G) est un couple entropie flux, on peut définir un flux d'entropie consistant pour lequel on a une inégalité d'entropie discrète. En effet,

$$U(u_i^{n+1}) = U \left(\frac{1}{\Delta x_i} \int_{x_{i-1/2}}^{x_{i+1/2}} u(x, \Delta t) \mathrm{d}x \right) \leq \frac{1}{\Delta x_i} \int_{x_{i-1/2}}^{x_{i+1/2}} U(u(x, \Delta t)) \mathrm{d}x$$

car U est convexe : c'est l'inégalité de Jensen qui généralise l'inégalité de convexité habituelle

$$U(\lambda u_1 + (1-\lambda)u_2) \leq \lambda U(u_1) + (1-\lambda)U(u_2)$$

quelques soient u_1, u_2 physiquement admissibles et $\lambda \in [0,1]$.

À partir de là, on procède comme dans (4.2) en remplaçant $\mathbf{f}$ par G et l'égalité $=0$ par ≤ 0. Le flux d'entropie consistant est

$$\mathcal{G}(u_G, u_D) = G(u'(0^{\pm}; u_D, u_G)).$$

Pour finir, on peut remarquer que le schéma peut aussi être construit comme suit. Tout d'abord,

$$u_i^{n+1} = \frac{1}{2}\left(u_i^{+,n+1} + u_i^{-,n+1}\right)$$

avec

$$u_i^{+,n+1} = \frac{2}{\Delta x}\int_{x_i}^{x_{i+1/2}} u(\frac{x}{\Delta t}; u_i, u_{i+1})\mathrm{d}x$$

$$u_i^{-,n+1} = \frac{2}{\Delta x}\int_{x_{i-1/2}}^{x_i} u(\frac{x}{\Delta t}; u_{i-1}, u_i)\mathrm{d}x.$$

En utilisant la relation (4.1), on peut réécrire $u_i^{+,n+1}$ comme

$$u_i^{+,n+1} = u_i^n - \lambda_i(\mathbf{f}_{i+1/2}^{+} - \mathbf{f}(u_i)) \tag{4.15}$$

où

$$\mathbf{f}_{i+1/2}^{+} = \mathbf{f}(u_i^n) + \frac{\Delta x}{\Delta t}\int_{x_i}^{x_{i+1/2}} \left(u'(\frac{x}{\Delta t}; u_i^n, u_{i+1}^n) - u_i^n\right)\mathrm{d}x \tag{4.16}$$

On réutilisera cette remarque un peu plus loin, pour définir le solveur de Roe.

Le défaut majeur de ce schéma est que le calcul de la solution du problème de Riemann peut être complexe et donc est coûteuse : il faut déterminer l'intersection de deux courbes. Qui plus est, quand la loi d'état est complexe, l'expression analytique de ces deux courbes est délicate à obtenir. On n'obtient généralement pas de forme « sympathique ». C'est principalement la complexité de la résolution du problème de Riemann qui a motivé la recherche d'autres schémas dont l'un des plus connus est le schéma de Roe.

Avant de conclure ce paragraphe, notons que ce schéma possède de nombreuses propriétés. L'une d'entre elles est que, partant d'états admissibles, on arrive toujours à des états admissibles. Pour fixer les idées, on va se placer dans le cadre de la mécanique des fluides. Ici, il faut que la masse volumique et la pression soient positives. Pour illustrer le problème, on va supposer une loi d'etat de gaz parfait,

$$p = (\gamma - 1)\left(E - \frac{1}{2}\rho u^2\right)$$

où E est l'énergie totale, u la vitesse du fluide et ρ sa masse volumique. On utilisera plus loin l'énergie interne spécifique ε reliée à l'énergie totale par

$$E = \rho\varepsilon + \frac{1}{2}\rho u^2.$$

Posons les notations

$$u_i^n = (\rho_i^n, m_i^n, E_i^n).$$

Si on définit la masse volumique, la vitesse et la pression dans la cellule C_i au temps t_n par

$$u_i^n = \frac{m_i^n}{\rho_i^n}$$
$$p_i^n = (\gamma - 1)\left(E_i^n - \frac{1}{2}\rho_i^n (u_i^n)^2\right)$$

Proposition 4.1. *On a toujours, sous la condition (4.11),*

$$\textit{quelques soient } i \textit{ et } n \geq 0, \rho_i^n \geq 0 \textit{ et } p_i^n \geq 0 \tag{4.17}$$

si les conditions initiales vérifient ces inégalités.

Démonstration. Supposons le résultat vrai au rang n. On a

$$u_i^{n+1} = \frac{1}{\Delta x_i} \int_{x_{i-1/2}}^{x_{i+1/2}} u(x, \Delta t) \mathrm{d}x$$

où u est solution du problème de Cauchy (4.10). On a donc nécessairement

$$\rho_i^{n+1} \geq 0 \text{ quelque soit } i \in \mathbb{Z}$$

car u est physiquement admissible. Ensuite

$$\Delta x_i E_i^{n+1} =$$

$$\int_{x_{i-1/2}}^{x_{i+1/2}} E(x, \Delta t)\ \mathrm{d}x = \int_{x_{i-1/2}}^{x_{i+1/2}} (\rho\varepsilon)(x, \Delta t)\ \mathrm{d}x + \frac{1}{2}\int_{x_{i-1/2}}^{x_{i+1/2}} \rho(x,t) u^2(x, \Delta t)\ \mathrm{d}x.$$

Pour simplifier les notation, on va abandonner la dépendance en $(x, \Delta t)$ sous le signe somme. D'après l'inégalité de Cauchy Schwarz (appliquée à $\sqrt{\rho}u$ et $\sqrt{\rho}$), on a

$$\left(\int_{x_{i-1/2}}^{x_{i+1/2}} \rho u\ \mathrm{d}x\right)^2 \leq \left(\int_{x_{i-1/2}}^{x_{i+1/2}} \rho\ \mathrm{d}x\right)\left(\int_{x_{i-1/2}}^{x_{i+1/2}} \rho u^2\ \mathrm{d}x\right).$$

On a donc

$$0 \leq \int_{x_{i-1/2}}^{x_{i+1/2}} (\rho\varepsilon)(x, \Delta t) \leq \int_{x_{i-1/2}}^{x_{i+1/2}} E(x, \Delta t)\mathrm{d}x - \frac{1}{2}\frac{\left(\int_{x_{i-1/2}}^{x_{i+1/2}} \rho u\ \mathrm{d}x\right)^2}{\int_{x_{i-1/2}}^{x_{i+1/2}} \rho\ \mathrm{d}x}.$$

En divisant par Δx_i et en multipliant par $\gamma - 1$, on obtient donc

$$p_i^{n+1} \geq (\gamma - 1)\frac{1}{\Delta x_i}\int_{x_{i-1/2}}^{x_{i+1/2}} (\rho\varepsilon)(x, \Delta t) \geq 0.$$

□

4.2.2 Le schéma de Roe

Dans le cas du problème scalaire (4.6), le flux est donné par

$$f(u^-, u^+) = \frac{1}{2}\big[f(u^+) + f(u^-) - |a(u^-, u^+)|(u^+ - u^-)\big]$$

avec

$$a(u^-, u^+) = \begin{cases} \dfrac{f(u^+) - f(u^-)}{u^+ - u^-} & \text{si } u^+ \neq u^- \\ f'(u^+) & \text{sinon} \end{cases}$$

Plutôt que d'aller plus avant dans les propriétés du schéma de Roe dans le cas scalaire (respect des discontinuités stationnaire par exemple), nous allons développer la discussion autour de ce schéma en partant directement d'un problème système. En effet, cela n'est pas plus compliqué, et cela permet peut–être de saisir les points clefs de la méthode. Enfin, on détaillera les calculs dans le cas de la mécanique des fluides,

Plutôt que de résoudre le problème de Cauchy (4.1) pour le problème de Riemann, on va chercher à exploiter la solution d'un problème de Riemann linéarisé

$$\begin{aligned} &\frac{\partial \widetilde{u}}{\partial t} + \overline{A}\frac{\partial \widetilde{u}}{\partial x} = 0 \quad x \in \mathbb{R}, t > 0 \\ &\widetilde{u}(x, 0) = u_0(x) \end{aligned} \tag{4.18}$$

où u_0 vaut u_D si $x > 0$ et u_G si $x < 0$. Le problème est double :

1. Quelle propriété doit satisfaire la matrice constante $\overline{A}$?
2. Quelles relations doit–il exister entre $\overline{A}$ et le problème réel ?

La réponse à la première question est simple : il faut que le problème linéarisé soit hyperbolique, c'est à dire que A soit diagonalisable à valeurs propres réelles. Il y a deux raisons à cela. D'une part, le problème de départ est lui-même hyperbolique, ensuite on désire que le problème approché soit bien posé : il faut donc que A soit diagonalisable à valeurs propres réelles.

La réponse à la seconde question est beaucoup moins simple. Tout d'abord, on peut souhaiter que les discontinuités soient exactement résolues par le problème linéarisé. Si u_D et u_D sont séparés par une discontinuité se déplaçant à la vitesse σ, on doit avoir simultanément

$$\begin{aligned} &\hat{\mathbf{F}}(u_D) - \hat{\mathbf{F}}(u_G) = \sigma(u_D - u_G) \\ &\overline{A}u_D - \overline{A}u_G = \sigma(u_D - u_G) \end{aligned} \tag{4.19}$$

et donc $\hat{\mathbf{F}}(u_D) - \hat{\mathbf{F}}(u_G) = \overline{A}(u_D - u_G)$ pour une discontinuité, relation qu'on étend à tous les cas.

Une autre approche consiste à souhaiter qu'en moyenne les solutions du problème de Riemann exact et du problème de Riemann approché soient identiques. Pour cela, on intègre (4.13) dans l'intervalle $[x_i, x_{i+1}]$, et on trouve

$$\frac{1}{\Delta x_i}\int_{x_i}^{x_{i+1}} u\left(\frac{x}{\Delta t}; u_G, u_D\right) \mathrm{d}x = \frac{u_G + u_D}{2} - \frac{\Delta t}{\Delta x_i}\left[\hat{\mathbf{F}}(u_D) - \hat{\mathbf{F}}(u_G)\right]. \tag{4.20}$$

La matrice $\overline{A}$ étant diagonalisable, on a le même résultat sur le système linéarisé :

$$\frac{1}{\Delta x_i}\int_{x_i}^{x_{i+1}} u^{app}\left(\frac{x}{\Delta t}; u_G, u_D\right) \mathrm{d}x = \frac{u_G + u_D}{2} - \frac{\Delta t}{\Delta x_i}\overline{A}(u_D - u_G).$$

où u^{app} est la solution du problème de Riemann (4.18). Si on impose que

$$\int_{x_i}^{x_{i+1}} \left(u^{app}\left(\frac{x}{\Delta t}; u_G, u_D\right) - u\left(\frac{x}{\Delta t}; u_G, u_D\right)\right) \mathrm{d}x = 0,$$

on en déduit à nouveau la relation (4.19).

Déterminons maintenant un flux numérique. La première idée est d'utiliser la solution exacte du problème linéarisé. Après calculs, on trouve

$$\hat{\mathbf{F}}(u_G, u_D) = \frac{1}{2}\big(\overline{A}(u_D + u_G) - |\overline{A}|(u_G - u_D)\big).$$

où $|\overline{A}|$ est définie par :

$$|\overline{A}| = R^{-1}|D|R$$

où $|D| = \text{diag } (|\lambda_i|)$. avec R, la matrice des vecteurs propres de $\overline{A}$ et $D = \text{diag } (|\lambda_i)$ la matrice diagonale formée par ses valeurs propres de telle sorte que l'on ait :

$$\overline{A} = R^{-1}DR$$

De même, on pose $\overline{A}^+ = \left(\overline{A} + |\overline{A}|\right)/2$ et $\overline{A}^- = \left(\overline{A} - |\overline{A}|\right)/2$. On peut reécrire $\hat{\mathbf{F}}$ comme

$$\hat{\mathbf{F}}(u_G, u_D) = \overline{A}^+ u_G + \overline{A}^- u_D. \tag{4.21}$$

On voit alors que l'une ou l'autre expression nécessite deux produits matrices vecteurs, ce qui est assez coûteux. On peut simplifier cette expression en remarquant que si $\tilde{\hat{\mathbf{F}}}$ est le flux numérique de matrice jacobienne (constante) $\overline{A}$, on a

$$\tilde{\hat{\mathbf{F}}}(u) = \overline{A}u.$$

On reécrit $\hat{\mathbf{F}}$ comme

$$\hat{\mathbf{F}}(u_G, u_D) = \frac{1}{2}\left(\tilde{\hat{\mathbf{F}}}(u_D) + \tilde{\hat{\mathbf{F}}}(u_G) - |\overline{A}|(u_G - u_D)\right).$$

Il est donc naturel de poser

$$\hat{\mathbf{F}}(u_G, u_D) = \frac{1}{2}\big(\hat{\mathbf{F}}(u_D) + \hat{\mathbf{F}}(u_G) - |\overline{A}|(u_G - u_D)\big) \tag{4.22}$$

ce qui revient à remplacer le flux « linéarisé » $\tilde{\hat{\mathbf{F}}}$ par le flux du problème $\hat{\mathbf{F}}$.

Une autre façon de construire le flux numérique est de partir des relations (4.15) et (4.16). En remplaçant dans (4.16) la solution exacte du problème de Riemann par u^{app}, on obtient le flux numérique

$$\hat{\mathbf{F}}(u_G, u_D) = \hat{\mathbf{F}}(u_G) - \frac{\Delta x}{\Delta t}\int_{x_i}^{x_{i+1/2}} \left(u^{\text{app}}(\frac{x}{\Delta t}; u_G, u_D) - u_G\right) \mathrm{d}x$$

Puisque

$$\frac{\Delta x}{\Delta t}\int_{x_i}^{x_{i+1/2}}\Big(u^{\mathrm{app}}(\frac{x}{\Delta t};u_G,u_D)-u_G\Big)\,\mathrm{d}x=\overline{A}(u_G)-\overline{A}^{+}u_G-\overline{A}^{-}u_D,$$

en utilisant la relation (4.19), on obtient l'expression (4.22).

Il reste à obtenir une expression simple de $\overline{A}$. On va en obtenir une valable pour un gaz parfait calorifiquement parfait, expression qui peut se généraliser à des lois d'état très complexes. Plus généralement, on montre que les équations d'Euler sont symétrisables en utilisant les variables entropiques (le gradient de $-\rho s$ où s est l'entropie spécifique physique). Utilisant ce résultat, on peut alors montrer qu'il existe toujours une linéarisation admissible mais elle est malheureusement inutilisable, voir [2]. C'est pourquoi, dans la pratique, on travaille au coup par coup.

Pour revenir au cas du gaz caloriquement parfait, il existe plusieurs constructions équivalentes. La plus rapide (mais la moins généralisable) est due à P.L. Roe [13]. La remarque est que tant u que $\hat{\mathbf{F}}(u)$ sont des fonctions quadratiques en Z, le « parameter vector », défini par

$$Z=(\sqrt{\rho},\sqrt{\rho}u,\sqrt{\rho}H)^{\mathrm{t}}=(Z_1,Z_2,Z_3)^{\mathrm{t}}$$

où H désigne l'enthaplie totale. En effet

$$u=\frac{1}{2}D(Z).Z \text{ et } \hat{\mathbf{F}}(u)=\frac{1}{2}R(Z).Z$$

Les variables conservatives sont

$$\begin{aligned}&\rho=Z_1^2\\&\rho u=Z_1Z_2\\&E=\frac{1}{\gamma}\left(Z_1Z_4+\frac{\gamma-1}{2}Z_2^2\right)\end{aligned}$$

et les flux et

$$\begin{aligned}&\rho u=Z_1Z_2\\&\rho u^2+p=Z_2^2+\frac{\gamma-1}{\gamma}\left(Z_1Z_4-\frac{1}{2}Z_2^2\right)\\&u(E+p)=Z_1Z_3.\end{aligned}$$

On en déduit que les matrices D et R sont

$$D=\begin{pmatrix}2Z_1&0&0\\ Z_2&Z_1&0\\ \dfrac{Z_4}{\gamma}&\dfrac{\gamma-1}{\gamma}Z_2&\dfrac{Z_1}{\gamma}\end{pmatrix}$$

et le jacobien R de $\hat{\mathbf{F}}$ par rapport à Z est

$$R=\begin{pmatrix}Z_2&Z_1&0\\ \dfrac{\gamma-1}{\gamma}Z_4&2Z_2-\dfrac{\gamma-1}{\gamma}Z_2&\dfrac{\gamma-1}{\gamma}Z_1\\ Z_1&0&Z_4\end{pmatrix}$$

On considère maintenant l'interpolation linéaire entre Z_G et Z_D. On a

$$\hat{\mathbf{F}}(u_G) - \hat{\mathbf{F}}(u_D) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}s}\hat{\mathbf{F}}(u(Z(s)))\mathrm{d}s$$

puis

$$\frac{\mathrm{d}}{\mathrm{d}s}\hat{\mathbf{F}}(u(Z(s))) = R(Z(s))\frac{dZ}{ds}.$$

Comme $\dfrac{\mathrm{d}Z}{\mathrm{d}s}$ est constante, on a

$$\hat{\mathbf{F}}(u_G) - \hat{\mathbf{F}}(u_D) = \left(\int_0^1 R(Z(s))ds\right)\frac{\mathrm{d}Z}{\mathrm{d}s} = R\left(\frac{Z_G + Z_D}{2}\right)\frac{\mathrm{d}Z}{\mathrm{d}s}.$$

De même,

$$u_G - u_D = \int_0^1 \frac{d}{ds}u(Z(s))\mathrm{d}s = \left(\int_0^1 D(Z(s))\mathrm{d}s\right)\frac{\mathrm{d}Z}{\mathrm{d}s} = D\left(\frac{Z_G + Z_D}{2}\right)\frac{\mathrm{d}Z}{\mathrm{d}s}$$

Si ρ_G et ρ_D sont strictement positifs, la matrice $D\left(\dfrac{Z_G + Z_D}{2}\right)$ est inversible (c'est une matrice triangulaire à coefficients strictement positifs sur la diagonale), on peut donc éliminer $\dfrac{\mathrm{d}Z}{\mathrm{d}s}$,

$$\hat{\mathbf{F}}(u_G) - \hat{\mathbf{F}}(u_D) = \overline{A}(u_G - u_D) \text{ avec } \overline{A} = R\left(\frac{Z_G + Z_D}{2}\right)\left[D\left(\frac{Z_G + Z_D}{2}\right)\right]^{-1}.$$

Pour avoir une expression simple de $\overline{A}$, il suffit de voir que pour un état u quelconque, on a toujours

$$\frac{\partial \hat{\mathbf{F}}}{\partial u} = R(Z)D(Z)^{-1}$$

où Z est le « parameter vector » correspondant à u. Si on pose

$$\begin{aligned}
\overline{\rho} &= \sqrt{\rho_G \rho_D} \\
\overline{u} &= \frac{\sqrt{\rho_G}u_G + \sqrt{\rho_D}u_D}{\sqrt{\rho_G} + \sqrt{\rho_D}} \\
\overline{H} &= \frac{\sqrt{\rho_G}H_G + \sqrt{\rho_D}H_D}{\sqrt{\rho_G} + \sqrt{\rho_D}}
\end{aligned} \tag{4.23}$$

on définit complètement un état moyen $\overline{u}$ où évaluer $\overline{A} = \dfrac{\partial \hat{\mathbf{F}}}{\partial u}(\overline{u})$.

En fait, la matrice jacobienne de $\hat{\mathbf{F}}$ ne dépend que de u et H (mais pas de ρ). Tous calculs faits, on a :

$$\overline{A} = \begin{pmatrix} 0 & 1 & 0 \\ \dfrac{\gamma - 3}{2}\overline{u}^2 & (\gamma - 2)\overline{u}^2 & (\gamma - 1)u \\ \overline{u}(\dfrac{\gamma - 1}{2}\overline{u}^2 - \overline{H}) & \overline{H} - (\gamma - 1)\overline{u} & (\gamma - 1) \end{pmatrix} \tag{4.24}$$

Grâce à cette relation, on voit que les valeurs propres de la matrice moyenne sont $\overline{u}$ et $\overline{u} \pm \overline{c}$ où

$$\overline{a}^2 = (\gamma - 1)\left(\overline{H} - \frac{1}{2}\overline{u}^2\right).$$

Vérifions que $\overline{a}^2 \geq 0$. On pose

$$\overline{H} = \overline{h} + \frac{1}{2}\overline{(u^2)}$$

où $\overline{h} = \dfrac{\sqrt{\rho_G}h_G + \sqrt{\rho_D}h_D}{\sqrt{\rho_D} + \sqrt{\rho_G}}$ désigne l'enthalpie spécifique moyenne et

$$\overline{(u^2)} = \frac{\sqrt{\rho_G}u_G^2 + \sqrt{\rho_D}u_D^2}{\sqrt{\rho_D} + \sqrt{\rho_G}}.$$

Un calcul simple montre que

$$\overline{(u^2)} - \overline{u}^2 = \frac{\sqrt{\rho_G\rho_D}}{\sqrt{\rho_G} + \sqrt{\rho_D}}(u_G - u_D)^2$$

On voit donc que $\overline{a}^2$ est la somme de deux nombres positifs, tant que la densité reste positive.

Malgré les apparences, le calcul du flux numérique est simple. Notons r_0, r_-, r_+ les vecteurs propres de $\overline{A}$,

$$r_0 = \begin{pmatrix} 1 \\ \overline{u} \\ \dfrac{\overline{u}^2}{2} \end{pmatrix} \quad r_- = \begin{pmatrix} 1 \\ \overline{u} - \bar{a} \\ \bar{H} - \bar{u}\bar{a} \end{pmatrix} \quad r_+ = \begin{pmatrix} 1 \\ \overline{u} + \bar{a} \\ \bar{H} + \bar{u}\bar{a} \end{pmatrix}$$

et $\mu_0 = \overline{u}$, $\mu_- = \bar{u} - \bar{a}$ et $\mu_+ = \bar{u} + \bar{a}$. En décomposant $u_G - u_D$, et en exploitant complètement l'expression des moyennes, on arrive a un résultat très simple

$$u_G - u_D = \alpha_0 r_0 + \alpha_- r_- + \alpha_+ r_+$$

avec

$$\alpha_0 = \Delta\rho - \frac{\Delta p}{\overline{a}^2}$$

$$\alpha_- = \frac{\Delta p}{\overline{a}^2} + \frac{\Delta\rho u}{\overline{a}}$$

$$\alpha_+ = \frac{\Delta p}{\overline{a}^2} - \frac{\Delta\rho u}{\overline{a}}$$

où, si f est une quantité quelconque, on a noté Δf la différence $f_G - f_D$.

Ces remarques étant faites, l'expression du flux est

$$\hat{\mathbf{F}}(u_G, u_D) = \frac{\hat{\mathbf{F}}(u_G) + \hat{\mathbf{F}}(u_D)}{2} - \sum_{\xi=0,+,-} |\mu_\xi|\alpha_\xi r_\xi. \tag{4.25}$$

Passons maintenant aux défauts du flux de Roe. Contrairement au flux de Godunov, même si le rapport $\lambda = \Delta t/\Delta x$ est très petit, la pression et la masse volumique ne restent pas nécessairement positives. Ce problème se rencontre, par exemple, lors de l'interaction de deux chocs forts se propageant en sens opposé, dans un tube à choc. Généralement, on résoud le problème en $\cdots$ changeant de solveur ou de flux numérique. Cependant, il existe une autre version de la matrice moyenne de Roe, très différente de celle présentée plus haut, qui permet de garantir la positivité de la densité et de la pression, voir [14] pour plus de détails.

Un autre défaut est qu'il ne faut pas espérer obtenir une inégalité d'entropie. Considérons une discontinuité séparant deux états u_D et u_G telle que

$$\hat{\mathbf{F}}(u_G) = \hat{\mathbf{F}}(u_D).$$

Dans ce cas,

$$\hat{\mathbf{F}}(u_G, u_D) = \hat{\mathbf{F}}(u_G) = \hat{\mathbf{F}}(u_D),$$

la discontinuité restera donc numériquement stationnaire. Si on suppose que les états gauche et droit sont tels que $\rho_D < \rho_G$ et $p_G < p_D$, alors, la discontinuité ne peut pas être admissible (voir le chapitre 3. Le schéma numérique fournit une solution qui ne respecte pas le deuxième principe de la thermodynamique. La raison profonde de ce phénomène est que $u_D - u_G$ est un vecteur propre de $\overline{A}$ associé à une valeur propre nulle : le schéma n'est plus dissipatif dans cette configuration.

Une solution à ce problème est de rajouter un peu de dissipation. Ceci peut se faire par la correction d'Harten–Yee. Au lieu d'évaluer $|\mu|$ dans (4.25), on évalue $\varphi_\varepsilon(\mu)$ où

$$\varphi_\varepsilon(x) = \begin{cases} |x| & \text{si } |x| \geq \varepsilon \\ \dfrac{x^2+\varepsilon^2}{2\varepsilon} & \text{sinon.} \end{cases}$$

Le paramètre ε est choisi en fonction du rayon spectral de $\overline{A}$,

$$\varepsilon = \alpha(|\bar{u}| + \bar{a})$$

où α dépend des applications : très faible en transonique, jusqu'à 0.25 en hypersonique. On voit que $\varphi_\varepsilon(x) \geq \frac{\varepsilon}{2}$, la dissipation numérique ne s'annulera jamais.

Notons que rien n'interdit d'employer une correction entropique différente suivant chaque mode propre. En particulier, si on cherche à capturer une couche limite, problème pour lequel le flux de Roe se révèle bien adapté grâce à sa propriété de capture exacte de discontinuité, il faut chercher à minimiser la quantité de dissipation que l'on met suivant le mode colinéaire à r_0.

4.2.3 Le flux de Lax–Friedrich

Dans le cas du flux de Roe, le paramètre de dissipation est donné par la matrice moyenne $\bar{A}$. On peut, dans certains cas, estimer que son calcul est trop coûteux, ou dans d'autres cas, il est trop difficile. Aussi, regardons ce qui se passe quand on la remplace par une matrice diagonale $\alpha\mathrm{Id}$.

Dans ce cas, le flux numérique est

$$\hat{f}_{j+1/2} = \frac{1}{2}\Big(f(u_j) + f(u_{j+1}) + \alpha_{j+1/2}(u_{j+1} - u_j)\Big).$$

On voit très facilement que ce flux sera L^∞ stable si on choisit

$$\alpha_{j+1/2} \geq \left|\frac{f(u_{j+1}) - f(u_j)}{u_{j+1} - u_j}\right|$$

et sous la contrainte CFL
$$\frac{\Delta t}{\Delta x}\max_j \alpha_{j+1/2}\leq \frac{1}{2}.$$

On pourra le vérifier en exercice.

Regardons maintenant à quelle condition éventuelle ce flux permet de vérifier des inégalités d'entropie. Pour cela, dans le cas scalaire, on sait qu'il suffit de vérifier ce qui se passe pour les entropie de Kruzkov. Le schéma s'écrit (après avoir posé $\lambda=\frac{\Delta t}{\Delta x}$)

$$\begin{aligned}
u_j^{n+1}=u_j^n-\lambda\bigg(&\big[\frac{f(u_{j+1})-f(u_j)}{u_{j+1}-u_j}+\alpha_{j+1/2}\big](u_{j+1}-u_j)\\
&+\big[\frac{f(u_j)-f(u_{j-1})}{u_j-u_{j-1}}-\alpha_{j-1/2}\big](u_j-u_{j-1})\\
&=(1-\lambda C_{j+1/2}-\lambda C_{j-1/2})u_j+C_{j+1/2}u_{j+1}+C_{j-1/2}u_{j-1}
\end{aligned}$$

avec
$$\begin{aligned}
C_{j+1/2}&=\frac{f(u_{j+1})-f(u_j)}{u_{j+1}-u_j}+\alpha_{j+1/2}\\
C_{j-1/2}&=\frac{f(u_j)-f(u_{j-1})}{u_j-u_{j-1}}-\alpha_{j-1/2}
\end{aligned}$$

On remarque que quelque soit $k\in\mathbb{R}$, si on pose, quelque soit $i\in\mathbb{Z}$, $v_i=u_i-k$, alors en prenant la valeur absolue et sous condition CFL,

$$\begin{aligned}
|v_j^{n+1}|\leq |v_j^n|-\lambda\bigg(&\big[\frac{f(u_{j+1})-f(u_j)}{u_{j+1}-u_j}+\alpha_{j+1/2}\big](|v_{j+1}|-|v_j|)\\
&+\big[\frac{f(u_j)-f(u_{j-1})}{u_j-u_{j-1}}+\alpha_{j-1/2}\big](|v_j|-|v_{j-1}|).
\end{aligned}$$

Ainsi, on voit que
$$|u_j^{n+1}-k|\leq |u_j^n-k|-\lambda\bigg(\hat{g}_{j+1/2}-\hat{g}_{j-1/2}\bigg)$$

avec
$$\begin{aligned}
\hat{g}_{j+1/2}=&g(u_j)+C_{j+1/2}(|u_{j+1}-k|-|u_j-k|)\\
&=\frac{1}{2}\bigg(g(u_j)+g(u_{j+1})+\alpha_{j+1/2}\big(|u_{j+1}-k|-|u_j-k|\big)\bigg)
\end{aligned}$$

avec $g(u)=\mathrm{sign}(u-k)(f(u)-f(k))$. La deuxième expression est obtenue par symétrie.

Dans le cas système, le flux est défini par
$$\hat{\mathbf{F}}(u^-,u^+)=\frac{1}{2}\big[\hat{\mathbf{F}}(u^+)+\hat{\mathbf{F}}(u^-)-\alpha(u^+-u^-)\big]$$

où
$$\alpha>\rho(\overline{A}(u^-,u^+)),$$

$\overline{A}(u^-,u^+)$ désignant la moyenne de Roe. Dans [15], on montre que ce schéma possède de nombreuses propriétés (respect dela positivite de la densité et de la pression par exemple).

4.2.4 Le Schéma de Lax Wendroff

On emploie ici une technique complètement différente. On va commencer par interpoler les données au moyen d'une interpolation linéaire en temps dans $[t_n, t_{n+1}]$ et linéaire continue en espace. Remarquons que dès lors, l'interprétation des données est différente des trois cas précédents : au lieu d'avoir une représentation constante par cellule de manière implicite, on a une représentation continue. On note u^h cette interpolation.

Puis, partant de

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0,$$

on intègre une première fois en temps,

$$u^{n+1}(x) - u^n(x) + \int_{t_n}^{t_{n+1}} \frac{\partial f(u)}{\partial x} dt.$$

Si on veut un schéma explicite, on va approcher l'intégrale à l'instant t_{n+1}, et on aura

$$\frac{u^{n+1}(x) - u^n(x)}{\Delta t} + \frac{\partial f(u^n)}{\partial x}$$

Cependant, on n'aura qu'une approximation d'ordre un en temps. On sait que

$$\frac{u^{n+1}(x) - u^n(x)}{\Delta t} = \frac{\partial u}{\partial t} + \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2} + o(\Delta t),$$

et

$$\frac{\partial^2 u}{\partial t^2} = -\frac{\partial}{\partial t}\left(\frac{\partial f(u^n)}{\partial x}\right)$$

Ainsi, si on approche de manière centrée les dérivées en espace, on obtient

$$u_j^{n+1} = u_j^n - \lambda\left(\hat{f}_{j+1/2} - \hat{f}_{j-1/2}\right)$$

avec

$$\hat{f}_{j+1/2} = \frac{1}{2}\left(f(u_j^n) + f(u_{j+1}^n) + d_{j+1/2}\right)$$

où $d_{j+1/2}$ est une approximation centrée de $\frac{\Delta t}{2\Delta x}\frac{\partial}{\partial t}\left(\frac{\partial f(u^n)}{\partial x}\right)$. De nombreux choix sont possibles. Par exemple $d_{j+1/2} = \frac{\Delta t}{2\Delta x} a_{j+1/2}(f(u_{j+1}) - f(u_j))$ avec $a_{j+1/2} \equiv f'(\frac{u_j + u_{j+1}}{2})$. On peut par exemple prendre

$$a_{j+1/2} = \begin{cases} f'(u_j) & \text{si } u_j = u_{j+1}, \\ \dfrac{f(u_{j+1}) - f(u_j)}{u_{j+1} - u_j} & \text{sinon} \end{cases}$$

Dans le cas système, il existe de nombreuses versions du schéma de Lax Wendroff. En voici une

$$\hat{\mathbf{F}}(u^-, u^+) = \frac{1}{2}\left(\hat{\mathbf{F}}(u^+) + \hat{\mathbf{F}}(u^-) + \frac{\Delta t}{2\Delta x}\overline{A}(u^-, u^+)\left[\hat{\mathbf{F}}(u^+) - \hat{\mathbf{F}}(u^-)\right]\right).$$

où $\overline{A}(u^-, u^+)$ désigne encore la linéarisation de Roe.

4.2.5 Schémas monotones et lipschitziens

Dans ce paragraphe, on ne considère que des problèmes scalaires. Considérons un schéma de la forme

$$u_j^{n+1} = u_j^n - \lambda\Big(\hat{f}_{j+1/2} - \hat{f}_{j-1/2}\Big)$$

où $\hat{f}_{j+1/2} = \hat{f}(u_{j+1}, u_j)$ et $\hat{f}_{j-1/2} = \hat{f}(u_j, u_{j-1})$. On dit que le schéma est monotone si quelque soit la suite v^0 avec $u_j^0 \leq v_j^0$, on a $u_j^n \leq v_j^n$ quelque soit $n \in \mathbb{N}$.

On a le résultat suivant

Lemme 4.2.1. *Une condition suffisante de monotonie est que le flux $\hat{f}$ soit croissant par rapport à son premier argument et décroissant par rapport au second argument.*

Démonstration. Voir [1] par exemple. □

Les versions scalaires des schémas de Godunov, Roe et lax Friedrichs sont monotone au sens du lemme précédent. En ce qui concerne le schéma de Godunov, c'est une conséquence du théorème de Kruzkov, en particulier de son point 4. En ce qui concerne les schéma de Roe et Lax Friedrichs, c'est la conséquence de simples calculs.

On dira que le flux $\hat{f}$ est lipschitzien s'il existe des constantes h_1 et h_2 telles que

$$\begin{array}{ll} \text{quelque soit } u, \text{ quelques soient } v, v', & \big|\hat{f}(u,v) - \hat{f}(u,v')\big| \leq h_2 \big|v - v'\big| \\ \text{quelque soit } v, \text{ quelques soient } u, u', & \big|\hat{f}(u,v) - \hat{f}(u',v)\big| \leq h_1 \big|u - u'\big| \end{array}$$

4.3 Le rôle de la représentation des données dans la méthode des volumes finis.

Dans cette partie, on va faire un travail formel sur un problème scalaire, mais les calculs sont identiques pour un système. On considère le problème (4.26)

$$\begin{array}{l} \dfrac{\partial u}{\partial t} + \text{ div } f(u) = \text{div } \big(\varepsilon \cdot \nabla u\big) \quad x \in \Omega \subset \mathbb{R}^d, t > 0 \\ u(x, t = 0) = u_0(x) \\ \text{ conditions aux limites} \end{array} \tag{4.26}$$

Pour simplifier l'exposé, on suppose qu'il n'y a pas de conditions aux limites, autrement dit on prend $\Omega = \mathbb{R}^d$. Considérons $\varphi \in \big(C_0^\infty(\mathbb{R}^d)\big)^p$ une fonction C^∞ à support compact. On obtient la formulation faible en intégrant par partie

$$\int_{\mathbb{R}^d \times \mathbb{R}^+} \left\{\frac{\partial \varphi}{\partial t} u(x,t) + \nabla\varphi(x,t) \cdot f(u)\right\} dxdt - \int_{\mathbb{R}^d \times \mathbb{R}^+} \nabla\varphi \cdot (\varepsilon \cdot \nabla u) dxdt$$
$$= \int_{\mathbb{R}^d} \varphi(x, 0) u_0(x) dx.$$

Soit $K \subset \mathbb{R}^d$ un compact, et t_1, t_2 deux instants > 0. On peut approcher la fonction caractéristique de $K \times [t_1, t_2]$ par une suite de fonctions de $C_0^\infty(\mathbb{R}^d)$, si bien qu'on obtient en passant à la limite formellement

$$\int_{K \times [t_1, t_2]} \left\{\frac{\partial \varphi}{\partial t} u(x,t) + \nabla\varphi(x,t) \cdot f(u)\right\} dxdt - \int_{K \times [t_1, t_2]} \nabla\varphi \cdot (\varepsilon \cdot \nabla u) dxdt$$

On obtient facilement que $\dfrac{\partial \varphi}{\partial t} \to \delta_{K\times\{t_1,t_2\}}$ au sens des distributions, tandis que si $\vec{n}$ est la normale unitaire extérieure à K, $\nabla\varphi \to \vec{n}\delta_{\partial K\times[t_1,t_2]}$. Ainsi on obtient

$$\int_K u(x,t_2)dx - \int_K u(x,t_1) + \int_{t_1}^{t_2}\int_{\partial K} f\cdot\vec{n}d\sigma - \int_{t_1}^{t_2}(K\nabla u)\cdot\vec{n}d\sigma \tag{4.27}$$

Cette relation est à la base de la méthode des volumes finis.

Dans la pratique, on triangule $\mathbb{R}^d$ en éléments «simples» (triangles/paralélogrammes en 2D, tétrahèdres/hexaèdres en 3D)

$$\mathbb{R}^d = \cup_i K_i.$$

A partir de là, il faut représenter les inconnues. Il y a classiquement deux approches

- on représente les inconnues par leur valeur moyenne dans chaque éléments K, c'est à dire u_i^n est une approximation de $\int_K iu(x,t_n)dx$. Ceci permet de définir une représentation globale de la solution au moyen d'une fonction constante dans chaque cellule de contrôle K_i,
- on représente les inconnues par leur «valeur» aux sommets de la triangulation. A partir de là, on interpole les données ce qui permet de définir une représentation globale de la solution.

La deuxième option sous–entend que le maillage soit conforme et généralement l'interpolation est continue.

On va regarder dans ce paragraphe comment s'exprime la première méthode et montrer que la seconde, pour des maillages triangulaires ou tétrahédriques, peut s'interpréter comme une méthode volume finis (i.e. on peut définir des volumes de contrôle) moyennant certains sacrifices.

La formulation (4.27) comporte, outre le terme temporel, le terme de flux convectif

$$\int_{t_1}^{t_2}\int_{\partial K} f\cdot\vec{n}d\sigma$$

et le terme de flux diffusif

$$\int_{t_1}^{t_2}(K\nabla u)\cdot\vec{n}d\sigma.$$

Dans un premier temps, on va regarder ce qui se passe en une dimension d'espace. Pour simplifier l'exposé, on va supposer le problème scalaire. Le cas des système se traite de manière analogue pour vu que le sous problème non visqueux soit hyperbolique, voir [2].

4.3.1 Dans la méthode des volumes finis en une dimension d'espace.

On se donne un «maillage» : $\{x_j\}_{j\in\mathbb{Z}}$ ou $\{x_{j+1/2}\}_{j\in\mathbb{Z}}$ suivant les cas (pour des problèmes de notation uniquement).

Premier cas : $K_j = [x_{j-1/2}, x_{j+1/2}]$.

Utilisant les résultats du théorème de Kruskov 2.2, on voit que si $t_2 - t_1$ n'est pas trop grand, l'influence des cellules «éloignées» de K_j, c'est à dire au delà de $[x_{j+1/2}, x_{j+3/2}]$ et $[x_{j-1/2}, x_{j+1/2}]$ est nulle. Il suffit donc de déterminer les flux aux interfaces de la cellule,les points $x_{j-1/2}$ et $x_{j+1/2}$. Cela peut se faire au moyen de flux numériques construits à partir de la résolution exacte ou approchée du problème de Riemann : on définit un flux $f_{j+1/2}$. L'expression de certain d'entre eux sera donné plus bas.

On regarde maintenant ;e terme diffusif. Il faut déterminer leur valeur en $x_{j\pm1/2}$. Les seules données dont on dispose sont des valeurs constantes. Si on ne fait rien de spécial, le gradient obtenu est nul !

Une technique employée est alors la suivante. On sait que si la fonction h est régulière, on a

$$\frac{1}{|K_j}\int_{K_j} h = dxh(x_j)$$

où x_j est le centre de gravité de K_j. On peut alors approximer les flux diffusifs au moyen de différences finies, mais la méthode (approximation des termes hyperboliques et diffusifs) n'est pas élégante.

Deuxième cas : interpolation des données.

On emploie pour commencer la méthode de Galerkin : les fonctions tests sont des fonctions continues linéaires par morceaux, on interpole les flux linéairement aussi. Ainsi, (en utilisant la compacité du support des fonctions de base φ_j), et après calcul, en ayant estimé les intégrales en temps $\int_{t_n}^{t_{n+1}} hdt$ par $\Delta th(t_n)$ pour simplifier[4]

$$\int_{x_{j-1}}^{x_{j+1}} \varphi_j u^{n+1}(x)dx - \int_{x_{j-1}}^{x_{j+1}} \varphi_j u^{n+1}(x)dx + f(u_{j+1}^n) - f(u_{j-1}^n) \\ + \frac{K(u_j^n - u_{j-1}^n)dx}{(x_j - x_{j-1})} - \frac{K(u_{j+1}^n - u_j^n)dx}{(x_{j+1} - x_j)} = 0.$$

On a supposé K constant pour simplifier. On effectue ensuite l'opération de mass lumping

$$\int_{x_{j-1}}^{x_{j+1}} \varphi_j u^{n+1}(x)dx \simeq (x_{j+1} - x_{j-1})u_j^{n+1}$$
$$\int_{x_{j-1}}^{x_{j+1}} \varphi_j u^{n}(x)dx \simeq (x_{j+1} - x_{j-1})u_j^{n},$$

on obtient un schéma de la forme

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{(x_{j+1} - x_{j-1})}(\hat{f}_{j+1/2} - \hat{f}_{j-1/2} + \hat{f}_{j+1/2}^v - \hat{f}_{j-1/2}^v)$$

avec

$$\hat{f}_{j+1/2} = \frac{f(u_{j+1}^n) + f(u_j^n)}{2}$$
$$\hat{f}_{j+1/2}^v = \frac{K(u_{j+1}^n - u_j^n)}{(x_{j+1} - x_j)}$$

Quand $\varepsilon = 0$, ce schéma est instable (c'est le schéma centré), aussi on remplace $\hat{f}_{j+1/2}$ par un flux numérique, par exemple un de ceux qui sont définis plus haut.

La encore, il y a une perte de «consistance» entre le flux visqueux et le flux convectif. Mais l'inconsitance n'est qu'une conséquence du manque de stabilité du schéma de Galerkin, pas un problème d'interprétation. On remarque que ce schéma peut cependant s'interpréter comme un schéma volume fini.

[4] on n'aura qu'un schéma d'ordre 1 en temps, mais ce défaut est facile à remédier.

4.3.2 Dans la méthode des volumes finis en plusieurs dimensions d'espace.

On regarde maintenant ce qui se passe en plusieurs dimensions d'espace.

Données constantes par cellules

On se donne un partitionnement de l'espace en cellules deux à deux disjointes et on représente les données par leur valeur moyenne. On note $\{K_i\}_{i=1,n_e}$ les cellules et

$$u_j^n \simeq \frac{\displaystyle\int_{K_j} u(x,t_n)dx}{|K_j|}.$$

La formulation faible conduit à

$$|K_j|\big(u_j^{n+1} - u_j^n\big) + \int_{t_n}^{t_{n+1}} \int_{\partial K_j} f \cdot \vec{n} d\sigma + \int_{t_n}^{t_{n+1}} \int_{\partial K_j} (\varepsilon \cdot \nabla u) \cdot \vec{n} d\sigma.$$

La frontière de K_j étant supposée polygonale, on pose (voir figure 4.3)

$$\partial K_j = \cup_\ell \Gamma_j^\ell$$

On note de plus l'ensemble des indices des cellules adjacentes à K_j par $\mathcal{V}$.[5] On a

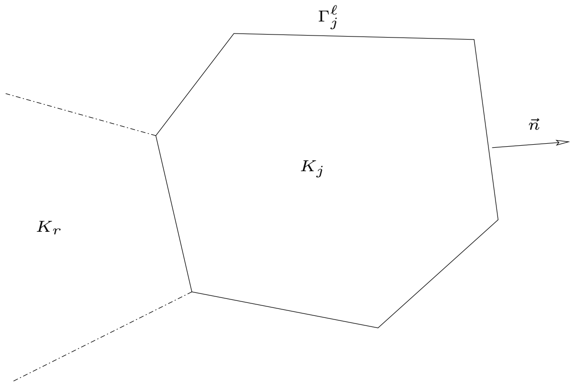


FIG. 4.3 – Eléments géométriques de la cellule K_j.

$$\int_{\partial K_j} f \cdot \vec{n} d\sigma = \sum_\ell \int_{\Gamma_j^\ell} f \cdot \vec{n} d\sigma.$$

Le segment Γ_j^ℓ sépare les cellules K_i et K_ℓ où $\ell \in \mathcal{V}$. On a seulement besoin de connaître le flux normal, ainsi on se ramène à un problème unidimensionnel, et donc ce flux peut être approché par l'un des flux décrits plus haut.

Par contre l'approximation des flux visqueux est plus délicate, du moins si on veut utiliser la même méthode qu'en une dimension d'espace. Il faudrait affecter au centre de gravité, et construire une interpolation. Cela est possible mais délicat.

[5] Pour simplifier, on a supposé sur la figure que le maillage était conforme, mais cette hypothèse n'est pas nécessaire.

Données par noeud.

On se donne un maillage conforme et un espace d'interpolation. Pour fixer les idées, on supposera le maillage triangulaire et une interpolation continue linéaire par triangle. On interpole linéairement les données et les flux, comme en une dimension d'espace. Notons u^h et F^h ces deux interpolations.

On utilise ensuite la méthode de Galerkin, on note φ_i la fonction de base au noeud M_i. On a

$$\int \varphi_i u^h(x,t_{n+1})dx - \int \varphi_i u^h(x,t_n)dx + \int_{t_n}^{t_{n+1}} \int \nabla\varphi_i \cdot F^h dx \\ + \int_{t_n}^{t_{n+1}} \int (\varepsilon \cdot \nabla u^h) \cdot \nabla\varphi_i dx$$

Le terme de diffusion est traité naturellement. Regardons maintenant le terme de convection. On a

$$\int \nabla\varphi_i \cdot F^h = \sum_{T \subset \mathrm{supp}\varphi_i} \int_T \nabla\varphi_i \cdot F^h.$$

Le terme $\nabla\varphi_i$ est constant dans T, et F^h est linéaire et donc

$$\int_T \nabla\varphi_i \cdot F^h = \frac{|T|}{3}\left(\sum_{j\in T} F_j^h\right) \cdot \nabla\varphi_i$$

Si $\vec{n}_i^T$ est le vecteur normal au coté du triangle T opposé au sommet M_i, on sait que

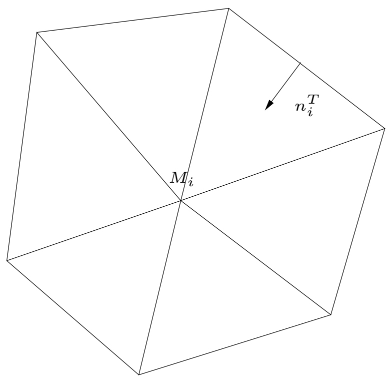


Fig. 4.4 – Eléments géométriques d'un triangle.

$$\nabla\varphi_i = \frac{\vec{n}_i^T}{|T|},$$

et donc

$$\sum_{T\ni i}\sum_{j\in T} F_j^h \cdot \frac{\vec{n}_j^T}{6} = \sum_{T\ni i}\sum_{j\in T} \frac{F_j^h + F_i^h}{2} \cdot \frac{\vec{n}_j^T}{3}$$

car $\sum_{T\ni i} \vec{n}_i^T = 0$, ainsi, en notant $\mathcal{V}(i)$ l'ensemble des voisins de M_i, on a (voir figure 4.5)

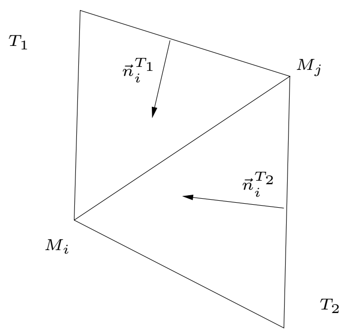


FIG. 4.5 – Eléments géométriques pour un segment.

$$\sum_{T\ni i}\sum_{j\in T} F_j^h \cdot \frac{\vec{n}_j^T}{6} = \sum_{j\in\mathcal{V}(i)} \frac{F_j+F_i}{2}\cdot\{\frac{n_i^{T_1}+n_i^{T_2}}{3}\}$$

Regardons maintenant le terme en temps. On effectue encore l'opération de condensation de masse

$$\int \varphi_i u^h \simeq u_i^h \int \varphi_i$$

et on voit que

$$\int \varphi_i = \sum_{T\ni i} \frac{|T|}{3}.$$

Remarque 4.3.1. *Quand le maillage est composé de triangle équilatéraux, on a*

$$\int \varphi_i u^h = u_i^h \int \varphi_i + O(h^2)$$

où h est le diamètre de T. Si le maillage est uniforme, l'opération de condensation de masse est précise à l'ordre deux.

On fait maintenant un peu de géométrie. On voit facilement à partir de la figure 4.6 que l'aire de $AIGJ$ est $\frac{|T|}{3}$ si bien qu'on peut interpréter $\sum_{T\ni i}\frac{|T|}{3}$ comme l'aire de la cellule qui entoure le noeud M_i formé des surfaces de la forme $AIGJ$. Elle est obtenue en joingnant les milieu des arêtes issues de M_i et les centre de gravité des triangles entourant M_i. Cette construction s'étend facilement en trois dimensions.

De même, on voit que le tier de la somme des normales opposées au sommet A vu de T_1 puis de T_2, les vecteurs $\vec{n}_A^{T_1}$ et $\vec{n}_A^{T_2}$ de la figure 4.6 est égal au vecteur orthogonal à GG' qu'on note $\vec{n}_{AA'}$.

On a donc

$$\sum_{j\in\mathcal{V}(i)} \frac{F_j+F_j}{2}\cdot\vec{n}_{ij} = \sum_{j\in\mathcal{V}(i)} \frac{F_j+F_j}{2}\cdot\frac{\vec{n}_i^{T_1}+\vec{n}_i^{T_2}}{3}.$$

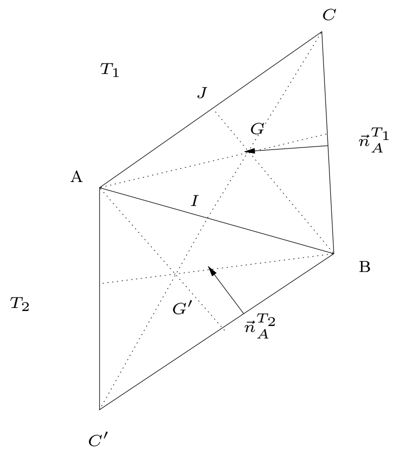


FIG. 4.6 – Géométrie et notation pour le schéma éléments finis/volumes finis.

4.4 Conclusion

Dans ce chapitre, on a évoqué plusieurs méthodes permettant d'approximer un problème hyperbolique scalaire. Dans certains cas, on a aussi donné quelques éléments sur les problème systèmes. Une méthode a été plus particulièrement détaillée : la méthode des volumes finis. Dans cette méthode, on approxime la solution par ses valeurs moyennes dans des «cellules» de contrôle qui recouvrent le domaine de calcul. Si en une dimension d'espace, le choix des cellules de contrôle ne joue pas vraiment de rôle, même dans le cas de problème de convection diffusion, il n'en est pas de même en plusieurs dimensions d'espace.

Dans ce chapitre, on s'est contenté de décrire des schémas d'ordre un en espace car notre but était de montrer quelques difficultés de la méthode, difficultés liées au couplace entre les termes diffusifs, les termes hyperbolique et le terme en temps. En ce sens, notre présentation est loin d'être exhaustive. On pourra consulter [1, 2] pour la construction et l'analyse de schéma d'ordre deux, et [16] pour une autre approche des termes diffusifs dans le cadre de la méthode des volumes finis.

Dans ce qui suit, on va étudier deux classes de méthode. La première est l'extension «naturelle» de la méthode des volumes finis cell–centered, la seconde une extension des schémas de type cell-vertex.

4.5 Notes bibliographiques

Herbin Gallouët, ENO, WENO, MUSCL.

Chapitre 5

La méthode de Galerkin discontinu.

La littérature est très vaste sur ce sujet. Dans tout ce chapitre, on va adopter le point de vue des travaux de Cockburn, Shu et leur collaborateurs [17, 18, 19, 20]. Leur but est de développer une méthode alliant les avantages des méthode éléments finis (souplesse dans le choix du maillage, formulation variationelle et analyse d'erreur) et les avantages des méthodes de type différences finis (contrôle des oscillations).

On part de la formulation faible de

$$\begin{aligned} &\frac{\partial u}{\partial t} + \operatorname{div} f(u) = 0 \quad x \in \Omega, t > 0 \\ &u(x,0) = u_0(x) \end{aligned} \tag{5.1}$$

Pour simplifier, on va supposer $u \in \mathbb{R}$ et $\Omega = \mathbb{R}^d$, $d = 1$ ou 2 suivant les cas. On supposera que le flux est différentiable, et on recherchera des solutions faibles entropiques.

La formulation faible de (5.1) est : quelque soit $\varphi \in C_0^\infty(\mathbb{R}^d \times \mathbb{R}^+)$,

$$\int_{\mathbb{R}^d \times \mathbb{R}^+} \left(\frac{\partial \varphi}{\partial t} u(x,t) + \nabla_x \varphi \cdot f(u(x,t)) \right) dx dt + \int_{\mathbb{R}^d} \varphi(x,0) u_0 dx = 0. \tag{5.2}$$

Dans un premier temps, on considère une triangulation $\mathcal{T}_h$ de $\mathbb{R}^d$. Les éléments K de $\mathcal{T}_h$ sont des éléments géométriques simples, typiquement des triangles, des parallélogrammes, ou plus généralement des polygones possédant un nombre finis de cotés. La triangulation n'est pas nécessairement conforme, mais l'intersection de deux éléments K et K' de $\mathcal{T}_h$ est d'intérieur vide. Enfin, $\mathbb{R}^d$ est recouvert par les éléments de $\mathcal{T}^h$.

Ensuite, on va chercher à approximer la solution faible entropique de (5.1) par u^h, fonction appartenant à V^h qui est définit comme suit.

L'idée de la méthode est de chercher à approcher les solutions de (5.2) par des fonctions qui sont dans chaque élément K est un polynôme de degré fixé k. Dans ce cours, on ne considèrera pas le cas où le degré du polynôme varie en fonction de la régularité attendue de la solution.

On notera $P^k(K)$ l'ensemble des fonctions polynômiales de degré inférieur ou égal à k dans K et

$$V_k^h = \{u \in L^1(\mathbb{R}^d) \cap BV(\mathbb{R}^d), u^h_{|K} \in P^k(K) \text{ quelque soit } K\}.$$

Ainsi les fonctions de V_k^h ne sont pas continues et $V_k^h \not\subset H^1(\mathbb{R}^d)$.

5.1 Le cas monodimensionnel non visqueux : description

Dans ce cas, les éléments K sont des intervalles. On les note $I_j = (x_{j-1/2}, x_{j+1/2})$ et $\mathbb{R} = \cup_j I_j$. L'espace $P^k(I_j)$ est l'espace des polynômes de degré $\leq k$.

Si on suppose la solution suffisament régulière et si φ a pour support I_j, on a

$$\int_{I_j} \varphi(x)\Big(\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x}\Big)dx = \int_{I_j} \varphi(x)\frac{\partial u}{\partial t} + \Big[\varphi(x) f(u(x,t))\Big]_{x_{j-1/2}}^{x_{j+1/2}} - \int_{I_j} \frac{\partial \varphi}{\partial x} f(u)dx$$

Ainsi on va chercher $u^h \in V_k^h$ tel que quelque soit $v^h \in V_k^h$,

$$\int_{I_j} v^h \frac{\partial u^h}{\partial t} dx + \Big[v^h\ f(u(x,t)\Big]_{x_{j-1/2}}^{x_{j+1/2}} - \int_{I_j} \frac{\partial v^h}{\partial x} f(u^h)dx.$$

On voit en particulier que si $k = 0$, on se ramène à une formulation volume fini «centrée cellule». Trois questions se posent :

1. Cette équation a-t-elle un sens ?
2. Si oui, admet–elle une solution unique ?
3. Cela définit–il une «bonne» méthode ?

Cette équation a–t–elle un sens ?

En fait, non. En effet, le terme

$$\Big[v^h\ f(u(x,t)\Big]_{x_{j-1/2}}^{x_{j+1/2}} = v^h(x_{j+1/2}) f(u(x_{j+1/2}, t) - v^h(x_{j-1/2}) f(u(x_{j-1/2}, t)$$

nécessite la définition de $f(u(x_{j\pm 1/2}, t)$. S'agit–il, pour $f(u(x_{j+1/2}, t)$, de $\lim_{x\rightarrow x_{j+1/2}, x<x_{j+1/2}} f(u^h(x,t))$ ou de $\lim_{x\rightarrow x_{j+1/2}, x>x_{j+1/2}} f(u^h(x,t))$? Il faut que la définition du terme de bord $f(u(x_{j+1/2}, t)$ soit intrinsèque.

Une meilleure approche est la suivante. On considère le problème de Cauchy au voisinage de $x_{j+1/2}$ pour des temps assez courts après l'instant t

$$\begin{aligned} &\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0 \qquad x \in I_j \cup I_{j+1} \\ &u(x, s=t) = \begin{cases} u^h_{|I_j}(x) & x < x_{j+1/2} \\ \\ u^h_{|I_{j+1}}(x) & x > x_{j+1/2} \end{cases} \end{aligned}$$

On voit que la condition intiale est dans $L^1 \cap BV \cap L^\infty$. Le théorème de Kruzkov fournit l'unique solution entropique u définie pour $s > t$. Utilisant les relations de Rankine Hugoniot, on voit que le flux $f(u)$ est *continu* en $x_{j+1/2}$, ainsi le choix naturel est $f(u(x_{j+1/2}, t^+) = \lim_{s\rightarrow t, s>t} f(u(x_{j+1/2}, s)$. On pose $f_{j+1/2} = f(u(x_{j+1/2}, t^+)$.

Le problème est que cette méthode est extrèmement complexe : elle demande de résoudre le problème de Riemann généralisé [21].

Aussi, on procède comme suit : on considère un flux numérique $\hat{f}$ (flux de Godunov, de Roe, Lax Friedrichs, etc) et on définit

$$\hat{f}_{j+1/2} = \hat{f}(u^-_{j+1/2}, u^+_{j+1/2})$$

où

$$u^+_{j+1/2} = \lim_{x\to x_{j+1/2}, x>x_{j+1/2}} u^h, \qquad u^-_{j+1/2} = \lim_{x\to x_{j+1/2}, x<x_{j+1/2}} u^h$$

Le schéma est donc : Trouver $u^h \in V_h^k$ tel que quelque soit $v^h \in P^k(I_j)$,

$$\int_{I_j} v^h \frac{\partial u^h}{\partial t} dx + v^h(x^-_{j+1/2})\hat{f}(u^-_{j+1/2}, u^+_{j+1/2}) - v^h(x^+_{j-1/2})\hat{f}(u^-_{j-1/2}, u^+_{j-1/2}) - \int_{I_j} \frac{\partial v^h}{\partial x} f(u^h) dx. \quad (5.3)$$

L'équation (5.3) définit–elle une méthode ?

On veut voir si le problème semi–discret (5.3) fournit une unique solution u^h. Il s'agit d'un problème d'approximation. En effet, (5.3) s'écrit formellement comme

$$M\frac{dU}{dt} = \delta F$$

où $U = (\dots, u_h^j, \dots)$ si u_h^j représente l'ensemble des degrés de liberté nécessaires pour définir u^h dans I_j. La matrice M est la matrice de masse du système et δF le second membre de (5.3).

Il suffit de vérifier que le système est inversible, l'existence de la solution découlera des théorèmes généraux sur les équations différentielles ordinaires. Bien sûr, dans la pratique le système d'ODE sera résolu numériquement au moyen d'une méthode discrète.

Regardons maintenant plus précisément la structure de la matrice de masse M. On écrit

$$u^h_{|I_j} = \sum_{\ell=0}^k a_\ell^j (x - x_j)^\ell$$

avec $x_j = \frac{x_{j+1/2}+x_{j-1/2}}{2}$. Considérons $k+1$ points distincts dans I_j notés $\{x_\ell^j\}_{\ell=0,k}$ distints de $x_{j\pm1.2}$. On définit les fonctions de base φ_ℓ^j aux noeuds x_ℓ^j, et on a

$$u^h_{|I_j} = \sum_{\ell=0}^k u^h_{|I_j}(x_\ell^j)\varphi_\ell^j.$$

On note $u_\ell^j = u^h_{|I_j}(x_\ell^j)$. Le schéma devient

$$\sum_{\ell=0}^k \left(\int_{I_j} \varphi_\ell^j \varphi_{\ell'}^j\right) \frac{du_\ell^j}{dt} + \left[\varphi_{\ell'}^j \hat{f}\right]_{x_{j-1/2}}^{x_{j+1/2}} - \int_{I_j} \frac{\partial \varphi_{\ell'}^j}{\partial x} f(u^h) = 0$$

La matrice de masse est donc une matrice bloc diagonale dont les blocs sont les matrices

$$M_j = \begin{pmatrix} \int_{I_j} (\varphi_1^j)^2 & \dots & \int_{I_j} \varphi_1^j \varphi_k^j \\ \vdots & \vdots & \vdots \\ \int_{I_j} \varphi_k^j \varphi_1^j & \dots & \int_{I_j} (\varphi_k^j)^2 \end{pmatrix}.$$

Le vecteur U est définit par

$$U = \begin{pmatrix} 0 \\ \vdots \\ U_j \\ \vdots \\ 0 \end{pmatrix}$$

avec

$$U_j = \begin{pmatrix} u_0^j \\ \vdots \\ u_k^j \end{pmatrix}.$$

On voit que

$$U_j^T M_j U_j = \int_{I_j} u^2$$

avec

$$u(x) = \sum_\ell u_\ell^j \varphi_\ell^j.$$

Ainsi $U_j^T M_j U_j \geq 0$ et cette quantité est nulle si et seulement $u_\ell^j = 0$ quelque soit ℓ.

On peut considérablement simplifier les calcults en considérant d'autres degrés de libertés : les points de Legendre associés à I_j. Des rappels sur les polynômes de Legendre L_n et les formules de quadratures associées sont effectués dans l'annexe A.

Dans l'intervalle $[-1, 1]$, le n ième polynôme de Legendre noté L_n est de degré n et est défini par la formule de Rodrigue

$$L_n(x) = \frac{(-1)^n}{2^n n!} \frac{d^n}{dx^n}\Big((1 - x^2)^n\Big).$$

Les polynômes de Legendre sont deux à deux orthogonaux, et plus précisément

$$\int_{-1}^{1} L_n(x) L_m(x) = \frac{2}{2n+1} \delta_n^m \tag{5.4}$$

où δ_n^m est le symbole de Kronecker. Enfi, voir l'annexe A, ils vérifient

$$L_m(1) = 1, \qquad L_m(-1) = (-1)^m.$$

On se ramène au cas d'un intervalle $(x_{j-1/2}, x_{j+1/2})$, au moyen du changement de variable $x \mapsto \frac{x - x_j}{\Delta x}$.

La formule (5.4) suggère de décomposer tout polynôme de degré k suivant la base des polynômes de legendre, et donc d'associer les degrés de liberté du problème aux polynôme de Legendre. Plus exactement, on va écrire

$$u^h_{|I_j} = \sum_{\ell=1}^{k} u_j^\ell \widetilde{L}_\ell$$

avec $\widetilde{L}_\ell(x) = L_\ell\left(\dfrac{x - x_i}{\Delta_i x}\right)$ et

$$u_\ell^j = \int_{I_j} u^h \widetilde{L}_\ell dx.$$

Le schéma devient alors

$$\frac{1}{2k+1}\Delta_j x \frac{du_k^j}{dt} - \hat{f}(u_h(x_{j+1/2}) - (-1)^k \hat{f}(u_h(x_{j-1/2})) + \int_{I_j} \frac{\partial \widetilde{L}_j}{\partial x} f(u^h)dx = 0, \tag{5.5}$$

avec

$$u_j^k(0) = \frac{2k+1}{\Delta_j x}\int_{I_j} u_0(x)L_k(x)dx.$$

Dans (5.5), le terme $\int_{I_j} \frac{\partial \widetilde{L}_j}{\partial x} f(u^h)dx$ est calculé avec la formule de quadrature (A.4) d'ordre $2k-1$.

Remarque 5.1.1 (Exemples). *Si on prend $k=2$, on a*

$$\widetilde{L}_0 = 1, \qquad \widetilde{L}_1(x) = 2\frac{x-x_i}{\Delta x}, \qquad \widetilde{L}_2 = \left(\frac{x-x_i}{\Delta x}\right)^2 - \frac{\Delta x^2}{12}.$$

On obtient alors

$$\begin{aligned}
&\frac{du_j^0}{dt} + \frac{1}{\Delta x}(\hat{f}_{j+1/2} - \hat{f}_{j-1/2}) = 0\\
&\frac{du_j^1}{dt} + \frac{1}{2\Delta x}(\hat{f}_{j+1/2} + \hat{f}_{j-1/2}) - \frac{1}{\Delta x^2}\int_{I_j} f(u^h)dx = 0\\
&\frac{du_j^2}{dt} + \frac{1}{6\Delta x}(\hat{f}_{j+1/2} - \hat{f}_{j-1/2}) - \frac{2}{\Delta x^3}\int_{I_j} f(u^h)\widetilde{L}_1 dx = 0
\end{aligned}$$

Le calcul du flux numérique nécessite la connaissance des états aux extrémités $x_{j\pm1/2}$ de I_j. En posant $u^{\pm}_{j+1/2} = u_j^{(0)} + \delta_j^{\pm}$, on obtient

Premier ordre	$k=1$	$\delta_j^{\pm} = 0$
Deuxième ordre	$k=2$	$\delta_j^{+} = 6u_j^1$ $\delta_j^{+} - 6u_j^1$
Troisième ordre	$k=3$	$\delta_j^{+} = 6u_j^1 + 30u_j^2$ $\delta_j^{-} = -6u_j^1 + 30u_j^2$

5.2 Analyse du cas monodimensionnel non visqueux

On suppose $\hat{f}$ lipschitzien.

5.2.1 Analyse de la précision

On veut montrer que le schéma semi–discret portant sur les $u_j^{(0)}$ est d'ordre p. Pour cela, on suppose la solution du problème régulière, et on analyse l'erreur de troncature du schéma. Le résultat est essentiellement une conséquence d'un lemme général sur l'interpolation. Ici, I est un intervalle.

Lemme 5.2.1. *Soit k et r deux entiers, $r \geq k+1$. On considère $\Pi\ : C^r(I) \longrightarrow P^k(I)$ un projecteur continu pour la norme de $C^r(I)$ qui respecte les éléments de $P^k(I) \subset C^r(I)$. Alors, quelque soit $u \in C^r(I)$,*

$$\max_I \left|u - \Pi(u)\right| \leq C(I)|I|^{k+1}||u^{(k)}||_\infty$$

où $C(I)$ est une constante qui ne dépend que de I.

Démonstration. Une bonne référence pour la démonstration, en fait la démonstration d'un résultat plus général, est [22]. □

Partant de

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0$$

et en intégrant sur I_j, on a

$$\frac{du_j^{(0)}}{dt} + \frac{1}{\Delta x_j}\big(f(u(x_{j+1/2}) - f(u(x_{j-1/2}))\big) = 0.$$

Etudions maintenant l'erreur de troncature

$$\varepsilon_j = \frac{du_j^{(0)}}{dt} + \frac{1}{\Delta x_j}\big(\hat{f}_{j+1/2} - \hat{f}_{j-1/2}\big)$$

On a, u^h désignant la projection L^2 de u sur $P^k(I)$,

$$\begin{aligned}
\varepsilon_j &= \frac{du_j^{(0)}}{dt} + \frac{1}{\Delta x_j}\big(\hat{f}_{j+1/2} - \hat{f}_{j-1/2}\big) \\
&= \Bigg(\frac{du_j^{(0)}}{dt} + \frac{1}{\Delta x_j}\big(\hat{f}_{j+1/2} - \hat{f}_{j-1/2}\big) \\
&\qquad - \frac{du_j^{(0)}}{dt} + \frac{1}{\Delta x_j}\big(f(u(x_{j+1/2}) - f(u(x_{j-1/2}))\big)\Bigg) \\
&= \frac{1}{\Delta x_j}\big(\hat{f}_{j+1/2} - f(u(x_{j+1/2})\big) - \frac{1}{\Delta x_j}\big(\hat{f}_{j-1/2} - f(u(x_{j-1/2})\big)
\end{aligned}$$

Ensuite,

$$\begin{aligned}
\hat{f}_{j+1/2} - f(u(x_{j+1/2}) &= \hat{f}\big(u^h_{|I_j}(x_{j+1/2}, u^h_{|I_{j+1}}(x_{j+1/2})\big) - \hat{f}\big(u(x_{j+1/2}, u(x_{j+1/2})\big) \qquad \text{consistance du flux numérique} \\
&= \Big[\hat{f}\big(u^h_{|I_j}(x_{j+1/2}, u^h_{|I_{j+1}}(x_{j+1/2})\big) - \hat{f}\big(\big(u^h_{|I_j}(x_{j+1/2}, u(x_{j+1/2})\big)\Big] \\
&+ \Big[\hat{f}\big(u^h_{|I_j}(x_{j+1/2}, u(x_{j+1/2})\big) - \hat{f}\big(\big(u(x_{j+1/2}, u(x_{j+1/2})\big)\Big]
\end{aligned}$$

Prenant la valeur absolue, on voit que les termes de la dernière somme sont respectivement majorés par $h_1\big|u^h_{|I_j}(x_{j+1/2} - u(x_{j+1/2})\big|$ et $L_2\big|u^h_{|I_j}(x_{j+1/2} - u(x_{j+1/2})\big|$. Ces deux termes sont, grâce au lemme 5.2.1, majorés par $L_1 C(I_j)\Delta x_j^{k+1}$ et $L_2 C(I_j)\Delta x_j^{k+1}$. Le maillage étant uniforme, on voit donc que

$$\varepsilon_j = O(\Delta x_j^k).$$

5.2.2 Analyse en variation totale

On ne s'intéresse qu'à $u_j^{(0)}$, i.e. la valeur moyenne de la solution dans I_j. L'analyse faite dans ce paragraphe conduit à «limiter» les arguments présents dans les flux numériques. Les autres quantités sont évaluées en utilisant les même quantités limitées dans les termes de bord. Les termes intérieurs sont évalués par formule de quadrature, donc sans modification des arguments.

L'évolution de $u_j^{(0)}$ est donnée par

$$\frac{du_j^{(0)}}{dt} = -\frac{1}{\Delta x_j}\Big\{\hat{f}_{j+1/2} - \hat{f}_{j-1/2}\Big\}$$

avec

$$\begin{aligned}\hat{f}_{j+1/2} &= \hat{f}(u_j^{(0)} + \delta_j^+, u_{j+1}^{(0)} + \delta_{j+1}^-)\\ \hat{f}_{j-1/2} &= \hat{f}(u_{j-1}^{(0)} + \delta_{j-1}^+, u_j^{(0)} + \delta_j^-).\end{aligned} \tag{5.6}$$

Une version discrète en temps de ce schéma est donné par

$$\big(u_j^{(0)}\big)^{n+1} = \big(u_j^{(0)}\big)^n - \lambda_j\Big\{\hat{f}_{j+1/2} - \hat{f}_{j-1/2}\Big\} \tag{5.7}$$

avec les flux $\hat{f}_{j\pm 1/2}$ comme dans (5.6) évalué à l'instant t_n. Dans (5.7), on a posé $\lambda_j = \frac{\Delta t}{\Delta x_j}$.

Dans la suite, on va supposer pour simplifier les notations que le pas d'espace est uniforme. Cependant, les modifications à apporter dans le cas non uniforme sont mineures.

On introduit la définition suivante.

Définition 5.2.2. *On dira que le schéma (5.7) est à variation totale décroissante si*

$$TV(u^{n+1}) := \sum_{j\in\mathbb{Z}} \big|\big(u_{j+1}^{(0)}\big)^{n+1} - \big(u_j^{(0)}\big)^{n+1}\big| \leq TV(u^n)$$

Généralement, un schéma numérique n'est pas à varriation totale décroissante ; il faut donc imposer un certain nombre de contraintes. L'intérêt de cette définition réside dans le fait suivant. L'espace $BV(\mathbb{R})$ définit par

$$BV(\mathbb{R}) = \{u \in L^1(\mathbb{R}) \text{ tel que quelque soit } \sup_{\varphi\in C_0^\infty(\mathbb{R}), ||\varphi||_\infty \leq 1} \int_{\mathbb{R}} \varphi'(x)u(x)dx < \infty\}$$

Pour une fonction quelconque de $BV(\mathbb{R})$, on note

$$TV(u) = \sup_{\varphi\in C_0^\infty(\mathbb{R}), ||\varphi||_\infty \leq 1} \int_{\mathbb{R}} \varphi'(x)u(x)dx.$$

s'injecte de manière compacte dans $L^1(\mathbb{R})$. En d'autres termes, si $(u_n)_{n\in\mathbb{N}}$ est une suite de $BV(\mathbb{R})$ telle qu'il existe $C >)$ indépendant de n avec

$$||u||_{BV} = ||u||_{L^1} + TV(u) < C,$$

il existe une sous suite et une fonction $u \in L^1(\mathbb{R})$ telle que cette sous suite converge vers u dans $L^(\mathbb{R})$. Ainsi, si un schéma est à variation totale décroissante, on peut identifier une sous suite convergente dans L^1.

Dans (5.7), on suppose désormais que le flux $\hat{f}$ est lipschitzien (voir §4.2.5. Ceci étant, on a le lemme suivant [1]

Lemme 5.2.3. *Soit $\theta \in \mathbb{R}$. Si le flux $\hat{f}$ est monotone (cf lemme 4.2.1), si les arguments dans $\hat{f}$ dans (5.7) sont tels que que quel que soit $j \in \mathbb{Z}$,*

$$-\theta \leq \frac{\Delta_+ u_j^-}{\Delta_+ u_j^{(0)}} \leq 1, \ -\theta \leq -\frac{\Delta_+ u_j^+}{\Delta_+ u_j^{(0)}} \leq 1$$

avec $\theta \geq 0$ et si λ vérifie

$$\lambda(h_1 + h_2) \leq \frac{1}{1+\theta}$$

alors le schéma est à variation totale décroissante.

Avant de démontrer ce lemme, nous indiquons l'interprétation géométrique de ce résultat. La figure 5.1 représente les divers arguments dans les flux utilisés dans (5.7) et dans le lemme 5.2.3. Le

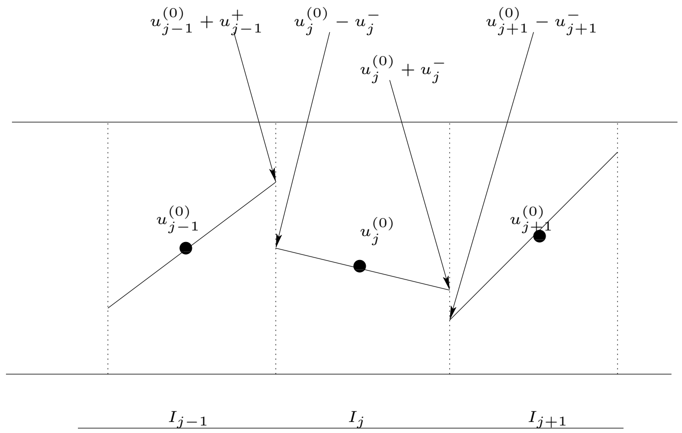


FIG. 5.1 – Représentation géométrique des arguments dans (5.7) et le lemme 5.2.3.

premier terme

$$\frac{\Delta_+ u_j^-}{\Delta_+ u_j^{(0)}} = \frac{u_{j+1}^- - u_j^-}{u_{j+1}^{(0)} - u_j^{(0)}}$$

représente le rapport des accroissement de u^h à l'intersection de deux cellules consécutives. Ce rapport ne doit pas être «trop» grand, et le pas de temps employé dans le schéma est relié au rapport maximal. L'inégalité

$$-\theta \leq -\frac{\Delta_+ u_j^+}{\Delta_+ u_j^{(0)}} \leq 1$$

s'interprète de la même manière.

Pour simplifier les notations, on va omettre l'exposant (0) dans les arguments du flux, étant donné qu'il n'y apas d'ambiguïté. La démonstration du lemme 5.2.3 utilise de manière essentielle le lemme d'Harten

[1] dans lequel on a posé pour une fonction f quelconque $\Delta_+ f_j = f_{j+1} - f_j$ et $\Delta_- f_j = f_j - f_{j-1}$

Lemme 5.2.4 (Lemme d'Harten). *Supposons que le schéma puisse s'écrire sous la forme*

$$u_j^{n+1} = u_j^n - \lambda\Big(C_{j+1/2}(u_{j+1}^n - u_j^n) - D_{j-1/2}(u_j^n - u_{j-1}^n)\Big).$$

Si quelque soit $j \in \mathbb{Z}$ on a

1. $C_{j+1/2} \geq 0$, $D_{j+1/2} \geq 0$,
2. $\lambda\Big(C_{j+1/2} + D_{j+1/2}\Big) \leq 1$

alors le schéma est à variation totale décroissante.

Démonstration. On écrit

$$\begin{aligned}
u_j^{n+1} &= u_j^n - \lambda\Big(C_{j+1/2}(u_{j+1}^n - u_j^n) - D_{j-1/2}(u_j^n - u_{j-1}^n)\Big) \\
u_{j+1}^{n+1} &= u_{j+1}^n - \lambda\Big(C_{j+3/2}(u_{j+2}^n - u_{j+1}^n) - D_{j+1/2}(u_{j+1}^n - u_j^n)\Big)
\end{aligned}$$

d'où

$$\begin{aligned}
u_{j+1}^{n+1} - u_j^{n+1} = &\Big(1 - \lambda\{C_{j+1/2} + D_{j+1/2}\}\Big)(u_{j+1}^n - u_j^n) \\
&+ \lambda C_{j+3/2}(u_{j+2} - u_{j+1}) \\
&+ \lambda D_{j-1/2}(u_j^n - u_{j-1}^n)
\end{aligned}$$

En prenant la valeur absolue, si les hypothèses du lemme son vérifiées, on obtient après sommation sur l'indice j,

$$\begin{aligned}
\sum_j \big|u_{j+1}^{n+1} - u_j^{n+1}\big| \leq &\sum_j \Bigg[\Big(1 - \lambda\{C_{j+1/2} + D_{j+1/2}\}\Big)\big|u_{j+1}^n - u_j^n)\big| \\
&+ \lambda C_{j+3/2}\big|(u_{j+2} - u_{j+1})\big| \\
&+ \lambda D_{j-1/2}\big|(u_j^n - u_{j-1}^n)\big|\Bigg] \\
= &\sum_{j\in\mathbb{Z}} \big|u_{j+1}^n - u_j^n\big|
\end{aligned}$$

d'où le résultat. □

Ceci étant, il suffit d'«exiber» des coefficients $C_{j+1/2}$ et $D_{j-1/2}$ qui permettent de réecrire le schéma (5.7) sous la forme employée dans le lemme d'Harten. Une telle écriture est loin d'être unique, *a priori*.

Démonstration du lemme 5.2.3. Pour simplifier les notations, étant donné qu'on n'a besoin de travailler que sur le second membre de (5.7), on oublie l'exposant n dans les calculs.

Un calcul simple mais laborieux montre que

$$\begin{aligned}
\hat{f}_{j+1/2} - \hat{f}_{j-1/2} &= \hat{f}(u_j^{(0)} + \delta_j^+, u_{j+1}^{(0)} + \delta_j^-) - \hat{f}(u_{j-1}^{(0)} + \delta_{j-1}^+, u_j^{(0)} + \delta_j^-) \\
&= C_{j+1/2}(u_{j+1}^{(0)} - u_j^{(0)}) - D_{j-1/2}(u_j^{(0)} - u_{j-1}^{(0)})
\end{aligned}$$

avec

$$C_{j+1/2} = -H_2\left(1 - \frac{\Delta_+ u_j^-}{\Delta_+ u_j^{(0)}}\right)$$

$$D_{j-1/2} = H_1\left(1 + \frac{\Delta_- u_j^+}{\Delta_- u_j^{(0)}}\right)$$

avec

$$H_1 = \frac{\hat{f}(u_j^{(0)} + \delta_j^+, u_{j+1}^{(0)} + \delta_{j+1}^-) - h(u_{j-1}^{(0)} + \delta_{j-1}^+, u_j^{(0)} + \delta_j^-)}{(u_j^{(0)} + \delta_j^+) - (u_{j-1}^{(0)} + \delta_{j-1}^+)}$$

$$H_2 = \frac{\hat{f}(u_j^{(0)} + \delta_j^+, u_{j+1}^{(0)} + \delta_{j+1}^-) - h(u_j^{(0)} + \delta_j^+, u_j^{(0)} - u_j^-)}{(u_{j+1}^{(0)} - u_{j=1}^-) - (u_j^{(0)} + u_j^-)}$$

Le flux $\hat{f}$ étant monotone, on voit que $C_{j+1/2} \geq 0$ et $D_{j-1/2} \geq 0$. D'autre part, le flux $\hat{f}$ étant lipschitien, $|H_1| \leq h_1$ et $|H_2| \leq h_2$, si bien que la condition

$$\lambda\big(C_{j+1/2} + D_{j+1/2}\big) \leq 1$$

du lemme d'Harten est vérifiée si

$$\lambda\big(h_1 + h_2\big) \leq \frac{1}{1+\theta}$$

$\square$

Remarque 5.2.5. *La condition*

$$\lambda\big(h_1 + h_2\big) \leq \frac{1}{1+\theta}$$

est en fait une condition de type CFL. Pour s'en convaincre, considérons par exemple le flux de Lax Friedrichs,

$$\hat{f} = \frac{1}{2}\Big(f(u) + f(v) - \alpha(u - v)\Big).$$

On a

$$\begin{aligned} \big|\hat{f}(u,v) - \hat{f}(u',v)\big| &= \frac{1}{2}\left|\frac{f(u) - f(u')}{u - u'} - \alpha\right| |u - u'| \\ \big|\hat{f}(u,v) - \hat{f}(u,v')\big| &= \frac{1}{2}\left|\frac{f(v) - f(v')}{v - v'} + \alpha\right| |v - v'| \end{aligned}$$

donc

$$h_1, h_2 \leq \frac{1}{2}\left(\max_{u \neq u'}\left|\frac{f(u) - f(u')}{u - u'}\right| + \alpha\right) \leq \max_{u \in [-||u^0||_\infty, ||u^0||_\infty]} |f'(u)|.$$

puisque si $\alpha \leq \max_{u \in [-||u^0||_\infty, ||u^0||_\infty]} |f'(u)|$, on sait que le schéma est monotone. On voit donc que la condition du lemme 5.2.3 est bien une condition de type CFL.

Comme on dit plus haut, les valeurs de $u_j^{(0)} + \delta_j^\mp$ fournies par (5.3) n'ont aucune raison de vérifier les hypothèses du lemme 5.2.3. On peut s'en convaincre en programmant la méthode ou en regardant de près les exemples 5.1.1. Il faut donc modifier les arguments $\delta_j^\pm$ dans les flux et introduire des limiteurs de pente.

Le limiteur minmod

Quand la solution est régulière, on voit que $\delta_j^\pm = O(h)$. On va modifier $\delta_j^\pm$ de manière à respecter cette propriété. Pour cela on introduit le limiteur minmod généralisé défini par

$$m(a_1, a_2, \ldots, a_m) = \begin{cases} s \min(|a_1|, |a_2|, \ldots, |a_m|) & \text{si } s = \operatorname{sign}(a_1) = \ldots \operatorname{sign}(a_m) \\ 0 & \text{sinon} \end{cases} \tag{5.8}$$

On modifie les arguments dans les flux par

$$\delta_j^\pm \to (\delta_j^\pm)^{mod} = m(\delta_j^\pm, \Delta_+ u_j^{(0)}, \Delta_- u_j^{(0)}).$$

En examinant tous les cas, on voit que

$$-1 \le \frac{\Delta_+ (\delta_j^+)^{mod}}{\Delta_+ u_j^{(0)}} \le 1, \qquad -1 \le -\frac{\Delta_+ (\delta_j^-)^{mod}}{\Delta_+ u_j^{(0)}} \le 1$$

et donc $\theta = 1$. On a résumé sur la Figure 5.2 le comportement typique du limiteur.

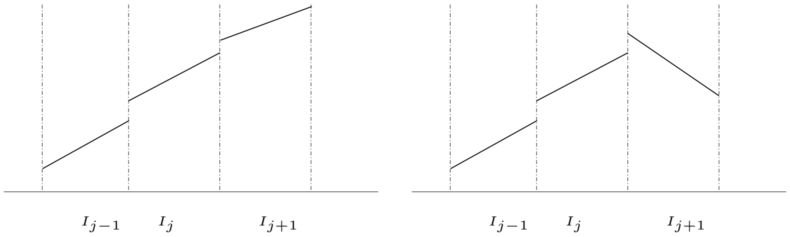


FIG. 5.2 – Comportement typique du limiteur. En (a), $(u_j^\pm)^{mod}$ est essentiellement égal aux valeurs initiales, tandis qu'en (b), elles sont essentiellement nulles. En d'autres terme, en (b), on se ramène à l'ordre un.

Remarquons que la précision (formelle) du schéma n'est pas affectée par cette construction. Dans le cas d'une fonction régulière, on a

$$\begin{aligned} u_j^{(0)} &= \frac{1}{\Delta x_j} \int_{I_j} u^h(x) dx = u(x_j) + O(\Delta x_j^2) \\ u_{j+1}^{(0)} &= u(x_j) + \frac{\partial u}{\partial x}(x_j) + O((\Delta x_j + \Delta x_{j+1})^2) \end{aligned}$$

donc

$$\Delta_\pm u_j^{(0)} = \frac{\partial u}{\partial x}(x_j) \frac{\Delta x_j + \Delta x_{j+1}}{2} + O(h^2)$$

pour un maillage régulier. D'autre part, $u_j^\pm = \frac{1}{2}\frac{\partial u}{\partial x}\Delta x_j + O(h^2)$, ce qui montre que la précision n'est pas affectée tant qu'on n'est pas à un extrémum.

5.2.3 Analyse de la stabilité entropique.

On rappelle que si $\dfrac{\partial u}{\partial t} + \dfrac{\partial f(u)}{\partial x} = 0$ est une loi de conservation, si U est une fonction convexe de u et G est telle que $U'f' = G'$ (dérivées par rapport à u), les solutions entropiques vérifient

$$\frac{\partial U}{\partial t} + \frac{\partial G}{\partial x} \leq 0.$$

On veut regarder le comportement de la méthode semi–discrète définie précédemment

$$\int_{I_j} \frac{\partial u^h}{\partial t} v^h - \int_{I_j} \frac{\partial v^h}{\partial x} f(u^h) + \int_{\partial I} \hat{f} v^h = 0$$

où $\hat{f}$ est un flux numérique, c'est à dire voir à quelle(s) condition(s), la solution u^h satisfait à une inégalité d'entropie discrète, par exemple de la forme

$$\int_{I_j} \frac{\partial U(u^h)}{\partial t} dx + \int_{\partial I_j} \hat{g} dx \leq 0$$

qui est la version variationnelle de l'inégalité d'entropie pour une fonction test $v^h = 1$.

Dans ce cours, on va regarder ce qui se passe pour $U(u) = \frac{u^2}{2}$, c'est à dire le comportement de la norme L^2 de la solution. D'autres commentaires peuvent être trouvés dans [19].

On a, pour tout v^h,

$$\int_{I_j} \frac{\partial u^h}{\partial t} v^h dx - \int_{I_j} f(u^h) \frac{\partial v^h}{\partial x} dx + \hat{f}_{j+1/2} v^-_{j+1/2} - \hat{f}_{j-1/2} v^+_{j-1/2} = 0$$

où $v^{\pm}_{j+1/2} = \lim_{x \to x_{j\pm1/2}, x \in I_j} v^h(x)$.

On choisit ensuite $v^h = u^h$, d'où

$$\frac{1}{2} \int_{I_j} \frac{\partial (u^h)^2}{\partial t} dx - \int_{I_j} f(u^h) \frac{\partial u^h}{\partial x} dx + \hat{f}_{j+1/2} u^-_{j+1/2} - \hat{f}_{j-1/2} u^+_{j-1/2} = 0 \tag{5.9}$$

Le flux d'entropie associé à $U(u) = \frac{u^2}{2}$ vérifie $uf' = G'$[2] donc G est une primitive de uf',

$$G = \int^u v f'(v) dv = u f(u) - \int^u f(v) dv.$$

On pose $g(u) = \displaystyle\int^u f(v) dv$.

On obtient ensuite

$$\frac{1}{2} \int_{I_j} \frac{\partial (u^h)^2}{\partial t} dx + \hat{G}_{j+1/2} - \hat{G}_{j-1/2} + A_j = 0$$

avec

$$\hat{G}_{j+1/2} = \hat{f}_{j+1/2} u^-_{j+1/2} - \int^{u^-_{j+1/2}} f(v) dv$$

[2] dérivées en u.

qui est consistant avec G et

$$A_j = -\hat{f}_{j-1/2}\big(u^+_{j-1/2} - u^-_{j-1/2}\big) + g(u^+_{j-1/2}) - g(u^-_{j-1/2}).$$

La fonction g étant une primitive de f, on a donc, par la formule de la moyenne,

$$g(u^+_{j-1/2}) - g(u^-_{j-1/2}) = \int_{u^-_{j-1/2}}^{u^+_{j-1/2}} f(v)dv = \big(u^-_{j-1/2} - u^+_{j-1/2}\big) f(\xi)$$

avec ξ compris entre $u^-_{j-1/2}$ et $u^+_{j-1/2}$. Finalement, on obtient $A_j = -\big(\hat{f}_{j-1/2} - f(\xi)\big)\ \big(u^-_{j-1/2} - u^+_{j-1/2}\big)$.

On introduit maintenant la définition d'un E-schéma (voir [23])

Définition 5.2.6. *On dit qu'un flux numérique est un E–Flux si quelque soit ξ compris entre u et v, $\big(\hat{f}(u,v) - f(\xi)\big)\ (u-v) \leq 0$*

Cette définition provient de l'écriture suivante du flux de Godunov

$$\hat{f}^{God}(u,v) = \max_{\xi\in[\min(u,v),\max(u,v)]} f(\xi).$$

On peut montrer qu'un E flux , donc un schéma de la forme

$$\frac{du_j}{dt} + \frac{(\hat{f}_{j+1/2} - \hat{f}_{j-1/2})}{\Delta x_j} = 0$$

avec $\hat{f}_{j+1/2} = \hat{f}(u_{j+1}, u_j)$ vérifie toujours toutes les inégalités d'entropies, autrement dit un E-flux définit un schéma qui est toujours plus dissipatif que le schéma de Godunov. En particulier, il est facile de voir que le flux de Lax Friedrichs est un E–flux. Ce n'est pas le cas du schéma de Roe. On peut cependant le modifier pour qu'il en soit ainsi.

Notons qu'ici le contexte n'est pas tout à fait identique puisque le schéma est différent. On a cependant

Proposition 5.2.7. *Si $\hat{f}$ est un E–flux, le schéma*

$$\int_{I_j} \frac{\partial u^h}{\partial t} v^h - \int_{I_j} \frac{\partial v^h}{\partial x} f(u^h) + \int_{\partial I} \hat{f} v^h = 0$$

vérifie l'inégalité suivante

$$\frac{1}{2}\int_{I_j} \frac{\partial (u^h)^2}{\partial t} dx + \hat{G}_{j+1/2} - \hat{G}_{j-1/2} \leq 0$$

Démonstration. Par construction, on a $A_j \geq 0$. □

Remarque 5.2.8.
1. *On a établi une inégalité d'entropie uniquement pour l'entropie quadratique. Ceci provient du terme en*

$$-\int_{I_j} \frac{\partial v^h}{\partial x} f(u^h)$$

de la formulation. Dans le cas d'un schéma volume fini, ce terme est nul.

2. *Par contre, l'inégalité a été obtenue sans utiliser explicitement la structure des arguments dans $\hat{f}$. On voit donc qu'on peut répéter le même calcul quand on emploie des arguments limités. Ceci est un peu miraculeux.*

5.3 Discretisation en temps.

DIRE que si on a une semi norme pour le schéma de base, la semi norme est respectée : TV, L infini Jusqu'à présent, on ne s'est pas préocupé de la discrétisation en temps. Un des intérêt de la méthode de Galerkin Discontinue, comme on l'a indiqué plus haut, est que la matrice de masse du schéma peut être rendue *diagonale*. Ceci suggère donc des schémas de type Runge–Kutta.

On peut construire des schémas de Runge–Kutta préservant certaines propriétés, en particulier celle de la variation totale. Ceci a été décrit dans [24, 25].

Considérons le problème

$$\frac{du}{dt} = L(u)$$

On considère une approximation $\mathcal{L}$ de L, en on considère la classe de schémas définis par

$$\begin{aligned} u^{(i)} &= \sum_{l=0}^{i-1} \left[\alpha_{il} u^{(l)} + \beta_{il} \Delta t \mathcal{L}(u^{(l)})\right] \\ u^{(0)} &= u^n, \qquad u^{(r)} = u^{n+1} \end{aligned} \qquad i = 1, \dots, r \tag{5.10}$$

On suppose que le schéma

$$w = u + \mathcal{L}(u)$$

est à variation totale décroissssante, i.e.

$$TV(w) \leq TV(u)$$

si $\Delta t \leq \Delta t_0$. A quelles condition sur les coefficients β_{il}, α_{il} et le pas de temps $\Delta t \leq CFL\ \Delta t_0$, le schéma (5.10) sera-t-il TVD ?

On a des réponses partielles à ce problème

1. Ordre 1 : il s'agit du schéma d'Euler qu'on suppose TVD sous la condition $\Delta t \leq \Delta t_0$,
$$u^{n+1} = u^n + \Delta \mathcal{L}(u^n).$$
2. Ordre 2. On prend
$$\begin{aligned} u^{(1)} &= u^n + \Delta t \mathcal{L}(u^n) \\ u^{(2)} &= \frac{1}{2} u^n + \frac{1}{2}\left[u^1 + \Delta t \mathcal{L}(u^1)\right] \\ u^{n+1} &= u^{(2)} \end{aligned}$$
et on a $CFL = 1$.
3. Ordre 3. On prend
$$\begin{aligned} u^{(1)} &= u^n + \Delta t \mathcal{L}(u^n) \\ u^{(2)} &= \frac{3}{4} u^n + \frac{1}{4}\left[u^{(1)} + \Delta t \mathcal{L}(u^{(1)})\right] \\ u^{(3)} &= \frac{1}{3} u^n + \frac{2}{3}\left[u^{(2)} + \Delta t \mathcal{L}(u^{(2)}\right] \\ u^{n+1} &= u^{(3)} \end{aligned}$$
avec $CFL = 1$ encore

Une étude systématique est effectuée dans [26]. La forme particulière de ces schéma de Runge–Kutta justifie le choix d'approximation temporelle effectué dans le lemme 5.2.3.

Il est facile de voir que si le schéma d'Euler vérifie pour une norme ou semi–norme $||\ .\ ||$ (typiquement $||\ ||_p$ ou la semi norme BV) que si le schéma d'Euler

$$u \mapsto v = u + \Delta t \mathcal{L}(u)$$

vérifie, pour $\Delta t \leq \Delta t_0$,

$$||v|| \leq ||u||$$

alors les schéma de Runge–Kutta décrits plus haut vérifient pour $\Delta t \leq \Delta t_0$ (CFL=1) la propriété

$$||u^{n+1}|| \leq ||u^n||.$$

Cette remarque justifie leur importance.

Remarque 5.3.1. *Il faut cependant noter que d'autres schémas, par exemple le schéma de Crank–Nicholson on des propriétés similaires [27]. Leur défaut est d'être plus complexe.*

Il faut aussi remarquer que ce type de schéma est bien adapté au caractère hyperbolique des équations : le pas de temps, via la condition CFL, est lié à la plus grande vitesse des ondes dans le problème. En revanche, quand on a des termes parabolique, comme dans la suite de ce chapitre, le pas de temps est régit par l'intermédiaire des termes hyperbolique et des termes visqueux. Ces derniers induisent des contraintes proportionnelles à ν/h^2, h étant une taille caractéristique des mailles et ν une valeur caractéristique du coefficient de diffusion : les pas de temps sont donc susceptibles de devenir très petit, auquel cas un schéma de type Runge–Kutta n'est pas forcément le plus adapté.

5.4 Le cas monodimensionnel visqueux : description

Ce paragraphe est tiré de [20]. On considère le problème

$$\begin{aligned} &\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} - \frac{\partial}{\partial x}\big(a(u)\frac{\partial u}{\partial x}\big) = 0 \quad t > 0 \\ &u(x, t = 0) = u_0(x). \end{aligned} \tag{5.11}$$

Afin d'éviter le problème des conditions aux limites, on suppose que la condition initiale est périodique de période 1. On travaillera dans $[0, 1]$ si nécessaire. On suppose aussi que le problème (5.11) est non dégénéré, i.e. il existe $\alpha > 0$ tel que $a(u) > \alpha > 0$ quelque soit $u \in \mathbb{R}$. On suppose que le flux et a sont suffisament régulier, et on sait donc dans ce cas qu'il existe une et une seule solution dans H^1 au problème (5.11), voir par exemple [1].

On veut approcher (5.11) au moyen de la méthode de Galerkin discontinue, i.e avec une représentation de la solution numérique au moyen de fonctions polynômiales par morceaux. Or ces fonctions ne sont pas dans H^1, il faut donc faire quelque chose.

On commence par réécrire le problème en posant $q = \sqrt{a(u)}\dfrac{\partial u}{\partial x}$,

$$\begin{aligned} &\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\Big(f(u) - \sqrt{a(u)}q\Big) = 0 \\ &q - \frac{\partial g(u)}{\partial x} = 0 \\ &u(x, t = 0) = u_0(x) \end{aligned} \tag{5.12}$$

avec $g(u) = \int^u \sqrt{a(s)}ds$. On pose

$$w = \begin{pmatrix} u \\ q \end{pmatrix}, \qquad s = \begin{pmatrix} 0 \\ q \end{pmatrix}, \qquad h(w) = \begin{pmatrix} f(u) - \sqrt{a}q \\ -g(u) \end{pmatrix}.$$

On nommera w_u, h_u (resp. w_q, h_q) les composantes en u (resp. en q) de w et h Utilisant ces variables, le problème devient

$$\frac{\partial w}{\partial t} + \frac{\partial h(w)}{\partial x} = S, \qquad w(t=0) \text{ donné.} \tag{5.13}$$

On va commencer par donner la formulation Galerkin discontinue dite «LDG» (pour Local Discontinuous Galerkin). Comme précédemment, $I_j = [x_{j-1/2}, x_{j+1/2}]$ et $V_h^k = \{v \in L^1, v_{|I_j} \in P^k(I_j)\}$.

Formellement, on multiplie (5.13) par $v - (v_u, v_q)$ et on intègre par partie,

$$\begin{aligned}
&\int_{I_j} \frac{\partial u}{\partial t} v_u(x)dx - \int_{I_j} h_u(w(x,t)) \frac{\partial v_u}{\partial x} dx \\
&\qquad + h_u(w(x_{j+1/2},t))v_u(x_{j+1/2}^-) - h_u(w(x_{j-1/2},t)v_u(x_{j-1/2}^+) = 0 \\
&\int_{I_j} qv_q - \int_{I_j} h_q(w)\frac{\partial v_q}{\partial u} dx \\
&\qquad + h_q(w(x_{j+1/2},t))v_q(x_{j+1/2}^-) - h_q(w(x_{j-1/2},t)v_q(x_{j-1/2}^+) = 0
\end{aligned} \tag{5.14}$$

Ensuite, on remplace les flux par des flux numériques

$$\hat{h}_{j+1/2} = \left(\hat{h}_u(w_{j+1/2}^h, w_{j-1/2}^h), \hat{h}_q(w_{j+1/2}^h, w_{j-1/2}^h)\right) = \left(\hat{h}_{u,j+1/2}, \hat{h}_{q,j+1/2}\right)$$

L'approximation du problème consiste à trouver $(u^h, q^h) \in V_h^k \times V_h^k$ telle que quelque soit $(v_u^h, v_q^h) \in V_h^k \times V_h^k$,

$$\begin{aligned}
&\int_{I_j} \frac{\partial u^h}{\partial t} v_u^h(x)dx - \int_{I_j} h_u(w^h(x,t)) \frac{\partial v_u^h}{\partial x} dx \\
&\qquad + \hat{h}_{u,j+1/2}v_u(x_{j+1/2}^-) - h_{u,j-1/2}v_u(x_{j-1/2}^+) = 0 \\
&\int_{I_j} qv_q - \int_{I_j} h_q(w)\frac{\partial v_q}{\partial u} dx \\
&\qquad + h_{q,j+1/2}v_q(x_{j+1/2}^-) - h_q(w(x_{j-1/2},t)v_q(x_{j-1/2}^+) = 0 \\
&\int_{I_j} u^h(x,0)v_u^h dx = \int_{I_j} u_0 v_u^h dx
\end{aligned} \tag{5.15}$$

On pose $[p] = p^+ - p^-$, $\overline{p} = \frac{1}{2}(p^+ + p^-)$, $p^\pm = p(x_{j+1/2}^\pm)$. Pour avoir la consistance avec la méthode quand le terme visqueux n'apparaît pas, on choisit $\hat{h}_{j+1/2} = \hat{h}(w_h(x_{j+1/2}^-, t), w_h(x_{j-1/2}^+, t)$ tel que :

1. $\hat{h}$ soit localement lipschitzien et consistant avec h,
2. le choix de $\hat{h}$ permette de résoudre *localement* q_h en fonction de u_h,
3. le choix de $\hat{h}$ permette de se ramener à un E–flux quand $a(u) = 0$,

4. ce choix assure la stabilité L^2 de la méthode.

Par analogie avec la décomposition des flux continus, on prend

$$\hat{h}(w^-, w^+) = \hat{h}_{\text{conv}}(w^-, w^+) + \hat{h}_{\text{diff}}(w^-, w^+)$$

avec

$$\hat{h}_{\text{conv}}(w^-, w^+) = \begin{pmatrix} \hat{f}(u^-, u^+) \\ 0 \end{pmatrix}$$

où $\hat{f}$ est un E–flux numérique lipschitzien et consistant, et

$$\hat{h}_{\text{diff}}(w^-, w^+) = \begin{pmatrix} -\dfrac{[g]}{[u]}\overline{q} \\ -\overline{g(u)} \end{pmatrix} - \mathbb{C}_{\text{diff}}[w]$$

et

$$\mathbb{C}_{\text{diff}} = \begin{pmatrix} 0 & c_{12} \\ -c_{12} & 0 \end{pmatrix}.$$

On suppose que la fonction c_{12} dépend de w^+ et w^-, est localement lipschitzienne, et s'annule quand $a = 0$.

Vérifions la consistance. Pour le flux convectif, c'est vrai par construction. Pour le flux visqueux, on a

$$\frac{[g]}{[u]} = \frac{\displaystyle\int_{u^-}^{u^+} \sqrt{a(s)}ds}{u^+ - u^-}$$

donc

$$\lim_{w^+ - w^- \to 0} \frac{[g]}{[u]} = \sqrt{a(u)}.$$

De même, $\lim_{w^+-w^-\to 0} \overline{q} = q$, $\lim_{w^+-w^-\to 0} \overline{g(u)} = g(u)$ et $\lim_{w^+-w^-\to 0} \mathbb{C}_{\text{diff}}[w] = 0$. La consistance est donc évidente, ainsi que le caractère lipschitzien.

Le seul point délicat est de voir si on peut effectivement calculer q_h en fonction de u_h, et de manière explicite.

On part de

$$\int_{I_j} q^h v_q^h = \int_{I_j} h_q(w^h) \frac{\partial v_q^h}{\partial x} dx - \hat{h}_{q,j+1/2} v_q^h(x_{j+1/2}^-) + \hat{h}_{q,j-1/2} v_q^h(x_{j-1/2}^+).$$

Tout d'abord,

$$\int_{I_j} h_q(w^h) \frac{\partial v_q^h}{\partial x} dx = -\int_{I_j} g(u^h) \frac{\partial v_q^h}{\partial x} dx, \tag{5.16}$$

donc q^h ne joue aucun rôle. Puis, tous calculs faits,

$$\begin{aligned} \hat{h}_{q,j+1/2} &= -\overline{g(u)}_{j+1/2} + c_{12}^{j+1/2}(u_{j+1/2}^+ - u_{j+1/2}^-) \\ \hat{h}_{q,j-1/2} &= -\overline{g(u)}_{j-1/2} + c_{12}^{j-1/2}(u_{j-1/2}^+ - u_{j-1/2}^-) \end{aligned} \tag{5.17}$$

et q^h ne joue encore aucun rôle. Finalement, on a la relation suivante

$$\begin{aligned}\int_{I_j} q_h v_q^h dx = \int_{I_j} g(u^h)\frac{\partial v_q^h}{\partial x} - \Big\{ -\overline{g(u)}_{j+1/2} + c_{12}^{j+1/2}(u_{j+1/2}^+ - u_{j+1/2}^-)\Big\} v_q^h(x_{j+1/2}^-) \\ + \Big\{ -\overline{g(u)}_{j-1/2} + c_{12}^{j-1/2}(u_{j-1/2}^+ - u_{j-1/2}^-)\Big\} v_q^h(x_{j-1/2}^+)\end{aligned} \tag{5.18}$$

Le second membre de (5.18) ne fait pas intervenir q^h. Il y a $k+1$ relations de la forme (5.18) puisque la dimension de V_h^k est $k+1$. Ce système couple l'intervale I_j avec ses voisins uniquement par le terme de flux, par contre la résolution du système est découplée. On obtient donc un système linéaire dont la matrice de masse est définie positive. Enfin, si on choisit de développer q^h en polynômes de Legendre, cette matrice de masse est diagonale, l'inversion du système est donc immédiate.

5.5 Cas monodimensionel visqueux : analyse d'erreur et de stabilité

5.5.1 Analyse de stabilité

Soit $T > 0$. La solution $w = (u, q)$ de (5.11) vérifie aussi

$$\mathcal{B}(w, v) = 0 \qquad \text{quelque soit } v \text{ régulière} \tag{5.19}$$

où

$$\begin{aligned}\mathcal{B}(w, v) = \int_0^T \int_0^1 \frac{\partial u}{\partial t}(x,t) v_u(x,t) dx dt + \int_0^T \int_0^1 q(x,t) v_q(x,t) dx dt \\ - \int_0^T \int_0^1 h(w(x,t))^t \frac{v}{x}(x,t) dx dt.\end{aligned} \tag{5.20}$$

On utilise le fait que –rappelons que $g(u) = \int^u \sqrt{a(s)} ds$–

$$\begin{aligned}h(w(x,t))^t \frac{v}{x}(x,t) &= \frac{\partial u}{\partial x}\big(f(u) - \sqrt{a(u)}q)\big) - \frac{\partial q}{\partial x} g(u) \\ &= -\frac{\partial g(u) q}{\partial x} + \frac{\partial \displaystyle\int^u f(s) ds}{\partial x}\end{aligned}$$

pour voir que

$$\mathcal{B}(w, w) = \frac{1}{2}\int_0^1 u^2(x, T) + \int_0^T \int_0^1 q^2(x,t) dx dt - \frac{1}{2}\int_0^1 u_0^2(x) dx \tag{5.21}$$

Puisque $\mathcal{B}(w, w) = 0$, on en déduit que

$$\frac{1}{2}\int_0^1 u^2(x, T) dx + \int_0^T \int_0^1 q^2(x,t) dx dt = \frac{1}{2}\int_0^1 u_0^2(x) dx. \tag{5.22}$$

C'est ce qu'on veut imiter au niveau discret.

Dans le cas discret, après intégration en temps, on peut réecrire le schéma comme

$$\mathcal{B}_h(w_h, v_h) = 0 \qquad \forall v_h(t) \in V_h^k \times V_h^k, \qquad \forall t \in [0, T]. \tag{5.23}$$

où

$$
\begin{aligned}
\mathcal{B}_h(w_h, v_h) = & \int_0^T \int_0^1 \frac{\partial u_h}{\partial t}(x,t)(v_h)_u(x,t)dxdt + \int_0^T \int_0^1 q_h(x,t)(v_h)_q(x,t)dxdt \\
& - \int_0^T \sum_{1\leq j\leq N} \hat{h}^t_{j+1/2}[v_h(t)]_{j+1/2}\, dt \\
& - \int_0^T \sum_{1\leq j\leq N} \int_{I_j} h(w(x,t))^t \frac{\partial v_h}{\partial x}(x,t)dxdt
\end{aligned}
\tag{5.24}
$$

Nous avons le

Lemme 5.5.1.

$$
\mathcal{B}(w_h, w_h) = \frac{1}{2}\int_0^1 u_h^2(x,T) + \int_0^T \int_0^1 q_h^2(x,t)dxdt + \Theta([w_h]) - \frac{1}{2}\int_0^1 u_h^2(x,0)dx
$$

où

$$
\Theta([w_h]) = \int_0^T \sum_{1\leq j\leq N} \{[w_h(t)]^t \mathbb{C}[w_h(t)]\}_{j+1/2}dt
$$

où $\mathbb{C} + \mathbb{C}^t \geq 0$ *si* $\hat{f}$ *est un E–flux, ainsi* $\Theta([w_h]) \geq 0$.

Démonstration. On a

$$
\mathcal{B}(w_h, w_h) = \frac{1}{2}\int_0^T u_h^2(x,T) + \int_0^T \int_0^1 q_h^2(x,t)dxdt + \int_0^T \Theta(t)dt - \frac{1}{2}\int_0^1 u_h^2(x,0)dx)
$$

avec

$$
\Theta(t) = \sum_{j=1}^N \left\{ \hat{h}(w_h)^t_{j+1/2}[w_h(t)]_{j+1/2} + \int_{I_j} h(w_h(x,t))^t \frac{\partial w_h}{\partial x}(x,t)dx \right\}.
$$

Par définition de h, et puisque g est une primitive de $\sqrt{a(\ .\ )}$, on a

$$
\begin{aligned}
h(w_h(x,t))^t \frac{\partial w_h}{\partial x}(x,t) &= \big(f(u_h) - \sqrt{a(u_h)}q_h\big)\frac{\partial u_h}{\partial x} - g(u_h)\frac{\partial q_h}{\partial x} \\
&= \frac{\partial}{\partial x}\bigg(\int^{u_h} f(s)ds - g(u_h)q_h \bigg) \\
&\equiv \frac{\partial}{\partial x} H(w_h(x,t)).
\end{aligned}
$$

On obtient donc

$$
\int_{I_j} h(w_h(x,t))^t \frac{\partial w_h}{\partial x}(x,t)dx = H(w_h(x^-_{j+1/2}) - H(w_h(x^+_{j-1/2}),
$$

et ainsi

$$
\Theta(t) = \sum_{j=1}^N \Big\{ \big[H(w_h)\big]_{j+1/2} - \hat{h}(w_h)^t_{j+1/2}\big[w_h\big]_{j+1/2}
$$

puisque $[a(x)]_{j+1/2} = a(x^+_{j+1/2}) - a(x^-_{j+1/2})$ (ceci explique le changement de signe), et finalement

$$\Theta(t) = \sum_{j=1}^{N} \Big\{ \big[H(w_h)\big] - \hat{h}(w_h)^t \big[w_h(t)\big] \Big\}_{j+1/2}$$

en utilisant le fait qu'il n'y a qu'un flux numérique par interface.

On utilise maintenant le fait que pour deux fonctions f et g quelconque, on a

$$[fg]_{j+1/2} = \overline{f}_{j+1/2}\, [g]_{j+1/2} + \overline{g}_{j+1/2}\, [f]_{j+1/2},$$

on obtient en détaillant

$$\begin{aligned}[H(w_h)]_{j+1/2} = {} & [\int_h^u f(s)ds] - [g(u_h)q_h] \\ & [\int_h^u f(s)ds] - [g(u_h)]\overline{q_h} - [q_h]\, \overline{g(u_h)}.\end{aligned}$$

Ainsi, on obtient

$$\begin{aligned}\Theta(t) = {} & \sum_{j=1}^{N} \Big\{ [\int_h^u f(s)ds] - [g(u_h)]\overline{q_h} - [u_h]\hat{h}_u \Big\}_{j+1/2} \\ & + \sum_{j=1}^{N} \Big\{ -[q_h]\, \overline{g(u_h)} - [q_h]\hat{h}_q \Big\}_{j+1/2}\end{aligned} \tag{5.25}$$

Puisque

$$\begin{aligned}\hat{h}_u &= \hat{f} - \frac{[g]}{[u]}\overline{q} - c_{12}[q] \\ \hat{h}_q &= -\overline{g} + c_{12}[u],\end{aligned}$$

on voit que

$$\Big\{ [\int_h^u f(s)ds] - [g(u_h)]\overline{q_h} - [u_h]\hat{h}_u \Big\}_{j+1/2} = \int_{u_h(x^-_{j+1/2})}^{u_h(x^+_{j+1/2})} \big[f(s) - \hat{f}\big(u_h(x^-_{j+1/2}), u_h(x^+_{j+1/2})\big)\big]\, ds - \Big\{ c_{12}[q] \Big\}_{j+1/2}$$

et

$$\Big\{ [q_h]\, \overline{g(u_h)} + [q_h]\hat{h}_q \Big\}_{j+1/2} = \Big\{ c_{12}[u] \Big\}_{j+1/2}$$

En sommant, on obtient

$$\Theta(t) = \sum_{j=1}^{N} \Big\{ [w_h]^t \mathbb{C}[w] \Big\}_{j+1/2}$$

avec

$$\mathbb{C}(u^+, u^-) = \begin{pmatrix} \int_{u_-}^{u^+} \big[f(s) - \hat{f}\big(u^-, u^+\big)\big] ds & -c_{12} \\ c_{12} & 0 \end{pmatrix}$$

qui est une matrice positive ($\mathbb{C} + \mathbb{C}^t \geq 0$) si $\hat{f}$ est un E–flux. $\square$

Remarque 5.5.2. *En fait, la forme de la dissipation $\mathbb{C}_{diss}$ et le choix du type de flux (E–flux) sont dictés par* (5.25).

5.5.2 Analyse d'erreur

On fait l'analyse pour le cas $f(u) = c\ u$ avec c constant, et $a(u) = a$ constant. On rappelle la semi norme de Sobolev

$$|u|_{H^m(J)} = \sqrt{\int_J \left(\frac{\partial^m u}{\partial x^m}\right)^2 dx}.$$

En suivant une technique classique et [20] en particulier, on commence par estimer l'erreur entre les solutions $w_l = (u_l, q_l)$ pour $l = 1, 2$, solutions des problèmes

$$\begin{array}{ll} \dfrac{\partial u_l}{\partial t} + \dfrac{\partial}{\partial x}\big(f(u_l) - \sqrt{a(u_l)}q_l\big) = 0 & \text{in } [0,T]\times[0,1] \\ q_l - \dfrac{\partial g(u_l)}{\partial x} & \text{in } [0,T]\times[0,1] \\ u_l(x, t=0) = u_{0,l}(x) & \text{in } [0,1] \end{array} \tag{5.26}$$

Ceci donnera un guide pour l'estimation d'erreur.

D'après (5.20), les fonctions u_l vérifient

$$\mathcal{B}(u_l, v) = 0 \qquad \text{quelque soit } v \text{ régulière quelque soit } t \in [0,T].$$

La forme $\mathcal{B}$ étant bilinéaire, on a, en posant $e = w_1 - w_2$,

$$\mathcal{B}(e, v) = 0$$

C'est là qu'on utilise la linéarité du flux f. D'après (5.21), on en déduit

$$\mathcal{B}(e,e) = \frac{1}{2}\int_0^T e_u^2(x,T)dx - \frac{1}{2}\int_0^T e_u^2(x,0)dx + \int_0^T\int_0^1 e_q^2(x,t)dxdt = 0.$$

Puisque $e_u(x,0) = u_{1,0}(x) - u_{2,0}(x)$, on en déduit

$$\frac{1}{2}\int_0^T e_u^2(x,T)dx \le \frac{1}{2}\int_0^T e_u^2(x,T)dx + \int_0^T\int_0^1 e_q^2(x,t)dxdt = \frac{1}{2}\int_0^T \big(u_{1,0}(x) - u_{2,0}(x)\big)^2 dx.$$

C'est la démarche que l'on veut imiter au niveau discret.

Dans le cas discret, utilisant (5.23), on a

$$\mathcal{B}_h(w_h, v_h) = 0 \qquad \text{quelque soit } v_h(t) \in V_h \times V_h,\ t \in [0,T]$$

Puis, en utilisant (5.14) et la consistance du flux, on a aussi

$$\mathcal{B}_h(w, v_h) = 0 \qquad \text{quelque soit } v_h(t) \in V_h \times V_h,\ t \in [0,T]$$

et on en déduit l'équation portant sur $e = w - w_h$

$$\mathcal{B}_h(e, v_h) = 0 \qquad \text{quelque soit } v_h(t) \in V_h \times V_h,\ t \in [0,T]. \tag{5.27}$$

On ne peut pas répéter l'argument «continu» car $e \notin V_h$. Notons que dans $\mathcal{B}_h$, le second argument est nécessairement dans V_h. Cependant, on va considérer $\Pi_h(e)$ où Π_h désigne la projection L^2 sur $V_h \times V_h$. Elle est définie par $\Pi_h(e) = (P_h(e_u(t)), P_h(e_q(t)))$, où $P_h(p)$ est défini par

$$\int_0^1 \Big(P_h(p) - p\Big) v_h(x)dx = 0 \qquad \text{quelque soit } v_h \in V_h.$$

On remarque que la condition initiale du schéma $u_h(t=0)$ vérifie $u_h(t=0) = P_h(u_0)$ par construction.

Puis, puisque $\Pi_h(w_h) = w_h$, on a

$$\mathcal{B}_h(\Pi_h(e), \Pi_h(e)) = \mathcal{B}_h(\Pi_h(e) - e, \Pi_h(e)) = \mathcal{B}_h(\Pi_h(w) - w, \Pi_h(e)),$$

et en utilisant le lemme 5.5.1, on a

$$\frac{1}{2}\int_0^1 \big[P_h(e_u(T))\big]^2 dx + \int_0^T \int_0^1 \big[P_h(e_q(t))\big]^2 dxdt + \Theta(P_h(e)) = \mathcal{B}_h(\Pi_h(w) - w, \Pi_h(e)). \tag{5.28}$$

La seule chose à faire est d'estimer le second membre, $\mathcal{B}_h(\Pi_h(w) - w, \Pi_h(e))$. Ceci se fait au moyen

1. d'erreur d'interpolation,
2. de l'inégalité de Young. Si $(\, . \, , \, . \,)$ est un produit scalaire et $\lambda > 0$, on a
$$\big|(a,b)\big| \leq \frac{1}{2\lambda}||a|| + \frac{\lambda}{2}||b||.$$
On va employer cela pour $(f,g) = \displaystyle\int_0^1 f(x)g(x)dx$.
3. du lemme de Gronwald [3] : si
$$f(x) \leq \int_0^x f(s)ds + A$$
alors
$$f(x) \leq A\,(\exp(x) - 1).$$

Le résultat est donné par le lemme suivant

Lemme 5.5.3. *On a*

$$\begin{aligned}\mathcal{B}_h(w, \Pi_h(e)) \leq{}& \frac{1}{2}\Theta(\Pi_h(e)) + \frac{1}{2}\int_0^T \int_0^1 [\Pi_h(e)]^2 dxdt \\ &+ \frac{1}{2}\Delta x^{2k} \int_0^T C_1(s)ds + \Delta x^k \int_0^T C_2(s)\left\{\sqrt{\int_0^1 [P_h(e_u(s))]^2 dx}\right\}ds\end{aligned}$$

avec

$$\begin{aligned}C_1(t) &= 2c_k^2\left\{\left(\frac{|c| + c_{11}}{c_{11}}\Delta x + 4|c_{12}|^2 d_k^2\right)|u(t)|^2_{k+1} + 4ad_k^2(\Delta x)^{2(\hat{k}-k)}|u(t)|^2_{\hat{k}+1}\right\} \\ C_2(t) &= \sqrt{8}c_k d_k\big\{\sqrt{a}|c_{12}||u(t)|_{k+2} + a(\Delta x)^{\hat{k}-k)}|u(t)|_{\hat{k}+2}\big\}\end{aligned}$$

dans lequel $|u(t)|_m$ *désigne la norme* L^2 *de la dérivée* m*–ième de* $u(t)$*,* c_{11} *est le coefficient de la première ligne/première colonne de* $\mathbb{C}$ *(qui dépend donc du E–flux choisi),* d_k *et* c_k *ne dépendent que de* k *et* $\hat{k} = k$ *sauf dans le cas d'un maillage uniforme où* $\hat{k} = k$*.*

[3] On remarque que si $F(x) = \displaystyle\int_0^x f(s)ds$, alors $F' \leq A + F$ et donc $F(x) \leq A(\exp(x) - 1)$

Ce lemme est conséquence des deux lemme suivants dont on trouvera la démonstration dans [20] page 2455 et dans [28]. Ce lemme est un résultat d'interpolation.

Lemme 5.5.4. *Si $p = \Pi_h(w) - w$ et $J_{j+1/2} = I_j \cup I_{j+1}$, on a*

$$\begin{aligned}
&\left|\overline{p_u}_{j+1/2}\right| \leq c_k \Delta x^{\hat{k}+1/2}\ |u|_{H^{\hat{k}+1}(J_{j+1/2})},\\
&\left|[p_u]_{j+1/2}\right| \leq c_k \Delta x^{k+1/2} |u|_{H^{k+1}(J_{j+1/2})}\\
&\left|\overline{p_q}_{j+1/2}\right| \leq c_k \sqrt{a}\ \Delta x^{\hat{k}+1/2}\ |u|_{H^{\hat{k}+2}(J_{j+1/2})},\\
&\left|[p_q]_{j+1/2}\right| \leq c_k \sqrt{a}\ \Delta x^{k+1/2} |u|_{H^{k+2}(J_{j+1/2})}.
\end{aligned}$$

Remarque 5.5.5. *Le terme $\sqrt{a}$ dans les équations 3 et 4 provient de la définition de $q = \sqrt{a}\dfrac{\partial u}{\partial x}$.*

Lemme 5.5.6 ([28]). *On a*

$$\left|[\Pi_h(p)]_{j+1/2}\right| \leq d_k \Delta x^{-1/2}\ ||\Pi_h(p)||_{L^2(J_{j+1/2})}.$$

Démonstration du Lemme 5.5.3. On pose $v_h = \Pi_h(e)$ pour simplifier les notations. En utilisant la définition de $\mathcal{B}_h$, on a

$$\begin{aligned}
\mathcal{B}_h(p, v^h) = &\int_0^T \int_0^1 \left(\frac{\partial p_u}{\partial t} v_u^h + q_u v_q^h\right) dxdt\\
&- \int_0^T \sum_{j=1}^N \hat{h}(p)^t_{j+1/2} [v^h(t)]_{j+1/2} dt\\
&- \int_0^T \sum_{j=1}^N \int_{I_j} h(p(x,t))^t \frac{\partial v^h}{\partial x}(x,t) dxdt\\
= &- \int_0^T \sum_{j=1}^N \hat{h}(p)^t_{j+1/2} [v^h(t)]_{j+1/2} dt
\end{aligned}$$

Ceci résulte de la définition de la projection L^2 et du fait que h est linéaire : c'est là et seulement là que l'on emploie la linéarité du flux.

En effectuant les même calculs que pour le lemme 5.5.1, on obtient

$$\begin{aligned}
\hat{h}(p)^t_{j+1/2} [v^h(t)]_{j+1/2} &= (c\overline{p_u} - c_{11}[p_u])[v_u]\\
&+(-\sqrt{a}\overline{p_q} - c_{12}[p_q])[v_u]\\
&+(-\sqrt{a}\overline{p_u} + c_{12}[p_u])[v_q]\\
&= \theta_1 + \theta_2 + \theta_3
\end{aligned}$$

Grâce aux lemmes (5.5.4) et (5.5.6), en notant J au lieu de $J_{j+1/2}$,

$$|\theta_1| \leq c_k\left(|c|\Delta x^{\hat{k}+1/2} + c_{11}\Delta x^{k+1/2}\right) |u|_{H^{k+1}(J)} \left|[v_u]\right| \leq c_k \Delta x^{\hat{k}+1/2}(|c| + c_{11}) \left|[v_u]\right|\ |u|_{H^{k+1}(J)}.$$

de même,

$$\begin{aligned}
|\theta_2| &\leq \left(c_k a \Delta x^{\hat{k}+1/2} |u|_{H^{\hat{k}+2}(J)} + |c_{12}|\sqrt{a}\ \Delta x^{k+1/2} |u|_{H^{k+2}(J_{j+1/2})}\right) \left|[v_u]\right|\\
&\leq c_k d_k \Delta x^k \left(a\Delta x^{\hat{k}-k} + \sqrt{a}|c_{12}|\right) |u|_{H^{k+2}(J_{j+1/2})}\ ||v_u||_{L^2(J_{j+1/2})}
\end{aligned}$$

$$|\theta_3| \leq c_k d_k \Delta x^k \left(\sqrt{a}|u|_{H^{k+1}(J_{j+1/2})} \Delta x^{\hat{k}-k} + |c_{12}|\ |u_{H^{k+1}(J_{j+1/2})}\right)\ ||v_q||_{L^2(J_{j+1/2})}$$

Puis, on écrit

$$\left|\theta_1\right| \leq \frac{1}{2}c_{11}\left|[v_u]\right|^2 + \frac{1}{2}\left(c_k^2\frac{|c|+c_{11}}{c_{11}}\Delta x + 4|c_{12}|d_k^2\right)|u(t)|^2_{H^{k+1}(J_{j+1/2})}\Delta x^{2k}$$

en employant l'inégalité de Young, et en faisant la même chose sur θ_2 et θ_3, on obtient

$$\hat{h}(p)^t[v_h(t)] \leq \frac{1}{2}c_{11}[v_u^h]^2 + \frac{1}{4}||v_q||^2_{L^2(J_{j+1/2})} + \frac{1}{4}C_{1,J}(t)\Delta x^{2k} + C_{2,J}\Delta x^k||v_u||_{L^2(J_{j+1/2})}$$

où $C_{1,J}$ et $C_{2,J}$ sont des fonctions dépendant de c_k, d_k et des semi normes de Sobolev de $u(t)$.

Utilisant ensuite le fait que

$$\mathcal{B}(p, v_h) \leq \int_0^T \sum_{j=1}^N \left|\hat{h}(p)^t_{j+1/2}\,[v_h(t)]_{j+1/2}\right| dt,$$

l'inégalité de Cauchy Schwarz, et le fait que

$$\begin{pmatrix}[u]\\ [q]\end{pmatrix}^t \mathbb{C}\begin{pmatrix}[u]\\ [q]\end{pmatrix} = c_{11}[u]^2,$$

on obtient le résultat. □

Remarque 5.5.7. *C'est à cause de la relation*

$$\begin{pmatrix}[u]\\ [q]\end{pmatrix}^t \mathbb{C}\begin{pmatrix}[u]\\ [q]\end{pmatrix} = c_{11}[u]^2$$

et de la présence de $\int_0^T v_q^2(s)ds$ *dans* $\mathcal{B}_h$ *que le traitement des termes* θ_2 *et* θ_3 *est différent de celui de* θ_1 *afin de faire «sortir»* $\Theta(\Pi_h(e))$.

On obtient alors le

Théorème 5.5.1. *Si* $e = w - w_h$*, on a*

$$\left\{\int_0^1 |e_u(x,T)|^2dx + \int_0^T\int_0^1 |e_q(x,t)|^2dxdt + \Theta([e])\right\}^{1/2} \leq C\Delta x^k$$

où C *dépend uniquement de* k*,* $|u|_{k+1}$ *et* $|u|_{k+2}$*.*

Dans le cas $a = 0$*,* $C = O(\Delta x^{1/2})$*. Dans le cas parabolique* $c = 0$*,* $C = O(\Delta x)$ *si* k *est pair, le maillage uniforme et* $\mathbb{C} = 0$*.*

Démonstration. On reprend les calculs partant du lemme précédent et utilisant la relation (5.28), et donc

$$\frac{1}{2}\int_0^1 \left[P_h(e_u(T))\right]^2dx + \int_0^T\int_0^1 \left[P_h(e_q(t))\right]^2dxdt + \Theta(P_h(e)) \leq \frac{1}{2}\Theta(\Pi_h(e)) + \frac{1}{2}\int_0^T\int_0^1 [\Pi_h(e)]^2dxdt$$
$$+\frac{1}{2}\Delta x^{2k}\int_0^T C_1(s)ds + \Delta x^k\int_0^T C_2(s)\sqrt{\int_0^1 [P_h(e_u(s))]^2dxds}$$

soit

$$\frac{1}{2}\int_0^1 \Big[P_h(e_u(T))\Big]^2 dx + \frac{1}{2}\int_0^T\int_0^1 \Big[P_h(e_q(t))\Big]^2 dxdt + \frac{1}{2}\Theta(P_h(e)) \le$$
$$+\frac{1}{2}\Delta x^{2k}\int_0^T C_1(s)ds + \Delta x^k \int_0^T C_2(s)\Big\{\sqrt{\int_0^1 [P_h(e_u(s))]^2 dx}\Big\}ds,$$

soit en posant

$$\begin{aligned}
f(t) &= \int_0^1 \Big[P_h(e_u(T))\Big]^2 dx + \int_0^T\int_0^1 \Big[P_h(e_q(t))\Big]^2 dxdt + \Theta(P_h(e)),\\
A &= \Delta x^{2k}\int_0^T C_1(s)ds\\
B(t) &= \Delta x^k C_2(t)
\end{aligned}$$

on en déduit

$$f(T) \le A + \int_0^T B(t)\sqrt{f(t)}dt \le A + \frac{1}{2}\int_0^T B^2(t)dt + \frac{1}{2}\int_0^T f(t)dt.$$

Dans cette inégalité, le choix de T est arbitraire, donc on peut voir cela comme une inéquation différentielle. Appliquant le lemme de Gronwald, on en déduit le résultat.

L'évaluation précise des constantes fait apparaître les facteurs Δx et $\Delta x^{1/2}$ qui apparaissent dans l'énoncé du résultat. □

5.6 Extension à plus d'une dimension d'espace

On considère $\mathbb{R}^d$, $d > 1$, qu'on recouvre au moyen d'une triangulation

$$\mathbb{R}^d = \Cup_j K_j$$

où les éléments K_j sont des domaines polyhédraux d'intérieur non vide : typiquement des triangles, quadrangles pour $d = 2$, des tétrahèdres/hexahèdres pour $d = 3$. Le maillage n'est pas supposé conforme, mais on suppose que si $i \neq j$, $K_i \cap K_j$ est soit vide soit d'intérieur vide.

Pour simplifier l'exposé, on va se limiter au cas d'un problème sans diffusion, le cas avec diffusion se traitant de manière similaire.

5.6.1 Cas d'un problème scalaire

On considèrera donc

$$\frac{\partial u}{\partial t} + \text{ div } (\mathbf{u}) = \frac{\partial \mathbf{u}}{\partial \mathbf{t}} + \sum_{\mathbf{j=1}}^{\mathbf{d}} \frac{\partial \mathbf{f_i}(\mathbf{u})}{\partial \mathbf{x_i}} = \mathbf{0}. \tag{5.29}$$

On notera encore $P^k(K_j)$ l'espace des polynômes de degré inférieur ou égal à k définis sur K_j, et on notera encore

$$V_h^k = \{u^h \in L^1 \cap BV(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d), \text{ quelque soit } j \in \mathbb{Z}, u^h_{|K_j} \in P^k(K_j)\}.$$

Pour des raisons analogues à celle évoquées plus haut, le schéma s'écrira

$$\int_{K_j} \frac{du^h}{dx} v^h - \int_{K_j} f(u^h)^t \nabla v^h dx + \int_{\partial K_j} \hat{h}(u^-, u^+, \vec{n}) d\sigma = 0 \tag{5.30}$$

où $\vec{n}$ désigne la normale extérieur à K_j, u^- désigne, pour $x \in \partial K_j$

$$u^-(x) = \lim_{y \to x, y \in K_j} u^h(y)$$

et u^+ désigne

$$u^+(x) = \lim_{y \to x, y \in \widetilde{K_j}} u^h(y),$$

$\widetilde{K_j}$ désignant l'élément de «l'autre coté» du bord de K_j au point x, voir figure 5.3.

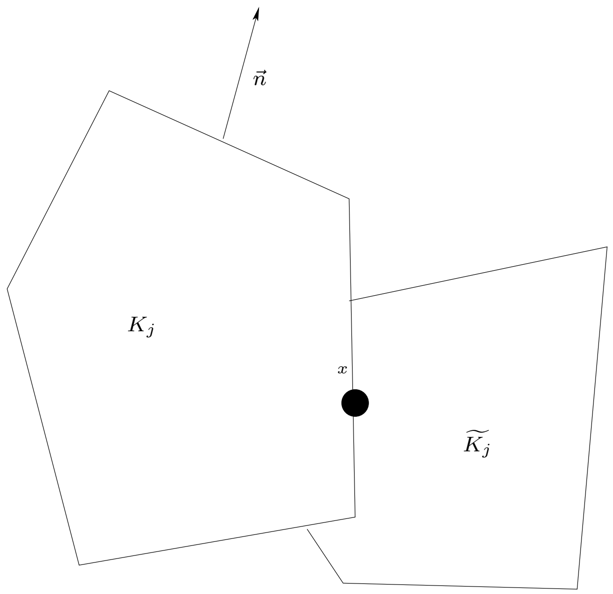


FIG. 5.3 – Eléments géométrique pour les schémas multidimensionnels.

Dans (5.30), on reconnait les trois éléments présents dans les schémas 1D :
- $\int_{K_j} \frac{du^h}{dx} v^h dx$ qui va donner la matrice de masse,
- $\int_{K_j} f(u^h)^t \nabla v^h dx$ qui nécessite une formule de quadrature,
- $\int_{\partial K_j} \hat{h}(u^-, u^+, \vec{n}) d\sigma$ qui nécessite une formule de quadrature sur chaque partie plane du bord, et un flux numérique $\hat{h}$.

Des formules de quadrature (pas nécessairement optimales) peuvent être trouvées dans Strang et Fix [28] et permettent de calculer la matrice de masse et $\int_{K_j} f(u^h)^t \nabla v^h dx$. **DONNER CA + POLYNOME DE DUBINER ANNEXE.** Etant donné que ∂K_j est formé de segments (en 2D) ou de facettes planes (en 3D), donc l'évaluation de cette dernière intégrale se ramène à des formules de quadrature, une fois que le flux $\hat{h}$ est défini. Celui–ci doit être consistant avec $(\mathbf{u})^{\mathbf{t}}\tilde{\mathbf{n}}$.

La méthode est de «projeter» (5.30) dans la direction $\vec{n}$ et de résoudre ou approcher le problème de Riemann suivant :

$$\frac{\partial u}{\partial t} + \frac{\partial (\mathbf{u})^{\mathbf{t}}\tilde{\mathbf{n}}}{\partial n} = 0$$
$$u(x,0) = \begin{cases} u^+ & \text{si } x^t\vec{n} < 0 \\ u^- & \text{si } x^t\vec{n} > 0 \end{cases}$$

Ceci peut se faire (par exemple) au moyen

– du flux de Roe

$$\hat{h}(u^-,u^+,\vec{n}) = \frac{1}{2}\big((\mathbf{u}^+)^{\mathbf{t}}\tilde{\mathbf{n}} + (\mathbf{u}^-)^{\mathbf{t}}\tilde{\mathbf{n}} + |\mathbf{a}(\mathbf{u}^+,\mathbf{u}^-,\tilde{\mathbf{n}})|\,(\mathbf{u}^+ - \mathbf{u}^-)\big)$$

où

$$a(u^+,u^-,\vec{n}) = \begin{cases} \dfrac{(\mathbf{u}^+)^{\mathbf{t}}\tilde{\mathbf{n}} - (\mathbf{u}^-)^{\mathbf{t}}}{u^+ - u^-} & \text{si } u^+ \neq u^- \\ \nabla f(u^+)^t\vec{n} & \text{sinon} \end{cases}$$

– du flux de Lax Friedrichs

$$\hat{h}(u^+,u^-,,\vec{n}) = \frac{1}{2}\big((\mathbf{u}^+)^{\mathbf{t}}\tilde{\mathbf{n}} + (\mathbf{u}^-)^{\mathbf{t}}\tilde{\mathbf{n}} + \alpha(\mathbf{u}^+ - \mathbf{u}^-)\big)$$

où

$$\alpha \geq \max_{u\in[\min(u^+,u^-),\max(u^+,u^-)]} ||\nabla f^t\vec{n}||.$$

Dans ce cas, les résultats du paragraphe 5.2.3 restent valables.

Les méthodes de stabilisation décrites dans le paragraphe 5.2.2 restent valables sans modification.

5.6.2 Passage au multidimensionnel

Il faut dire quelque part que le schéma (semi discret) a la propriété de résidu.

5.6.3 Cas des équations d'Euler

Dans ce cas, il faut préciser le flux numérique $\hat{h}$ et indiquer comment transposer les méthodes de stabilisation décrites dans le paragraphe 5.2.2.

Commençons par le flux. On montre facilement l'invariance des équations d'Euler par rotation. Plus précisément, en effectuant une rotation de manière à transformer le repère canonique de $\mathbb{R}^d$ en un repère orthonormé dont le premier vecteur est $\vec{n}$, soit $\{\vec{n},\{\tau_j\}_{j=2,d}\}$, on peut transformer les équations d'Euler en un système de deux sous–système. Le premier n'est rien d'autre que les équations d'Euler portant sur $(\rho,\rho\vec{u}^t\vec{n},E')$ où $E' = \rho\varepsilon + \frac{1}{2}\big(\vec{u}^t\vec{n})^2\big)$, et l'autre système est un système convectant $\vec{u'} = \vec{u} - \vec{u}^t\vec{n}\ \vec{n}$ à la vitesse $\vec{u}^t\vec{n}$. Ainsi, on peut «découpler» les phénomènes dans la direction $\vec{n}$ de ceux dans l'hyperplan orthogonaux. On peut donc considérer un schéma monodimensionnel tel que l'un de ceux décrits dans le chapitre 4 pour les systèmes, et un schéma de convection dans l'hyperplan orthogonal, puis effectuer une rotation inverse.

Par exemple, le flux de Roe devient,

$$F(u^-, u^+, \vec{n}) = \frac{1}{2}\big(F(u^+)^t\vec{n} + F(u^-)^t\vec{n} + \left|\overline{A}(u^-, u^+, \vec{n})\right| (u^+ - u^-)\big)$$

où la matrice moyenne $\overline{A}(u^-, u^+, \vec{n})$ est obtenue en linéarisant le flux $F(u)^t\vec{n}$. Plus de détails peuvent être trouvés dans [2].

Le flux de Lax Friedrichs devient

$$F(u^-, u^+, \vec{n}) = \frac{1}{2}\big(F(u^+)^t\vec{n} + F(u^-)^t\vec{n} + \alpha\ (u^+ - u^-)\big)$$

où

$$\alpha \geq \max\big(|\vec{u^+}^t\vec{n}| + c(u^+), |\vec{u^-}^t\vec{n}| + c(u^-)\big)$$

5.7 Illustrations numériques

Notes bibliographiques

Papiers de Rebay et al. Probleme du choix des constantes, lié à la stabilisation nonlinéaire de la méthodes. Autres méthodes DGs (Wheeler et col.) Cas adaptatif (Schwab)

Chapitre 6

Méthodes de Galerkin stabilisé

Dans le chapitre précédent, on s'est intéressé à l'approximation variationnelle au moyen de fonctions localement régulières mais possédant des discontinuités.

Dans ce chapitre, on va s'intéresser à leur approximation au moyen de fonctions continues. On va se contenter de regarder ce qui se passe pour des problèmes scalaires de la forme

$$\begin{aligned} &\sigma u + \vec{\lambda}^t \nabla u - \varepsilon \Delta u = f && \text{dans } \Omega \\ &u = g && \text{sur } \Gamma = \partial\Omega. \end{aligned} \tag{6.1}$$

avec $\sigma > 0$. Dans (6.1), Ω est un domaine borné de $\mathbb{R}^d$, $\vec{\lambda}$ est une fonction régulière bornée dans Ω et $\varepsilon > 0$. Plus généralement, on considèrera

$$\begin{aligned} &\sigma u + \text{ div } \mathbf{f}(u) - \varepsilon \Delta u = f && \text{dans } \Omega \\ &u = g && \text{sur } \Gamma = \partial\Omega. \end{aligned} \tag{6.2}$$

avec $\mathbf{f}$ régulière et $\sigma \geq 0$. Le cas de problèmes d'évolution et des systèmes sera évoqué dans les notes bibliographiques à la fin de ce chapitre.

Si $\vec{n}(x)$ désigne le vecteur normal extérieur à Γ en $x \in \Gamma$, on note le bord sortant $\Gamma_+ = \{x \in \Gamma, \vec{n}(x)^t\vec{\lambda} \geq 0\}$ et $\Gamma_- = \{x \in \Gamma, \vec{n}(x)^t\vec{\lambda} < 0\}$ le bord entrant. On a une définition analogue (utilisant $\nabla \mathbf{f}(u)$) dans le cas du problème (6.2).

Deux méthodes seront détaillées : la méthode streamline diffusion de Johnson, Hughes et col., et les méthodes distribuant le résidu qui sont plus récentes. La première classe de méthodes est construite afin de satisfaire certains contraintes pour les normes L^2 et H^1. La seconde classe se focalise sur le comportement en norme L^∞.

6.1 La méthode streamline diffusion

Cette partie est inspirée de [29]. On supposera que

$$\sigma - \frac{1}{2} \text{ div } \vec{\lambda} \geq \sigma_0 \tag{6.3}$$

où σ_0 est une constante strictement positive. Sous la condition (6.1), le théorème de Lax–Milgram donne l'existence et l'unicité d'une solution dans $H^1_0(\Omega)$ (si $\varepsilon > 0$) ou $H^1(\Omega) \cap \{u_{\Gamma^-} = 0\}$ si $\varepsilon = 0$[1]. On notera cette solution u.

[1] Cette analyse impose $\sigma > 0$ afin d'obtenir une inégalité de coercivité dans H^1 ou H^1_0.

On considère une famille de triangulations régulières *conforme* de Ω. Chaque triangulation $\mathcal{T}_h$ est adaptée aux conditions aux limites. Pour $k \in \mathbb{N}^*$, on introduit

$$V_h = \{v \in H^1(\Omega), v_{|K} \in P_k(K), \forall K \in \mathcal{T}_h\}.$$

On introduit $V_h^0 = \{v \in V_h, v_\Gamma = 0\}$.

La méthode de Galerkin standard provient de la formulation variationnelle de (6.1) : il s'agit de trouver $u^h \in V_h$ telle que quelque soit $v^h \in V_h^0$,

$$\varepsilon \int_\Omega (\nabla u^h)^t \nabla v^h dx + \int_\Omega (\vec{\lambda}^t \nabla u^h + \sigma u^h) v^h dx = \int_\Omega f v^h \tag{6.4}$$

soit encore

$$\varepsilon(\nabla u^h, \nabla v^h) + (\vec{\lambda}^t \nabla u^h + \sigma u^h, v^h) = (f, v^h).$$

Quand $\varepsilon = 0$, le problème est hyperbolique, et la formulation se devient : trouver $u^h \in V_h$ telle que quelque soit $v^h \in V_h$,Il manque un bout

$$(\vec{\lambda}^t \nabla u^h + \sigma u^h, v^h) - \langle u^n, v^h \rangle_- = (f, v^h) - \langle g, v^h \rangle_-.$$

On sait que la méthode de Galerkin n'est pas très satisfaisante du moins quand ε est petit. On obtient, pour cette méthode, des erreurs de la forme

$$||u - u^h|| + |u - u^h| \leq Ch^k ||u||_{k+1}$$

pour lesquelles on a aucun contrôle du gradient de u^h.

En effet, dans le cas monodimensionnel par exemple, la méthode de Galerkin se réduit à un schéma centré non dissipatif pour le quel le seul contrôle que l'on ait porte sur la norme L^2 de la solution. On a aucune information sur ∇u. Or, on sait que le modèle (6.1) possède des couches limites (voir par exemple le chapitre 2) : la solution possède des gradients pouvant être important. Le seul moyen de procéder alors est de rafiner le maillage afin que la variation de la solution numérique dans une maille soit faible. Ceci est difficilement envisageable dans la pratique.

Il est donc important de posséder une méthode qui permette de contrôle à la fois u et son gradient.

Plusieurs méthodes existent, comme par exemple le fait d'ajouter un terme de la forme $h(\nabla u^h, \nabla v^h)$ à l'équation. Ceci permet de résoudre les problèmes de stabilité, au détriment de la qualité de l'approximation : elle est du premier ordre quelque soit le degré k !

Une meilleure méthode est de modifier (6.1) en introduisant un terme de diffusion dans la direction $\vec{\lambda}$. Plus précisément on considère le problème variationnel : trouver $u^h \in V_h^0$ tel que

$$(\vec{\lambda}^t \nabla u^h + \sigma u^h - f, v^h) + h(\vec{\lambda}^t \nabla u^h + \sigma u^h - f, \vec{\lambda} \nabla v^h) + \varepsilon(\nabla u^h, \nabla v^h) = 0. \tag{6.5}$$

Dans la suite ce ce paragraphe, on regarde ce qui se passe pour $\varepsilon = 0$, puis on étudie le cas $\varepsilon > 0$.

6.1.1 La méthode streamline diffusion pour $\varepsilon = 0$.

On définit

$$\mathcal{B}_h(w,v) = (\vec{\lambda}^t \nabla w + \sigma w, v + h\vec{\lambda}^t \nabla v) - \langle w, v\rangle \tag{6.6}$$

et

$$L(v) = (f, v + h\vec{\lambda}^t \nabla v) - \langle g, v\rangle_-$$

Si $e = u - u^h$ désigne l'erreur de la méthode, on a, quelque soit $v \in V_h$,

$$\mathcal{B}_h(e,v) = 0. \tag{6.7}$$

Il suffit de constater que $\mathcal{B}_h(u, v_h) = 0$: le terme supplémentaire entre (6.1) et (6.5) est nul pour u, par construction.

On va déterminer le comportement de e. Pour cela on introduit la norme

$$|||v||| = \left(h||\vec{\lambda}^t \nabla u||^2 + \sigma_0 ||v||^2 + \frac{1}{4}|v|^2\right)^{1/2}.$$

La grande différence avec la norme L^2, qui est celle qu'il faut employer pour la méthode de Galerkin standard, est la présence du terme $h||\vec{\lambda}^t \nabla u||^2$ qui permet d'avoir un contrôle sur le gradient de u^h, uniquement dans la direction $\vec{\lambda}$. On ne peut pas espérer plus puisque, généralement comme on le voit par la méthode des caractéristiques, la solution peut posséder des discontinuités dans la direction orthogonale à $\vec{\lambda}$. Ce sont des discontinuités qui sont propagées par le champs $\vec{\lambda}$ à partir du bord. Il faut bien sûr que la condition au bord g possède de telle discontinuités ; ceci est tout à fait compatible avec le fait que $g \in L^2(\Gamma^-)$.

On a le lemme suivant :

Lemme 6.1.1. *La forme quadratique $\mathcal{B}_h$ est définie dans $H^1(\Omega)$. Elle est coercive dans pour la norme $|||\ .\ |||$.*

Démonstration. La forme $\mathcal{B}_h$ est clairement définie dans $H^1(\Omega)$. Si $v \in H^1(\Omega)$, on a le résultat en utilisant la formule de Green l'inégalité de Young et la condition (6.3). □

On obtient alors le résultat suivant

Théorème 6.1.1. *Si u^h satisfait (6.5) et $u \in H^{k+1}(\Omega)$ est la solution de (6.1) avec $\varepsilon = 0$, il existe C qui ne dépend que de Ω tel que*

$$|||u - u^h||| \leq Ch^{k+1/2}\ ||u||_{k+1}$$

Avant de montrer ce résultat, rappelons un certain nombre de résultats classiques dont on peut trouver la démonstration dans [28] ou dans [30] par exemple. Dans ces références, on montre qu'il existe une constante C telle que quelque soit $u \in H^{k+1}(\Omega)$, il existe $\bar{u}^h \in V_h$ pour lequel

$$\begin{aligned} &||u - \bar{u}^h|| + h||u - u^h|_1 \leq C\ h^{k+1}\ ||u||_{k+1} \\ &|u - u^h| \leq Ch^{k+1/2}\ ||u||_{k+1}. \end{aligned} \tag{6.8}$$

Démonstration du théorème 6.1.1. Soit $\bar{u}_h$ définit comme dans (6.8). On écrit $e = u - u_h = (u - \bar{u}^h) - (u^h - \bar{u}^h) = \nu_h - e^h$. Grâce à (6.7), on a

$$\mathcal{B}_h(e, e^h) = 0$$

car $e^h \in V^h$. On obtient donc, grâce à l'inégalité de coercivité sur $\mathcal{B}_h$,

$$\begin{aligned} c\,|||e||| &\leq \mathcal{B}_h(e,e) = \mathcal{B}_h(e,\nu^h) - \mathcal{B}(e,e^h) = \mathcal{B}(e,e^h) \\ &= (\vec{\lambda}^t\nabla e, \nu^h) + h(\vec{\lambda}^t\nabla e, \vec{\lambda}^t\nabla\nu^h) + (\sigma e, \nu^h) + h(\sigma e, \vec{\lambda}^t\nabla\nu^h) - \langle e, \nu^h\rangle_-. \end{aligned}$$

Puis, utilisant l'inégalité de Young

$$ab \leq \frac{\lambda}{2}a^2 + \frac{1}{2\lambda}b^2$$

pour $\lambda > 0$ bien choisi plusieurs fois, on a

$$(\vec{\lambda}^t\nabla e, \nu^h) \leq \frac{1}{4}ch||\vec{\lambda}^t\nabla e||^2 + \frac{4}{ch}||\nu_h||^2$$

$$h(\vec{\lambda}^t\nabla e, \vec{\lambda}^t\nabla\nu^h) \leq \frac{h}{4}||\vec{\lambda}^t\nabla e||^2 + 4h||\vec{\lambda}^t\nabla\nu^h||^2$$

$$(\sigma e, \nu^h) \leq \frac{1}{4}c\overline{\sigma}||e||^2 + \frac{1}{c\overline{\sigma}}||\nu^h||^2$$

$$h(\sigma e, \vec{\lambda}^t\nabla\nu^h) \leq \frac{1}{4}c\overline{\sigma}||e||^2 + \frac{h^2}{c\overline{\sigma}}||\vec{\lambda}^t\nabla\nu^h||^2$$

$$-\langle e, \nu^h\rangle_- \leq \frac{1}{8}c|e|^2 + \frac{4}{c}|\nu^h|^2$$

soit, en utilisant encore (6.8) et pour C convenable,

$$(\vec{\lambda}^t\nabla e, \nu^h) \leq \frac{1}{4}ch||\vec{\lambda}^t\nabla e||^2 + Ch^{2k+1}||u||^2_{k+1}$$

$$h(\vec{\lambda}^t\nabla e, \vec{\lambda}^t\nabla\nu^h) \leq \frac{h}{4}||\vec{\lambda}^t\nabla e||^2 + Ch^{2k+1}||u||^2_{k+1}$$

$$(\sigma e, \nu^h) \leq \frac{1}{4}c\overline{\sigma}||e||^2 + Ch^{2k+1}||u||^2_{k+1}$$

$$h(\sigma e, \vec{\lambda}^t\nabla\nu^h) \leq \frac{1}{4}c\overline{\sigma}||e||^2 + Ch^{2k+1}||u||^2_{k+1}$$

$$-\langle e, \nu^h\rangle_- \leq \frac{1}{8}c|e|^2 + Ch^{2k+1}||u||^2_{k+1},$$

soit, en combinant le tout avec la définition de $|||\,.\,|||$,

$$|||e|||^2 \leq Ch^{2k+1}\;||u||^2_{k+1}.$$

□

6.1.2 La méthode streamline diffusion pour $\varepsilon > 0$.

On va commencer par légèrement modifier (6.4) afin d'avoir le même résultat d'approximation que dans le théorème 6.1.1. Le problème est la présence du terme de diffusion

6.2 Schémas distribuant le résidu.

Dans cette partie, on va considérer le problème stationnaire

$$\begin{array}{ll} \text{div } \mathbf{f}(u) = f & x \in \Omega \\ u = 0 & x \in \Gamma_- \end{array} \tag{6.9}$$

où $\Gamma_- = \{x \in \partial\Omega, \nabla_u(\mathbf{u})^{\mathbf{t}}\tilde{\mathbf{n}}(\mathbf{x}) < \mathbf{0}\}$, $\vec{n}(x)$ désignant le vecteur normal unitaire extérieur à $\partial\Omega$ en x. On ne s'intéressera en fait qu'à (6.9) pour simplifier l'exposé, le cas plus général

$$\begin{array}{ll} \text{div } \mathbf{f}(u) = \varepsilon\Delta u + g & x \in \Omega \\ u = 0 & x \in \partial\Omega := \Gamma \end{array} \tag{6.10}$$

et, dans le cas où $\varepsilon = 0$ sera évoqué dans les notes biliographiques. De même, tout l'exposé sera effectué en supposant deux dimensions d'espace, l'extension à la dimension trois ne posant pas de problème conceptuel majeur.

Dans cette partie, on va se contenter de décrire ces schémas, très peu de résultats théoriques de convergence existent.

On considère une triangulation conforme de Ω formée de triangles en 2 dimensions, de tétrahèdres en 3 dimensions. *On ne considèrera que ce type d'éléments.* Les éléments de la triangulation $\mathcal{T}_h$ sont notés $\{T_{jt}\}_{jt=1,n_t}$, Les sommets de $\mathcal{T}_h$ sont $\{M_{is}\}_{is=1,ns}$. En principe, nous devons indexer les sommets du triangle T dans la liste $\{M_{is}\}_{is=1,ns}$, c'est à dire M_{i_1}, M_{i_2}, M_{i_3}. Cependant, quand il n'y a pas d'ambiguïté, nous les identifierons par leur indice dans la liste des noeuds i_1, i_2, i_3 ou encore par $1, 2, 3$. Dans tout l'exposé les fonctions employées seront supposées exister dans

$$V_h = \{u \text{ continu dans } \Omega, u_T \in P^1(T), \text{ quelque soit } T \in \mathcal{T}_h\}.$$

Afin d'approximer (6.9), nous considérons le schéma suivant, dit schéma distribuant le résidu

$$|C_i|\frac{u_i^{n+1} - u_i^n}{\Delta t} + \sum_{T, M_i \in T} \Phi_i^T = 0 \tag{6.11}$$

où u_i^n est une approximation de $u(M_i, t_n)$ solution de (6.9), $|C_i|$ est la mesure de la cellule de contrôle duale[2]associée au noeud M_i, Δt est le pas de temps, et Φ_i^T représente le résidu «envoyé» par l'élément T au noeud M_i. Les résidus doivent satisfaire la relation de conservation suivante

$$\sum_{j=1,3} \Phi_{i_j}^T = \int_T \text{div } \mathbf{f}^h dx := \Phi^T. \tag{6.12}$$

Dans (6.12), $\mathbf{f}^h$ est une interpolation continue du flux $\mathbf{f}$ qui est supposée converger dans L^1_{loc} vers $\mathbf{f}$. Dans [31] et dans le paragraphe 6.2.1, on montre que sous les conditions classiques du théorème de Lax Wendroff, la solution limite de (6.11) est une solution faible de (6.10).

[2]construite en joignant les milieux des segments issus de M_i aux centres de gravités des triangles contenant M_i, voir figure 6.1. Dans cette figure, la cellule duale C_i entoure M_i.

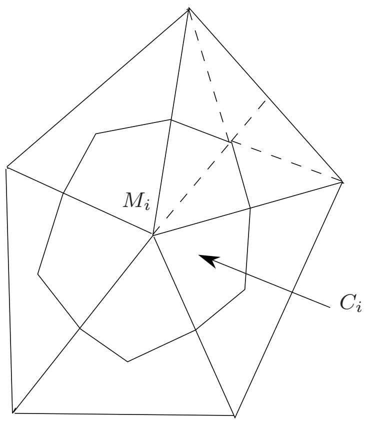


FIG. 6.1 – Cellule duale C_i associée au sommet M_i.

Un premier exemple de schéma distribuant le résidu est donné par les schémas de type volumes finis sur des maillages triangulaires, voir la figure 6.2 pour les notations. Ces schéma ont étés décrits au paragraphe 4.3.2, et plus particulièrement au paragraphe «données par noeuds». La formulation distributive de ces schémas n'est qu'une réecriture que nous décrivons maintenant.

Si Q représente la matrice de dissipation du schéma – nous faisons le calcul pour un Q–schéma, voir [2]– le flux numérique correspondant à l'arêter $[M_i, M_j]$ est

$$\begin{aligned}\hat{f}(u_i, u_j, \vec{n}_{ij}) &= \frac{1}{2}\Big(\mathbf{f}.\vec{n}^1_{ij}(u_j) + \mathbf{f}.\vec{n}^1_{ij}(u_i) - Q(u_i, u_j, \vec{n}^1_{ij}).(u_i - u_j)\Big) \\ &+ \frac{1}{2}\Big(\mathbf{f}.\vec{n}^2_{ij}(u_j) + \mathbf{f}.\vec{n}^2_{ij}(u_i) - Q(u_i, u_j, \vec{n}^2_{ij}).(u_i - u_j)\Big).\end{aligned} \tag{6.13}$$

Dans (6.13), nous avons employé la notation $\mathbf{f}.\vec{n}$ pour $n_x\ \mathbf{f}_x(u) + n_y\ \mathbf{f}_y(u)$ ou n_x et n_y sont les composantes de $\vec{n}$ et $\mathbf{f}_x$, $\mathbf{f}_y$ sont les composantes de $\mathbf{f}$. Puisque le bord de C_i est fermé, le schéma est identique au schéma distribuant le résidu (6.11) ou nous avons posé

$$\begin{aligned}\Phi_i^{T_1} &= \frac{1}{2}\Big(\mathbf{f}.\vec{n}^1_{ij}(u_j) - \mathbf{f}.\vec{n}^1_{ij}(u_i) - Q(u_i, u_j, \vec{n}^1_{ij}).(u_i - u_j)\Big) \\ &+ \frac{1}{2}\left(\mathbf{f}.\vec{n}^1_{ik}(u_k) - \mathbf{f}.\vec{n}^1_{ik}(u_i) - Q(u_i, u_k, \vec{n}^1_{ik}).(u_i - u_k)\right)\end{aligned} \tag{6.14}$$

et défini une expression analogue pour le triangle T_2. Dans (6.14), les indices j et k sont les numéros des sommets de T_1 qui sont différents de M_i. Nous obtenons finalement

$$\sum_{M_i \in T_1} \Phi_i^{T_1} = \frac{1}{2} \sum_{M_i \in T_1} \mathbf{f}(u_i).\vec{n}_i$$

où $\vec{n}_j$ est la normale entrante à la face opposée au noeud M_j dans le triangle T_1. Dans le cas présent, l'approximation $\mathbf{f}^h$ nécessaire dans la relation (6.12) n'est rien d'autre que l'interpolation linéaire de $\mathbf{f}$ dans T.

D'autres exemples, tels que le schéma SUPG de Hughes ou le schéma streamline décrit au paragraphe 6.1 s'écrivent naturellement dans le même formalisme. En effet, dans le cas de (6.1), si on suppose une représentation linéaire par triangle de la solution et si φ_i représente la fonction de base P^1 au sommet M_i,

$$\Phi_i^T = \int_T \Big(\vec{\lambda}^t \nabla u^h + \sigma u^h - g\Big)\varphi_i \, dx + h \int_T \Big(\vec{\lambda}^t \nabla u^h + \sigma u^h - g\Big) \, \vec{\lambda} \nabla \varphi_i dx$$

et donc,

$$\Phi^T = \int_T \Big(\vec{\lambda}^t \nabla u^h + \sigma u^h - g\Big) dx.$$

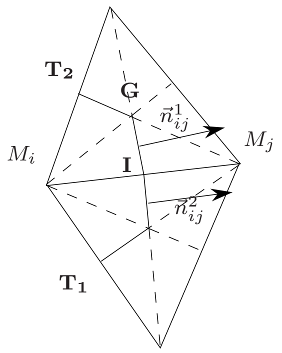


FIG. 6.2 – Eléments géometriques pour le schéma volume fini. $\vec{n}_{ij}^1$ est orthogonal à $[G, I]$ et de même norme. Même chose pour $\vec{n}_{ij}^2$.

D'autres exemples, que nous ne détaillerons pas, et qui entre dans ce formalisme, sont les schémas volumes finis centré–sommet de Ni Ni [32] et Morton *et coll.* [33, 34].

6.2.1 Un théorème de type Lax-Wendroff pour les schémas distribuant le résidu

On suppose une famille de triangulations $\mathcal{T}_h$ régulières : pour chaque triangulation, le minimum des angles des triangles T est minoré par $\alpha_0 > 0$. En d'autres termes, le maximum des rapports entre les cercles inscrits et les cercles circonscrits est majoré par une constante C qui ne dépend que de la famille de triangulation.

On rappelle la définition de V_h, espace des fonctions continues linéaire dans chaque triangle de $\mathcal{T}_h$ et on définit l'espace $\mathcal{X}_h$ des fonctions constantes par cellules de contrôle

$$\mathcal{V}_h = \{u_h; u_h \in C^0(\mathbb{R}^d)^m, u_{h|T} \in (\mathcal{P}_1)^m, \forall\, T \in \mathcal{T}_h\}$$

$$\mathcal{X}_h = \{u_h; u_{h|C} \in (\mathcal{P}_0)^m, \forall\, C \in \mathcal{C}_h\} \ \ .$$

On note $u_i \in \mathbb{R}^m$ l'ensemble des degrés de liberté aux noeuds M_i qui permettent de décrire u_h à la fois dans $\mathcal{V}_h$ et $\mathcal{X}_h$. Par exemple, si $N_i(x)$ est la fonction de base linéaire par triangle au noeud M_i pour la triangulation $\mathcal{T}_h$ telle que $N_i(M_j) = \delta_{ij}$, alors si $u_h \in \mathcal{V}_h$

$$u_h(x) = \sum_{M_i \in \mathcal{T}_h} N_i(x)\, u_i \ \ .$$

De même, si $\chi_i(x)$ est la fonction caractéristique de la cellule duale $C_i \in \mathcal{C}_h$,

$$\chi_i(x) = \begin{cases} 1 & x \in C_i \\ 0 & x \notin C_i \end{cases}$$

alors, quelque soit $\widetilde{u}_h \in \mathcal{X}_h$ on a

$$\widetilde{u}_h(x) = \sum_{M_i \in \mathcal{T}_h} \chi_i(x)\, u_i \quad .$$

Considérons le schéma (6.11).

$$|C_i| \frac{u_i^{n+1} - u_i^n}{\Delta t} + \sum_{T, M_i \in T} \Phi_i^T = 0$$

Dans les applications pratiques, les variables conservées sont des fonctions régulières de u, c'est à dire $u = H(u)$. Un ecemple est donné par les variables physiques $u = (\rho, \mathbf{u}, p)$. Un autre exemple est donné par le paramètre vecteur de Roe $u = (\sqrt{\rho}, \sqrt{\rho}\mathbf{u}, \sqrt{\rho}H)$ [3] où $H = \dfrac{E+p}{\rho}$ est l'enthalpie totale. Un dernier exemple est $u = \nabla_u\big(\rho s\big)$, l'ensemble des variables entropiques.

Les variables nodales u_i^n sont en bijection (au moyen de $u(\mathbf{u})$ et $\mathbf{u}(u)$) avec u_i^n, les degrés de liberté dans $\mathcal{V}_h$ et $\mathcal{X}_h$ à l'instant $t_n \equiv n\Delta t, n \in [0, N]$. De plus, les résidus $\mathbf{\Phi}_{i,T}^n$ et u_i^n sont supposées satisfaire les conditions suivantes :

Assumption 1 (H1). *Soit $\mathcal{T}_h$ une triangulation régulière. Pour $C \in \mathbb{R}$ et n fixé mais quelconque, il existe $C'(C) \in \mathbb{R}$ qui dépend de $\mathcal{T}_h$ telle que $\forall\ u_h^n \in \mathcal{V}_h$ et $||u_h^n||_{L^\infty(\mathbb{R}^d)^m} \le C$*

$$\big|\big|\mathbf{\Phi}_{i,T}^n\big|\big| \le C'\ h^{d-1} \sum_{M_j \in T} \big|\big|u_j^n - u_i^n\big|\big|, \quad \forall T \in \mathcal{T}_h \text{ et } \forall M_i \in T \ . \tag{6.15}$$

Il s'agit d'une hypothèse de continuité sur les résidus définis sur T en fonction des valeurs nodales $u_i^n, M_i \in T$. En particulier, si u_h^n est constant dans T alors $\mathbf{\Phi}_{i,T}^n = 0$.

Assumption 2 (H2). *Quelque soit $u_h^n \in \mathcal{V}_h$ et n fixé, on a la relation de conservation suivante*

$$\mathbf{\Phi}_T^n = \sum_{i=1}^{d+1} \mathbf{\Phi}_{i,T}^n = \int_T \mathit{div}\ \mathbf{f}^i(u_h^n)\, dx. \tag{6.16}$$

Remarque 6.2.1. **(i)** *Puisque $u_h \in \mathcal{V}_h$ est continue, quelque soit l'arête $\Gamma_{jk} = \{x \mid \partial T_j \cap \partial T_k \neq \emptyset\}$, le flux normal à l'arête est continu.*

(ii) *Pour $u^n \in \mathcal{V}_h$ et n fixé avec $||u_h^n||_{L^\infty(\mathbb{R}^d)} \le C$, il existe une constante $C'(C)$ telle que pour $T \in \mathcal{T}_h$*

$$||\mathbf{\Phi}_T^n|| \le \frac{C'}{h} \sum_{M_i, M_j \in T} ||u_j^n - u_i^n|| \quad . \tag{6.17}$$

[3] La notation classique pour le vecteur paramètre est $\mathbf{Z}$.

(iii) *Enfin, quelque soit la suite $(u_h^n)_h$ telle que (u_h^n) est bornée dans $L^\infty(\mathbb{R}^d \times \mathbb{R}^+)^m$ indépendamment de h et N et qui converge dans $L^2_{loc}(\mathbb{R}^d \times \mathbb{R}^+)^m$ vers u, on a*

$$\lim_{h\to 0} ||\mathbf{f}(u_h) - \mathbf{f}(u)||_{L^1_{loc}(\mathbb{R}^d\times\mathbb{R}^+)^m} = 0, \quad i = 1, 2, \ldots, d \ . \tag{6.18}$$

Remarquons que puisque $\mathbf{u} \mapsto H(\mathbf{u})$ est régulière et $\mathbf{u}_h$ est bornée, les hypotheses sur $\mathbf{u}_h$ sont équivalentes à des hypothèses équivalentes sur u_h définies par les valeurs nodales u_i.

Théorème 6.2.1. *Considérons $u_0 \in L^\infty(\mathbb{R}^d)^m$ et $\tau > 0$. Soit u_i une approximation nodale définie par (6.11) ou u_i est obtenu via $u_i \equiv u(u_i)$. Supposons que le schéma satisfasse les hypothèses (1) et (2), et qu'il existe une constante C qui ne dépend que de u_0, ainsi que des fonctions $u \in L^2(\mathbb{R}^d \times \mathbb{R}^+)^m$ et u_h telles que $u_h \in \mathcal{V}_h$*

$$\sup_h \sup_{x,t} ||u_h(x,t)|| \le C, \quad \lim_{h\to 0} ||u_h - u||_{L^2_{loc}(\mathbb{R}^d\times\mathbb{R}^+)^m} = 0 \ . \tag{6.19}$$

Alors $\mathbf{u}(x,t)$ est une solution faible de (6.10), c'est à dire que quelque soit φ régulière à support compact

$$\int_{\mathbb{R}^d\times\mathbb{R}^+} \left(\frac{\partial \varphi}{\partial t} u(x,t) + \sum_{i=1}^d \frac{\partial \varphi}{\partial x_i}(x,t) \cdot \mathbf{F}_i(u(x,t)) \right) dx\, dt + \int_{\mathbb{R}^d} \varphi(x,0)\, u_0(x)\, dx = 0. \tag{6.20}$$

La démonstration est inspirée de [35] et [36].

6.2.2 Démonstration du théorème 6.2.1

La preuve du théorème 6.2.1 est conséquence d'une suite de lemmes classiques mais qui sont adaptés aux schémas distributifs. Pour simplifier, on suppose que τ est un multiple entier de Δt, i.e. $\tau = N\,\Delta t$, bien que la généralisation à une valeur arbitraire de τ soit simple.

Lemme 6.2.2. *Soit $\mathcal{Q} = \cup T$ un domaine borné de $\mathbb{R}^d$ et $\tau > 0$. Considérons $(u_h)_h$ une suite telle que $u_h(\ .\ , n\Delta t) \in \mathcal{V}_h$ quelque soit $n \le N$. Supposons qu'il existe une constante C indépendante de h et N et une fonction $u \in L^2(\mathcal{Q} \times [0,\tau])$ telle que*

$$\sup_h \sup_{x,t} ||u_h(x,t)|| \le C, \quad \lim_{h\to 0} ||u_h - u||_{L^2_{loc}(\mathcal{Q}\times[0,\tau])^m} = 0 \ . \tag{6.21}$$

Alors on a

1. $\lim_{h\to 0} \sum_{n=0}^N \Delta t \sum_{\forall T\in\mathcal{Q}} |T| \sum_{M_i, M_j \in T} \left|\left| u_i^n - u_j^n \right|\right| = 0.$
2. $\lim_{h\to 0} \sum_{n=0}^N \Delta t \sum_{\forall T\in\mathcal{Q}} |T| \sum_{M_i, M_j \in T} \left|\left| u_i^n - u_j^n \right|\right|^2 = 0.$
3. $\lim_{h\to 0} h\ ||\nabla_x u_h||_{L^2(\mathcal{Q}\times[0,\tau])^m} = 0.$
4. *Il existe C' indépendante de h et n telle que $h||\nabla_x u_h||_{L^\infty(\mathcal{Q}\times[0,\tau])^m} \le C'$.*

Démonstration. Pour démontrer ce résultat, il suffit de considérer des fonctions à valeurs réelles.

L'idée de la démonstration repose sur l'observation que $\sum_{M_i,M_j\in T}||v_i - v_j||$ peut être réecrite comme une moyenne sur toutes les permutations possibles σ de $\{1,2,\cdots,d+1\}$ agissant sur $\sum_{M_i}\left|\left|v_i - v_{\sigma(i)}\right|\right|$. Puisqu'il n'y a qu'un nombre fini de permutations, il suffit de prouver que pour σ quelconque

$$\lim_{h\to 0}\sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}|T|\sum_{M_i\in T}\left|\left|v_i^n - v_{\sigma(i)}^n\right|\right| = 0 \ .$$

Quelques soient T et $I^n =]t_{n-1}, t_n[$, on construit les deux fonctions suivantes qui sont constantes par cellules

$$u_h(x,t)|_{T\times I^n} = \sum_{M_i\in T}\chi_{C_i\cap T}(x)\, u_i^n \tag{6.22}$$

et

$$\widetilde{u}_h(x,t)|_{T\times I^n} = \sum_{M_i\in T}\chi_{C_i\cap T}(x)\, u_{\sigma(i)}^n \tag{6.23}$$

où $\sigma(i)$ est une permutation de i et $\chi_{C_i\cap T}$ est la fonctions caractéristiques de $C_i\cap T$ où C_i est la cellule duale au noeud M_i. Cela définit deux fonctions sur $\mathcal{Q}\times[0,\tau]$ qui sont bornées indépendamment de h et N. De plus, on a les identités suivantes pour T et $p\geq 0$ arbitraires

$$|T|\sum_{M_i,M_j\in T}\|u_i^n - u_j^n\|^p = (d+1)\int_T\|u_h - \widetilde{u}_h\|^p\, dx \tag{6.24}$$

où le facteur "$d+1$" provient de la définition des cellules duales sur $\mathbb{R}^d$. En intégrant en temps

$$\sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}|T|\sum_{M_i,M_j\in T}\|u_i^n - u_j^n\| = \int_0^\tau\int_{(\cup T)\in\mathcal{Q}}\|u_h - \widetilde{u}_h\|\, dx\, dt \ . \tag{6.25}$$

La suite $(u_h)_h$ est bornée, donc il existe une fonction $u'\in L^\infty(\mathcal{Q}\times[0,\tau])^m$ telle que $u_h\to u'$ pour la topologie faible $*$. Grâce aux hypothèses précédentes, $u_h\to u$ dans L^2_{loc} ce qui implique que $u=u'$ puisque $\mathcal{Q}\times[0,\tau]$ est bornée et $C_0^\infty(\mathcal{Q}\times[0,\tau])$ est dense dans $L^1(\mathcal{Q}\times[0,\tau])$. De même, il existe une fonction $\widetilde{u}\in L^\infty(\mathcal{Q}\times[0,\tau])^m$, telle que $\widetilde{u}_h\to\widetilde{u}$ dans la topologie faible $*$. Il faut maintenant montrer que $\widetilde{u}=u$ et donc finalement $\widetilde{u}=u=u'$. Pour cela, soit $\varphi(x,t)\in C_0^\infty(R^d\times\mathbb{R}^+)$, intégrons $\varphi\, u_h$ dans $\mathcal{Q}\times[0,\tau]$, et en utilisant la définition de $\widetilde{u}$

$$\begin{aligned}\int_0^\tau\int_{\mathcal{Q}}\varphi\, u_h\, dx\, dt &= \sum_{n=0}^{n}\Delta t\sum_{\forall T\in\mathcal{Q}}\sum_{M_i\in T}u_i^n\int_T\varphi\,\chi_{C_i\cap T}\, dx\, dt\\ &= \int_0^T\int_{\mathcal{Q}}\varphi\,\widetilde{u}_h\, dx\, dt + \sum_{n=0}^{n}\Delta t\sum_{M_i\in T}u_i^n\left(\int_T\varphi\,\chi_{C_i\cap T}\, dx - \int_T\varphi\,\chi_{C_{\sigma^{-1}(i)}\cap T}\, dx\right)\end{aligned} \tag{6.26}$$

où $\sigma^{-1}(i)$ est la permutation inverse telle que $\sigma(\sigma^{-1}(i))=i$. En employant la définition des cellules duales C_i,

$$\int_T\chi_{C_i\cap T}\, dx = \int_T\chi_{C_{\sigma^{-1}(i)}}\, dx = |C_i\cap T|, \quad i=1,2,\ldots,d+1 \ . \tag{6.27}$$

En employant la formule de la moyenne, on peut trouver $\overline{x}_i$ et $\overline{x}'_i \in T$ telle que

$$\int_T \varphi\, \chi_{C_i \cap T}\; dx = |C_i \cap T|\, \varphi(\overline{x}_i), \quad \int_T \varphi\, \chi_{C_{\sigma^{-1}(i)}}\; dx = |C_i \cap T|\, \varphi(\overline{x}'_i) \ . \tag{6.28}$$

Puisque $\|\nabla\varphi\|$ est borné sur $\mathcal{Q} \times [0,T]$ et V_h est bornée,

$$\left|\left|\int_0^T \int_{\mathcal{Q}} \varphi\, u_h\; dx\, dt - \int_0^\tau \int_{\mathcal{Q}} \varphi\, \widetilde{u}_h\; dx\, dt\right|\right| \leq Ch \tag{6.29}$$

où C est indépendant de h et N. Donc on a $\widetilde{u} = u$ et finalement $\widetilde{u} = u = u'$.

Soit v_h, $\widetilde{v}_h$, et v'_h les composantes des fonctions u_h, $\widetilde{u}_h$, et u'_h. En employant la même technique, on voit que (v_h^2) et $(\widetilde{v}_h^2)$ ont la même limite faible $*$. On montre maintenant que cette limite est v^2. Une fois de plus, grâce à la densité de $C_0^\infty(\mathcal{Q} \times [0,\tau])$ dans $L^1(\mathcal{Q} \times [0,\tau])$ et le fait que v_h^2 est bornée indépendamment de h et N, on prend des fonctions tests $\varphi \in C_0^\infty(\mathcal{Q} \times [0,\tau])$. La fonction φ est bornée dans $\mathcal{Q} \times [0,\tau]$ et $v_h \to v \in L^2_{loc}(\mathcal{Q} \times [0,\tau])$, ainsi

$$\int_{\mathcal{Q}\times[0,\tau]} \varphi\, |v - v_h|^2\; dx\, dt \to 0, \tag{6.30}$$

et donc

$$\int_{\mathcal{Q}\times[0,\tau]} \varphi\, v^2\; dx\, dt - 2\int_{\mathcal{Q}\times[0,\tau]} \varphi\, v\, v_h\; dx\, dt + \int_{\mathcal{Q}\times[0,\tau]} \varphi\, v_h^2\; dx\, dt \to 0 \ . \tag{6.31}$$

Grâce à l'inégalité de Cauchy-Schwarz et $\varphi\, v \in L^1(\mathcal{Q} \times [0,\tau])$, le second terme converge vers

$$\int_{\mathcal{Q}\times[0,\tau]} \varphi\, v^2\; dx\, dt \ . \tag{6.32}$$

Ainsi, $v_h^2 \to v^2 \in L^\infty$ faible$-*$. On peut choisir $\varphi = 1$, et puisque $\widetilde{v}_h^2 \to v^2$ dans L^∞ faible $-*$, on obtient

$$\int_{\mathcal{Q}\times[0,\tau]} |\widetilde{v}_h - v|^2\; dx\, dt \to 0 \tag{6.33}$$

et pour finir

$$\int_{\mathcal{Q}\times[0,\tau]} |\widetilde{v}_h - v_h|^2\; dx\, dt \to 0 \ . \tag{6.34}$$

En interprétant cette relation sous la forme (6.24) on obtient la limite (1) du lemme 6.2.2. Le point (2) du lemme 6.2.2 en est une conséquence facile : $\mathcal{Q} \times [0,\tau]$ est borné, ainsi $L^2(\mathcal{Q} \times [0,\tau]) \subset L^1(\mathcal{Q} \times [0,\tau])$. Dans le but d'obtenir le point (3) du lemme 6.2.2, considérons la figure 6.3 qui représente un simplexe. Le vecteur $\vec{n}^i$ est le vecteur normal extérieur à la facette opposée au sommet M_i dont la norme est la mesure de cette facette. si bien que $\sum_{i=1}^{d+1} \vec{n}^i = 0$. En employant ces notations, on a

$$\nabla_x u_{h|T} = \frac{1}{(d+1)\,|T|} \sum_{j=1}^{d+1} \vec{n}^j\, u_j = \frac{1}{(d+1)\,|T|} \sum_{j=2}^{d+1} \vec{n}^j\, (u_j - u_1) \ . \tag{6.35}$$

Puis en intégrant en temps et en espace

$$\sum_{n=0}^{N} \int_{I^n} \int_{\mathcal{Q}} ||\nabla_x u_h^n||^2\; dx\, dt = \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} |T|\, \left|\left|(\nabla_x u_h^n)_{|T}\right|\right|^2 \leq C \frac{1}{h^2} \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \sum_{M_i, M_j \in T} |T|\, \left|\left|u_i^n - u_j^n\right|\right|^2$$

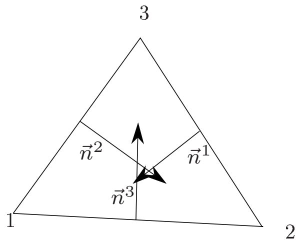


FIG. 6.3 – Description des normales $\vec{n}^i$, pour un simplexe de $\mathbb{R}^2$.

car les gradients sont constants dans un simplexe et puisque la triangulation est régulière. La limite (3) du lemme 6.2.2 s'obtient en applicant le résultat (2) du même lemme. Pour obtenir l'inégalité (4) du lemme 6.2.3, on part encore de l'équation 6.35 en supposant que u_j soit borné

$$h\,||\nabla_x u_h||_{L^\infty(\mathcal{Q}\times[0,\tau])^m} \leq \frac{h}{(d+1)\,|T|}\max_{1\leq j\leq d+1}|\vec{n}^j|\max_{1\leq j\leq d+1}u_j \leq C\frac{h}{(d+1)\,|T|}\max_{1\leq j\leq d+1}|\vec{n}^j| \tag{6.36}$$

qui est borné supérieurement par une constante independante de h et N, toujours pour une triangulation régukière. Ceci achève la preuve. □

Lemme 6.2.3. *Spoit $\varphi(x,t) \in C_0^1(\mathbb{R}^d \times \mathbb{R}^+)$. Sous les hypothèses du théorème 6.2.1, on a*

$$\sum_{n=0}^{N}\sum_{M_i\in\mathcal{T}_h}|C_i|\,\varphi(M_i,t_n)\,(u_i^{n+1}-u_i^n)+\int_{\mathbb{R}^d\times\mathbb{R}^+}\frac{\partial\varphi}{\partial t}(x,t)\,\mathbf{u}_h(x,t)\;dx\,dt+\int_{\mathbb{R}^d}\varphi(x,0)\,\mathbf{u}_0(x)\;dx\to 0 \tag{6.37}$$

quand $h\to 0$.

La preuve est classique, voir par exemple [1, 37].

Lemme 6.2.4. *Soit $\varphi(x,t)\in C_0^1(\mathbb{R}^d\times\mathbb{R}^+)$. Supposons que u_h satisfasse aux hypothèses du théorème 6.2.1. On a alors quand $h\to 0$*

$$\lim_{h\to 0}\sup_h\left|\left|\sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}\pi_h\varphi(x_T,t_n)\sum_{M_i\in T}\mathbf{\Phi}_{i,T}^n+\int_{\mathbb{R}^d\times\mathbb{R}^+}\sum_{i=1}^{d}\frac{\partial\varphi}{\partial x_i}(x,t)\,\mathbf{F}^i(u_h(x,t))\;dx\,dt\right|\right|=0 \tag{6.38}$$

où $C(\mathcal{T}_h,\mathbf{F})$ est une constante qui ne dépend que de $\mathcal{T}_h$ et $\|\nabla_u^{k+1}\mathbf{F}_{,u}\|$.

Démonstration. Choisissons $\varphi(x,t)$ telle que $\mathrm{supp}(\varphi)\subset\mathcal{Q}\times[0,\tau]$. On rappelle que $\mathbf{\Phi}_T^n=\sum_{i=1}^{d+1}\mathbf{\Phi}_{i,T}^n$ représente une approximation $\displaystyle\int_{\partial T}\mathbf{f}\cdot\vec{n}$. Un calcul direct donne

$$\begin{aligned}\sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}\pi_h\varphi(x_T,t_n)\left(\sum_{i=1}^{d+1}\mathbf{\Phi}_{i,T}^n\right) &= \sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}\int_T\pi_h\varphi(x_T,t_n)\sum_{i=1}^{d}\mathbf{F}_{,x_i}^i(u_h)\;dx \\ &= \sum_{n=0}^{N}\Delta t\sum_{\forall T\in\mathcal{Q}}\int_T\pi_h\varphi(x,t_n)\sum_{i=1}^{d}\mathbf{F}_{,x_i}^i(u_h)\;dx\end{aligned} \tag{6.39}$$

$$+ \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \int_T (\pi_h\varphi(x_T, t_n) - \pi_h\varphi(x, t_n)) \sum_{i=1}^{d} \mathbf{F}^i_{,x_i}(u_h)\ dx \tag{6.40}$$

On applique ensuite la formule de Green sur chauque simplexe apparaissant dans la somme de droite de (6.40)

$$\sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \int_T \pi_h\varphi(x, t_n) \sum_{i=1}^{d} \mathbf{F}^i_{,x_i}(u_h)\ dx = \sum_{n=0}^{N} \int_{I^n} \sum_{\forall T \in \mathcal{Q}} \int_T \pi_h\varphi(x, t_n) \sum_{i=1}^{d} \mathbf{F}^i_{,x_i}(u_h)\ dx\, dt \tag{6.41}$$

$$= -\sum_{n=0}^{N} \int_{I^n} \sum_{\forall T \in \mathcal{Q}} \int_T \sum_{i=1}^{d} \frac{\partial \pi_h\varphi}{\partial x_i} \mathbf{F}^i(u_h)\ dx\, dt + \sum_{n=0}^{N} \int_{I^n} \sum_{\forall T \in \mathcal{Q}} \int_{\partial T} \pi_h\varphi(x, t_n) \sum_{i=1}^{d} \mathbf{F}^i(u_h) \cdot n_i\ dt \tag{6.42}$$

On rappelle que $\pi_h\varphi$ et $\mathbf{F}$ sont tous les deux des fonctions bornés et continues. Grâce à la remarque 6.2.1 (i) et au fait que φ est à supports compact, on voit que le membre de droite de la somme (6.42) tends vers 0 uniformément. On regarde ensuite le membre de droite de l'équation (6.42), et on observe que

$$\left\| \sum_{n=0}^{N} \int_{I^n} \sum_{\forall T \in \mathcal{Q}} \int_T \left(\sum_{i=1}^{d} \frac{\partial \pi_h\varphi}{\partial x_i} \mathbf{F}^i(u_h) - \sum_{i=1}^{d} \frac{\partial \varphi}{\partial x_i} \mathbf{F}^i(u) \right) dx\, dt \right\| \leq C_0 \sum_{n=0}^{N} \int_{I^n} \sum_{\forall T \in \mathcal{Q}} \int_T ||\mathbf{F}(u_h) - \mathbf{F}(u)||\ dx\, dt + C_1 \sum_{n=0}^{N} \int_{I^n} \sum_{\forall T \in \mathcal{Q}} \int_T ||\nabla_x \pi_h\varphi - \nabla_x\varphi||\ ||\mathbf{F}(u_h)||\ dx\, dt \tag{6.43}$$

le premier terme de droite de (6.43) est égal à $||\mathbf{F}(u_h) - \mathbf{F}(u)||_{L^1(\mathcal{Q}\times[0,\tau])}$ et converge vers 0 quand $h \to 0$, voir la remarque 6.2.1 (iii). Puisque $\mathbf{V}_h$ est bornée et puisque $\mathbf{F}$ is continue, $\mathbf{F}(u_h)$ est borné par une constante. Le second terme du membre de droite de. (6.43) est borné par $||\nabla_x \pi_h\varphi - \nabla_x\varphi||_{L^1(\mathcal{Q}\times[0,\tau])}$ qui converge vers 0 quand $h \to 0$. Ainsi, on obtient

$$\lim_{h\to 0} \left\| \sum_{n=0}^{N} \int_{I^n} \sum_{\forall T \in \mathcal{Q}} \int_T \left(\sum_{i=1}^{d} \frac{\partial \pi_h\varphi}{\partial x_i} \mathbf{F}^i(u_h) - \sum_{i=1}^{d} \frac{\partial \varphi}{\partial x_i} \mathbf{F}^i(u) \right) dx\, dt \right\| = 0 \tag{6.44}$$

et par conséquent

$$\lim_{h\to 0} \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \int_T \pi_h\varphi(x, t_n) \sum_{i=1}^{d} \mathbf{F}^i_{,x_i}(u_h)\ dx = -\sum_{n=0}^{N} \int_{I^n} \sum_{\forall T \in \mathcal{Q}} \int_T \sum_{i=1}^{d} \frac{\partial \varphi}{\partial x_i} \mathbf{F}^i(u)\, dx\, dt\ . \tag{6.45}$$

Considérons mainteneant le deuxième terme de la somme du membre de droite de. (6.40), D'après la remarque 6.2.1 (ii) on obtient

$$\left\| \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \int_T (\pi_h\varphi(x_T, t_n) - \pi_h\varphi(x, t_n)) \sum_{i=1}^{d} \mathbf{F}^i_{,x_i}(u_h^n)\ dx \right\| \leq \Sigma$$

où

$$\Sigma \equiv C \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \int_T \left| \frac{\pi_h \varphi(x_T, t_n) - \pi_h \varphi(x, t_n)}{h} \right| \sum_{M_i, M_j \in T} \left\| u_j^n - u_i^n \right\| \, dx \ . \tag{6.46}$$

Puisque $||\nabla_x \pi_h \varphi||$ est supposé borné par une constante, on obtient

$$\int_T \left| \frac{\pi_h \varphi(x_T, t_n) - \pi_h \varphi(x, t_n)}{h} \right| \, dx = \int_T \left| \nabla_x \left(\pi_h \varphi\right) \cdot \frac{x_T - x}{h} \right| \, dx \leq C \, h^d$$

où C est indépendant de h. En utilisant cette inégalité, on obtient

$$\left\| \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \int_T \left| \pi_h \varphi(x_T, t_n) - \pi_h \varphi(x, t_n) \right| \sum_{i=1}^{d} \mathbf{F}^i_{,x_i}(u_h^n) \, dx \right\| \leq C h^d \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \sum_{M_i, M_j \in T} \left\| u_j^n - u_i^n \right\| \tag{6.47}$$

et donc

$$\lim_{h \to 0} \left\| \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \int_T \left(\pi_h \varphi(x_T, t_n) - \pi_h \varphi(x, t_n) \right) \sum_{i=1}^{d} \mathbf{F}^i_{,x_i}(u_h^n) \, dx \right\| = 0 \tag{6.48}$$

En assemblant les différents termes que l'on vient d'analyser quand $h \to 0$ et en employant le lemme 6.2.2 on achève la preuve du lemme 6.2.4. $\square$

Démonstration. (Théorème 6.2.1) Multiplions l'équation . (6.11) par $\varphi(M_i, t_n) \, |C_i|$, où $\varphi(x, t)$ est une fonction test de $C_0^1 \left(\mathbb{R}^d \times [0, +\infty[\right)$, telle que $\mathrm{supp}(\varphi) \subset \mathcal{Q} \times [0, \tau]$. En sommant sur les indices n et i , on obtient

$$\sum_{n=0}^{N} \sum_{M_i \in \mathcal{T}_h} |C_i| \, \varphi(M_i, t_n) \, \left(u_i^{n+1} - u_i^n\right) + \sum_{n=0}^{N} \Delta t \sum_{M_i \in \mathcal{T}_h} \sum_{T; M_i \in T} \varphi(M_i, t_n) \, \mathbf{\Phi}^n_{i,T} = 0 \ . \tag{6.49}$$

D'après le lemme 6.2.3,

$$\lim_{h \to 0} \sum_{n=0}^{N} \sum_{M_i \in \mathcal{T}_h} |C_i| \, \varphi(M_i, t_n) \, \left(u_i^{n+1} - u_i^n\right) = - \int_{\mathbb{R}^d \times \mathbb{R}^+} \frac{\partial \varphi}{\partial t} \, \mathbf{u}(u_h(x, t)) \, dx - \int_{\mathbb{R}^d} \varphi(x, 0) \, \mathbf{u}(u_0(x)) \, dx \ .$$

Le terme en espace est réécrit comme

$$\begin{aligned} \sum_{n=0}^{N} \Delta t \sum_{M_i \in \mathcal{T}_h} \sum_{T; M_i \in T} \varphi(M_i, t_n) \, \mathbf{\Phi}^n_{i,T} &= \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \sum_{M_i \in T} \pi_h \varphi(x_T, t_n) \, \mathbf{\Phi}^n_{i,T} \\ &+ \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \sum_{M_i \in T} \left(\varphi(M_i, t_n) - \pi_h \varphi(x_T, t_n) \right) \mathbf{\Phi}^n_{i,T} \end{aligned} \tag{6.50}$$

où encore une fois $\pi_h \varphi(x_T, t_n)$ désigne la valeur médiane de la fonction linéaire φ pour t_n constant. En examinant le premier terme du membre de droite de . 6.50, et en employant le lemme 6.2.4, on a

$$\lim_{h \to 0} \sup_h \left\{ \sum_{n=0}^{N} \Delta t \sum_{\forall T \in \mathcal{Q}} \sum_{M_i \in T} \pi_h \varphi(x_T, t_n) \mathbf{\Phi}^n_{i,T} + \int_{\mathbb{R}^d \times \mathbb{R}^+} \sum_{i=1}^{d} \frac{\partial \varphi}{\partial x_i}(x, t) \, \mathbf{F}^i(u_h(x, t)) \, dx \, dt \right\} = 0$$

Ensuite, examinons le second terme de la somme de droite de . 6.50. Puisque $\|D\varphi\|$ est borné et en tenant compte de l'hypothèse (1)

$$||\mathbf{\Phi}_{i,T}^n|| \leq C'\, h^{d-1} \sum_{M_j\in T} ||u_j^n - u_i^n||\,, \tag{6.51}$$

on obtient

$$\begin{aligned} \left|\left|\sum_{n=0}^{N} \Delta t \sum_{\forall T\in\mathcal{Q}} \sum_{M_i\in T} \left(\varphi(M_i,t_n) - \pi_h\varphi(x_T,t_n)\right) \mathbf{\Phi}_{i,T}^n\right|\right| \quad &\leq \quad Ch\sum_{n=0}^{N}\Delta t \sum_{\forall T\in\mathcal{Q}} \sum_{M_i\in T} \|\mathbf{\Phi}_{T,i}^n\| \\ &\leq \quad Ch^d \sum_{n=0}^{N}\Delta t \sum_{\forall T\in\mathcal{Q}} \sum_{M_i,M_j\in T} \|u_j^n - u_i^n\| \end{aligned} \tag{6.52}$$

En conséquence, grâce au lemme 6.2.2, quand $h\to 0$,

$$\left|\left|\sum_{n=0}^{N} \Delta t \sum_{\forall T\in\mathcal{Q}} \sum_{M_i\in T} \left(\varphi(M_i,t_n) - \pi_h\varphi(x_T,t_n)\right) \mathbf{\Phi}_{i,T}^n\right|\right| \to 0. \tag{6.53}$$

Ceci achève la preuve du théorème 6.2.1. □

Quelque exemples pour les problèmes scalaires.

Nous considérons le problème

$$\begin{array}{ll} \vec{\lambda}\cdot\nabla u = 0 & x\in\Omega, t>0 \\ u = g & \text{on } \Gamma_- \end{array} \tag{6.54}$$

où Γ_- est le bord entrant de $\Gamma = \partial\Omega$. Si l'inconnue u est interpolée lin'irement sur chaque élément, le résidu total Φ^T est donné par

$$\Phi^T = \sum_{j=1}^{3} k_j u_j$$

où

$$k_j = \frac{1}{2}\vec{\lambda}\cdot\vec{n_j}.$$

On remarque que $\displaystyle\sum_{j=1}^{3} k_j = 0$. Ici, nous avons identifié les sommets de T avec leus indices $j = 1,2,3$ car il n'y a pas d'ambiguïté possible. De même, on supprime l'indice T dans $\Phi_i^T \equiv \Phi_i$.

Dans la suite, nous posons $\mathbf{F} = u\vec{\lambda}$, si bien que (6.54) se réécrit comme $\operatorname{div}\mathbf{F} = 0$ en tenant compte des conditions aux limites.

6.2.3 Principes de construction

La propriété de résidu et la condition de préservation de la linéarité.

Nous rappelons une condition simple qui permet de garantir que le schéma est effectivement du second ordre. à l'état stationnaire (et uniquement à l'état stationnaire). Cette condition a été introduite dans [38], par exemple. Elle implique aussi que le schéma n'est que du premier ordre dans la phase instationnaire de l'algorithme, c'est à dire tant que l'on n'a pas atteint l'état stationnaire numérique. L'avantage de ceci est que le schéma est nécessairement robuste : tant qu'on est loin de l état stationnaire, les résultats ne sont que du premier ordre en espace. Par contre l'inconvénient est qu'il faut avoir une *très* bonne convergence de l'algorithme, au moins jusqu'à un résidu local de l'ordre de l'erreur du schéma.

A l'état stationnaire, le schéma (1.3) vérifie

$$\text{quelque soit } M_i, \sum_{T, M_i \in T} \Phi_i^T = 0.$$

Ainsi, pour toute fonction test à support compact $\varphi \in C^1(\mathbb{R}^2)^4$, nous avons

$$\sum_{M_i} \varphi_i \cdot \left(\sum_{T, M_i \in T} \Phi_i^T \right) = 0.$$

En posant $\varphi_G^T = \frac{\varphi_1+\varphi_2+\varphi_3}{3}$, on obtient, grâce à (2.1)

$$\sum_T \varphi_G^T \int_T \text{div } \mathbf{f}^h dxdy + \sum_T \sum_{M_i \in T} \left(\varphi_i - \varphi_G^T\right) \ . \ \Phi_i^T = 0,$$

c'est à dire, en utilisant la linéarité de φ^h,

$$\int_\Omega \varphi^h \text{ div } \mathbf{f}^h dxdy + \sum_T \sum_{M_i \in T} \left(\varphi_i - \varphi_G^T\right) \ . \ \Phi_i^T = 0. \tag{6.55}$$

Si l'interpolant $\mathbf{f}^h$ est du second ordre, on remarque que

$$\int_\Omega \varphi^h \text{ div } \mathbf{f}^h dxdy = \mathcal{O}(h^2)$$

si le maillage est assez régulier, la solution régulière et φ est C^1 à support compact. Dans ce cas,

$$\begin{aligned}
\int_\Omega \varphi^h \text{ div } \mathbf{f}^h dxdy &= \int_\Omega \varphi^h \text{ div } \left(\mathbf{f}^h - \mathbf{f}\right) dxdy \\
&= \int_\Omega \nabla \varphi^h \cdot \left(\mathbf{f}^h - \mathbf{f}\right) dxdy \\
&= \mathcal{O}(h^2).
\end{aligned}$$

Ici, h représente le maximum des diamètres externes des triangles T. Ceci montre que, si on fait l'hypothèse que l'interpolation du flux est du second ordre, on peut obtenir l'ordre deux à l'état stationnaire si l'équation équivalente (6.55) du schéma satisfait

$$\sum_{M_i} \left(\varphi_i - \varphi_G^T\right) \ . \ \left(\sum_{T, M_i \in T} \Phi_i^T \right) = \mathcal{O}(h^2) \tag{6.56}$$

quand les résidus sont évalués pour une solution exacte régulère.

Une manière d'obtenir (6.56) est d'avoir, pour une solution régulière quelconque de (6.10), $\Phi_i^T(W) = \mathcal{O}(h^3)$ grâce à (6.55). Ceci est évident d'après (6.55) car (i) le nombre de sommets dans un domaine borné est $\mathcal{O}(h^{-2})$ dans le cas d'un maillage régulier, (ii) l'équation (6.56) nécessite $(\varphi_i - \varphi_G)\Phi_i^T = \mathcal{O}(h^4)$ ce qui est vrai car on sait que $||\varphi_i - \varphi_G|| \leq ||\nabla\varphi^h||_\infty h$ and $||\nabla\varphi^h||_\infty$est borné pour toute fonction test φ.

Cette condition est clairement vraie pour les schémas de type SUPG et le schéma stream–line diffusion car les résidus s'écrivent

$$\Phi_i^T = \beta_i^T \Phi^T$$

où β_i^T est uniformément borné.

De plus, si W est une solution régulière de (6.10), nous avons $\Phi^T = \displaystyle\int_T \text{div } \mathbf{f}^h(W)dx = \mathcal{O}(h^3)$. In fact, we have

$$\begin{aligned} \int_T \text{div } \mathbf{f}^h(W)dx &= \int_T \text{div } \Big(\mathbf{f}^h(W) - \mathbf{f}(W)\Big)\, dx \\ &= \int_{\partial T} \Big(\mathbf{f}^h(W) - \mathbf{f}(W)\Big) \cdot \vec{n} dl \end{aligned}$$

En supposant que $\mathbf{f}^h$ est une approximation du second ordre de $\mathbf{f}$, et s'achant que la longueur ∂T est $\mathcal{O}(h)$, on obtient

$$\Phi^T = \int_T \text{div } \mathbf{f}^h(W)dx = \mathcal{O}(h^3).$$

Ces arguments montrent clairement que *(i)* ces schémas sont bien du second ordre à l'état staionnaire, et uniquement à l'état stationnaire, *(ii)* l'approximation du flux $\mathbf{f}^h$ doit être du second ordre dans la même norme. La condition

$$\Phi^T \to 0 \text{ implies } \Phi_i^T \to 0 \text{ quelque soit } i \text{ est uniformément}$$

est appelée la *propriété de préservation de la linéarité* (LP en anglais). On la rencontre pour les schémas de type SUPG Les schémas non statbilités possedent une propriété similaire.

Schémas préservant la monotonie.

En pratique, tous les schémas distribuant le résidu que l'on connait s'écrivent sous la forme

$$\Phi_i^T = \sum_{M_j \in T, M_i \neq M_i} c_{ij}^T (u_i - u_j). \tag{6.57}$$

By (1.3) and (6.57), we get

$$u_i^{n+1} = \sum_{M_j \text{voisin de } M_i} d_{ij} u_j^n$$

with

$$\begin{aligned} d_{ii} &= 1 - \frac{\Delta t}{|C_i|} \sum_{T, M_i \in T} \sum_{M_j \in T, M_i \neq M_j} c_{ij}^T \\ d_{ij} &= \sum_{T, M_i, M_j \in T} \frac{\Delta t}{|C_i|} c_{ij}^T \qquad\qquad \text{if } M_i \neq M_j. \end{aligned}$$

Le schéma préserve la monotonie si $d_{ij} \geq 0$ (leur somme vaut 1). Cette condition est vraie si

$$
\begin{aligned}
&c_{ij}^T \geq 0 && \text{for all } i, j \\
&\frac{\Delta t}{|C_i|} \sum_{T, M_i \in T} \sum_{M_j \in T, M_i \neq M_j} c_{ij}^T \leq 1.
\end{aligned}
$$

La condition $c_{ij}^T \geq 0$ est appelée la condition de monotonie (un peu imporprement il est vrai). On sait qu'il n'est pas possible d'avoir un schéma dépendant linéairement de la solution (c'est à dire que les coefficients . c_{ij}^T sont indépendants de la solution) qui préserve à la fois la monotonie et la linéarité [38]. Les deux conditions ne peuvent être vraies (préserve linéarité et monotonie) que si le schéma est non linéaire, même pour un problème linéaire. Ce type de condition se rencontre dans le cas des schémas de type TVD. La construction de tels schémas est l'objet du prochain paragraphe.

Le théorème de Godunov

En une dimension d'espace, il est bien connu que les schémas linéaires monotones sont au plus du premier ordre. Ce résultat, connu comme le thérème de Godunov, a un prolongement pour les schémas multidimensionnels sur maillage non structuré. On a le

Théorème 6.2.2 (Godunov). *Considérons le problème de convection $\vec{\lambda} \cdot \nabla u = 0$ avec des conditions aux limites sur le bord entrant. On suppose que $\vec{\lambda}$ est constant. Un schéma préservant la monotonie de la forme (6.57) où les coefficients C_{ij} sont indépendants des $(u_j)_{j \in T}$ ne peut pas préserver la linéarité.*

Démonstration. Considérons un schéma linéaire distribuant le résidu de la forme, qui sur chaque triangle T est de la forme

$$\Phi_i^T = \sum_{j=1}^{3} c_{ij}(u_i - u_j).$$

On suppose que le schéma est linéaire et préserve la monotonie : $c_{ij} \geq 0$. Puisqu'il préserve la monotonie,

$$\sum_{j=1}^{3} k_j u_i = 0 \rightarrow \sum_{j=1}^{3} c_{ij}(u_i - u_j) = 0 \text{ for } i = 1, 2, 3.$$

Ainsi, le vecteur (u_1, u_2, u_3) appartient au noyau des trois formes linéaires

$$\Phi(u_1, u_2, u_3) = \sum_{j=1}^{3} k_j u_i = 0,$$

$$\Phi_i^T(u_1, u_2, u_3) = \sum_{j=1}^{3} c_{ij}(u_i - u_j) \quad i = 1, 2, 3.$$

De plus, Ker $\Phi_i^T \subset$ Ker $\Phi \subset \mathbb{R}^3$. Puisque ces espaces sont de dimension deux, ils sont égaux. Il existe donc des coefficients α_i tels que

$$\Phi_i^T = \alpha_i \Phi.$$

Ainsi tous les coefficients c_{ij} ne peuvent pas être positifs car $\sum_j k_j = 0$. ☐

6.2.4 Construction de schémas du second ordre, exemples.

On commence par donner trois exemples de schémas préservant la monotonie. Le premier est le schéma de Rusanov, ou plutôt son extension au distributif,

$$\Phi_i^T = \frac{1}{3}\Big(\Phi - \alpha \sum_{j\neq i}(u_i - u_j)\Big) \tag{6.58}$$

avec $\alpha \geq \max_i |k_i|$. Clairement, $\sum\limits_j \Phi_j^T = \Phi$: la propriété de conservation est garantie.

Un autre exemple est le schéma N (pour Narrow) de Roe [38]. Il s'écrit comme

$$\Phi_i^T = k_i^+ \big(u_i - \widetilde{u}\big) \tag{6.59}$$

où $\widetilde{u}$ est construit afin de garantir la propriété de conservation, c'est à dire

$$\widetilde{u} = \big(\sum_j k_j^-\big)^{-1}\big(\sum_j k_j^- u_j\big).$$

Le scalaire $n := \big(\sum\limits_j k_j^-\Big)^{-1}$ est toujours défini à moins que $\vec{\lambda} = 0$[4]. Ce schéma peut être vu comme une méthode des caractéristiques conservative. Il est monotone sous une condition de type CFL car

$$\Phi_i^T = \sum_j k_i^+ n k_j^- (u_i - u_j),$$

et donc $c_{ij} = k_i^+ n k_j^- \geq 0$.

Un dernier exemple est donné par le très classique schéma décentré réécrit sous forme distributive suivant la relation (6.14). Aucun de ces schéma ne préserve la linéarité.

Il est important de construire des schémas préservant à la fois la monotonie et la linéarité. Cependant, d'après le théorème de Godunov, ces schémas sont nécessairement non linéaire même dans le cas d'une EDP linéaire. Il y a plusieurs constructions possibles. On peut citer le schéma PSI de R. Struijs [39] qui fut le premier de ce type, la construction de D. Sidilkover du même schéma [40]. La construction systématique que nous présentons dans ce qui suit est l'une des nombreuses possibles, mais c'est celle qui semble le plus simple.

L'approche par projection

Pour commencer, supposons que Φ_i^T, la fraction de Φ_i^T «envoyé» au i^{e}sommet, s'écrive

$$\Phi_i^T = \sum_{j, M_j\in T} c_{ij}^T (u_i - u_j).$$

Les coefficients c_{ij}^T mpeuvent être indépendants des données (lschémas linéairess) ou peuvent en dépendre (schémas non linéaires). Le schéma préservera la monotonie si $c_{ij} \geq 0$ sous la condition

[4] Comme pour chaque i, $\Phi_i \to 0$ uniformément si $\vec{\lambda} \to 0$, il n'y a pas de problème de définition dans le cas $\vec{\lambda} = 0$.

de type CFL

$$\Delta t \left[\frac{\displaystyle\sum_{T, M_i \in T} \sum_{M_j \in T} c_{ij}^T}{|C_i|} \right] \leq 1 \text{ for any } M_i.$$

est vérifiée.

Il est possible de construire des schémas préservant la monotonie pour des problèmes stationnaires mais ceux–ci sont nécessairement non linéaires. Une manière systématique de procéder est de partir d'un schémas monotone, linéaire et donc du premier ordre, puis de le modifier afin d'obtenir un schéma d'ordre élevé.

Pour simplifier les notations, toutes les sommations porteront sur les indices des noeuds d'un triangle donné ou concerneront les degrés de liberté de l'élément considéré. De même, on omettra l'exposant T dans les résidus puisuqe tous les calculs se réfèreront à l'élément générique T sans ambiguïté possible.

On dénote par Φ_i^M la fraction du résidu total envoyé au sommet i au moye d'un schéma monotone, et par Φ_i^H la fraction envoyée par un schéma non linéaire à construire. On demande que

$$\begin{array}{ll} \displaystyle\sum_j \Phi_j^H = \sum_j \Phi_j^M = \Phi & \text{Conservation} \\ \Phi_j^H \Phi_j^M \geq 0 & \text{Préservation de la monotonie.} \end{array}$$

On obtient un schéma du second ordre si

$$\begin{array}{ll} \Phi_j^H = \mathcal{O}(\Phi) & \\ \text{and} & \\ \Phi = \mathcal{O}(h^3) & \text{à l'état stationnaire pour une solution régulière.} \end{array}$$

Deux conditions se révèlent importantes d'après l'analyse du paragraphe 6.2.3 ; en particulier, il faut que l'approximation du flux soit du second ordre. Sans cela, il n'ya aucun espoir pour attendre l'ordre deux.

On introduit

$$\beta_j = \frac{\phi_j^M}{\phi}, \qquad \hat{\beta}_j = \frac{\phi_j^H}{\phi}$$

Les conditions mentionnées plus haut peuvent être reformulée comme des contraintes sur les poids β_j à savoir

$$\begin{array}{ll} \displaystyle\sum_{j=1}^N \beta_j = \sum_{j=1}^N \hat{\beta}_j = 1 & \text{Conservation} \\ \beta_j \hat{\beta}_j \geq 0 & \text{Prservationdelamonotonie} \\ \hat{\beta}_j \text{ est borné} & \text{Ordre élevé si } \phi^T \text{ est suffisament petit.} \end{array}$$

Notre technique impose trois contraintes sur une application qui envoie les poids $\{\beta_j\}$ correspondant à un schéma monotone (qui ne peut qu'être du premier ordre) sur les trois poids $\{\hat{\beta}_j\}$ correspondant à un schéma qui préserve à la fois la monotonie et qui est d'ordre maximal. Cette méthode peut être appliquée à un nombre arbitraire N de degrés de liberté par triangle. Cette application envoie le vecteur de $\mathbb{R}^N$ $\{\beta_j\}$ sur le vecteur $\{\hat{\beta}_j\} \in \mathbb{R}^N$ et ne peut pas être linéaire en vertue du théorème de Godunov. Il existe de nombreuses applications de ce type, ce sont les analogues véritablement multidimensionnels des limiteurs classiques décrits dans [2] par exemple.

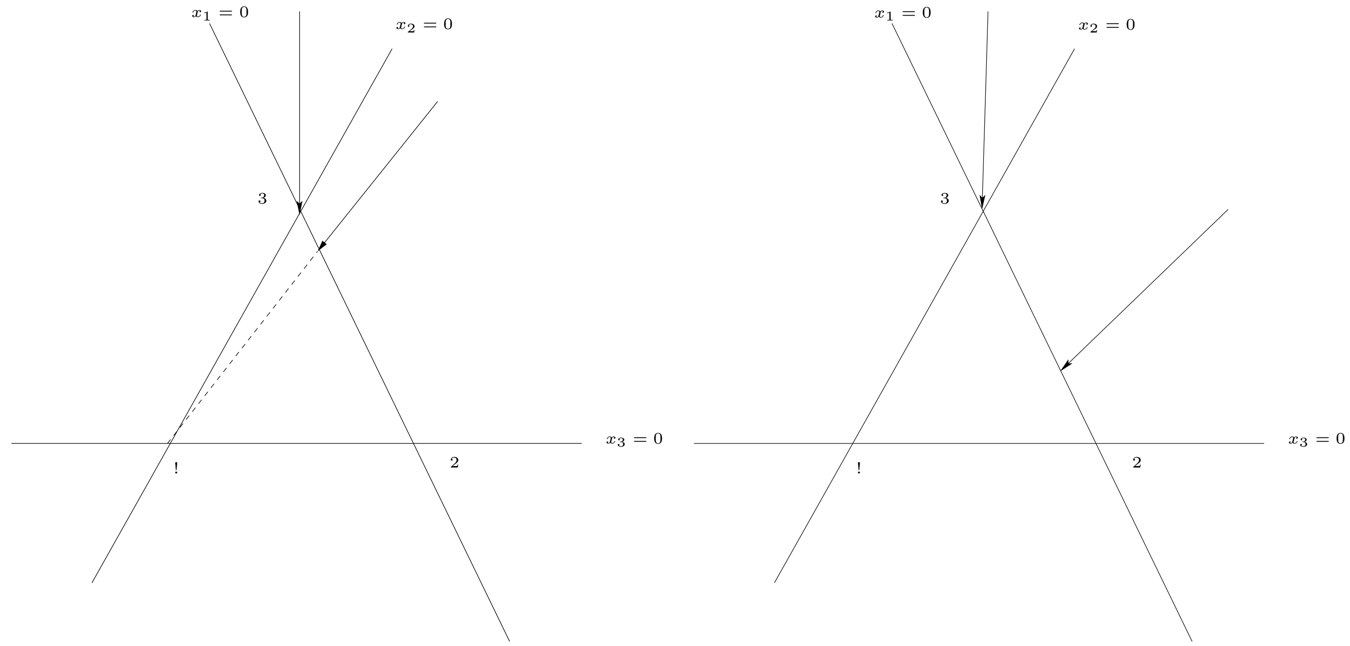
Ici, on décrit simplement la technique et deux exemples dans le cas $N = 3$. L'idée est d'interpréter les vecteurs $\beta = (\beta_1, \beta_2, \beta_3)$ et $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ comme les coordonnées barycentriques par rapport à un triangle de référence qui est équilatéral[5] d'un point d'un plan abstrait.

Il faut que les poids β_j soient bornés, on va imposer que, quelque soit j, $0 \leq \hat{\beta}_j \leq 1$: le point β appartient au triangle de référence ou sur son bord. Une contrainte plus faible est d'imposer quei β appartient à un voisinage du triangle de référence.

Si les poids du schéma monotones appartiennent au triangle de référence, alors il est naturel de prendre simplement $\hat{\beta} = \beta$. Si β n'apprtient pas au triangle de référence, une solution est de «projeter» β sur le bord du triangle. Par exemple, on peut prendre

$$\hat{\beta}_j = \frac{\beta_j^+}{\sum_j \beta_j^+} \qquad \text{with } x^+ = \max(0, x) \tag{6.60}$$

La fonction définie par (6.60) est représentée sur la figure 6.4 à gauche.



Limitation par (6.60) Limitation par projection orthogonale

FIG. 6.4 –

Une autre solution au problème est de choisir $\hat{\beta}$ comme étant le point du bord du triangle de référence qui est le plus proche de β. Ceci est réalisé analytiquement par

1. Si β_1, β_2, β_3 sont positifs, on pose $\hat{\beta}_j = \beta_j$,
2. sinon

[5]On choisit un triangle équilatéral afin que tous les points de T jouent le même rôle.

(a) si β_1 est négatif, on pose $\beta_2^* = \beta_2 + \frac{\beta_1}{2}$ et $\beta_3^* = \beta_3 + \frac{\beta_1}{2}$

   i. si $\beta_2^* \leq 0$, $\hat{\beta}_1 = \hat{\beta}_2 = 0$, $\hat{\beta}_3 = 1$

   ii. si $\beta_3^* \leq 0$, $\hat{\beta}_1 = \hat{\beta}_3 = 0$, $\hat{\beta}_2 = 1$

   iii. si $\beta_2^* \geq 0$ et $\beta_3^* \geq 0$, $\hat{\beta}_1 = 0$, $\hat{\beta}_2 = \beta_2^*$ et $\hat{\beta}_3 = \beta_3^*$

(b) Si β_2 est négatif, on applique la logique précédente en changeant les indices selon $1 \rightarrow 2$, $2 \rightarrow 1$, $3 \rightarrow 3$.

(c) Si β_2 est négatif, on applique la logique précédente en changeant les indices selon $1 \rightarrow 3$, $2 \rightarrow 2$, $3 \rightarrow 1$

Cette solution est représentée sur la figure 6.4 à droite.

On remarque que comme $\hat{\beta}_j$ est homogène de degré 1 en x_j, ces fonctions peuvent être programmées sans division.

Décrivons une autre solution pour laquelle la zone du plan reste invariante contient strictement le triangle de référence. On considère $a > 0$. Dans la figure 6.5, les droites C et D sont les bisectrices des angles qui les contiennent. En utilisant les coordonnées barycentriques, un point de C est définit par $x_2 = x_1$, un point de D par $x_2 = x_3$. La zone invariante est définie par la réunion des zones

- $x_1, x_2, x_3 \geq 0$ (triangle de référence),
- $-a \leq x_1 \leq 0$ et $x_2, x_3 \geq 0$,
- $-a \leq x_2 \leq 0$ et $x_1, x_3 \geq 0$,
- $-a \leq x_3 \leq 0$ et $x_2, x_1 \geq 0$,
- $x_1 \leq 0$, $x_2 \leq 0$ et $0 \leq x_3 \leq 1 + a$,
- $x_2 \leq 0$, $x_3 \leq 0$ et $0 \leq x_1 \leq 1 + a$,
- $x_1 \leq 0$, $x_3 \leq 0$ et $0 \leq x_2 \leq 1 + a$.

Le limiteur est alors défini par

1. Si on est dans la zone hachurée de la figure 6.5, on pose $\hat{\beta} = \beta$.
2. Si $x_2, x_3 \geq 0$ et $x_1 \leq 1 + a$,

   (a) si $x_3 \leq x_2$,

      i. si $x_3 \leq \frac{1+a}{2}$, $\hat{\beta} == -a, 1 - x_3 + a, x_3)$: cela correspond au point B de la figure 6.5,

      ii. sinon $\hat{\beta} = (-a, \frac{1+a}{2}, \frac{1+a}{2})$.

   (b) si $x_2 \leq x_3$,

      i. si $x_2 \leq \frac{1+a}{2}$, $\hat{\beta} = (-a, x_2, 1 - x_2 + a)$.

      ii. sinon $\hat{\beta} = (-a, \frac{1+a}{2}, \frac{1+a}{2})$. Cela correspond au point A de la figure.

3. si $x_1, x_2 \leq 0$ et $x_3 > 1 + a$,

   (a) si $x_1 \leq x_2$,

      i. si $x_2 \geq -\frac{a}{2}$, $\hat{\beta} = (-(a + x_2), x_2, 1 + a)$,

      ii. sinon $\hat{\beta} = (-\frac{a}{2} - \frac{a}{2}, 1 + a)$,

   (b) si $x_1 > x_2$,

      i. si $x_1 \geq -\frac{a}{2}$, $\hat{\beta} = (x_1, -(a + x_1), 1 + a)$,

      ii. sinon $\hat{\beta} = (-\frac{a}{2}, -\frac{a}{2}, 1 + a)$

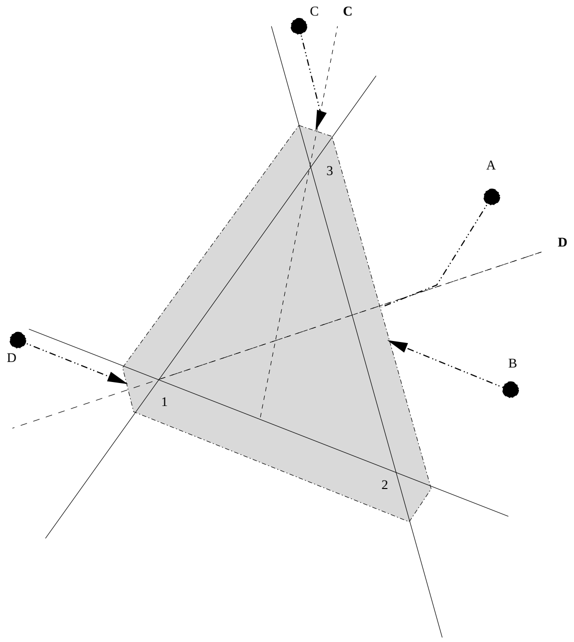


FIG. 6.5 – Illustration du limiteur où la zone invariante contient strictement le triangle de référence. La zone invariante est représentée en grisé.

6.2.5 Exemples d'applications

On applique les constructions qui précèdent au système suivant

$$\begin{array}{ll} \dfrac{1}{2}\dfrac{\partial u^2}{\partial x} + \dfrac{\partial u}{\partial y} = 0 & (x,y) \in [0,1]^2 \\ u(x,0) = 1.5 - x & u(0,y) = 1.5 \\ u(1,y) = 0.5 & \end{array} \tag{6.61}$$

La solution est donnée par

$$u(x,y) = \begin{cases} -0.5 & \text{si } y \geq 0.5 \text{ et } -2(x - 3\frac{3}{4}) + (y - \frac{1}{2}) \leq 0 \\ 1.5 & \text{si } y \geq 0.5 \text{ et } -2(x - 3\frac{3}{4}) + (y - \frac{1}{2}) > 0 \\ \max\left(-0.5, \min\left(1.5, \dfrac{x - 3/4}{y - 1/2}\right)\right) & \text{si } y < 0.5. \end{cases} \tag{6.62}$$

Linéarisation conservative.

L'équation aux dérivées partielle (EDP) (6.61) est non linéaire. La plupard des schémas décrits plus haut se généralise sans difficulté au cas non linéaire. Dans le cas du schéma de Rusanov, c'est clair. Les schémas de type volume fini sont définis classiquement et réinterprétés de manière distributive. Seul le cas du schéma N est un peu délicat, nous décrivons la procédure en détail.

La définition du schéma N fait intervenir de manière essentielle une vitesse de convection qui est employée pour déterminer si le triangle est une cible ou deux cibles. De plus cette vitesse intervient pour définir le résidu total. Le lien entre (6.61) et une EDP linéaire s'effectue au mouen d'une linéarisation conservative [39].

Il faut déterminer pour chaque triangle une moyenne $\bar{u}$ telle que

$$\int_T \left(\frac{1}{2}\frac{\partial u^2}{\partial x} + \frac{\partial u}{\partial y}\right) dxdy = \bar{u} \int_T \frac{\partial u}{\partial x} dxdy + \int_T \frac{\partial u}{\partial y} dxdy. \tag{6.63}$$

Dans le membre de gauche de (6.63), on interpole u linéairement. Une solution évidente est $\bar{u} = \frac{1}{3} \sum_{j=1,3} u_{i_j}$. Dans le cas d'un problème plus général comme

$$\frac{\partial f(u)}{\partial x} + \frac{\partial g(u)}{\partial y} = 0$$

où les flux f et g sont C^1 au moins, on cherche $\overline{\lambda}_x$ et $\overline{\lambda}_y$ tels que

$$\int_T \left(\frac{\partial f(u^h)}{\partial x} + \frac{\partial g(u^h)}{\partial y}\right) dxdy = \int_T \left(f'(u^h)\frac{\partial u^h}{\partial x} + g'(u^h)\frac{\partial u^h}{\partial y}\right) dxdy = \overline{\lambda}_x \int_T \frac{\partial u^h}{\partial x} dxdy + \overline{\lambda}_y \int_T \frac{\partial u^h}{\partial y} dxdy$$

dont une solution évidente est, puisque u^h est linéaire donc ∇u^h est constant sur T,

$$\begin{aligned} \overline{\lambda}_x &= \frac{1}{|T|} \int_T f'(u^h) dxdy \\ \overline{\lambda}_y &= \frac{1}{|T|} \int_T g'(u^h) dxdy. \end{aligned} \tag{6.64}$$

Si les intégrales doivent être calculées au moyen d'une formule de quadrature, il est fondamental que les formules de quadratures soient consitante avec une intégrale sur le bord, autrement dit, si on a

$$\overline{\lambda}_x = \frac{1}{|T|} \overline{\int}_T f'(u^h) dxdy \text{ et } \overline{\lambda}_y = \frac{1}{|T|} \overline{\int}_T g'(u^h) dxdy,$$

il faut que

$$\overline{\lambda}_x \int_T \frac{\partial u^h}{\partial x} dxdy + \overline{\lambda}_y \int_T \frac{\partial u^h}{\partial y} dxdy = \overline{\overline{\int}}_{\partial T} \big(f(u^h) n_x + g(u^h) n_y\big) d\partial T$$

où $\overline{\int}$ et $\overline{\overline{\int}}$ représente des formules de quadrature numérique. L'origine de cette difficulté est à rechercher dans l'hypothèse (1), (6.16) du théorème 6.2.1 puisqu'il faut que l'approximation des flux soit continue au travers des bords des triangles.

Il n'est pas toujours possible de trouver de telles formules de quadratures consistantes. En effet, le terme

$$\overline{\overline{\int}}_{\partial T} \big(f(u^h) n_x + g(u^h) n_y\big) d\partial T$$

ne fait intervenir que des points du bord de T, tandis qu'*a priori*, les termes

$$\overline{\int}_T f'(u^h) dxdy \text{ et } \overline{\int}_T g'(u^h) dxdy$$

feront nécessairement intervenir des points intérieurs à T si on veut que les formules soient exactes pour des polynômes de degré 1. En effet, on sait que pour obtenir un schéma d'ordre deux, une condition naturelle est que

$$\overline{\overline{\int}}_{\partial T} \big(f(u^h)n_x + g(u^h)n_y\big)d\partial T = \mathcal{O}(h^3)$$

quand on interpole une solution exacte suffisament régulière.

Afin de tourner cette difficulté, Csik et al ont, dans [41], proposé une alternative aux formules (6.64). On considère un état moyen $\bar{u}$, par exemple

$$\bar{u} = \frac{u_1 + u_2 + u + 3}{3},$$

et on définit

$$\begin{aligned}\overline{\lambda}_x &= f'(\bar{u})\\ \overline{\lambda}_y &= g'(\bar{u}).\end{aligned} \tag{6.65}$$

Clairement, la relation

$$\int_T \left(\frac{\partial f(u^h)}{\partial x} + \frac{\partial g(u^h)}{\partial y}\right)dxdy = \overline{\lambda}_x \int_T \frac{\partial u^h}{\partial x}dxdy + \overline{\lambda}_y \int_T \frac{\partial u^h}{\partial y}dxdy$$

est violée. Ainsi, si on définit les résidus du schéma N par

$$\Phi_i = k_i + (u_i - \widetilde{u}) \tag{6.66a}$$

avec

$$\widetilde{u} = n\Big(\sum_j k_j^- u_j\Big)$$

et

$$k_i = \overline{\lambda}_x n_x^i + \overline{\lambda}_y n_y^i, \tag{6.66b}$$

la relation de conservation est violée.

L'astuce de [41] est de définir $\widetilde{u}$ par

$$\widetilde{u} = n\left(\sum_{j=1}^{3} k_j^- u_j - \Phi_T\right) \tag{6.66c}$$

où

$$\Phi_T = \int_{\partial T} \Big(f(u^h)n_x + g(u^h)n_y\Big)d\partial T \tag{6.66d}$$

est calculé par une formule de quadrature qui ne fait donc intervenir que des points du bord. Ainsi les hypothèses du théorème 6.2.1 peuvent être adaptées afin de garantir la conservation. Le schéma N de [41] est défini par les relations (6.66a), (6.66b), (6.66c) et (6.66d). Il permet de de résoudre complètement le problème de la conservation, mais le schéma construit n'est pas rigoureusement positif. Cependant, toutes les applications numériques qui ont été réalisées avec ce schéma montrent son caractère parfaitement non oscillant.

Dans [31], on présente un schéma de type N qui lui préserve la monotonie par construction mais qui n'est pas rigoureusement conservatif. Il est obtenu en effectuant une quadrature numérique de (6.64) (avec des poids positifs). La conservation n'est obtenue dans la limite d'un nombre de point de quadrature grand. Dans la pratique, on observe que le nombre de point de quadrature nécessaire pour une capture correcte des discontinuités est relativement modéré. La solution de [41] semble néanmoins supérieure car plus simple et très efficace.

Discrétisation

Nous revenons au problème (6.61). Le résidu total Φ_T

$$\Phi = \bar{u} \int_T \frac{\partial u}{\partial x} dxdy + \int_T \frac{\partial u}{\partial y} dxdy$$

est est distribué aux trois sommets d eT au moyen des schémas de Rusanov, du schéma N ou du schéma décentré monodimensionnel d'ordre un. Plus précisément, nous considèrerons deux version de ce schéma. La première emploie le schéma de Murman–Roe adapté à (6.61) et réécrit sous forme distributive. La seconde version est construite à partir de la linéarisation (6.63). Nous représentons le maillage sur la figure 6.6 ainsi que la solution obtenue par un schéma du second ordre sur maillage non structuré de type ENO décrit dans [42]. Dans la suite, on notera par L–Rusanov (resp. L–N[6],

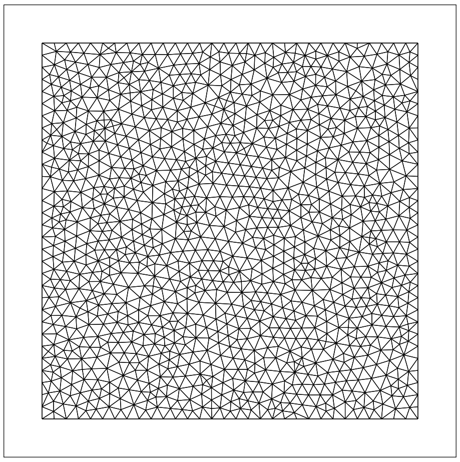


FIG. 6.6 – Mesh

L–Roe–1D et L–Roe-2D) le schéma construit à partir du schéma de Rusanov (resp. le schéma N, Roe–1D et Roe–2D) en utilsant le limiteur défini par (6.60). Les autres limiteurs décrits dans le paragraphe 6.2.4 ont aussi été utilisés sans différences notables sur la qualité des résultats et le comportement du schéma. Le solutions sont représentées sur les figures (6.7) et (6.8).

Différentes sections des solutions sont représentées : au travers de la discontinuité pour la figure 6.9 et au travers de la détente pour la figure 6.10.

On voit que la qualité des résultats est *toujours* meilleure pour les schémas distribuant le résidu quand on les compare au schéma ENO du second ordre. Ceci est une conséquence de la propriété

[6]Habituellement, on le denomme par schéma PSI. On a pris la dénomination L–N pour avoir des notations cohérentes.

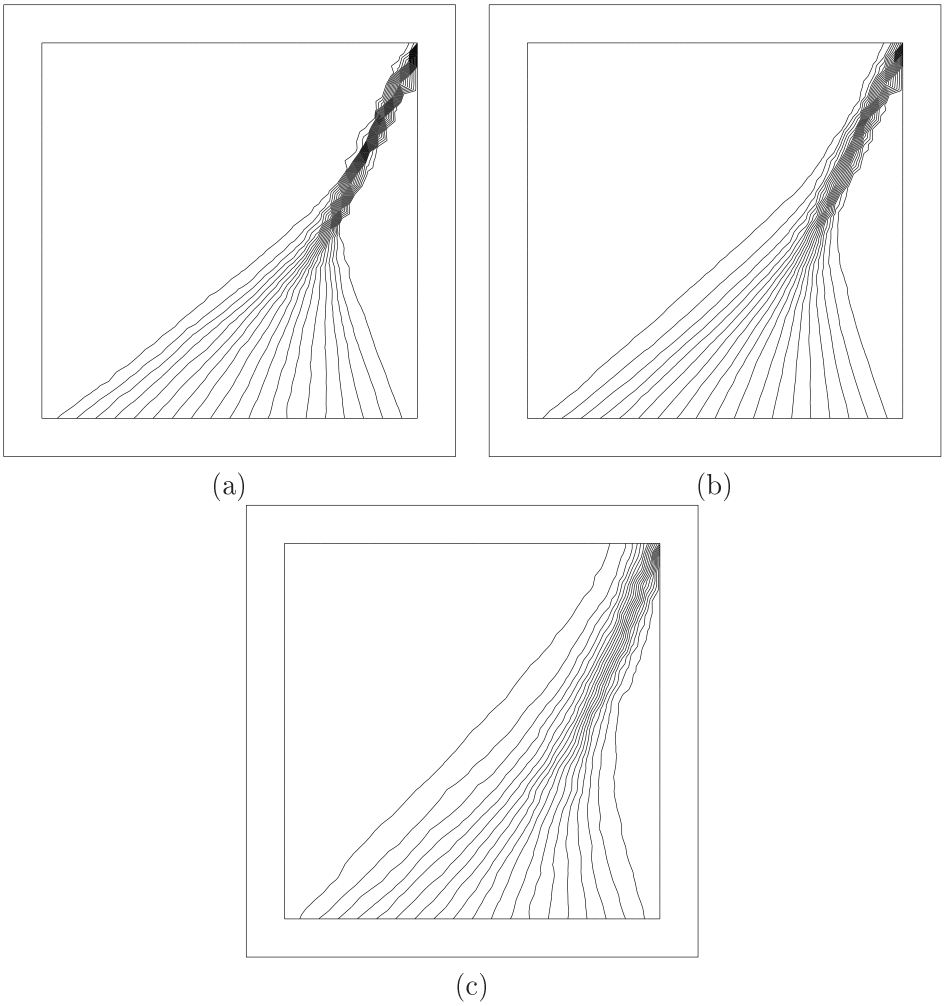
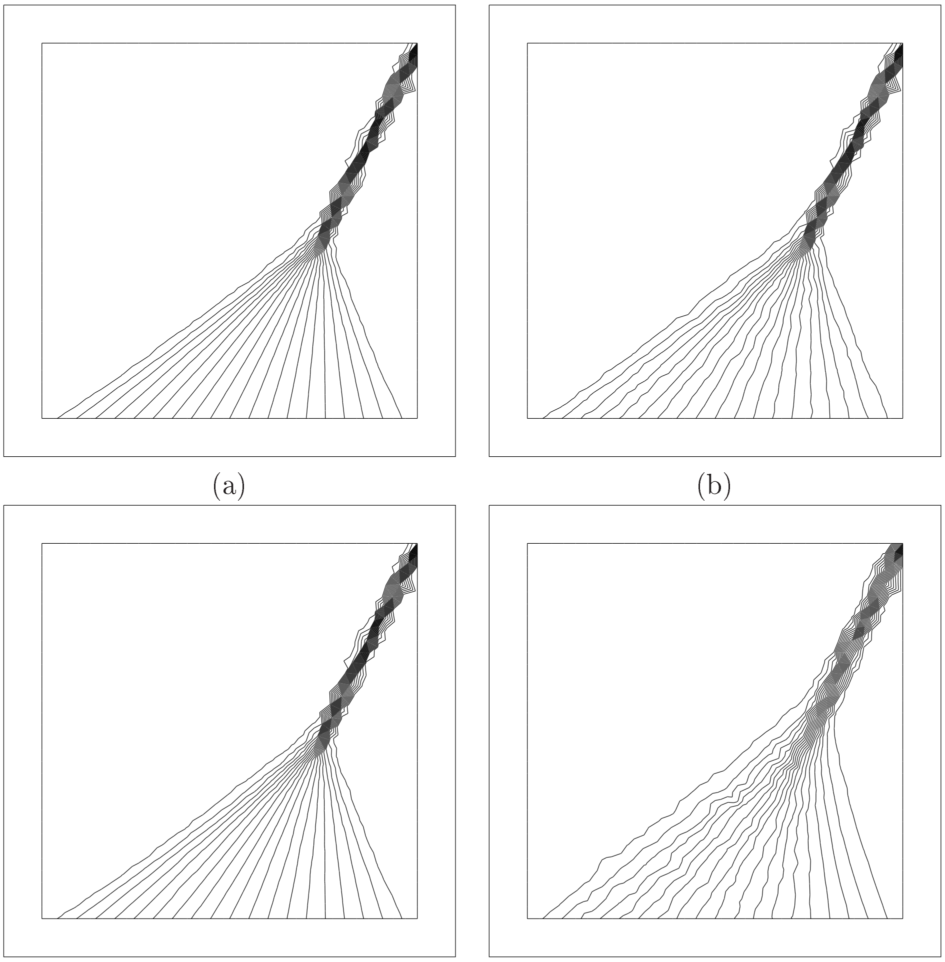
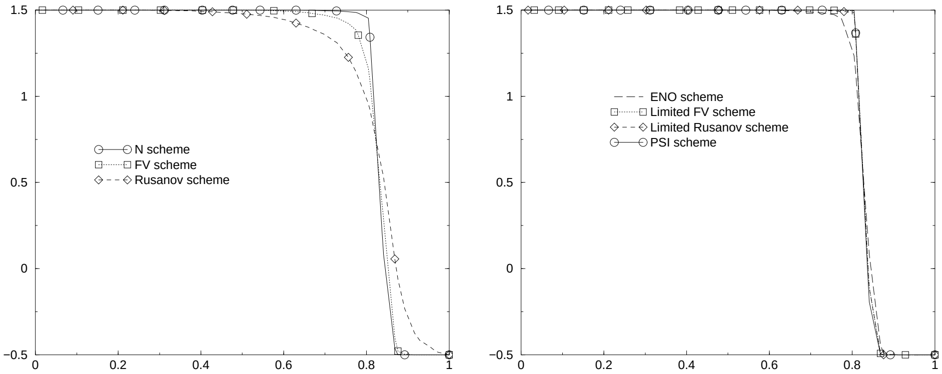


FIG. 6.7 – (a) : Schéma N, (b) Schéma décentré, (c) Schéma de Rusanov



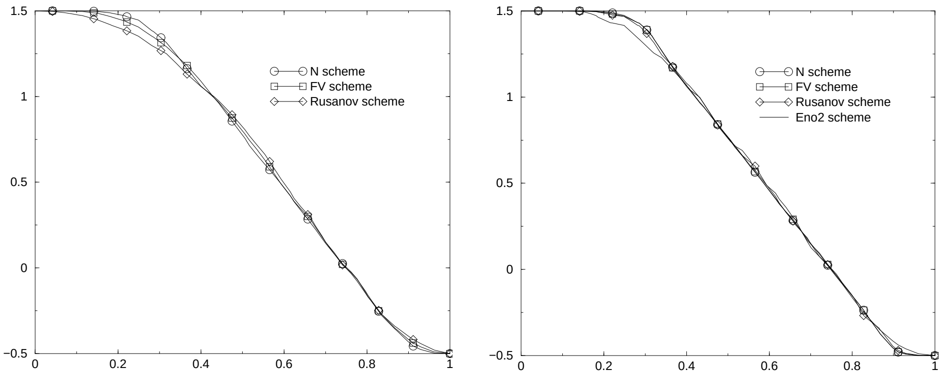
(a) (b) (c) (d)

Fig. 6.8 – schémas (a) L–N, (b) : L–Rusanov , (c) L–Upwind, (d) ENO du second ordre.



Schémas du premier ordre du second ordre

Fig. 6.9 – Coupe à travers le choc.



Schémas du premier ordre — du second ordre

FIG. 6.10 – Coupe à travers la détente.

de résidu qui est partagée par tous les schémas du second ordre distribuant le résidu. Cependant, l'étude de ces figure montre aussi une hierarchie parmi ces schémas. Le schéma L–N donne les meilleurs résultats, les résultats sont pratiquement similaires pour les schémas décentrés. Le schéma L–Rusanov donne les moins bon résultatts, surtout dans la détente. On note que dans tous les cas, le choc est résolu en 1 point.

6.2.6 Remèdes aux difficultés

Revenons au schéma L–Rusanov. En étudiant la figure 6.8, on voit que les résultats sont un peu oscillants dans la détente. Ce n'est pas un problème de stabilité numérique, puisque le schéma est stable en norme L^∞. En étudiant la figure 6.10, on voit que au lieu d'avoir un comportement linéaire à y constant (voir (6.62)), on a une série de plateaux. C'est cela qui dégrade la qualité de la solution. Une étude de convergence en maillage montrerait que la solution est en fait d'ordre 1 alors qu'elle est quasiment d'ordre 2 pour les autres schémas. Enfin, la solution est obtenue par une méthode itérative (schéma d'Euler) qui ne converge pas. Après un bon départ, le résidu itératif stagne. Ceci indique que la méthode de limitation est probablement trop compressive. Nous revenons sur ce problème dans le paragraphe 6.2.6.

6.2.7 Quadrangles

6.2.8 Convergence du schéma

Introduction

Soit $\mathbf{f} \;:\mathbb{R} \rightarrow \mathbb{R}^N$ une fonction de classe C^1 au moins. On recherche la solution entropique $u \;: \mathbb{R}^N \times \mathbb{R}^+ \rightarrow \mathbb{R}$ du problème

$$\begin{aligned} &\frac{\partial u}{\partial t} + \operatorname{div} \mathbf{f}(u) = 0 \quad x \in \mathbb{R}^N, t > 0 \\ &u(x,0) = u_0(x). \end{aligned} \tag{6.67}$$

On considère $\{\mathcal{T}_h\}_{h\in\mathbb{R}^+}$ une famille de triangulation régulière de $\mathbb{R}^N$. Les nœuds de $\mathcal{T}_h$ sont notés $\{M_i\}_{i=1,\dots,n_s}$ et ses éléments sont $\{T_j\}_{j=1,\dots,n_e}$. On note $\mathcal{V}(i)$ l'ensemble des voisins M_j de M_i pour lesquels $[M_i, M_j]$ est un segment d e$\mathcal{T}_h$. Les cellules du maillage dual sont notées $\mathcal{C}_i$, $i = 1, \dots, n_s$. Pour simplifier l'exposé, on supposera que les éléments de $\mathcal{T}_h$ sont des triangles, que $N = 2$ et on identifiera le nœud M_j et son indice j.

On considère alors le schéma suivant

$$\begin{aligned} &|\mathcal{C}_i| \frac{u_i^{n+1} - u_i^n}{\Delta t} + \sum_{T \ni M_i} \Phi_i^T = 0 \\ &u_i^0 = \inf_{x \in \mathcal{C}_i} u_0(x). \end{aligned} \tag{6.68}$$

Les résidus Φ_i^T dépendent de u_i^n et des valeurs de u^n dans un voisinage de M_i. Ils vérifient les relations de conservation

$$\sum_{j \in T} \Phi_j^T = \int_{\partial T} \left[\mathbf{f}(u_h^n)\right]^h d\partial T \tag{6.69}$$

où u_h^n désigne l'interpolant des données $\{u_j^n\}_{j,n}$ définit par

$$u_h(\mathbf{x}, t) = u_h^n(x), \qquad t \in [t_n, t_{n+1}[$$

et u_h^n représente l'interpolant P_1 de $\{u_j^n\}_j$. On notera aussi par φ_i la fonction de forme P_1 au sommet M_i.

Ici, on supposera, et c'est toujours le cas dans la pratique, que Φ_i^T est la forme

$$\Phi_i^T = \sum_{j \in T} c_{ij}^T(u_h)(u_j - u_i). \tag{6.70}$$

On voit facilement que la forme (6.70) combinée à la relation

$$\sum_{i,j \in T} \big(c_{ij}(u_h) - c_{ji}(u_h)\big) = \int_T \nabla_u \mathbf{f}(u_h) \cdot \nabla \varphi_i d\mathbf{x} \tag{6.71}$$

entraîne la relation de conservation (6.69). On supposera enfin que les coefficients c_{ij} vérifient

$$c_{ij}(u_h) \geq 0. \tag{6.72}$$

L'originalité par rapport à ce qui se fait pour l'étude des schémas volumes finis monotone est double. D'une part, on ne peut pas, en général, écrire le schéma (6.68) sous la forme

$$|\mathcal{C}_i| \frac{u_i^{n+1} - u_i^n}{\Delta t} + \sum_{j \in \mathcal{V}(i)} \mathcal{F}(u_i^n, u_j^n, \vec{n}_{ij}) = 0$$

où u_j^n est l'état voisin de u_i^n de l'autre coté de l'interface de normale $\vec{n}_{ij}$. Dans cette écriture, on fait un bilan de flux sur le pourtour de la cellule duale $\mathcal{C}_i$, et les flux (normaux) dépendent de deux états (les états de part et d'autre de l'interface) et de na normale locale. Une écriture similaire pour le schéma (6.68), c'est à dire un bilan sur le pourtour de la cellule duale, nécessite trois état, voire quatre, par segment d'interface. L'autre originalité est que lorsque l'on fait le bilan d'entropie, celui–ci ne s'exprime pas sous forme de flux. De plus, l'écriture d'un «résidu d'entropie» n'est pas toujours claire. Notons qu'il existe dans la littérature une référence où la convergence d'un schéma distributifs est prouvée. Il s'agit de [43]. La technique employée est complètement différente. Celle de [43] ne s'applique qu'à un flux $\mathbf{f}$ linéaire en u alors que notre méthode s'applique à des flux non linéaires pour des problèmes convexes on non convexes. Par rapport à [43], nous devons restreindre la classes des initialisations numériques, mais notre façon de procéder est celle qu'on rencontre en pratique dans les applications numériques.

Ce texte est organisé comme suit. Après avoir rappelé quelques points de base concernant les schémas distribuant le résidu, on donne des conditions nécessaire de convergence d'une suite de solutions numériques construite explicitement. La condition principale est celle de monotonie. En cela on suit les idées de [44]. On donne ensuite quelques exemples de schémas v

Convergence

On étudie la convergence du schéma (6.68) sous les conditions (6.69) et (6.71). On a tout d'abord le lemme suivant

Lemme 6.2.5. *Sous les conditions* (6.72) *et*

$$\Delta t \ \max_i \frac{\displaystyle\sum_{T\ni M_i}\sum_{j\in T} c_{ij}^T}{|\mathcal{C}_i|} \leq 1, \tag{6.73}$$

on a quelque soit j et n tel que $n\Delta t \leq T$,

$$|u_j^n| \leq ||u_0||_\infty.$$

Remarque 6.2.6. *Si le schéma est monotone, la condition* (6.73) *ne dépend que de la condition initiale. Autrement dit, le pas de temps ne dépend que de u_0 et du maillage. On attendra $t = T$ en un nombre fini de pas de temps.*

Démonstration. On écrit le schéma sous la forme

$$\begin{aligned} u_i^{n+1} &= u_i^n - \frac{\Delta t}{|\mathcal{C}_i|}\sum_{T\ni M_i}\Phi_i^T \\ &= \sum_{j\in\mathcal{V}(i)} d_j(u_h^n)u_j^n \end{aligned}$$

où

$$\begin{aligned} d_i &= 1 - \frac{\Delta t}{|\mathcal{C}_i|}\sum_{T\ni M_i}\sum_{j\in T} c_{ij}^T \\ d_j &= \frac{\Delta t}{|\mathcal{C}_j|}\sum_{T\ni M_i, T\ni M_j} c_{ij}^T \text{ si } i\neq j. \end{aligned}$$

Clairement d_i et d_j sont positifs sons les conditions du lemme, ce qui achève la preuve. □

On construit maintenant une suite de solution u_h de la manière suivante. Chaque triangle est découpé en quatre sous–triangles comme sur la figure 6.11. La triangulation qui en résulte notée $\mathcal{T}_{h/2'}$ est ubiforme. On construit suivant (6.68) une suite $u_{h/2}$ qui approxime (6.67). Clairement $u_{0,h} \leq u_{0,h/2}$. On obtient alors le théorème

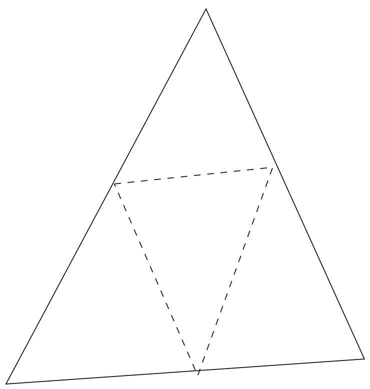


FIG. 6.11 – Découpage d'un triangle.

Théorème 6.2.3. *Si le schéma est monotone sur chaque maillage ($u_i^0 \leq v_i^0 \Rightarrow u_i^n \leq v_i^n$ quelque soit $n \in \mathbb{N}$) alors la suite u_h converge vers la solution entropique de (6.67).*

La preuve se fait en plusieurs lemmes.

Lemme 6.2.7. *Sous les hypothèses du théorème 6.2.3, la suite $\{u_{h/2^k}\}_{k\in\mathbb{N}}$ converge simplement vers une fonction mesurable notée u. Quelque soit h, si u_0 est à support compact, u_h est à support compact. Si u_0 est à support compact, u est en support compact et la convergence a aussi lieu dans L^2_{loc}.*

Démonstration. Le fait que u_h soit à support compact est évident, c'est une conséquence de la condition CFL (6.73). D'après le Lemme 6.2.5, quelque soit h, $t \leq T$ et $\mathbf{x} \in \mathbb{R}^2$, on a $|u^(x,t)| \leq ||u_0||_\infty$. De plus, la suite $u_{h/2^k}$ est croissante. Donc quelque soit $(\mathbf{x}, t)$, la suite $u_{h/2^k}(\mathbf{x}, t)$ est convergente. La limite $u(x,t) = \lim k \to +\infty u_{h/2^k}(\mathbf{x}, t)$ définit une fonction mesurable. $\square$

Lemme 6.2.8. *Soit $k \in \mathbb{R}$, on pose $v_i^n = |u_i^n - k|$. Sous les conditions du théorème (6.2.3), on a*

$$|\mathcal{C}_i|\frac{v_i^{n+1} - v_i^n}{\Delta t} + \sum_{T \ni M_i} \Psi_i^T(v) \leq 0$$

où

$$\Psi_i^T(v) = \sum_{j\in T} c_{ij}(u_h^n)(v_j^n - v_i^n).$$

Démonstration. On part de l'écriture

$$u_i^{n+1} \sum_j \in \mathcal{V}(i) d_j^n u_j^n.$$

Puisque $\sum_{j\in\mathcal{V}(i)} d_j = 1$ et $d_j \geq 0$ sous les conditions du lemme, on a

$$u_i^{n+1} - k = \sum_{j\in\mathcal{V}(i)} d_j(u_j^n - k)$$

et donc

$$v_i^{n+1} \leq \sum_{j \in \mathcal{V}(i)} d_j(u_h^n) v_j^n.$$

En utilisant à nouveau la définition des coefficients d_j, on obtient alors le résultat. $\square$

On a aussi le lemme suivant

Lemme 6.2.9. *Let $\mathcal{Q} = \cup T$ denote a bounded domain of $\mathbb{R}^d$ and $\tau > 0$ a bounded time. Further, let $(u_h)_h$ denote a sequence such that $u_h(\, . \, , n\Delta t) \in \mathcal{V}_h$ for any $n \leq N$. Assume there exists a constant C independent of h and N and a function $u \in L^2(\mathcal{Q} \times [0, \tau])$ such that*

$$\sup_h \sup_{x,t} ||u_h(x,t)|| \leq C, \quad \lim_{h \to 0} ||u_h - u||_{L^2_{loc}(\mathcal{Q} \times [0,\tau])^m} = 0 \ . \tag{6.74}$$

Under these assumptions, the following limits and bound are obtained

1. $\lim_{h\to 0} \sum_{n=0}^N \Delta t \sum_{\forall T \subset \mathcal{Q}} |T| \sum_{M_i, M_j \in T} \left|\left|u_i^n - u_j^n\right|\right| = 0.$
2. $\lim_{h\to 0} \sum_{n=0}^N \Delta t \sum_{\forall T \subset \mathcal{Q}} |T| \sum_{M_i, M_j \in T} \left|\left|u_i^n - u_j^n\right|\right|^2 = 0.$
3. $\lim_{h\to 0} h \ ||\nabla_x u_h||_{L^2(\mathcal{Q}\times[0,\tau])^m} = 0.$
4. *There exists C' independent of h and n such that $h||\nabla_x u_h||_{L^\infty(\mathcal{Q}\times[0,\tau])^m} \leq C'$.*

Preuve du théoreme 6.2.3. On finit la preuve du théorème. Soit ϕ une fonction ≥ 0 C^∞ à support compact dans $\mathbb{R}^2 \times \mathbb{R}^+$. On note $\phi_i^n = \phi(M_i, t_n)$ et ϕ_h la fonction interpolant ϕ.

En multipliant (6.68) par ϕ_i^n, sommant sur i, permutant la somme en i et T, puis sommant en temps, on a

$$\begin{aligned}\sum_{n\in\mathbb{N}} \Delta t \sum_T \sum_{j\in T} \phi_j^n \frac{v_j^{n+1} - v_j^n}{\Delta t} \frac{|T|}{3} \\ + \sum_{n\in\mathbb{N}} \Delta t \sum_T \sum_{j\in T} \phi_j^n \Psi_j^T \leq 0.\end{aligned} \tag{6.75}$$

Ici, on a utilisé la définition de l'aire de la cellule duale. Regardons d'abord le premier terme. Par application du procédé de sommation d'Abel on a

$$\begin{aligned}\sum_n \Delta t \sum_T \sum_{j\in T} \phi_j^n \frac{|T|}{3} \frac{v_j^{n+1} - v_j^n}{\Delta t} \\ \sum_n \sum_T (\phi_j^{n+1} - \phi_j^n) v_j^n + \sum_T \phi_j^0 v_j^0 \frac{|T|}{3},\end{aligned}$$

et on finit en appliquantle lemme 6.2.7, le fait que $\phi \in C_0^\infty$, le fait que $\dfrac{\partial \phi}{\partial t} v$ ($v = |u - k|$) est à support compact et enfin le théorème de convergence dominée. On en déduit donc que le premier terme de (6.75) converge vers

$$\int_{\mathbb{R}^+} \int_{\mathbb{R}^2} \frac{\partial \phi}{\partial t} v d\mathbf{x} dt + \int_{\mathbb{R}^2} \phi(\mathbf{x}, 0) v_0 d\mathbf{x}.$$

On regarde maintenant le second terme. On a

$$\sum_{j\in T}\phi_j^n\Psi_j^n = \overbrace{\phi_G^n\sum_j \Psi_j^n}^{I} + \overbrace{\sum_{l,j} c_{jl}(\phi_j-\phi_G)(v_l^n-v_j^n)}^{II}\ .$$

On regarde I

$$\begin{aligned}\sum_j \Psi_j^n &= \sum_j v_j^n\Big(\sum_l (c_{jl}-c_{lj})\Big)\\ &\sum_j k_j v_j^n\\ &= \int_T \nabla_u \mathbf{f}(u^h)\cdot\nabla v_h\\ &= \int_T \nabla_u \mathbf{g}(u_h)\end{aligned}$$

où $G(u)=\mathrm{sign}(u-k)\big(\mathbf{f}(u)-\mathbf{f}(k)\big)$.

On a donc

$$II = \int_{\mathbb{R}^+}\int_{\mathbb{R}^2}(\dots) + \text{Reste}$$

et

$$I + II \leq -\text{Reste}$$

Grâce au théorème de convergence dominée, $I+II$ converge vers

$$\int_{\mathbb{R}^+}\int_{\mathbb{R}^2}\Big(\frac{\partial\phi}{\partial t}v+\nabla\phi\cdot\mathbf{g}(u)\Big)d\mathbf{x}dt + \int_{\mathbb{R}^+}\phi(x,0)v_0(x)d\mathbf{x}$$

et le reste tends vers 0 par application du lemme 6.2.9. □

Exemples

Remarquons que si on suppose le résidu (6.70) possède des coefficients $c_{ij}^T(u_h)$ différentiables, et si ces coefficients ne dépendent que des valeurs de u_j^n dans T (ce qui sera le cas en pratique), la condition de monotonie est conséquence de la condition suffisante

$$\frac{\partial\Phi_i^T}{\partial u_i}\leq 0, \qquad \frac{\partial\Phi_i^T}{\partial u_j}\geq 0$$

soit

$$\sum_{p\neq i}\frac{\partial c_{ip}}{\partial u_i}u_p - \sum_{p\neq i}c_{ip}\leq 0 \tag{6.76a}$$

$$-\Big(\sum_{p\neq i,p\neq j}\frac{\partial c_{ip}}{\partial u_j}\Big)u_i + c_{ij}\geq 0 \tag{6.76b}$$

On va examiner ces conditions pour trois exemples : les schémas de type volumes finis, le schéma de Lax–Friedrichs et le schéma N.

Considérons $\{u_i\}$ et $\{v_i\}$ telles que $u_i \leq v_i$ quelque soit i. On notera $w_i = u_i - v_i$. Tout d'abord on remarque que

$$u_i - \frac{\Delta t}{|C_i|} \sum T \ni M_i \Phi_i^T = \sum_{T \ni i} \frac{|T|}{|C_i|} \left(u_i - \frac{\Delta t}{|T|} \Phi_i^T \right)$$

si bien que si on montre que si $u_j \leq v_j$ alors

$$u_i - \frac{\Delta t}{|T|} \Phi_i^T(u) \leq v_i - \frac{\Delta t}{|T|} \Phi_i^T(v) \tag{6.77}$$

le schéma est clairement monotone.

Schémas de type volume finis

Schéma de Lax–Friedrichs Par analogie avec le cas monodimensionnel et par abus de langage, on appelle schéma de Lax–Friedrichs le schéma défini par le résidu

$$\Phi_i^T = \frac{1}{3} \left(\Phi^T + \alpha_T \sum_{j \neq i} (u_j - u_i) \right). \tag{6.78}$$

Dans ce cas, l'écriture (6.70) est garantie avec

$$c_{ij} = k_j + \alpha, \qquad k_j = \int_T \nabla_{\mathbf{u}} \mathbf{f}(u_h) \cdot \nabla \varphi_i d\mathbf{x}.$$

Le schéma possède un principe du maximum si $\alpha \geq \max_j |k_j|$. Examinons à quelle condition ce schéma peut être monotone.

Regardons la condition (6.77).

$$\begin{aligned} &\left(u_i - \frac{\Delta t}{|T|} \Phi_i^T(u) \right) - \left(v_i - \frac{\Delta t}{|T|} \Phi_i^T(v) \right) \\ &= w_i - \frac{\Delta t}{|T|} \left(\int_{\partial T} \big(\mathbf{f}(u^h) - \mathbf{f}(v^h) \big) \cdot \vec{n} + \alpha \sum_{j \neq i} (w_j - w_i) \right) \end{aligned}$$

Notons par $[i, j]$ le segment entre les points i et j, voir figure 6.12 On a

$$\int_{\partial T} \big(\mathbf{f}(u^h) - \mathbf{f}(v^h) \big) \cdot \vec{n} = \int_{[1,2]} + \int_{[2,3]} + \int_{[3,1]}$$

et on regarde de qui se passe sur chaque arête de T. Il suffit d'examiner ce qui se passe sur $[1, 2]$. On a

$$\int_{[1,2]} \big(\mathbf{f}(u^h) - \mathbf{f}(v^h) \big) \cdot \vec{n} = w_1 \int_{[1,2]} \nabla_u \mathbf{f}(Z_{12}^h) \vec{n}_{12} \varphi_1 + w_2 \int_{[1,2]} \nabla_u \mathbf{f}(Z_{12}^h) \varphi_2$$

où Z_{12}^h est bien choisi (formule de la moyenne). Ainsi,

$$\begin{aligned} D &= \left(u_i - \frac{\Delta t}{|T|} \Phi_i^T(u) \right) - \left(v_i - \frac{\Delta t}{|T|} \Phi_i^T(v) \right) \\ &= w_i - \frac{\Delta t}{|T|} \left(\int_{\partial T} \big(\mathbf{f}(u^h) - \mathbf{f}(v^h) \big) \cdot \vec{n} + \alpha \sum_{j \neq i} (w_j - w_i) \right) \\ &= w_i - \frac{\Delta t}{|T|} \left(\sum_j a_j^i w_j + \alpha \sum_{j \neq i} (w_i - w_j) \right) \end{aligned}$$

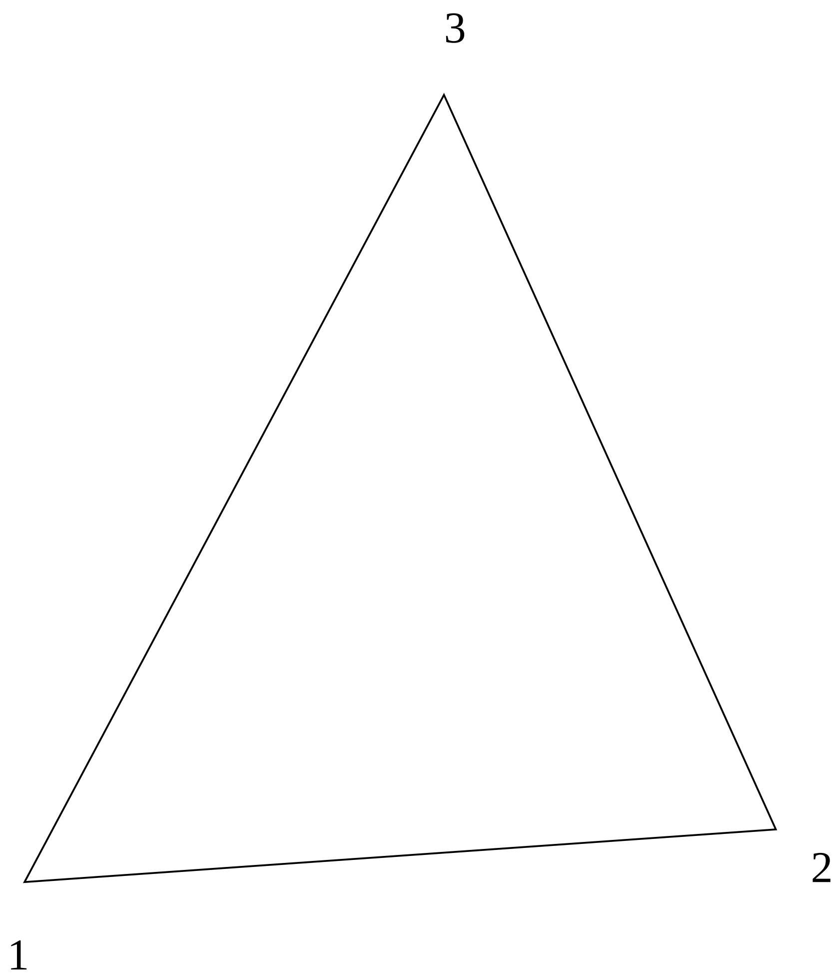


FIG. 6.12 –

où (par exemple)

$$a_1^i = \int_{[1,2]} \nabla_u \mathbf{f}(Z_{12}^h)\vec{n}_{12}\varphi_1 + \int_{[3,1]} \nabla_u \mathbf{f}(Z_{31}^h)\vec{n}_{31}\varphi_1.$$

On se ramène donc à une forme

$$D = \sum_{j \in T} \beta_j w_j$$

avec

$$\beta_j = \begin{cases} 1 - \dfrac{\Delta t}{|T|}\left(a_i^i + 3\alpha\right) & \text{si } j = i \\ -a_j^i + \alpha & \text{sinon} \end{cases}$$

Si α est assez grand, le schéma est monotone.

Schéma N On écrit le résidu comme

$$\Phi^T = \int_T \text{ div } \mathbf{f}(u^h) d\mathbf{x} = \sum_j k_j u_j$$

où on pose

$$k_j = \int_T \nabla_u \mathbf{f}(u^h) \cdot \nabla \varphi_j.$$

Le schéma N est définit par

$$\Phi_i = k_i^+(u_i - \widetilde{u})$$

et $\widetilde{u}$ est définit par

$$\Phi = \sum_j \Phi_j = (\sum_j k_j^+)u_i - \sum k_j^- u_j$$

et donc

$$\widetilde{u} = \frac{\sum_j k_j^- u_j}{\sum_j k_j^-}$$

Après calculs, on a

$$\Phi_j^N = \sum_j k_i^+ n k_j^-(u_i - u_j)$$

où $n = ($
$Sum_j k_j^-)^{-1}$.

Expérimentalement le schéma semble monotone, mais le prouver ....

Notes bibliographiques

Cas instationnaire. Mezine, Mario. Linearisation Abgr Barth et surtout Mario. Quads. Papier de Szepessy. Références Hughes Mallet.

Le cas du problème

$$\frac{\partial u}{\partial t} + \vec{\lambda}^t \nabla u - \varepsilon \Delta u = h$$

peut être traité à partir de (6.10). En effet, si on discrétise d'abord en temps par

$$\frac{u^{n+1} - u^n}{\Delta t} + \vec{\lambda}^t \nabla u^{n+1} - \varepsilon \Delta u^{n+1} = f^{n+1}$$

, on se ramène à la forme (6.10) en ayant posé

$$\alpha = \frac{1}{\Delta t}, \qquad f = h + \frac{u^n}{\Delta t}.$$

Une méthode analogue peut être adoptée pour le problèmes non linéaire.

Ordre très élevé

Annexe A

Quadrature numériques, polynômes orthogonaux

Dans ce chapitre, indépendant du reste du texte, on fait quelques rappels sur les polynômes orthogonaux et les formules de quadratures. Ce qui suit est fortement inspiré de [45, 46, 47] et est très classique.

A.1 Position du problème

Soit $[a,b]$ un intervalle compact et w une fonction intégrable positive sur $[a,b]$. Etant donné de n points $\{x_i\}_{i=1,\dots,n}$ deux à deux distincts, on sait qu'il existe une formule d'ordre au moins $n-1$. Elle est construite en considérant l'interpolant de Lagrange sur ces nœuds et est donnée par

$$\int_a^b f(t)dt \simeq \sum_{i=1}^n \lambda_i f(x_i)$$

où, si ℓ_i représente le polynôme tel que $\ell_i(x_j) = \delta_i^j$,

$$\lambda_i = \int_a^b \ell_i(t)w(t)dt. \tag{A.1}$$

Par construction, cette formule d'intégration numérique est d'ordre $n-1$: elle est exacte pour tous les polynômes de degré $\leq n-1$ (dont l'ensemble est noté $\mathbb{P}_{n-1}$).

La question qui se pose est la suivante. Comment déterminer ces n nœuds et les poids λ_i de manière à ce que la formule soit de degré le plus haut possible, c'est à dire que

$$\int_a^b f(x)w(x)dx - \sum_{i=1}^n \lambda_i f(x_i)$$

s'annule sur $\mathbb{P}_k$ avec k le plus grand possible. En fait seuls les points x_i sont à déterminer, les λ_i étant donnés par (A.1).

Si k est l'ordre maximal, on a forcément $k \geq n$ car on sait que $k > n$. Si $P \in \mathbb{P}_{k-1}$, on doit avoir

$$\int_a^b P(x)w(x)dx = \sum_{j=1}^n \lambda_j P(x_j).$$

Considérons le polynôme de degré n, noté Q, et défini par

$$Q(x) = \Pi_{j=1}^{n}(x - x_j).$$

On fait la division euclidienne de P par Q,

$$P = Q\ M + r,$$

où le degré de r est $\leq n-1$ et celui de M est $\leq k-1-n$. On a donc

$$\int_a^b P(x)w(x)dx = \sum_{j=1}^{n} \lambda_j P(x_j) = \sum_{j=1}^{n} \lambda_j r(x_j).$$

Ceci montre qu'en fait $k \geq 2n = 1$ car tout polynôme de $\mathbb{P}_{2n-1}$ s'écrit de la forme $P = QM + r$ avec $\partial(M) \leq n-1$. Si $k > 2n-1$, on peut considérer $P = Q^2$ qui est positif. On a

$$\int_a^b Q^2(x)w(x)dx = 0$$

car la formule doit être exacte, et $\displaystyle\int_a^b Q^2(x)w(x)dx > 0$ car Q et w ne sont pas nuls, d'où la contradition. Ainsi, le degré maximal atteint avec n points de quadrature est nécessairement inférieur ou égal à $2n-1$.

De plus, ces calculs montrent que

$$\int_a^b Q(x)M(x)w(x)dx = 0 \tag{A.2}$$

quelque soit M polynôme de degré $n-1$ car la formule est d'ordre k.

On est donc conduit à étudier la forme bilinéaire définie sur $C^0(a,b)$ par

$$(f,g) \in C^0(a,b)^2 \mapsto \int_a^b f(x)g(x)w(x)dx := (f,g). \tag{A.3}$$

A.2 Polynômes orthogonaux

La relation (A.2) dit que Q est orthogonal à $\mathbb{P}_{n-1}$. On introduit donc la définition suivante :

Définition A.2.1. *On appelle polynôme orthogonaux relativement à un poids w positif et intégrable sur un intervalle $[a,b]$ la suite suite des polynômes $P_0, P_1, \ldots P_n, \ldots$ ayant les propriétés suivantes*

1. *Quel que soit n, P_n est de degré n, le coefficient de plus haut degré vaut 1,*
2. *Quel que soit n, quel que soit $P \in \mathbb{P}_k$ avec $k \leq n-1$,*

$$\int_a^b P_n(x)P(x)w(x)dx = 0.$$

Montrons d'abord que de telles suites existent

Lemme A.2.2. *Que que soit le poids $w > 0$ intégrable sur $[a, b]$, il existe une unique suite de polynômes orthogonaux qui satisfont à*

$$P_n = x^n - \sum_{i=0}^{n-1} c_{in} P_i$$

avec

$$c_{in} = \frac{(x^n, P_i)}{(P_i, P_i)}.$$

Démonstration. On a $P_0 = 1$. Supposons connus les $n-1$ polynômes orthogonaux. On a nécessairement (si existence)

$$P_n = x^n - \sum_{j=0}^{n-1} \alpha_j P_j$$

On détermine α_j en écrivant que $(P_n, P_j) = 0$ si $j < n$.

Une autre façon de faire le calcul est d'appliquer la méthode d'orthogonalisation de Gramm Schmidt à $\{1, x, \ldots, x^n\}$. □

Donnons une propriété qualitative des polynômes orthogonaux qui permet de construire les formules de quadrature : les zéros sont deux à deux distincts et tous dans $[a, b]$. Plus précisément

Théorème A.2.1. *Soit w un poids intégrable et strictement positif dans $[a, b]$. Alors, quelque soit n, toutes les racines du n ième polynôme orthogonal P_n sont simples et sont dans $]a, b[$.*

Démonstration. Soient x_1, $x_2, \ldots, x_j$ les racines de P_n appartenant à $]a, b[$, et comptées avec leur multiplicité. On a $j \leq n$ et peut être nul. Supposons que $j < n$. Comme les coefficients de P_n sont réels, P_n va changer de signe à toutes les racines de multiplicité impaire. Posons, si $j > 0$,

$$Q(x) = \Pi_{k=1}^{j}(x - x_k).$$

Si $j = 0$, on prend $Q = 1$. Le produit $P_n Q$ ne change pas de signe dans $]a, b[$, et Q est de degré $\leq n-1$, donc

$$\int_a^b P_n(x) Q(x) w(x) dx = 0.$$

Comme l'expression $P_n(x)Q(x)w(x)$ est positive, elle doit être nulle ce qui est impossible : les racines de P_n sont donc toutes dans $[a, b]$. Il faut montrer maintenant qu'elles sont simples.

Soit x_1 supposée multiple. Alors $P_n(x) = p(x)(x - x_1)^2$ et P_n et p sont de même signe. Comme p est de degré au plus $n-2$, et donc

$$\int_a^b P_n(x) p(x) w(x) dx = 0$$

ce qui est absurde.

Les racines sont donc simples. □

A.3 Formules de quadrature

On finit par donner la réponse à la question posée :

Théorème A.3.1. *L'unique formule à n points d'ordre maximale est la formule par interpolation construite en prenant pour nœuds les zéros du n ième polynôme orthogonal par rapport au poids w. Elle est d'ordre $2n-1$. La formule ainsi déterminée est dite gaussienne.*

Démonstration. On part de (A.3). Cette formule exprime que les points d'interpolations $x_1, \ldots, x_n$, s'ils existent, sont tels que

$$Q(x) = \Pi_{j=1}^{n}(x - x_j)$$

est orthogonal à tout polynôme de degré $n-1$. Au vu de ce qui vient d'être dit sur les polynômes orthogonaux, on voit que nécessairement les x_i sont les zéros d'un polynôme orthogonal pour le poids w. C'est forcément P_n, et on a l'existence et l'unicité.

Son ordre est $2n-1$: Soit k l'ordre de cette formule. Nécessairement $k \geq 2n-1$ car si P est de degré $2n-1$, on effectue la division euclidienne avec P_n, $P = P - np + r$ et comme précédement, r est de degré $n-1$ donc

$$\int_a^b P(x)w(x)dx = \int_a^b r(x)w(x)dx = \sum_{j=1}^{n} \lambda_j r(x_j) = \sum_{j=1}^{n} \lambda_j P(x_j)$$

la première égalité étant conséquence de l'orthogonalité de P_n et de de p car son degré est au plus $n-1$.

Si $k > 2n-1$, on prend $P(x) = P_n^2(x)$, et on a nécessairement

$$\sum_{j=0}^{n} \lambda_j P_n^2(x_j) = 0$$

donc

$$\int_a^b P_n^2(x)w(x)dx = 0$$

ce qui est absurde. □

On a de plus le signe des coefficients λ_i :

Proposition A.3.1. *Si λ_i est le poids associée au nœud x_i, alors $\lambda_i > 0$.*

Démonstration. Le polynôme ℓ_i étant de degré $n-1$, le polynôme ℓ_i^2 est de degré $2(n-1) < 2n-1$. Ainsi,

$$\int_a^b \ell_i^2(x)w(x)dx = \sum_{j=1}^{n} \lambda_j \ell_i^2(x_j) = \lambda_i.$$

, Comme $\displaystyle\int_a^b \ell_i^2(x)w(x)dx > 0$, ceci conclue la preuve. □

ERREUR DE QUADRATURE, formules composée Dans le cas d'un intervalle $(x_{j-1/2}, x_{j+1/2})$, on se ramène à $(-1,1)$ au moyen du changement de variable $x \mapsto \frac{x-x_j}{\Delta x}$. Ainsi l'erreur de la formule de quadrature est

$$\left| \int_{x_{j-1/2}}^{x_{j+1/2}} \Phi(x) - \Delta x \sum_{j=1}^{m} \omega_j \Phi(x_j) \right| \leq C_{2m} \Delta x^{2m} \sup_{x \in [x_{j-1/2}, x_{j+1/2}]} |\Phi^{(2m)}(x)|. \tag{A.4}$$

A.4 Exemple du poids $w = 1$ et $[a,b] = [-1,1]$: les polynômes de Legendre

On a dans ce cas

$$P_n(x) = \sqrt{n + \frac{1}{2}} \frac{1}{2^n n!} \frac{d^n}{dx^n} \Big[(x^2-1)^n \Big].$$

Démonstration. On pose pour simplifier les notations

$$R_n(x) = \frac{d^n}{dx^n} \Big[(x^2-1)^n \Big]$$

qui est un polynôme de degré n. On vérifie ensuite que si $p < n$, il existe un polynôme $r_p(x)$ tel que

$$\frac{d^p}{dx^p} \Big[(x^2-1)^n \Big] = r_p(x)(x^2-1)^{n-p}.$$

Cette relation est vraie si $p = 0$. Supposons la vraie pour $p-1$. Alors

$$\frac{d^p}{dx^p} \Big[(x^2-1)^n \Big] = \frac{d}{dx} \Big[r_{p-1}(x)(x^2-1)^{n-p+1} \Big]$$
$$\Big(r'_{p-1}(x)(x^2-1) + 2x(n-p+1) r_{p-1}(x) \Big)(x^2-1)^{n-p}.$$

A partir de là, et supposant $m \leq n$, on intègre par partie

$$\int_{-1}^{1} R_n(x) R_m(x) dx = \frac{d^{n-1}}{dx^{n-1}}[(x^2-1)^n] \frac{d^m}{dx^m}[(x^2-1)^m] \Bigg|_{x=-1}^{x=1}$$
$$- \int_{-1}^{1} \frac{d^{n-1}}{dx^{n-1}}[(x^2-1)^n] \frac{d^{m+1}}{dx^{m+1}}[(x^2-1)^m] dx$$

Le premier terme s'annule en ± 1 grâce au résultat préliminaire, et de proche en proche, on a

$$\int_{-1}^{1} R_n(x) R_m(x) dx$$
$$= (-1)^p \int_{-1}^{1} \frac{d^{n-p}}{dx^{n-p}}[(x^2-1)^n] \frac{d^{m+p}}{dx^{m+p}}[(x^2-1)^m] dx$$

si $p \leq n$. Si $m < n$, on prend $p = m+1$ et la dérivée d'ordre $m+p = 2m+1$ de $(x^2-1)^m$ est nulle d'où

$$\int_{-1}^{1} R_n(x)R_m(x)dx = 0.$$

Si $m = n$, on prend $p = n$ et

$$\int_{-1}^{1} R_n(x)R_m(x)dx = (2n)! \int_{-1}^{1} (1-x^2)^n dx = (2n)! I_n.$$

En intégrant par partie, on voit que

$$\int_{-1}^{1} (1-x^2)^n dx = x(1-x^2)^n \Big|_{-1}^{1} + 2n \int_{-1}^{1} x^2(1-x^2)^{n-1} dx$$

et donc par récurence,

$$I_n = 2n(I_n - I_{n-1})$$

soit

$$I_n = \frac{2n}{2n+1} I_{n-1}$$

et donc

$$I_n = \frac{2n(2n-2)\dots 2}{(2n+1)(2n-1)\dots 3.1} 2 = \frac{2^{n+1} n!}{(2n+1)(2n)!/2^n n!} = \frac{2^{2n+1}(n!)^2}{(2n+1)(2n)!}.$$

Finalement,

$$\int_{-1}^{1} R_n(x)^2 dx = \frac{2^{2n+1}(n!)^2}{(2n+1)}$$

Ceci montre que les P_n forme une famille orthonormée, le coefficient dominant de P_n est strictement positif. □

Définition A.4.1. *On appelle polynôme de Legendre les polynômes L_n donnés par*

$$L_n(x) = \frac{1}{n! 2^n} \frac{d^n}{dx^n} \Big[(x^2-1)^n \Big].$$

On a le corollaire suivant

Corollaire A.4.2. *On a les relations suivantes*

$$\int_{-1}^{1} L_n^2(x) dx = \frac{2}{2n+1}. \tag{A.5}$$

Démonstration. Voir la preuve précédente. □

Les poids sont donnés par les formules explicites simples

$$\lambda_k = \frac{2}{n L_n'(x_k) L_{n-1}(x_k)} \tag{A.6}$$

Preuve de (A.6). On note $Q_k(x) = \frac{\Pi_{j=1}^n (x-x_j)}{x-x_k}$ et on écrit

$$L_n(x) = a_{0,n}\Pi_{j=1}^n (x-x_j).$$

On voit que $Q_k(x_k) = L_n'(x_k)$ et puisque $\ell_k = Q_k/Q_k(x_k)$,

$$\lambda_k = \frac{1}{Q_k(x_k)} \int_{-1}^{1} Q_k(x)dx,$$

on a

$$\lambda_k = \frac{1}{L_n'(x_k)} \int_{-1}^{1} \frac{L_n(x)}{x-x_k}dx.$$

Etant donné que $\frac{L_{n-1}(x)-L_{n-1}(x_k)}{x-x_k}$ est un polynôme de degré $n-2$,

$$\int_{-1}^{1} L_n(x)\frac{L_{n-1}(x)-L_{n-1}(x_k)}{x-x_k}dx = 0,$$

soit

$$\lambda_k = \frac{1}{L_n'(x_k)L_{n-1}(x_k)} \int_{-1}^{1} \frac{L_n(x)L_{n-1}(x)}{x-x_k}dx.$$

Puisque $\frac{L_n(x)}{x-x_k}$ est de degré $n-1$, on a

$$\frac{L_n(x)}{x-x_k} = \frac{a_{0,n}}{a_{0,n-1}}L_{n-1}(x) + r(x)$$

où r est de degré inférieur ou égal à $n-2$. Ainsi,

$$\begin{aligned}
\lambda_k &= \frac{a_{0,n}}{a_{0,n-1}} \frac{1}{L_n'(x_k)L_{n-1}(x_k)} \int_{-1}^{1} L_{n-1}^2(x)dx \\
&= \frac{a_{0,n}}{a_{0,n-1}} \frac{1}{L_n'(x_k)L_{n-1}(x_k)} \frac{2}{n} \\
&=
\end{aligned}$$

□

Donnons quelques exemples.

1. Si $n=1$, on a $L_1(x) = x$, d'où
$$\int_{-1}^{1} f(x)dx \simeq 2f(0).$$
C'est la formule du point milieu qui est précise à l'ordre 1.
2. Si $n=2$, $L_2(x) = 3/2\,x^2 - 1/2$, d'où
$$\int_{-1}^{1} f(x)dx \simeq f\Big(-\frac{1}{\sqrt{3}}\Big) + f\Big(\frac{1}{\sqrt{3}}\Big)$$
qui est précise à l'ordre 3.

3. Si $n=3$, $L_3(x) = /2\,x^3 - 3/2\,x$ d'où

$$\int_{-1}^{1} f(x)dx \simeq \frac{5}{9} f\left(-\frac{\sqrt{15}}{5}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\frac{\sqrt{15}}{5}\right).$$

   qui est précise à l'ordre 5
4. Si $n=4$, $L_4(x) = \frac{35}{8}\,x^4 - \frac{15}{4}\,x^2 + 3/8$ d'où

$$\begin{aligned}\int_{-1}^{1} f(x)dx \simeq{}& \frac{1}{180}\left(-5+3\sqrt{30}\right)\sqrt{30}\left[f\left(-\frac{\sqrt{525+70\sqrt{30}}}{35}\right) + f\left(\frac{\sqrt{525+70\sqrt{30}}}{35}\right)\right] \\ &+ \frac{1}{180}\left(5+3\sqrt{30}\right)\sqrt{30}\left[f\left(-\frac{\sqrt{525-70\sqrt{30}}}{35}\right) + f\left(\frac{\sqrt{525-70\sqrt{30}}}{35}\right)\right]\end{aligned}$$

   qui est d'ordre 7,

Pour plus d'information sur ces polynômes et les formules de quadratures associées, en particulier quand leur ordre devient grand, on pourra consulter [48], chapitre 2.

POLYNOMES DE LEGENDRE

$$L_m(1) = 1, \qquad L_m(-1) = (-1)^m.$$

Bibliographie

[1] E. Godlewski and P.A. Raviart. *Hyperbolic systems of conservation laws.* Mathématiques et Applications. Ellipses, Paris, 1991.

[2] E. Godlewski and P.A. Raviart, editors. *Numerical approximation of systems of conservation laws*, volume 118 of *Applied Mathematical Sciences.* Springer, 1996.

[3] D. Serre. *Systèmes de lois de conservation I : hyperbolicité, entropies, ondes de choc.* Diderot éditeur Arts et Sciences, 1996.

[4] D. Serre. *Systèmes de lois de conservation II : structures géométriques, oscillations et problèmes mixtes.* Diderot éditeur Arts et Sciences, 1996.

[5] B. Cockburn. Devising discontinuous Galerkin methods for non–linear hyperbolic conservation laws. *J. Comput. and Appl. Math..*, 128 :187–204, 2001.

[6] P.D. Lax. *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*, volume 11 of *Conf. Board Math. Sci. Regional Conference Series in Applied Mathematics.* SIAM, 1972.

[7] W. Rudin. *Real and Complex analysis.* Tata McGraw, New Delhi, 1980.

[8] R. Menikoff and B.J. Plohr. The Riemann problem for fluids of real materials. *Rev. Mod. Phys.*, 61(1) :75–130, 1989.

[9] H. Guillard and R. Abgrall. *Modélisation numérique des fluides compressibles*, volume 5 of *Series in applied mathematics.* Gauthier–Villars, 2001.

[10] T.J. Hughes, L.P. Franca, and M. Mallet. A new finite element formulation for CFD : I symmetric forms of the compressible Euler and Navier Stokes equations and the second law of thermodynamics. *Comp. Meth. Appl. Mech. Engrg*, 54 :223–234, 1986.

[11] A. Harten. On the symmetric form of systems of conservation laws with entropy. *J. Comput. Phys.*, 49 :357–393, 1983.

[12] M.S. Mock. Systems of conservation laws of mixed type. *J. Diff. Eqns.*, 37 :70–88, 1980.

[13] P.L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.*, 43 :357–372, 1981.

[14] Bruno Dubroca. Solveur de Roe positivement conservatif. (Positively conservative Roe's solver for Euler equations). *Comptes rendus de l'Académie des Sciences*, Série I, 1999.

[15] B. Perthame and C.-W. Shu. On positivity preserving finite volume schemes for Euler equations. *Numer. Math.*, 73(1) :119–130, 1996.

[16] Robert Eymard, Thierry Gallouët, and Raphaèle Herbin. Finite volume methods. In *Ciarlet, P. G. (ed.) et al., Handbook of numerical analysis. Vol. 7 : Solution of equations in* $\mathbb{R}^n$ *(Part 3).*

Techniques of scientific computing (Part 3). Amsterdam : North-Holland/ Elsevier. 713-1020 . 2000.

[17] B. Cockburn and C.W. Shu. TVD Runge Kutta local projection discontinuous Galerkin Finite element method for conservation laws II : general framework. *Math. of Comp.*, 52(186) :411–435, 1989.

[18] B. Cockburn, S. Hou, and C.W. Shu. The Runge–Kutta local projection Discontinuous Galerkin finite element method for conservation laws IV : the multidimensional case. *Math. of Comp.*, 54(190) :545–581, 1990.

[19] G. Jiand and C.W. Shu. On a cell enrtopy for Discontinuous Galerkin Methods. *Math. of Comp.*, 62(206) :531–538, 1994.

[20] B. Cockburn and C.W. Shu. The Local Discontinuous Galerkin method for time–dependant convection diffusion systems. *SIAM J. Numer. Anal,*, 35(6) :2440–2463, 1998.

[21] Matania Ben-Artzi and Joseph Falcovitz. An upwind second-order scheme for compressible duct flows. *SIAM J. Sci. Stat. Comput.*, 7 :744–768, 1986.

[22] P.G. Ciarlet and P.A. Raviart. General Lagrange and Hermite interpolation in R^n with applications to finite element methods. *Arch. Ration. Mech. Anal.*, 46 :177–199, 1972.

[23] S. Osher. Rieman solvers, the entropy condition and difference approximations. *SIAM J. Numer. Anal.*, 21 :217–235, 1984.

[24] C.W. Shu. Total–variation diminishing time discretisations. *SIAM J. Sci. Comput.*, 9 :1073–1084, 1988.

[25] C.W. Shu and S. Osher. Efficient implementation of essentially non–oscillatory schock capturing schemes. *J. Comput. Phys.*, 77 :439–471, 1988.

[26] S. Gottliebd and C.W. Shu. Total variation diminishing Runge–Kutta schemes. *Math. of Comp.*, 221 :73–85, 1998.

[27] C. Bolley and M. Crouzeix. Conservation de la positivité lors de la discrétisation des problèmes d'évolution parabolique. *RAIRO Numerical Analysis*, 12(3) :237–245, 1978.

[28] P. Ciarlet. *The finite element method for elliptic problems.* North Holland, Amsterdam, 1975.

[29] C. Johnson, U. Nävert, and J. Pitkäranta. Finite element methods for linear hyperbolic problems. *Comput. Methods Appl. Mech. Engrg.*, 45(1-3) :285–312, 1984.

[30] A. Ern and J.L. Guermond. *Éléments finis : théorie, applications, mise en œuvre*, volume 36 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2002.

[31] R. Abgrall, K. Mer, and B. Nkonga. A lax–wendroff type theorem for residual schemes. In M. Haffez and J.J Chattot, editors, *Innovative methods for numerical solutions of partial differential equations*, pages 243–266. World Scientific, 2002.

[32] R.-H. Ni. A multiple grid scheme for solving the Euler equations. *AIAA J.*, 20 :1565–1571, 1981.

[33] K.W. Morton and E. Süli. Finite volume methods and their analysis. *IMA Journal of Numerical Analysis*, 11 :241–260, 1991.

[34] P.I. Crumpton, J.A. MacKenzie, and K.W. Morton. Cell vertex algorithms for the compressible Navier-Stokes equations. *J. Comp. Phys.*, 109 :1–15, 1993.

[35] D. Kröner, M. Rokyta, and M. Wierse. A Lax-Wendroff type theorem for upwind finite volume schemes in 2-d. *East-West J. Numer. Math.*, 4(4) :279–292, 1996.

[36] R. Abgrall and K. Mer. Un théorème de type Lax–Wendroff pour les schémas distributifs. Technical Report 98010, Mathématiques Appliquées de Bordeaux, March 1998.

[37] D. Kröner. *Numerical Schemes for Conservation Laws.* John Wiley and B.G. Teubner, New York, USA, 1997.

[38] H. Deconinck, R. Struijs, G. Bourgeois, and P.L. Roe. Compact advection schemes on unstructured meshes. VKI Lecture Series 1993–04, Computational Fluid Dynamics, 1993.

[39] H. Deconinck, P.L. Roe, and R. Struijs. A multidimensional generalisation of Roe's difference splitter for the Euler equations. *Computer and Fluids*, 22(2/3) :215–222, 1993.

[40] P.L. Roe and D. Sidilkover. Optimum positive linear schemes for advection in two and three dimensions. *SIAM J. Numer. Anal.*, 29(6) :1542–1568, 1992.

[41] Á. Csík, M. Ricchiuto, and H. Deconinck. A conservative formulation of the multidimensional upwind residual distribution schemes for general nonlinear conservation laws. *J. Comput. Phys.*, 179(2) :286–312, 2002.

[42] R. Abgrall. An essentially non-oscillatory reconstruction procedure on finite element type meshes, application to compressible flows. *Comput. Methods Appl. Mech. Engrg*, 116 :95–101, 1994.

[43] B. Perthame. Convergence of n-schemes for linear advection equations. Technical Report 94036, Laboratoire d'Analyse Numérique, Paris VI, 1994.

[44] M. Breuss. The implicit upwind method for 1–d scalar conservation laws with continuous fluxex. *SIAM J. Numer. Anal.*, to appear.

[45] M. Crouzeix and A.L. Mignot. *Analyse numérique des équations différentielles.* Collection Mathématiques Appliquées pour la maîtrise. Masson, 1984.

[46] M. Schatzman. *Analyse Numérique.* Enseignement des Mathématiques. Masson/InterEdition, 1991.

[47] C. Guilpin. *Manuel de calcul numérique appliqué.* EDP Sciences, 1999.

[48] C. Bernardi, Y. Maday, and F. Rapetti. *Discrétisations variationnelles de problèmes aux limites*, volume 45 of *Mathématiques et Applications.* Springer Verlag, 2004.