THÈSE

présentée à

L'ÉCOLE NATIONALE DES PONTS ET CHAUSSÉES

pour obtenir le titre de

DOCTEUR

Spécialité : Mathématiques Appliquées

Etude de schémas d'ordre élevé en volumes finis pour des problèmes hyperboliques. Application aux équations de Maxwell, d'Euler et aux écoulements diphasiques dispersés.

par Sophie Depeyre

Soutenue le 14 Janvier 1997 devant la commission composée de :

MM.	Bernard	LARROUTUROU	Directeur
	Alain	DERVIEUX	Rapporteurs
	Olivier	PIRONNEAU	
Mme	Loula	FÉZOUI	Examinateurs
MM.	Benoît	PERTHAME	
	Lionel	SAINSAULIEU	
	Jean-Marc	TALBOT	

A mes parents bien aimés,
A Annick,
A Philippe,
A Jésus Tout-Puissant,
pour leur Amour.

"Et moi je vous dis : Demandez, on vous donnera. Cherchez, vous trouverez. Frappez, on vous ouvrira. Car quiconque demande reçoit. Qui cherche trouve. A qui frappe, on ouvrira."

Evangile selon Saint Luc, XI,9

REMERCIEMENTS

Je voudrais tout d'abord exprimer ma profonde gratitude à Monsieur Bernard LARROUTUROU, pour la confiance qu'il m'a accordée en m'accueillant au sein du CERMICS et en acceptant de diriger mes recherches. Ses précieux conseils, ses encouragements ainsi que sa gentillesse naturelle, m'ont permis d'aboutir dans ce travail, malgré l'éloignement et les lourdes tâches professionnelles qui lui incombent.

Ma sincère reconnaissance va à Madame Loula FÉZOUI pour sa disponibilité, son encadrement et les discussions enrichissantes que nous avons eues, concernant les domaines de l'électromagnétisme et de la mécanique des fluides. Je la remercie pour l'attention particulière qu'elle a porté à la rédaction de ce mémoire.

Je remercie vivement Messieurs Alain DERVIEUX et Olivier PIRONNEAU qui, malgré leurs occupations ont accepté la lourde tâche d'être rapporteur.

Mes remerciements vont également à Messieurs Benoît PERTHAME et Jean-Marc TALBOT, qui ont bien voulu porter un jugement sur mon travail, et qui m'honorent de leur présence à la soutenance.

J'ai eu la chance de travailler avec Monsieur Lionel SAINSAULIEU, lors d'un stage de cinq mois effectué au CERMICS, avant ma thèse. J'ai pu au cours de ce stage, puis pendant ma thèse, bénéficier de sa grande expérience des écoulements diphasiques, et je le remercie de s'être intéréssé de près à mon travail.

Je tiens aussi à remercier chaleureusement les personnes avec qui j'ai directement collaboré pour certains travaux de cette thèse, notamment Serge PIPERNO, Didier ISSAUTIER ainsi que Romuald CARPENTIER.

Enfin, je remercie tous les membres du CERMICS-INRIA Sophia-Antipolis ou projet CAIMAN, qui ont su créer une ambiance de travail chaleureuse, familiale et vivante.
Nathalie GLINSKY-OLIVIER, Armel de LA BOURDONNAYE, et Robert RIVIÈRE se sont toujours montrés disponibles et m'ont apporté leur soutien et leur expérience pendant ces trois années.
Bien sûr, je n'oublie pas de saluer mes "compagnons de thèse", Marco, Stéphanie, Cédric, Jipé (alias Jean-Pierre), Francesco (alias François), Frédéric, ainsi que les "nouveaux arrivés" Malika et Mihai, sans lesquels ce travail n'aurait pas été aussi agréable.
Mes pensées vont aussi à nos secrétaires, Martine, Christine, Pascale et Brigitte, que je remercie pour leur aide et leur gentillesse.

Je remercie enfin toutes les personnes qui de près ou de loin m'ont soutenue et encouragée tout au long de cette thèse.

Table des matières

I INTRODUCTION GÉNÉRALE. **1**
I.1 Présentation des modèles étudiés. 1
I.1.1 Les équations de Maxwell. 2
I.1.2 Les équations d'Euler. 4
I.1.3 Un modèle "simplifié" d'écoulement diphasique. 7
I.2 Description de la résolution numérique. 8
I.3 Présentation du plan de la thèse. 11

PREMIÈRE PARTIE **15**

II ÉTUDE DE L'ÉQUATION D'ADVECTION. **17**
II.1 Introduction. 18
II.2 Présentation des méthodes numériques 18
II.2.1 Equation d'advection 18
II.2.2 Schéma d'ordre un en maillage rectangulaire 19
II.2.3 Maillage triangulaire régulier 23
II.2.4 Extension aux ordres supérieurs 25
II.3 Equations équivalentes 27
II.3.1 Méthode de calcul 28
II.3.2 Equations équivalentes des schémas. 30
II.3.3 Schémas "sans diffusion numérique" 33
II.4 Etude de stabilité des schémas 34
II.4.1 Schémas précis à l'ordre un 34
II.4.2 Schémas d'ordre plus élevé 35
II.4.3 Schémas d'ordre quatre 42
II.5 Illustration numérique 44
II.6 Conclusion. 49
II.7 Annexe A. 51
II.7.1 Schémas sans matrice de masse 51

II.7.2 Schémas avec matrice de masse . . . 53
II.8 Annexe B. . . . 54
II.9 Annexe C . . . 58

III PRÉSENTATION DU SYSTÈME DE MAXWELL ET ÉTUDE DE STABILITÉ. 63
III.1 Introduction. . . . 64
III.2 Maxwell system. . . . 65
III.2.1 Electromagnetic field equations. . . . 65
III.2.2 Conservative formulation and hyperbolic character. . . . 65
III.3 Numerical approximation. . . . 67
III.3.1 Spatial formulation. . . . 67
III.3.2 First-order upwind scheme. . . . 68
III.3.3 High order approximation. . . . 68
III.3.4 Time integration. . . . 69
III.4 Stability analysis. . . . 70
III.4.1 First-order accurate schemes. . . . 70
III.4.2 Higher order schemes. . . . 77
III.4.3 Domaines de stabilité pour les schémas centrés. . . . 81
III.5 Illustration numérique. . . . 83

IV UNE NOUVELLE FORMULATION DU SYSTÈME DE MAXWELL. 89
IV.1 Introduction. . . . 90
IV.2 Nouvelle formulation du système de Maxwell. . . . 91
IV.2.1 Présentation de la méthode. . . . 91
IV.2.2 Adimensionnement. . . . 93
IV.2.3 Formulation faible. . . . 94
IV.2.4 Traitement des conditions aux limites. . . . 96
IV.2.5 Intégration en temps. . . . 97
IV.3 Etude de stabilité des schémas. . . . 97
IV.3.1 Schémas précis à l'ordre un. . . . 98
IV.3.2 Schémas précis d'ordre supérieur. . . . 100
IV.4 Equations équivalentes. . . . 106
IV.4.1 Analyse et comparaison des termes d'erreur des schémas d'ordre un. 106
IV.4.2 Analyse et comparaison des termes d'erreur des β-schémas. . . . 108
IV.5 Résultats numériques . . . 110
IV.6 Conclusion . . . 117

V CONCLUSION. **119**

<u>DEUXIÈME PARTIE</u> **125**

VI ÉTUDE DE MODÈLES NON LINÉAIRES ET APPROXIMATION NUMÉRIQUE. **127**
VI.1 Introduction. 127
VI.2 L'équation de Burgers. 128
VI.2.1 Approximation d'ordre un. 129
VI.2.2 Extension à un ordre supérieur. 130
VI.3 Le système des équations d'Euler. 130
VI.3.1 Généralités. 130
VI.3.2 Hyperbolicité. 131
VI.4 Approximation spatiale. 132
VI.4.1 Calcul des flux. 132
VI.4.2 Traitement des conditions aux limites. 135
VI.5 Approximation d'ordre supérieur en temps et en espace. 136
VI.6 Schémas implicites. 137
VI.6.1 Linéarisation des flux convectifs: 138
VI.6.2 Linéarisation sur le bord $\delta\Omega$: 138
VI.6.3 Méthode de résolution. 139
VI.6.4 Convergence vers la solution stationnaire. 139

VII UN NOUVEAU LIMITEUR DE FLUX POUR LES ÉQUATIONS D'EULER. **141**
VII.1 Introduction 142
VII.2 Existing high-order TVD schemes 143
VII.2.1 The linear advection equation. 144
VII.2.2 The Burgers equation. 147
VII.3 New limiters for hyperbolic scalar equations 147
VII.3.1 Limiters for upwind fluxes 147
VII.3.2 Limiters for centered fluxes 149
VII.3.3 Numerical comparison of the limited schemes. 152
VII.4 Extension to two-dimensional Euler equations 153
VII.4.1 MUSCL scheme for unstructured triangular meshes 154
VII.4.2 Effective limitation for the vector W 155
VII.4.3 Preliminary results on a shock tube 155
VII.4.4 Two-dimensional numerical results 160

VII.5 Conclusion 165

VIII UNE MÉTHODE COUPLÉE POUR LES ÉCOULEMENTS DIPHASIQUES. 169

VIII.1 Introduction. 170
VIII.2 Présentation du modèle. 171
VIII.2.1 Le modèle original. 173
VIII.2.2 Un modèle simplifié. 174
VIII.3 Approximation numérique. 175
VIII.3.1 Hyperbolicité du système. 177
VIII.3.2 Méthode de résolution. 178
VIII.3.3 Formulation variationnelle. 179
VIII.3.4 Calcul des flux. 180
VIII.3.5 Approximation d'ordre supérieur en espace. 181
VIII.3.6 Traitement des conditions aux limites. 181
VIII.3.7 Intégration en temps. 182
VIII.3.8 Résolution par un schéma implicite. 187
VIII.4 Simulation numérique. 192
VIII.4.1 Cas instationnaire. 192
VIII.4.2 Cas stationnaire. 204
VIII.5 Annexe D. 219
VIII.5.1 Matrice jacobienne des équations d'Euler. 219
VIII.5.2 Matrice jacobienne du sous-système des gouttes. 219
VIII.5.3 Matrice jacobienne du terme source. 220

IX CONCLUSION. 221

Chapitre I

INTRODUCTION GÉNÉRALE.

Les équations ou les systèmes d'équations aux dérivées partielles hyperboliques décrivent un grand nombre de problèmes physiques, qui mettent en jeu des phénomènes de vibration et de propagation d'ondes. Les applications sont nombreuses, et recouvrent des domaines variés comme l'acoustique, l'électromagnétisme, l'élasticité, la mécanique des fluides, la mécanique des solides, etc ... Nous parlerons essentiellement d'electromagnétisme et de mécanique des fluides.

I.1 Présentation des modèles étudiés.

Nous considérons la classe des systèmes hyperboliques de lois de conservation, c'est-à-dire les systèmes qui s'écrivent (en une dimension d'espace) :

$$\frac{\partial}{\partial t}\mathbf{W}(x,t) + \frac{\partial}{\partial x}\mathbf{F}\left(\mathbf{W}(x,t)\right) = 0 \tag{I.1}$$

où $\mathbf{W} : I\!R\mathrm{x}I\!R^{+} \longrightarrow I\!R^{m}$ est le vecteur des quantités conservées de dimension m (m entier strictement positif), appelé aussi variable d'état, et où $\mathbf{F} : I\!R^{m} \longrightarrow I\!R^{m}$ représente la fonction de flux du système. Cette forme découle de lois physiques, comme nous le verrons plus loin, elle exprime localement les principes de conservation de la mécanique [14].

Sous cette forme conservative, le système (I.1) est dit hyberbolique (resp. strictement hyperbolique) si et seulement si, pour tout $\mathbf{W} \in I\!R^{m}$, le jacobien $\dfrac{\partial \mathbf{F}(\mathbf{W})}{\partial \mathbf{W}}$ est diagonalisable avec valeurs propres et vecteurs propres réels (resp. avec m valeurs propres réelles distinctes).

Le caractère hyperbolique d'un système a une interprétation physique : il signifie que les ondes et l'énergie associée se propagent en temps fini suivant des directions particulières.

Pour résoudre (I.1), il est nécessaire d'adjoindre des informations sur les conditions initiales et sur les conditions aux limites, lorsque l'on considère un domaine spatial borné, comme c'est toujours le cas en pratique. Le problème le plus simple est le problème aux valeurs initiales, ou problème de Cauchy, qui s'écrit :

$$\begin{cases} \dfrac{\partial}{\partial t}\mathbf{W}(x,t) + \dfrac{\partial}{\partial x}\mathbf{F}\left(\mathbf{W}(x,t)\right) = 0 \\ \mathbf{W}(x,0) = \mathbf{W}_0(x) \end{cases} \tag{I.2}$$

Dans cette thèse, nous distinguerons les problèmes hyperboliques linéaires (c'est-à-dire lorsque la fonction $\mathbf{F}$ est linéaire en $\mathbf{W}$) des problèmes hyperboliques non linéaires. Pour chacune de ces catégories, nous étudierons tout d'abord le cas *scalaire*, et plus particulierement l'équation modèle d'advection et l'équation modèle de Burgers, puis nous passerons à l'étude de *systèmes* hyperboliques, comme les équations de Maxwell, les équations d'Euler, et enfin un modèle d'écoulement diphasique dispersé.
L'équation d'advection et l'équation de Burgers ont l'avantage d'être des problèmes simples à étudier, de plus leurs solutions ressemblent aux solutions de modèles physiques plus compliqués, comme les équations de la dynamique des gaz. Leur étude servira de base à l'analyse des systèmes cités précédemment, que nous décrivons brièvement ici.

I.1.1 Les équations de Maxwell.

L'étude des phénomènes électromagnétiques consiste à déterminer les quatre champs de vecteurs $\mathbf{E}$, $\mathbf{B}$, $\mathbf{D}$, $\mathbf{H}$ solutions du système tridimensionnel :

$$\begin{cases} \dfrac{\partial \mathbf{D}}{\partial t} - rot(\mathbf{H}) = -\mathbf{j} \\ \dfrac{\partial \mathbf{B}}{\partial t} + rot(\mathbf{E}) = 0 \\ div(\mathbf{D}) = \rho \\ div(\mathbf{B}) = 0 \end{cases} \tag{I.3}$$

Les quantités $\mathbf{E}$, $\mathbf{B}$, $\mathbf{D}$, $\mathbf{H}$ sont des fonctions du temps et de l'espace définies sur $I\!R^3$ x $I\!R$ et à valeurs dans $I\!R^3$.
Ce système, appelé système des équations de Maxwell, est obtenu à partir des lois de conservation intégrales suivantes [2, 1, 9] :

- la loi de Faraday qui relie la force électromotrice à la variation de flux d'induction.
- le théorème d'Ampère, qui permet de calculer le champ magnétique engendré par un courant.
- la loi de Gauss, qui définit la charge électrique.
- la loi de Gauss, qui postule l'absence de charge magnétique.

Nous verrons que les deux dernières équations de (I.3) sont redondantes lorsque l'on considère un problème de Cauchy où elles sont vérifiées à l'instant initial. C'est pourquoi nous les omettrons dans la suite.
Au système (I.3) sont ajoutées la loi de conservation de la charge électrique, qui relie les densités de charge ρ et de courant $\mathbf{j}$:

$$\frac{\partial \rho}{\partial t} + div(\mathbf{j}) = 0 \tag{I.4}$$

ainsi que les relations constitutives des matériaux :

$$\left\{ \begin{array}{l} \mathbf{D} = \epsilon \mathbf{E} \\ \mathbf{B} = \mu \mathbf{H} \end{array} \right. \tag{I.5}$$

qui permettent de lier les inductions électrique et magnétique $\mathbf{D}, \mathbf{B}$ aux champs électrique et magnétique $\mathbf{E}, \mathbf{H}$. Nous nous placerons toujours dans un milieu homogène, linéaire, isotrope, comme le vide; dans ce cas la permittivité ϵ et la perméabilité μ sont constantes.

Le système (I.3) est un système d'équations aux dérivées partielles hyberbolique, conservatif et linéaire. En effet, la matrice jacobienne associée a trois valeurs propres réelles distinctes de multiplicité double : en une dimension d'espace, elles s'écrivent $(c, -c, 0)$, où c désigne la vitesse de la lumière dans le milieu de propagation.

Il existe une autre formulation, appelée formulation harmonique ou fréquentielle, qui s'obtient à partir d'une transformée de Fourier de (I.3). Cette deuxième formulation, dominée par la technique des équations intégrales, a été très utilisée dans les décennies passées, pour des applications concernant la furtivité radar ou le calcul de l'écho radar d'un véhicule [1]. Cependant, cette méthode semble mal adaptée aux géométries complexes des problèmes actuels, et la formulation temporelle, que nous considérons, s'avère plus efficace, lorsque les applications nécessitent une modélisation très fine des phénomènes électromagnétiques [10, 4, 1].

Si on considère le problème de Cauchy pour le système de Maxwell dans le vide :

$$\left\{ \begin{array}{ll} \text{Trouver } (\mathbf{E}(\mathbf{x},t), \mathbf{H}(\mathbf{x},t)) \text{ tels que} & \\ \epsilon \dfrac{\partial \mathbf{E}}{\partial t} - rot(\mathbf{H}) = -\mathbf{j} & \mathbf{x} \in I\!R^3,\ t > 0 \\ \mu \dfrac{\partial \mathbf{H}}{\partial t} + rot(\mathbf{E}) = 0 & \\ \mathbf{E}(x,0) = \mathbf{E}_0(x) & \\ \mathbf{H}(x,0) = \mathbf{H}_0(x) & \end{array} \right. \tag{I.6}$$

La résolution de (I.6) rentre dans le cadre des systèmes de Friedichs [3], et l'unicité de la solution est assurée en considérant les solutions d'énergie finie. Nous avons en particulier le résultat suivant (voir [3, 10]) :

Proposition I.1.1 *On suppose* $(\mathbf{E}_0, \mathbf{H}_0) \in (L^2(I\!R^3))^3 \, \mathrm{x} \, (L^2(I\!R^3))^3$, *et* $\mathbf{j} \in (L^2(I\!R^3 \mathrm{x}[0,T]))^3$. *Toute solution du problème (I.6) vérifie :*

$$||\mathbf{E}||_{L^\infty(L^2(I\!R^3)\mathrm{x}[0,T])} \leq C \left[||\mathbf{E}_0||_{L^2(I\!R^3)} + ||\mathbf{H}_0||_{L^2(I\!R^3)} + ||\mathbf{j}||_{L^1(L^2(I\!R^3)\mathrm{x}[0,T])} \right]$$

où C est une constante strictement positive.

Ce résultat reste valable pour le problème de Cauchy (I.6) muni de conditions aux limites de type métal parfaitement conducteur [1, 3].

Nous passons maintenant au cas de systèmes hyperboliques conservatifs non linéaires.

I.1.2 Les équations d'Euler.

Le système des équations d'Euler, qui décrit l'écoulement d'un fluide non visqueux, s'obtient en écrivant la conservation de la masse, de la quantité de mouvement et de l'énergie totale. Il s'écrit, dans le cas d'un écoulement tridimensionnel :

$$\left\{ \begin{array}{l} \dfrac{\partial \rho}{\partial t} + div(\rho \mathbf{V}) = 0 \\ \dfrac{\partial \rho \mathbf{V}}{\partial t} + div(\rho \mathbf{V} \otimes \mathbf{V}) + \nabla p = 0 \\ \dfrac{\partial E}{\partial t} + div\left((E+p)\mathbf{V}\right) = 0 \end{array} \right. \tag{I.7}$$

ρ représente la masse volumique du fluide, $\mathbf{V}$ le champ de vitesse, p la pression et E l'énergie totale par unité de volume. Le système (I.7) est aussi complété par des lois d'états. En particulier, pour un gaz parfait polytropique, nous avons les lois suivantes :

$$p = \frac{\rho R T}{M}, \quad E = \rho e + \frac{\rho |\mathbf{V}|^2}{2}, \quad e = C_v T,$$

où R est la constante universelle des gaz, M la masse molaire du gaz considéré, e désigne l'énergie interne spécifique du gaz et C_v la chaleur massique à volume constant.
Ces équations sont les plus simples de la dynamique des gaz, toutefois elles rendent compte des phénomènes les plus importants.

Le système (I.7) est hyperbolique; il est strictement hyperbolique dans le cas monodimensionnel. Il a alors trois valeurs propres simples $(u-c, u, u+c)$, où u désigne la vitesse de l'écoulement et c la vitesse du son dans le gaz qui s'écrit :

$$c = \sqrt{\frac{\gamma p}{\rho}}$$

$\gamma = C_p/C_v$ est le rapport entre les chaleurs spécifiques à pression et volume constant, et dans le cas de l'air, $\gamma = 1.4$.

En toute généralité, le système des équations d'Euler peut se réécrire de la manière suivante :

$$\frac{\partial \mathbf{W}}{\partial t} + \sum_{i=1}^{N} \frac{\partial \mathbf{F}_i(\mathbf{W})}{\partial x_i} = 0 \tag{I.8}$$

où N représente la dimension en espace, $\mathbf{W}$ est le vecteur d'état des variables conservatives, il s'écrit :

$$\begin{pmatrix} \rho \\ (\rho u_i)_{i=1,\dots,N} \\ E \end{pmatrix}$$

Les fonctions de flux $\mathbf{F}_i$ sont des fonctions de $\Omega \subset \mathbb{R}^{N+2}$ dans $\mathbb{R}^{N+2}$, et Ω désigne l'espace des solutions admissibles (correspondant à une densité et à une pression positives).

Contrairement au cas linéaire, le problème (I.8) muni de conditions initiales et de conditions aux limites n'admet pas en général de solutions classiques (c'est-à-dire de solutions de classe C^1). En particulier, une condition initiale régulière peut engendrer une solution discontinue. C'est pourquoi on introduit la notion de solutions faibles permettant la recherche d'une solution $\mathbf{W}$ dans un espace qui contient $\mathcal{C}^1(\mathbb{R}^N \mathrm{x} \mathbb{R}_*^+)^3$ pour le problème sans bord en espace. On a alors, sans que cela soit prouvé dans certains cas, existence des solutions mais non unicité. Pour sélectionner la “bonne” solution, on impose une condition supplémentaire appelée “inégalité d'entropie”.

Soit la fonction :

$$\begin{aligned} S : \Omega &\longrightarrow \mathbb{R} \\ \mathbf{W} &\longmapsto S(\mathbf{W}). \end{aligned}$$

On dit que S est une entropie pour le système (I.8) si S est une fonction convexe, et s'il existe des fonctions $(q_i)_{i=1,\dots,N}$ appelées flux d'entropie, telles que :

$$\begin{aligned} q_i : \Omega &\longrightarrow \mathbb{R} \\ \frac{\partial q_i}{\partial \mathbf{W}} &= \frac{\partial S}{\partial \mathbf{W}} \frac{\partial \mathbf{F}_i}{\partial \mathbf{W}}. \end{aligned}$$

On dit qu'une solution faible de (I.8) est entropique si :

$$\frac{\partial S(\mathbf{W})}{\partial t} + \sum_{i=1}^{N} \frac{\partial q_i(\mathbf{W})}{\partial x_i} \leq 0, \tag{I.9}$$

pour toute entropie S.

Ce critère est en fait inspiré de la physique, d'après le second principe de la thermodynamique, seuls sont admissibles les chocs tels que l'entropie S des particules fluides croît à la traversée du choc. Ces chocs sont d'ailleurs appelés “chocs entropiques”. On peut alors prouver, dans certains cas, l'existence et l'unicité de solutions entropiques. En particulier, dans le cas scalaire, l'unicité est obtenue grâce à un résultat dû à Kruzkov. Pour des lois de conservation scalaires, toute fonction convexe est une entropie, mais une solution w est une solution entropique dès qu'elle vérifie la condition d'entropie (I.9) pour le couple entropie-flux d'entropie de Kruzkov :

$$\begin{cases} S(w) = |w-k|, k \in I\!R \\ q_i(w) = sgn(w-k)(f_i(w) - f_k(w)) \quad i = 1, ..., N \end{cases}$$

Pour un gaz parfait polytropique, un couple entropie-flux d'entropie possible est :

$$\begin{cases} S(\mathbf{W}) = -\rho \; log\left(\dfrac{p}{\rho^\gamma}\right) \\ q_i(\mathbf{W}) = -S \; u_i \end{cases}$$

La résolution du problème de Riemann, qui correspond à l'expérience du tube à choc, joue un rôle important pour l'étude d'un système hyperbolique, elle est en particulier à la base du schéma de Godunov.
Il s'agit d'un problème de Cauchy, unidimensionnel où la condition initiale présente une discontinuité :

$$\begin{cases} \dfrac{\partial \mathbf{W}}{\partial t} + \dfrac{\partial \mathbf{F}(\mathbf{W})}{\partial x} = 0 \\ \mathbf{W}(x,0) = \begin{cases} \mathbf{W}_L & \text{si } x < 0 \\ \mathbf{W}_R & \text{si } x > 0 \end{cases} \end{cases} \tag{I.10}$$

La solution générale du problème de Riemann pour un système strictement hyperbolique d'ordre N est constitué de $N+1$ états constants séparés par N ondes, chacune de ces ondes étant soit une onde de raréfaction, soit un choc entropique, soit une discontinuité de contact. Cette solution est unique et il est possible de la construire pour un couple $(\mathbf{W}_L, \mathbf{W}_R)$ donné, cependant la procédure est relativement longue et coûteuse dans le cas d'un système [6].

Dans le cas des équations d'Euler, nous avons le résultat suivant, démontré dans [12] :

Proposition I.1.2 *Sous l'hypothèse*

$$\mathbf{W}_R - \mathbf{W}_L < \frac{2}{\gamma - 1}(c_L + c_R),$$

le problème de Riemann (I.10) admet une unique solution auto-similaire,

$$\mathbf{W}(x,t) = \mathbf{W}(\frac{x}{t}, 1), \quad t > 0,$$

constituée de quatre états constants $\mathbf{W}_L$, $\mathbf{W}^1$, $\mathbf{W}^2$, $\mathbf{W}_R$, les états $\mathbf{W}_L$ et $\mathbf{W}^1$ étant séparés par une 1-onde de raréfaction ou un 1-choc entropique, les états $\mathbf{W}^1$ et $\mathbf{W}^2$ étant séparés par une 2-discontinuité de contact, les états $\mathbf{W}^2$ et $\mathbf{W}_R$ étant séparés par une 3-onde de raréfaction ou un 3-choc entropique.

Comme nous le verrons plus loin, la résolution explicite du problème de Riemann a permis de définir des méthodes numériques efficaces pour la dynamique des gaz.

I.1.3 Un modèle “simplifié” d'écoulement diphasique.

L'étude de ce modèle d'écoulement diphasique va nous conduire en fait au problème plus général des systèmes hyperboliques conservatifs comportant un terme source raide. C'est de cela dont nous allons parler, mais introduisons tout d'abord le modèle diphasique considéré.

On considère un écoulement diphasique dispersé, constitué d'un nuage de particules (soit des gouttelettes de liquide, soit des particules solides) dans un écoulement de gaz. Ce type d'écoulement joue un rôle important dans les applications industrielles, comme les moteurs Diesel, les moteurs à essence à injection directe, les moteurs cryogéniques de fusées ainsi que les moteurs à poudre utilisés dans les boosters de fusées.

Il existe deux approches bien distinctes pour modéliser un écoulement diphasique : l'approche eulérienne ou l'approche cinétique. Dans le modèle eulérien, le nuage de gouttes est représenté comme un fluide continu de même que la phase gazeuse. Au contraire, l'approche cinétique associe au nuage de gouttes une fonction densité de probabilité qui rend compte de la position, de la vitesse, du rayon, de la température de la goutte au temps t. Si le modèle est plus précis et prend en compte un plus grand nombre de phénomènes physiques, il s'avère toutefois beaucoup plus cher [16].
Ici, nous avons considéré un modèle eulérien, où l'écoulement est caractérisé par un champ de vecteurs $\mathbf{W}(\mathbf{x}, t)$ ($\mathbf{x} \in \Omega \subset I\!R^3, 0 \leq t \leq T_0$) qui représente la densité de masse, la vitesse et l'énergie spécifique du gaz et des gouttes. Ce modèle est composé de deux systèmes indépendants associés à chacune des phases, et ils sont couplés par un terme source.

Le modèle original [17, 16] s'écrit en une dimension d'espace sous la forme condensée suivante :

$$\frac{\partial \mathbf{W}}{\partial t} + \mathbf{A}(\mathbf{W})\frac{\partial F(\mathbf{W})}{\partial x} - \frac{\partial}{\partial x}\left(\mathbf{D}(\mathbf{W})\frac{\partial \mathbf{W}}{\partial x}\right) = \frac{1}{\epsilon}\mathbf{R}(\mathbf{W}) \tag{I.11}$$

où $\mathbf{A}(\mathbf{W})$ est la matrice de convection du système, $\mathbf{D}(\mathbf{W})$ représente la matrice de diffusion et $\frac{1}{\epsilon}\mathbf{R}(\mathbf{W})$ modélise les termes de traînée qui ont pour effet la relaxation vers 0 de l'écart des vitesses entre les deux phases. Le paramètre ϵ a la dimension d'un temps de relaxation et il est proportionnel au carré du rayon des gouttes.

Le système (I.11) est obtenu à l'aide d'un processus de moyennes stochastiques des équations de Navier-Stokes écrites au niveau microscopique. Ces moyennes portent sur un ensemble d'expériences ayant les mêmes caractéristiques microscopiques, mais où la position de chaque goutte peut différer d'une expérience à l'autre. Cette modélisation permet de définir la fraction volumique du gaz α comme suit : si ρ_l désigne la masse volumique des gouttes, nous avons : $1 - \alpha = \dfrac{\text{masse des gouttes par unité de volume}}{\rho_l}$.

Les termes de diffusion dans (I.11) sont petits et nous pouvons les omettre. Le système du premier ordre extrait de (I.11) est hyperbolique, cependant les valeurs propres de la matrice $\mathbf{A}(\mathbf{W})$ ne peuvent pas être obtenues analytiquement. De plus, il est non conservatif. En fait, L. Sainsaulieu a montré dans [18], que le système est "faiblement non conservatif" au sens où les termes non conservatifs apparaissent comme des perturbations de termes conservatifs. Ces termes proviennent de la force de gradient de pression qui s'exerce sur les gouttes, ainsi que du travail de cette force.
En définissant les solutions ondes de choc de ce système non-conservatif comme les limites de profils visqueux solutions de (I.11) lorsque la viscosité tend vers 0, et en analysant ces solutions [17], L. Sainsaulieu a introduit un système plus simple, qui est maintenant hyperbolique et conservatif. C'est ce nouveau modèle que nous considèrerons, et nous verrons au chapitre VIII qu'il est alors possible d'obtenir une expression analytique des valeurs propres.

Avec ce modèle plus simple, la principale difficulté réside dans le traitement du terme source, qui devient raide lorsque le rayon des gouttes est petit. Nous avons dans ce problème deux échelles en temps qui interviennent, l'une étant reliée à la convection, l'autre au paramètre de relaxation ϵ. Pour des temps caractéristiques petits devant ϵ, les termes de convection sont prépondérants, l'influence des termes de traînée reste limitée et les solutions de (I.11) se rapprochent de celles obtenues pour un écoulement monophasique. Par contre, lorsque les temps caractéristiques sont de l'ordre de ϵ ou plus grands, les termes de traînée deviennent importants et la dynamique du système (I.11) est différente de celle obtenue lorsqu'on considère un écoulement de gaz pur. Comme nous allons le voir, les méthodes habituelles de résolution numérique introduisent un amortissement non physique des ondes et nous avons cherché à construire une méthode mieux adaptée.

I.2 Description de la résolution numérique.

Différents choix de méthodes numériques sont envisageables pour résoudre ces problèmes, les plus classiques sont les méthodes de type éléments finis et celles de type volumes finis.
Historiquement, la méthode des éléments finis a été développée en premier, dans le but de

résoudre des problèmes non linéaires, mais aussi pour des problèmes paraboliques comme la résolution de l'équation de convection-diffusion. Nous pouvons citer à titre d'exemple, les éléments finis de Lesaint-Raviart [13], la méthode "Streamline Upwind Petrov Galerkin" de Hughes [8] et la méthode "Streamline Diffusion" introduite par Johnson [11]. Dans les deux dernières méthodes, la fonction test tient compte du décentrement, et dans la dernière méthode, un terme de diffusion artificielle est introduit pour permettre que les solutions se comportent mieux au voisinage des discontinuités. Pour toutes ces méthodes, la convergence vers la solution entropique est assurée (voir par exemple [20]).
La méthode de volumes finis reste actuellement la plus utilisée dans le milieu industriel. Moins coûteuse, elle s'avère bien adaptée à la résolution de problèmes hyperboliques conservatifs, puisqu'elle fait appel à des calculs de flux aux interfaces des volumes de contrôle, et elle permet facilement la construction de schémas numériques conservatifs. De plus, la méthode d'éléments finis demande souvent une trop grande régularité de la solution pour les problèmes considérés tandis que la méthode des volumes finis qui repose sur une formulation faible du système, permet de traiter tout type de discontinuité dans les champs. Bien que cette méthode ait été développée spécialement pour résoudre des problèmes non linéaires, comme les équations de la dynamique des gaz, elles ont montré leur efficacité dans le cas de systèmes linéaires, en particulier pour les équations de Maxwell [19, 15, 1]. Dans ce cas, les schémas ne sont pas d'un coût excessif et leurs expressions sont beaucoup plus simples que dans le cas non linéaire.

Pour tous nos problèmes, nous nous plaçons dans un cadre bidimensionnel; accessoirement, nous nous limiterons à l'étude en une dimension d'espace, par exemple pour la résolution numérique de l'équation de Burgers. Pour discrétiser le domaine de calcul, nous considérons deux types de maillages : tout d'abord, les maillages structurés en rectangles, qui se mettent en oeuvre très facilement et qui sont peu coûteux en ce qui concerne le stockage informatique. Cependant ces maillages qui ont été les premiers utilisés et qui le sont encore à l'heure actuelle prennent mal en compte la géométrie du domaine, en particulier aux bords, où la discrétisation "en marches d'escalier" s'est avérée insuffisante. L'autre catégorie de maillages concerne les maillages en triangles, structurés ou non, qui permettent de mailler plus facilement les géométries complexes et qui rendent possible le raffinement local du maillage lorsque cela est nécessaire (contrairement aux maillages en rectangles où cela est impossible). Cependant, ces dernières méthodes entraînent un coût supplémentaire.

Nous utilisons une méthode mixte volumes finis/éléments finis [5] sur des maillages en triangles et en rectangles. Pour des problèmes linéaires, nous considérerons des maillages structurés en rectangles et en triangles. Dans le cas non linéaire, nous nous limitons à des maillages en triangles structurés, ou non. La méthode de résolution numérique utilise

des fonctions de flux décentrées ou centrées, et nous recherchons des schémas hautement précis en temps et en espace.
Dans le cas linéaire, les schémas décentrés d'ordre un en espace sont identiques. En particulier, il n'y a pas de choix parmi les fonctions de flux décentrées pour résoudre le système de Maxwell en milieu homogène [1]. Pour les problèmes non linéaires, de nombreux schémas sont envisageables. La méthode historique pour résoudre les équations de la dynamique des gaz en une dimension d'espace est la méthode de Godunov. Elle est basée sur la résolution de problèmes de Riemann exacts et nous la décrivons au chapitre VI pour l'équation de Burgers. En ce qui concerne les équations d'Euler ou notre modèle d'écoulement diphasique, notre choix s'est porté, pour des raisons de qualité des résultats/coût de calcul, sur une résolution approchée de problèmes de Riemann, par un solveur de Roe.
Pour atteindre un ordre supérieur en espace, nous combinons cette méthode avec une approche de type MUSCL (Monotonic Upwind Schemes for Conservation Laws) introduite par Van Leer [21]. Cette méthode a été initialement développée par la NASA dans les années 80 pour la simulation d'écoulements externes. Cependant, pour des raisons de précision et de stabilité, nous calculons les nouvelles valeurs aux interfaces des cellules à l'aide d'une combinaison convexe de gradients hermitiens centrés, en introduisant un paramètre de décentrage β. Les schémas ainsi obtenus sont connus sous le nom de β-schémas. Avec cette approche, nous pouvons obtenir dans le cas linéaire des schémas d'ordre trois et quatre en espace, pour des maillages structurés en rectangles et en triangles.

Dans le cas d'une équation non linéaire, la notion de variation totale permet de prouver des résultats d'existence et d'unicité pour le problème continu et d'étudier la convergence de la solution approchée vers la solution exacte quand le pas d'espace tend vers 0. Dans le cas continu, elle est définie par $TV(w) = \int_{\mathbb{R}} |\frac{dw}{dx}|dx$ et dans le cas discret par $TV(w_i) = \sum_{i\in\mathbb{Z}} |w_i - w_{i-1}|$. Nous avons vu que la valeur absolue est une entropie dans le cas scalaire, c'est pourquoi la variation totale de la solution w décroît au cours du temps. Les schémas qui ont cette propriété ont été appelés schémas TVD (Total Diminishing Schemes) par Harten [7]. Cependant l'extension à un ordre spatial supérieur à un s'accompagne en général de la perte du caractère TVD du schéma. C'est pourquoi on associe au calcul des pentes de la technique MUSCL, l'utilisation de limiteurs de pentes. Il existe une littérature abondante concernant ce sujet, et nous en discuterons au chapitre VII.

Dans le cas linéaire où nous considérons les équations de Maxwell en régime transitoire, les méthodes numériques mises en oeuvre seront complètement explicites, et hautement précises en temps et en espace. La précision en temps sera réalisée à l'aide de méthodes de type Runge-Kutta d'ordre trois et quatre.
En ce qui concerne les écoulements de gaz pur ou les écoulements diphasiques, nous

envisageons des calculs de solutions instationnaires et stationnaires. Dans ce cas, nous utiliserons des méthodes implicites linéarisées d'ordre élevé en espace.

I.3 Présentation du plan de la thèse.

La thèse est divisée en deux parties bien distinctes, l'une concerne les problèmes linéaires, l'autre le cas non linéaire. Chacune des deux parties comprend trois chapitres ainsi qu'un chapitre de conclusion. Nous détaillons maintenant le contenu des chapitres principaux.

Dans le chapitre II, nous étudions et construisons plusieurs méthodes numériques décentrées de type volumes finis ou éléments finis pour la résolution numérique de l'équation d'advection sur des maillages rectangulaires ou triangulaires. Une étude détaillée de la précision et de la stabilité de ces méthodes permet notamment de dégager plusieurs méthodes nouvelles, précises à l'ordre quatre en temps et en espace.

Nous appliquons dans le chapitre III, la méthode mixte de type éléments finis/volumes finis à la résolution des équations de Maxwell. Nous présentons une étude de stabilité pour les schémas utilisant des maillages en rectangles et en triangles. En particulier, une condition nécessaire et suffisante de stabilité est montrée pour le schéma décentré d'ordre un en maillage rectangulaire. Nous comparons les domaines de stabilité à ceux obtenus à l'aide de schémas décalés, très utilisés en électromagnétisme. Nous présentons également des résultats numériques destinés à comparer l'efficacité et la précision de β-schémas d'ordre trois et quatre en temps et en espace.

Dans le chapitre suivant, nous proposons une nouvelle formulation des équations de Maxwell afin de mieux vérifier numériquement les relations de divergence $div\mathbf{B} = 0$, $div\mathbf{D} = \rho$, qui sont redondantes dans le modèle continu. Nous montrons en établissant les équations équivalentes des schémas précédents et à l'aide d'une étude de stabilité, pourquoi la nouvelle formulation prend mieux en compte ces relations.

Nous passons ensuite à l'étude de problèmes non linéaires. Au chapitre VI, nous détaillons l'approximation numérique de l'équation de Burgers et des équations d'Euler par des schémas explicites hautement précis en temps et en espace utilisant des solveurs de type Godunov et de type Roe. Nous présentons également une méthode de résolution des équations d'Euler par un schéma implicite linéarisé qui utilise la méthode de Jacobi comme méthode de résolution du système linéaire.

Par la suite, nous cherchons à rendre nos schémas TVD. Pour cela, nous détaillons, dans le chapitre VII, quelques techniques de limitation connues, puis nous construisons deux nouveaux limiteurs pour des schémas d'ordre trois et quatre. Nous discutons de l'efficacité de ces limiteurs avec l'expérience du tube à choc de Sod et aussi avec un calcul stationnaire d'écoulement transsonique autour d'un profil d'aile NACA 0012.

Dans le chapitre VIII, nous considérons le modèle "simplifié" hyperbolique et conservatif d'écoulement diphasique pour lequel nous construisons une méthode de volumes finis. La méthode jusque là utilisée pour traiter des termes sources raides est une méthode de pas fractionnaires. Cette méthode donne de bons résultats tant que le pas de temps Δt et le rapport $\Delta t/\epsilon$ restent petits [16].
Cependant, si nous voulons contrôler le pas de temps par une condition de CFL relative au système hyperbolique homogène, les solutions numériques présentent alors un mauvais comportement lorsque ϵ tend vers 0. Contrairement à ce que l'on observe pour certains modèles de combustion, si on considère le cas de la propagation d'ondes sonores dans un milieu diphasique, la vitesse de propagation des ondes est correcte mais la méthode de pas fractionnaires introduit un amortissement non physique des ondes sonores : en particulier, l'amortissement numérique ne tend pas vers 0 avec ϵ, contrairement à la théorie [17].
Aussi proposons-nous une "méthode couplée" qui traite conjointement la partie convective et le terme source.

Les références bibliographiques sont données à la fin de chaque partie, sauf pour les chapitres I et VII où elles figurent en fin de chapitre.

Bibliographie

[1] CIONI J. P., *Résolution numérique des équations de Maxwell instationnaires par une méthode de volumes finis*, Thèse de Doctorat de l'Université de Nice-Sophia-Antipolis (1995).

[2] CIONI J.P., FÉZOUI L., STÈVE H., *Approximation des équations de Maxwell par des schémas décentrés en éléments finis*, Rapport de recherche INRIA no.1601 (1992).

[3] DAUTRAY R., LIONS J.L. *Analyse mathématique et calcul numérique*, Vol. 1, pp 68-127, Masson (1987).

[4] DENIS J. M., VIRETTE L., LAUNEY R., *Use of a finite difference time-domain code for computing radar signatures*, La Recherche Aérospatiale, Vol 5, pp. 343-364, (1994).

[5] DERVIEUX A., *Steady Euler simulations using unstructured meshes*, Von Karman Institute Lectures Series 85-04, (1995).

[6] GODUNOV S. K., Mat. Sb. 47, 271, (1947).

[7] HARTEN A., *High resolution schemes for hyperbolic conservation laws*, Mathematics of Computation, (1982).

[8] HUGHES T.J.R., BROOKS A., *A multidimensional upwind scheme with no cross-wind diffusion*, AMD-Vol 34, Finite Element Methods for Convection Dominated Flows, ed. Hughes (1979).

[9] JACKSON J. D., *Classical Electrodynamics*, seconde édition, John Wiley & Sons, New-York, (1975).

[10] JOLY P., *Equations de Maxwell et ondes électromagnétiques : quelques aspects mathématiques et numériques*, Support de Cours, INRIA-Rocquencourt, (1989).

[11] JOHNSON C., NAVERT U., PITKARANTA J., *Finite element methods for linear hyperbolic problemes*, Computer Methods in Applied Mechanics and Engineering 45, pp. 285-312 (1984).

[12] LARROUTUROU B., *Modélisation mathématique et numérique pour les sciences de l'ingénieur*, Cours de Majeure de l'Ecole Polytechnique (1995).

[13] LESAINT P., *Sur la résolution des systèmes hyperboliques du premier ordre par des méthodes d'éléments finis.*, Thèse de L'Université Paris VI (1975).

[14] LEVÈQUE R. J., *Numerical Methods for Conservation Laws*, Birkhäuser, Basel (1990).

[15] LÖHNER R., AMBRIOSANO J., *A finite element solver for the Maxwell equations*, GAMNI-SMAI Conference on Numerical Methods for the solutions of Maxwell equations, Paris (1989).

[16] SAINSAULIEU L., *Contribution à la modélisation mathématique et numérique des écoulements diphasiques constitués d'un nuage de particules dans un écoulement de gaz.*, Habilitation à diriger les recherches de l'Université Paris VI (1995).

[17] SAINSAULIEU L., *Modélisation, analyse mathématique et numérique d'écoulements diphasiques constitués d'un brouillard de gouttes*, Thèse de Doctorat de l'Ecole Polytechnique (1991).

[18] SAINSAULIEU L., *Finite-volume approximation of two-phase fluid flows based on an approximate roe-type riemann solver*, Rapport CERMICS no 10, (1992).

[19] SHANKAR V., HALL W.F., MOHAMMADIAN A.H., *A time-domain differential solver for electromagnetic scattering problems*, Proceedings on the IEEE, Vol 77, pp. 709-721, No 5 (1989).

[20] SZEPESSY A., *Convergence of the streamline diffusion finite element method for conservation laws.*, Thèse de L'Université de Göteborg, Suède (1989).

[21] VAN LEER B., *Flux vector splitting for the Euler equations*, Lecture Notes in Physics, Vol 170, pp 405-512 (1982)

PREMIÈRE PARTIE :

CAS LINÉAIRE.

Chapitre II

ÉTUDE DE L'ÉQUATION D'ADVECTION.

Réalisé avec Bernard Larrouturou*, Romuald Carpentier**

* Ecole Polytechnique, F-91128 Palaiseau Cedex, France
**CERMICS-INRIA, 06902 Sophia-Antipolis Cedex, France

Ce chapitre est une version un peu plus étendue du rapport Cermics (N^o 95-41) intitulé "Méthodes numériques décentrées d'ordre élevé en deux dimensions d'espace".

II.1 Introduction.

Dans le but de construire des schémas numériques précis en temps et en espace pour la résolution de systèmes hyperboliques en plusieurs dimensions spatiales (tels que les systèmes des équations d'Euler ou des équations de Maxwell), nous allons étudier en détail la précision et la stabilité d'une classe de méthodes numériques d'ordre deux, trois ou quatre opérant sur des maillages plans, triangulaires ou rectangulaires.

Les méthodes d'approximation spatiale considérées seront de type volumes finis et éléments finis. En particulier, certaines des méthodes étudiées ci-dessous sont construites à partir d'une formulation en éléments finis usuelle, P1 ou Q1, que l'on stabilise ensuite par addition d'un terme de viscosité numérique du type de celui utilisé en volumes finis. Sur des maillages triangulaires, une telle méthode revient à des vrais volumes finis, alors que, sur des maillages rectangulaires, le schéma obtenu est nouveau. Bien que plus lourde, cette méthode devrait s'avérer plus précise. La précision en temps étant également importante pour des problèmes instationnaires, nous utiliserons une intégration temporelle explicite multi-pas (Runge-Kutta).

Nous utilisons comme problème modèle l'équation d'advection bidimensionnelle. Les différents schémas étudiés sont présentés dans la section suivante pour cette équation. Pour chacun de ces schémas, nous présenterons ensuite l'étude détaillée de leur précision et de leur stabilité ; en particulier, nous employons une nouvelle méthode pour le calcul de l'équation équivalente de ces schémas numériques, qui était jusque là hors de portée. Après quelques illustrations numériques, nous dégagerons dans la conclusion les principales propriétés de ces méthodes numériques.

II.2 Présentation des méthodes numériques

II.2.1 Equation d'advection

On considère l'équation d'advection linéaire bidimensionnelle :

$$\begin{cases} \partial_t u + \vec{V}.\vec{\nabla} u = 0 \quad \text{pour } (x,y,t) \in I\!R^2 \times [0,+\infty[\ , \\ u(x,y,t) = u_0(x,y) \qquad \text{sur} \quad I\!R^2 \ , \end{cases} \tag{II.1}$$

avec $\vec{V}$ vecteur vitesse constant :

$$\vec{V} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} .$$

La solution initiale $u_0(x,y)$ est supposée périodique, de période 1 en x et en y, ce qui permet de ramener l'étude au domaine $\Omega =]0,1[\times]0,1[$. On peut donc réécrire (II.1) sous la forme:

$$\begin{cases} \partial_t u + c_1 \partial_x u + c_2 \partial_y u = 0 & \text{pour } (x,y,t) \in \Omega \times [0,+\infty[\ , \\ u(x,y,t) = u_0(x,y) & \text{sur } \ \Omega \ . \end{cases} \tag{II.2}$$

On fera parfois apparaître l'angle d'advection θ en écrivant:

$$c_1 = c \cos\theta \ , \quad c_2 = c \sin\theta \ , \tag{II.3}$$

avec $c > 0$.

II.2.2 Schéma d'ordre un en maillage rectangulaire

Décrivons maintenant les méthodes numériques que nous allons étudier, en commençant par leur version la plus simple, c'est-à-dire à l'ordre un, et en considérant d'abord le cas d'un maillage rectangulaire.

Considèrons donc un maillage rectangulaire régulier de pas d'espace Δx et Δy. Nous allons écrire sur ce maillage une méthode de volumes finis et une méthode d'*éléments finis stabilisés* pour la résolution de l'équation (II.2).

Discrétisation spatiale

Commençons par écrire la méthode de volumes finis, classique, que nous utiliserons.

– **Formulation en volumes finis**

Autour de chaque noeud I, de coordonnées $(j\Delta x, k\Delta y)$, on construit une cellule $\mathcal{C}_I$ en utilisant les médianes des rectangles de sommet I. Les cellules ainsi obtenues sont des rectangles d'aire $\Delta x \ \Delta y$.

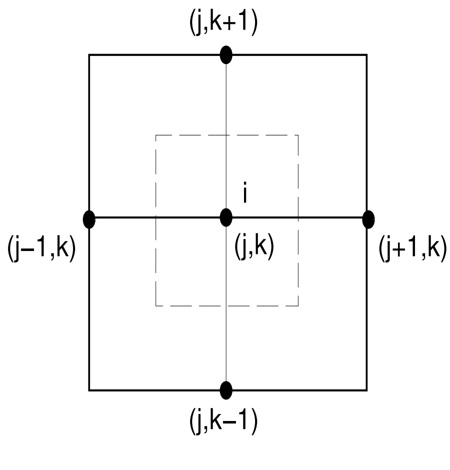


FIG. II.1 – *Cellule C_I*

Avec ces cellules, la formulation d'une méthode de volumes finis pour la résolution de (II.2) s'écrit :

$$\frac{d}{dt}\int\int_{C_I} u\,dxdy + \int\int_{C_I}(c_1\partial_x u + c_2\partial_y u)\,dxdy = \frac{d}{dt}\int\int_{C_I} u\,dxdy + \int_{\partial C_I}(c_1 u\, n_x + c_2 u\, n_y)\,d\sigma = 0\ ,$$

où $\vec{n} = (n_x, n_y)$ est la normale extérieure unitaire sur ∂C_I.

Posant $\mathcal{F}(u,\vec{n}) = c_1 u\, n_x + c_2 u\, n_y$ et utilisant l'hypothèse de périodicité, on peut alors écrire :

$$\int_{\partial C_I} \mathcal{F}(u,\vec{n})\,d\sigma = \sum_{J\in K(I)} \int_{\partial C_{IJ}} \mathcal{F}(u,\vec{n})\,d\sigma\ ,$$

où $K(I)$ est l'ensemble des sommets voisins du sommet I et $\partial C_{IJ} = \partial C_I \cap \partial C_J$.

Avec une approximation décentrée, le terme de flux $\int_{\partial C_{IJ}} \mathcal{F}(u,\vec{n})\,d\sigma$ entre les cellules C_I et C_J est approché par le flux numérique :

$$\Phi_{IJ} = (c_1\nu_{IJ}^x + c_2\nu_{IJ}^y)\frac{(u_I+u_J)}{2} - \mid c_1\nu_{IJ}^x + c_2\nu_{IJ}^y \mid \frac{(u_J - u_I)}{2}\ , \tag{II.4}$$

avec:

$$\vec{\nu}_{IJ} = \int_{\partial C_{IJ}} \vec{n}\,d\sigma\ .$$

Le schéma numérique décentré d'ordre un obtenu à partir de ce flux spatial est un schéma à cinq points.

Dans la suite, nous noterons parfois l'expression (II.4) sous la forme:

$$\Phi_{IJ} = \tilde{\Phi}(u_I, u_J, \vec{\nu}_{IJ})\ , \tag{II.5}$$

où:

$$\tilde{\Phi}(u,v,\vec{\nu}) = (c_1\nu^x + c_2\nu^y)\frac{(u+v)}{2} - \mid c_1\nu^x + c_2\nu^y \mid \frac{(v-u)}{2} \tag{II.6}$$

désigne la fonction de flux numérique.

On obtient alors facilement l'expression du flux spatial total $\Phi = \sum_{J\in K(I)} \Phi_{IJ}$ pour la cellule C_I. Le schéma est écrit en repérant le noeud I par ses indices (j,k).

$$\begin{aligned}\Phi &= \frac{1}{2}c_1\,\Delta y\,(u_{j+1,k} - u_{j-1,k}) + \frac{1}{2}\,c_2\Delta x\,(u_{j,k+1} - u_{j,k-1})\\ &+\frac{1}{2}\mid c_1\Delta y\mid(2u_{jk} - u_{j+1,k} - u_{j-1,k}) + \frac{1}{2}\mid c_2\Delta x\mid(2u_{jk} - u_{j,k+1} - u_{j,k-1})\end{aligned}$$

– **Formulation en éléments finis stabilisés**

Une autre approche possible pour l'approximation spatiale est d'utiliser une formulation en éléments finis $Q1$, que nous stabiliserons ensuite par addition d'un terme de viscosité numérique, comme proposé dans [17].

Considérons donc les fonctions de base des éléments finis Q1; nous noterons Ψ_I la fonction de base associée au noeud I, qui vérifie $\Psi_I(J) = \delta_{IJ}$ pour tout noeud J. On recherche alors une solution u qui se décompose (à un instant $n\Delta t$) dans cette base, sous la forme $u = \sum_J u_J \Psi_J$, et l'équation discrète associée au noeud I fait intervenir le terme:

$$\int\int_{S_I} (c_1 u_x + c_2 u_y)\Psi_i \, dxdy = \sum_J \vec{\mathcal{F}}_J . \int_{S_I} \vec{\nabla}\Psi_J \Psi_I \, dxdy \ , \tag{II.7}$$

avec:

$$\vec{\mathcal{F}}_J = \begin{pmatrix} c_1 \ u_J \\ c_2 \ u_J \end{pmatrix} \ ,$$

S_I désignant le support de la fonction de base Ψ_I.

Posons $\vec{M_{IJ}} = \int_{S_I} \vec{\nabla}\Psi_J \Psi_I \, dxdy$. Deux propriétés de la matrice de vecteurs $\vec{M_{IJ}}$ ont été montrées dans [17] :

Lemme II.2.1 *[17] Pour tout noeud I, on a :*

$$\vec{M_{II}} = \vec{0} \ , \quad \sum_J \vec{M_{IJ}} = \vec{0} \ . \bullet$$

Rappelons la démonstration. Tout d'abord, on a:

$$\vec{M_{II}} = \int_{S_I} \vec{\nabla}\Psi_I \Psi_I \, dxdy = \int_{\partial S_I} \frac{\Psi_I^2}{2} \vec{n} \, ds = \vec{0} \ . \tag{II.8}$$

Ensuite, notant $\mathcal{G} = \sum_J \Psi_J$, on voit facilement que $\mathcal{G}$ est constante et égale à 1 en tout point, si bien que, pour I fixé :

$$\sum_J \vec{M_{IJ}} = \sum_J \int_{S_I} \vec{\nabla}\Psi_J \Psi_I \, dxdy = \int_{S_I} \left(\sum_J \vec{\nabla}\Psi_J \right) \Psi_I \, dxdy = \int_{S_I} \vec{\nabla}\mathcal{G}\Psi_I \, dxdy = \vec{0} \ . \bullet$$

Suivant encore [17], posons $\vec{n_{IJ}} = 2\vec{M_{IJ}}$ et notons Φ_I le terme (II.7). Le lemme permet d'écrire ce terme sous la forme:

$$\Phi_I = \sum_{J \neq I} \frac{\vec{\mathcal{F}}_I + \vec{\mathcal{F}}_J}{2} . \vec{n_{IJ}} \ , \tag{II.9}$$

et on reconnaît ici la partie centrée de l'expression du flux numérique pour une approximation en volumes finis dans laquelle les vecteurs $\vec{n_{IJ}}$ joueraient le rôle des normales $\vec{\nu_{IJ}}$. Puisque le "flux numérique" (II.9), totalement centré, conduirait tel quel à un schéma instable, il est nécessaire pour obtenir un schéma d'ordre un stable d'ajouter un terme de viscosité numérique, que nous écrivons en calquant l'approximation en volumes finis. Ainci, le le flux numérique total de la méthode que nous appellerons "éléments finis stabilisés" s'écrit finalement:

$$\Phi = \sum_{J\neq I}(c_1 n^x_{IJ} + c_2 n^y_{IJ})\frac{(u_I + u_J)}{2} - \mid c_1 n^x_{IJ} + c_2 n^y_{IJ} \mid \frac{(u_J - u_I)}{2} \ . \tag{II.10}$$

Ce flux a donc la même forme que (II.4), mais on utilise maintenant les "normales" $\vec{n_{IJ}} = 2\vec{M_{IJ}} = \int_{S_I} \vec{\nabla}\Psi_J \Psi_I \, dxdy$ issues de la formulation en éléments finis Q1. En particulier, une différence importante avec l'approximation en volumes finis présentée précédemment concerne le nombre de noeuds qui interviennent dans l'expression: en effet, le support des fonctions de base Q1-Lagrange fait intervenir huit voisins et le schéma en éléments finis stabilisés est donc un schéma à neuf points. Bien que l'expression (II.10) soit du type "volumes finis", ce schéma ne correspond pas à de vrais volumes finis, c'est-à-dire à une partition du plan en cellules C_I.

Il est facile de calculer les vecteurs "normaux" $\vec{n_{IJ}}$ et d'en déduire l'expression du flux total Φ au noeud $I = (j,k)$ pour la méthode des éléments finis stabilisés;

$$\begin{aligned}\Phi = & \frac{1}{3}c_1\Delta y\,(u_{j+1,k} - u_{j-1,k}) + \frac{1}{3}\,c_2\Delta x\,(u_{j,k+1} - u_{j,k-1}) \\ & +\frac{1}{3} \mid c_1\Delta y \mid \,(2u_{jk} - u_{j+1,k} - u_{j-1,k}) + \frac{1}{3} \mid c_2\Delta x \mid \,(2u_{jk} - u_{j,k+1} - u_{j,k-1}) \\ & +\frac{1}{12}(c_1\Delta y + c_2\Delta x)\,(u_{j+1,k+1} - u_{j-1,k-1}) + \frac{1}{12}(c_1\Delta y - c_2\Delta x)\,(u_{j+1,k-1} - u_{j-1,k+1}) \\ & +\frac{1}{12} \mid c_1\Delta y + c_2\Delta x \ \mid (2u_{jk} - u_{j+1,k} - u_{j-1,k}) + \frac{1}{12} \mid c_1\,\Delta y + c_2\Delta x \ \mid (2u_{jk} - u_{j,k+1} - u_{j,k-1})\end{aligned} \tag{II.11}$$

Approximation du terme temporel

Pour les schémas précis à l'ordre un, nous considérons deux formulations différentes pour l'intégration temporelle, que nous écrivons rapidement maintenant.

– **<u>Formulation sans matrice de masse</u>**

La technique classique pour le schéma en volumes finis consiste à considérer u constant sur la cellule C_I. On obtient alors un schéma d'ordre un en temps en écrivant:

$$\frac{d}{dt}\int_{C_I} u\,dxdy = aire(C_I)\,\frac{u_I^{n+1} - u_I^n}{\Delta t} = \Delta x\Delta y\frac{u_I^{n+1} - u_I^n}{\Delta t} \ .$$

Pour le schéma en éléments finis, on doit approcher la quantité:

$$\int_{S_I} \frac{u^{n+1} - u^n}{\Delta t}\Psi_I\,dxdy = \sum_J \left(\int_{S_I} \Psi_J\Psi_I\,dxdy\right)\left(\frac{(u_J^{n+1} - u_J^n}{\Delta t}\right) \ . \tag{II.12}$$

La technique la plus simple consiste alors à “diagonaliser la matrice de masse” en remplaçant le terme (II.12) par le terme:

$$\left(\frac{u_I^{n+1}-u_I^n}{\Delta t}\right)\left(\sum_J \int_{S_I} \Psi_J \Psi_I \, dxdy\right) = \left(\frac{u_I^{n+1}-u_I^n}{\Delta t}\right)\left(\int_{S_I} \Psi_I \, dxdy\right) = \Delta x \Delta y \frac{u_I^{n+1}-u_I^n}{\Delta t} \; ; \tag{II.13}$$

on se ramène donc à la même approximation temporelle qu'en volumes finis.

– **<u>Formulation avec matrice de masse.</u>**

Une autre méthode consiste à calculer entièrement la matrice de masse $\left(\int_{S_I} \Psi_J \Psi_I \, dxdy\right)$ qui apparaît dans (II.12). En repérant encore le noeud I par les indices (j, k) de ligne et de colonne, et en notant $\delta u = u^{n+1} - u^n$, on obtient alors comme expression du terme en temps :

$$\frac{\Delta x \Delta y}{36 \Delta t}(16\,\delta u_{jk} + 4\,\delta u_{j+1k} + 4\,\delta u_{j-1k} + 4\delta u_{jk+1} + 4\delta u_{jk-1} + \delta u_{j+1k-1} + \delta u_{j+1k+1} \\ + \delta u_{j-1k-1} + \delta u_{j-1k+1})$$

Dans la suite, il nous a semblé également intéressant d'étudier un schéma mixte volumes finis/éléments finis dans lequel on introduit comme ci-dessus la matrice de masse des éléments finis Q1-Lagrange dans la discrétisation des termes en temps tout en conservant l'approximation spatiale en volumes finis des flux convectifs.

II.2.3 Maillage triangulaire régulier

Nous pouvons aussi écrire les schémas précédents sur des maillages triangulaires (éventuellement non structurés). Pour les besoins de l'analyse cependant, nous allons étudier ici le cas d'une triangulation structurée et régulière, obtenue à partir d'un maillage cartésien en divisant chaque maille rectangulaire en deux triangles.

Approximation spatiale

– **<u>Schéma en volumes finis</u>**

Comme précédemment, nous pouvons construire autour de chaque noeud I une cellule C_I, qui est maintenant délimitée par les médianes des mailles triangulaires (comme par exemple dans [19]; voir la Figure II.2).

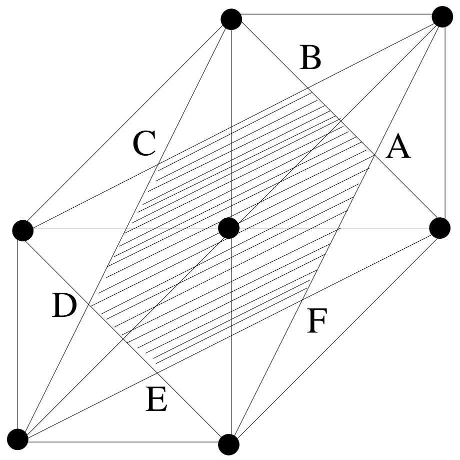


FIG. II.2 – *Cellule C_I*

Le flux spatial total s'écrit toujours $\Phi = \sum_{J \in K(I)} \Phi_{IJ}$, où les flux Φ_{IJ} ont encore la forme (II.4), avec les normales $\vec{\nu}_{IJ} = \int_{\partial C_I \cap \partial C_J} \vec{n} d\sigma$. On obtient alors :

$$\begin{aligned} \Phi = & \frac{1}{6}(c_1 \Delta y + c_2 \Delta x)\,(u_{j+1,k+1} - u_{j-1,k-1}) + \frac{1}{6}\,(-c_1 \Delta y + 2c_2 \Delta x)\,(u_{j,k+1} - u_{j,k-1}) \\ & + \frac{1}{6}(2c_1 \Delta y - c_2 \Delta x)\,(u_{j+1,k} - u_{j-1,k}) + \frac{1}{6} \mid c_1 \Delta y + c_2 \Delta x \mid \;(2u_{jk} - u_{j+1,k+1} - u_{j-1,k-1}) \\ & + \frac{1}{6} \mid -c_1 \Delta y + 2c_2 \Delta x \mid \;(2u_{jk} - u_{j,k+1} - u_{j,k-1}) + \frac{1}{6} \mid 2c_1 \Delta y - c_2 \Delta x \mid (2u_{jk} - u_{j+1,k} - u_{j-1,k}) \end{aligned} \tag{II.14}$$

On aboutit ainsi à un schéma numérique à sept points (voir Figure II.2).

– **<u>Schéma en éléments finis stabilisés</u>**

Bien sûr, nous pouvons aussi appliquer pour le maillage triangulaire la construction des éléments finis stabilisés. Cependant, il a été montré dans [17] que, en choisissant comme fonctions Ψ_J les fonctions de base P1-Lagrange des éléments finis triangulaires, cette méthode redonne exactement le flux (II.14) ! En d'autres termes, on a alors :

$$\vec{n}_{IJ} = 2 \int_{S_I} \vec{\nabla} \Psi_J \Psi_I \, dxdy = \vec{\nu}_{IJ} = \int_{\partial C_{IJ}} \vec{n} \, ds \;, \tag{II.15}$$

si bien que le schéma en volumes finis précédent peut tout aussi bien s'interpréter comme un schéma en éléments finis stabilisés. C'est pourquoi nous l'appellerons parfois ci-dessous le schéma en volumes finis/éléments finis triangulaires.

Approximation du terme temporel

Comme en maillage rectangulaire, deux approches sont possibles pour l'intégration temporelle.

- **Schéma sans matrice de masse**

A l'ordre un, le terme temporel s'écrit alors simplement :

$$\Delta x \Delta y \, \frac{u_I^{n+1} - u_I^n}{\Delta t} \ .$$

- **Schéma avec matrice de masse**

Avec la matrice de masse des fonctions P1-Lagrange, on obtient, pour le noeud I (en posant $\delta u = u^{n+1} - u^n$) :

$$\frac{\Delta x \Delta y}{12 \Delta t}(6\, \delta u_{jk} + \delta u_{j+1k} + \delta u_{j-1k} + \delta u_{jk+1} + \delta u_{jk-1} + \delta u_{j+1k+1} + \delta u_{j-1k-1}) \qquad \text{(II.16)}$$

II.2.4 Extension aux ordres supérieurs

Approximation spatiale d'ordre supérieur

La méthode **M.U.S.C.L** (Monotonic Upwind Schemes for Conservation Laws), introduite par B. van Leer [31] et appliquée aux éléments finis par L. Fezoui [19] permet une extension à l'ordre deux des flux d'espace. On augmente la précision en élevant le degré de l'interpolation de l'inconnue u dans chaque cellule, utilisant une solution linéaire par cellules au lieu d'une solution constante par cellules. Cela nécessite une approximation des gradients de la solution aux noeuds du maillage.

Plus précisément, nous allons introduire pour cette approximation des "pentes" un paramètre de décentrage β, dont la valeur déterminera la précision et la stabilité du schéma.

Présentons d'abord ce "β-schéma" dans le cadre unidimensionnel, où il a été introduit par J.A. Désidéri *et al.* dans [18]. Pour l'équation d'advection monodimensionnelle :

$$\begin{cases} \partial_t u + c\, \partial_x u = 0 \ , & \text{avec } c > 0 \text{ constant} \ , \\ u(x,t) = u_0(x) & \text{sur } \ I\!R \ , \end{cases} \qquad \text{(II.17)}$$

on construit le β-schéma avec une combinaison des pentes centrées et décentrées au noeud i :

$$u_i^{n+1} = u_i^n - \frac{c\Delta t}{2\Delta x}((1-\beta)(u_{i+1}^n - u_{i-1}^n) + \beta(3u_i^n - 4u_{i-1}^n + u_{i-2}^n)) \ . \qquad \text{(II.18)}$$

On obtient pour $\beta = 0$ un schéma centré (en espace), pour $\beta = \frac{1}{2}$ un schéma "demi-décentré" et pour $\beta = 1$ un schéma totalement décentré. La précision et la stabilité de ce β-schéma, étudiées dans [18], seront détaillées plus loin.

Revenons au problème bidimensionnel. Comme dans l'approche MUSCL, nous allons garder la même fonction de flux numérique $\tilde{\Phi}$ que dans (II.4)-(II.5) et écrire maintenant, pour le schéma en volumes finis:

$$\Phi_{IJ} = \tilde{\Phi}\left(u_{IJ}, u_{JI}, \vec{\nu}_{IJ}\right) \ ,$$

où u_{IJ} et u_{JI} sont les états interpolés à l'interface $\partial C_I \cap \partial C_J$ dans les cellules C_I et C_J respectivement. Ces valeurs seront évaluées par les formules:

$$u_{IJ} = u_I + \frac{1}{2}[(1-2\beta)(u_J - u_I) + 2\beta \vec{\nabla} u_I . \vec{IJ}] \ ,$$

$$u_{JI} = u_J - \frac{1}{2}[(1-2\beta)(u_J - u_I) + 2\beta \vec{\nabla} u_J . \vec{IJ}] \ ,$$

où $\vec{\nabla} u_I$ et $\vec{\nabla} u_J$ désignent des approximations centrées du gradient de u dans les cellules C_I et C_J. Ainsi, selon notre choix de la valeur de β, nous utilisons une valeur plus ou moins décentrée du gradient pour l'interpolation dans chaque cellule. En une dimension, il est facile de voir que l'on retrouve bien le β-schéma (II.18) en prenant $\nabla u_i = \dfrac{u_{i+1} - u_{i-1}}{2\Delta x}$ dans les expressions précédentes. En deux dimensions, on retrouvera l'approche MUSCL (avec des pentes centrées) en prenant $\beta = \dfrac{1}{2}$.

Il reste à préciser comment nous évaluerons le gradient centré $\vec{\nabla} u_I$, selon les trois méthodes utilisées (volumes finis rectangulaires, éléments finis stabilisés rectangulaires ou volumes finis/éléments finis triangulaires).

1. Schéma en volumes finis rectangulaires

   A partir de la formule de Green $\int\int_{C_I} \vec{\nabla} u \, dxdy = \int_{\partial C_I} u\vec{n} \, ds$, nous écrirons:

$$\vec{\nabla} u_I = \frac{1}{aire(C_I)} \sum_{J \in K(I)} \left(\frac{u_I + u_J}{2} \right) \vec{\nu}_{IJ} \ . \tag{II.19}$$

2. Schéma en éléments finis stabilisés rectangulaires

   On utilisera dans ce cadre l'approximation suivante :

$$\vec{\nabla} u_I = \frac{\displaystyle\int_{S_I} \sum_J u_J \vec{\nabla} \Psi_J \, dxdy}{\displaystyle\int_{S_I} dxdy} \ , \tag{II.20}$$

   où les fonctions Ψ_J sont les fonctions de base Q1, et S_I le support de la fonction Ψ_I.

3. Schéma en volumes finis/éléments finis triangulaires

   En maillage triangulaire, on utilise aussi une approximation par éléments finis du gradient, calculée avec l'expression (II.20) dans laquelle on emploie les fonctions de base P1 des éléments triangulaires (on obtiendrait d'ailleurs un résultat analogue avec l'expression (II.19) en y utilisant les normales $\vec{\nu}_{IJ}$ des cellules associées au maillage triangulaire).

On aboutit ainsi à un β-schéma à neuf points dans le cas des volumes finis rectangulaires, à vingt-cinq points pour le schéma éléments finis rectangulaires et à dix-neuf points pour le schéma en maillage triangulaire.

Intégration en temps

Nous souhaitons aussi choisir une intégration temporelle précise puisque nous nous intéressons à un problème instationnaire. Pour cela, nous utilisons la méthode de Runge-Kutta explicite pour obtenir des schémas d'ordre deux, trois et quatre en temps. Rappelons l'algorithme Runge-Kutta d'ordre N (pour $1 \leq N \leq 4$):

$$\begin{cases} u^0 = u^n \ , \\ M\, u^{(l)} = M\, u^{(0)} - \dfrac{\Delta t}{(N+1-l)} \Phi(u^{(l-1)}) \quad l = 1, 2, ..., N \ , \\ u^{n+1} = u^{(N)} \ , \end{cases} \tag{II.21}$$

où $t^n = n\Delta t$ et où Φ représente le flux spatial total et M la matrice de masse entière ou diagonalisée.

II.3 Equations équivalentes

Le schéma en éléments finis stabilisés en maillage rectangulaire est le plus coûteux des trois schémas étudiés. Cela est dû au fait que l'on calcule une contribution du flux avec chaque voisin du noeud I. De plus, conserver la matrice de masse dans l'approximation en temps augmente le temps de calcul (à cause de l'inversion de cette matrice). Nous cherchons cependant à savoir si, malgré leur coût supérieur, ces schémas en éléments finis stabilisés avec matrice de masse s'avèrent plus précis et plus efficaces. C'est pourquoi nous allons analyser en détail la précision de tous les schémas introduits précédemment en calculant leurs équations équivalentes.

Les équations équivalentes, introduites par Warming and Hyett [32] permettent une analyse détaillée et précise des termes d'erreur de troncature; notamment des erreurs de dispersion et de dissipation. Elles donnent aussi des critères de stabilité des schémas. Pour obtenir ces équations, nous allons appliquer la méthode décrite dans [4], qui simplifie considérablement les calculs dans le cas linéaire : cette méthode reste la même quelque soit le schéma étudié (implicite ou explicite, schéma à plusieurs niveaux comme ceux de Runge-Kutta) et quelle que soit la dimension d'espace.

II.3.1 Méthode de calcul

Rappelons sans démonstration la méthode de calcul de l'équation équivalente, en renvoyant à [4] pour les détails. Considérons l'équation aux dérivées partielles d'évolution suivante, en deux dimensions d'espace:

$$u_t = \sum_{0 \le k,m \le K} \gamma_{k,m} \frac{\partial^{k+m} u}{\partial x^k \partial y^m} \ , \tag{II.22}$$

approchée par un schéma numérique avec matrice de masse, qui, à l'ordre un en temps, s'écrit:

$$\sum_{k,m} B_{k,m} \frac{u^{n+1}_{j+k,l+m} - u^n_{j+k,l+m}}{\Delta t} = \sum_{k,m} A_{k,m}(\Delta x, \Delta y) u^n_{j+k,l+m} \ . \tag{II.23}$$

Si l'on utilise un schéma de Runge-Kutta d'ordre N comme dans (II.21), alors l'équation équivalente de la méthode numérique s'écrit:

$$u_t = \sum_{0 \le k,m} \alpha_{k,m}(\Delta t, \Delta x, \Delta y) \frac{\partial^{k+m} u}{\partial x^k \partial y^m} \ , \tag{II.24}$$

où $\sum_{0 \le k,m} \alpha_{k,m}(\Delta t, \Delta x, \Delta y) X^k Y^m$ est le développement en série entière de la fonction:

$$\mathcal{F}(X,Y) = \frac{\log\left(1 + \sum_{p=1}^{N} \frac{\Delta t^p g_{\Delta x,\Delta y}(X,Y)^p}{p! \ h_{\Delta x,\Delta y}(X,Y)^p}\right)}{\Delta t} \ , \tag{II.25}$$

les fonctions $g_{\Delta x,\Delta y}(X,Y)$ et $h_{\Delta x,\Delta y}(X,Y)$ étant définies par:

$$g_{\Delta x,\Delta y}(X,Y) = \sum_{0 \le k,m} A_{k,m}(\Delta x, \Delta y) e^{k\Delta x X} e^{m \Delta y Y} \tag{II.26}$$

$$h_{\Delta x,\Delta y}(X,Y) = \sum_{0 \le k,m} B_{k,m}(\Delta x, \Delta y) e^{k\Delta x X} e^{m \Delta y Y} \ . \tag{II.27}$$

Pour examiner l'influence de la matrice de masse, explicitons maintenant l'écriture de l'équation équivalente pour les schémas de Runge-Kutta. Pour un flux spatial donné, en supposant le rapport $\frac{\Delta y}{\Delta x}$ constant, on écrit la fonction $g_{\Delta x,\Delta y}$ sous la forme $g_{\Delta x,\Delta y}(X,Y) =$

$g_0(X,Y) + \sum_{p\geq 1} \Delta x^p g_p(X,Y)$ avec $g_p(X,Y)$ fonctions polynomiales en X, Y. Notons que, dans ce développement, le premier terme g_0 est imposé par la consistance du schéma (par exemple $g_0 = -c_1 X - c_2 Y$ pour tout schéma consistant avec l'équation d'advection (II.2)), et que le terme suivant g_1 est nul lorsque le flux spatial est précis à l'ordre deux.

On obtient alors à l'ordre quatre les équations équivalentes suivantes pour un schéma Runge-Kutta N $(1 \leq N \leq 4)$, dans le cas des schémas sans matrice de masse:

$$\begin{aligned}\mathcal{F}_{(N=1)} = g_0 + \Delta x g_1 - \frac{\Delta t}{2} g_0^2 + \Delta x^2 g_2 - \Delta t \Delta x g_0 g_1 + \frac{\Delta t^2}{3} g_0^3 + \Delta x^3 g_3 - \frac{\Delta t \Delta x^2}{2} g_1^2 \\ -\Delta t \Delta x^2 g_0 g_2 + \Delta t^2 \Delta x g_0^2 g_1 - \frac{\Delta t^3}{4} g_0^4 + O(\Delta t, \Delta x)^4 \ ,\end{aligned} \tag{II.28}$$

$$\mathcal{F}_{(N=2)} = g_0 + \Delta x g_1 + \Delta x^2 g_2 - \frac{\Delta t^2}{6} g_0^3 + \Delta x^3 g_3 - \frac{\Delta t^2 \Delta x}{2} g_0^2 g_1 + \frac{\Delta t^3}{8} g_0^4 + O(\Delta t, \Delta x)^4 \ , \tag{II.29}$$

$$\mathcal{F}_{(N=3)} = g_0 + \Delta x g_1 + \Delta x^2 g_2 + \Delta x^3 g_3 - \frac{\Delta t^3}{24} g_0^4 + O(\Delta t, \Delta x)^4 \ , \tag{II.30}$$

$$\mathcal{F}_{(N=4)} = g_0 + \Delta x g_1 + \Delta x^2 g_2 + \Delta x^3 g_3 + O(\Delta t, \Delta x)^4 \ . \tag{II.31}$$

Dans le cas de schémas avec matrice de masse, on donne l'expression de la fonction h pour les deux types de maillages considérés:

$$h(X,Y)_{rectangle} = 1 + \frac{1}{6} X^2 \Delta x^2 + \frac{1}{6} Y^2 \Delta y^2 + O(\Delta x, \Delta y)^4 \ , \tag{II.32}$$

$$h(X,Y)_{triangle} = 1 + \frac{1}{6} X^2 \Delta x^2 + \frac{1}{6} Y^2 \Delta y^2 + \frac{1}{6} XY \Delta x \Delta y + O(\Delta x, \Delta y)^4 \ , \tag{II.33}$$

ce qui s'écrit aussi formellement: $h(X,Y) = 1 + h_2 \Delta x^2 + O(\Delta x^4)$.
On obtient alors les expressions suivantes:

$$\begin{aligned}\mathcal{F}_{(N=1)} = g_0 + \Delta x g_1 - \frac{\Delta t}{2} g_0^2 + \Delta x^2 g_2 - \Delta x^2 h_2 g_0 - \Delta t \Delta x g_0 g_1 + \frac{\Delta t^2}{3} g_0^3 + \Delta x^3 g_3 \\ -h_2 g_1 \Delta x^3 - \frac{\Delta t \Delta x^2}{2} g_1^2 - \Delta t \Delta x^2 g_0 g_2 + \Delta t \Delta x^2 g_0^2 h_2 + \Delta t^2 \Delta x g_0^2 g_1 - \frac{\Delta t^3}{4} g_0^4 + O(\Delta t, \Delta x)^4 \ ,\end{aligned} \tag{II.34}$$

$$\begin{aligned}\mathcal{F}_{(N=2)} = g_0 + \Delta x g_1 + \Delta x^2 g_2 - \Delta x^2 h_2 g_0 - \frac{\Delta t^2}{6} g_0^3 + \Delta x^3 g_3 - h_2 g_1 \Delta x^3 - \frac{\Delta t^2 \Delta x}{2} g_0^2 g_1 \\ + \frac{\Delta t^3}{8} g_0^4 + O(\Delta t, \Delta x)^4 \ ,\end{aligned} \tag{II.35}$$

$$\mathcal{F}_{(N=3)} = g_0 + \Delta x g_1 + \Delta x^2 g_2 - \Delta x^2 h_2 g_0 + \Delta x^3 g_3 - h_2 g_1 \Delta x^3 - \frac{\Delta t^3}{24} g_0^4 + O(\Delta t, \Delta x)^4 \ , \tag{II.36}$$

$$\mathcal{F}_{(N=4)} = g_0 + \Delta x g_1 + \Delta x^2 g_2 - \Delta x^2 h_2 g_0 + \Delta x^3 g_3 - h_2 g_1 \Delta x^3 + O(\Delta t, \Delta x)^4 \ . \tag{II.37}$$

Sur ces expressions, on remarque que les termes d'"erreur temporelle" (faisant intervenir Δt) sont les mêmes quel que soit le schéma spatial considéré dès que le flux spatial est précis à l'ordre deux (puisqu'alors le terme g_1 est nul et ces termes ne dépendent que du polynôme g_0 ; bien sûr, ces termes d'erreur temporelle disparaissent jusqu'à l'ordre quatre lorsqu'on emploie un schéma d'ordre quatre en temps). On voit aussi que la prise en compte de la matrice de masse modifie les termes d'erreur de dispersion troisième et de dissipation quatrième dans le cas d'un schéma spatial d'ordre un, et seulement les termes de dispersion troisième pour des schémas d'ordre deux en temps et en espace.

II.3.2 Equations équivalentes des schémas.

On écrit maintenant les équations équivalentes des schémas étudiés, avec ou sans matrice de masse, pour $1 \leq N \leq 4$, avec une approximation des flux spatiaux d'ordre un puis d'ordre deux. Pour plus de clarté, on adoptera doréravant les notations suivantes: **VF R** désignera le schéma en volumes finis rectangulaires, **EF R** le schéma en éléments finis stabilisés rectangulaires et **VF/EF T** le schéma en volumes finis/éléments finis triangulaires.

Pour ne pas alourdir les expressions des termes d'erreur on suppose ici que $c_1 = c_2$ dans (II.2) et on se place dans le cas particulier $\Delta x = \Delta y = h$. On note "en gras" les termes d'erreur affectés par la présence de la matrice de masse.

On pourra se référer à l'Annexe C qui donne les expressions des équations équivalentes pour $(c_1, c_2, \Delta x, \Delta y)$ quelconques.

Schémas d'ordre un.

– Schéma **VF R** sans matrice de masse.

$$\begin{aligned} u_t + c_1(u_x + u_y) = {} & (-\frac{\Delta t}{2}c_1^2 + \frac{|c_1|}{2}h)\,(u_{xx} + u_{yy}) + (-\frac{\Delta t^2}{3}c_1^3 + \frac{c_1^2}{2}\Delta t\, h - \mathbf{c_1}\frac{\mathbf{h^2}}{\mathbf{6}})\,(u_{xxx} + u_{yyy}) \\ & -\Delta t c_1^2 u_{xy} + (-\Delta t^2 c_1^3 + \frac{\Delta t\, h}{2}|c_1|c_1)\,(u_{xxy} + u_{xyy}) + O(\Delta t, h)^3 \end{aligned} \tag{II.38}$$

– Schéma **VF R** avec matrice de masse.

$$\begin{aligned} u_t + c_1(u_x + u_y) = {} & (-\frac{\Delta t}{2}c_1^2 + \frac{|c_1|}{2}h)\,(u_{xx} + u_{yy}) + (-\frac{\Delta t^2}{3}c_1^3 + \frac{c_1^2}{2}\Delta t\, h)\,(u_{xxx} + u_{yyy}) \\ & -\Delta t c_1^2 u_{xy} + (-\Delta t^2 c_1^3 + \frac{\Delta t\, h}{2}|c_1|c_1 + \mathbf{c_1}\frac{\mathbf{h^2}}{\mathbf{6}})\,(u_{xxy} + u_{xyy}) + O(\Delta t, h)^3 \end{aligned} \tag{II.39}$$

– Schéma **EF R** sans matrice de masse.

$$\begin{aligned} u_t + c_1(u_x + u_y) = {} & (-\frac{\Delta t}{2}c_1^2 + \frac{|c_1|}{2}h)\,(u_{xx} + u_{yy}) + (-\frac{\Delta t^2}{3}c_1^3 + \frac{c_1|c_1|}{2}\Delta t\, h - \mathbf{c_1}\frac{\mathbf{h^2}}{\mathbf{6}})\,(u_{xxx} + u_{yyy}) \\ & +(-\Delta t c_1^2 + \frac{|c_1|\,h}{3})u_{xy} + (-\Delta t^2 c_1^3 + \frac{5\Delta t\, h}{6}|c_1|c_1 - \mathbf{c_1}\frac{\mathbf{h^2}}{\mathbf{6}})\,(u_{xxy} + u_{xyy}) + O(\Delta t, h)^3 \end{aligned} \tag{II.40}$$

– Schéma **EF R** avec matrice de masse.

$$\begin{aligned} u_t + c_1(u_x+u_y) &= (-\frac{\Delta t}{2}c_1^2 + \frac{|c_1|}{2}h)\,(u_{xx}+u_{yy}) + (-\frac{\Delta t^2}{3}c_1^3 + \frac{c_1|c_1|}{2}\Delta t\, h)\,(u_{xxx}+u_{yyy}) \\ &+(-\Delta t c_1^2 + \frac{|c_1|\,h}{3})u_{xy} + (-\Delta t^2 c_1^3 + \frac{5\Delta t\, h}{6}|c_1|c_1)\,(u_{xxy}+u_{xyy}) + O(\Delta t, h)^3 \end{aligned} \tag{II.41}$$

– Schéma **VF/EF T** sans matrice de masse.

$$\begin{aligned} u_t + c_1(u_x+u_y) &= (-\frac{\Delta t}{2}c_1^2 + \frac{|c_1|}{2}h)\,(u_{xx}+u_{yy}) + (-\frac{\Delta t^2}{3}c_1^3 + \frac{c_1|c_1|}{2}\Delta t\, h - \mathbf{c_1}\frac{\mathbf{h^2}}{\mathbf{6}})\,(u_{xxx}+u_{yyy}) \\ &+(-\Delta t c_1^2 + \frac{2|c_1|\,h}{3})u_{xy} + (-\Delta t^2 c_1^3 + \frac{7\Delta t\, h}{6}|c_1|c_1 - \mathbf{c_1}\frac{\mathbf{h^2}}{\mathbf{3}})\,(u_{xxy}+u_{xyy}) + O(\Delta t, h)^3 \end{aligned} \tag{II.42}$$

– Schéma **VF/EF T** avec matrice de masse.

$$\begin{aligned} u_t + c_1(u_x+u_y) &= (-\frac{\Delta t}{2}c_1^2 + \frac{|c_1|}{2}h)\,(u_{xx}+u_{yy}) + (-\frac{\Delta t^2}{3}c_1^3 + \frac{c_1|c_1|}{2}\Delta t\, h)\,(u_{xxx}+u_{yyy}) \\ &+(-\Delta t c_1^2 + \frac{2|c_1|\,h}{3})u_{xy} + (-\Delta t^2 c_1^3 + \frac{7\Delta t\, h}{6}|c_1|c_1)\,(u_{xxy}+u_{xyy}) + O(\Delta t, h)^3 \end{aligned} \tag{II.43}$$

On remarque que les termes d'erreur "non croisés" (portant sur les directions x ou y) sont identiques pour tous les schémas. Seuls les termes d'erreur de diffusion et de dispersion "croisés" (u_{xy}, u_{xxy}, u_{xyy}) diffèrent selon le schéma considéré.
La matrice de masse a pour effet d'annuler les termes de dispersion spatiale sauf dans le cas du schéma **VF R** où apparaissent des termes croisés d'ordre trois.

Schémas d'ordre élevé (β-schémas).

– Schéma **VF R** sans matrice de masse.

$$u_t + c_1(u_x+u_y) = -\frac{c_1}{6}(1-3\beta)h^2\,(u_{xxx}+u_{yyy}) - \frac{\beta}{4}|c_1|h^3(u_{xxxx}+u_{yyyy}) + O(\Delta t, h)^4 \tag{II.44}$$

– Schéma **VF R** avec matrice de masse.

$$\begin{aligned} u_t + c_1(u_x+u_y) = \frac{\mathbf{c_1}\boldsymbol{\beta}}{\mathbf{2}}h^2\,(u_{xxx}+u_{yyy}) - \frac{\beta}{4}|c_1|h^3(u_{xxxx}+u_{yyyy}) + \frac{\mathbf{c_1 h^2}}{\mathbf{6}}\,(u_{xyy}+u_{xxy}) \\ +O(\Delta t, h)^4 \end{aligned} \tag{II.45}$$

– Schéma **EF R** sans matrice de masse.

$$u_t + c_1(u_x + u_y) = -\frac{c_1}{6}(1-3\beta)h^2\,(u_{xxx} + u_{yyy} + u_{xxy} + u_{xyy}) - \frac{\beta}{4}|c_1|h^3(u_{xxxx} + u_{yyyy} + u_{xxxy} + u_{xyyy}) - \frac{2\beta}{3}|c_1|h^3 u_{xxyy} + O(\Delta t, h)^4 \tag{II.46}$$

– Schéma **EF R** avec matrice de masse.

$$u_t + c_1(u_x + u_y) = \frac{\mathbf{c_1}\boldsymbol{\beta}}{\mathbf{2}}h^2\,(u_{xxx} + u_{yyy} + u_{xxy} + u_{xyy}) - \frac{\beta}{4}|c_1|h^3(u_{xxxx} + u_{yyyy} + u_{xxxy} + u_{xyyy}) - \frac{2\beta}{3}|c_1|h^3 u_{xxyy} + O(\Delta t, h)^4 \tag{II.47}$$

– Schéma **VF/EF T** sans matrice de masse.

$$u_t + c_1(u_x + u_y) = -\frac{c_1}{6}(1-3\beta)h^2\,(u_{xxx} + u_{yyy} + 2u_{xxy} + 2u_{xyy}) - \frac{\beta}{4}|c_1|h^3(u_{xxxx} + u_{yyyy}) - \frac{11\beta}{18}|c_1|h^3(u_{xxxy} + u_{xyyy}) - \frac{8\beta}{9}|c_1|h^3 u_{xxyy} + O(\Delta t, h)^4 \tag{II.48}$$

– Schéma **VF/EF T** avec matrice de masse.

$$u_t + c_1(u_x + u_y) = \frac{\mathbf{c_1}\boldsymbol{\beta}}{\mathbf{2}}h^2\,(u_{xxx} + u_{yyy} + 2u_{xxy} + 2u_{xyy}) - \frac{\beta}{4}|c_1|h^3(u_{xxxx} + u_{yyyy}) - \frac{11\beta}{18}|c_1|h^3(u_{xxxy} + u_{xyyy}) - \frac{8\beta}{9}|c_1|h^3 u_{xxyy} + O(\Delta t, h)^4 \tag{II.49}$$

Comme à l'ordre un, seuls les termes d'erreur "non croisés" sont les mêmes pour les trois schémas (avec ou sans matrice de masse). On remarque que le schéma **VF R** est le seul à n'avoir aucun terme d'erreur croisé (avec ou sans matrice de masse). Les expressions (II.47) et (II.49) révèlent que le schéma **EF R** se comporte mieux en général que le schéma **VF/EF T**: les termes d'erreur de dispersion et de dissipation sont plus petits pour le schéma **EF R**.

La matrice de masse modifie les termes de dispersion puisqu'on passe d'un facteur $(\frac{\beta}{2} - \frac{1}{6})$ sans matrice de masse à un facteur $\frac{\beta}{2}$ avec matrice de masse. Le signe de ces termes dispersifs change pour $\beta \in [0,\frac{1}{3}]$ et reste le même pour $\beta \in [\frac{1}{3},1]$; ces termes d'erreur sont réduits en valeur absolue lorsque $\beta \in [0,\frac{1}{6}]$ et amplifiés pour $\beta \in [\frac{1}{6},1]$. Remarquons que pour le schéma **VF R**, la matrice de masse donne des termes d'erreur supplémentaires, des termes croisés de dispersion.

Schémas d'ordre trois ou quatre

L'étude des équations équivalentes montre que deux valeurs de β sont intéressantes pour minimiser les erreurs de dispersion et de dissipation. On donne dans le tableau II.1 les différentes précisions en espace obtenues selon ces deux valeurs du paramètre de décentrage: $\beta = \frac{1}{3}$ et $\beta = 0$.

Schémas		$\beta=\frac{1}{3}$	$\beta=0$
VF R	sans masse	ordre 3	ordre 2
	avec masse	ordre 2	ordre 2
EF R	sans masse	ordre 3	ordre 2
	avec masse	ordre 2	ordre 4
EF/VF T	sans masse	ordre 3	ordre 2
	avec masse	ordre 2	ordre 4

TAB. II.1 – *Précision spatiale des β-schémas.*

Sans matrice de masse, le choix $\beta = \frac{1}{3}$ permet d'annuler les termes de dispersion d'ordre deux; avec cette valeur et $N \geq 3$ on obtient donc des schémas sans matrice de masse d'ordre trois. Avec matrice de masse, le choix $\beta = 0$ permet d'annuler les termes de dispersion d'ordre deux et les termes de dissipation d'ordre trois des schémas **EF R** et **EF/VF T**; *avec $\beta = 0$ et $N \geq 4$, les schémas* **EF R** *et* **EF/VF T** *avec matrice de masse sont donc d'ordre quatre en temps et en espace.* Notons que le schéma **VF R** avec matrice de masse est d'ordre deux, cela provient de la présence de termes de dispersion croisés indépendants de β dans l'équation équivalente (II.45).

Dans le cas des schémas en volumes finis (rectangles et triangles), ces résultats généralisent ceux obtenus par J. A. Désidéri et al. [18] et S. Lantéri [25] pour l'étude monodimensionnelle: sans matrice de masse, $\beta = \frac{1}{3}$ annule le terme de dispersion u_{xxx}, β=0 annule le terme de dissipation u_{xxxx}; avec matrice de masse, β=0 annule à la fois les termes u_{xxx} et u_{xxxx}.

II.3.3 Schémas "sans diffusion numérique"

Un autre moyen de mettre au point des schémas d'ordre quatre en temps et en espace est proposé dans [5]. Ce moyen consiste à remplacer la fonction de flux numérique (II.6) par une fonction de flux numérique centrée :

$$\tilde{\Phi}(u, v, \vec{\nu}) = (c_1 \nu^x + c_2 \nu^y) \frac{(u+v)}{2}, \tag{II.50}$$

et à conserver par ailleurs le β-schéma. Avec $\beta = \frac{1}{3}$ et RK4, sans matrice de masse, on obtient alors des schémas **VF R**, **EF R** et **EF/VF T** d'ordre quatre en temps et en espace. On peut le montrer à l'aide des équations équivalentes (II.44),(II.46),(II.48) puisque lorsqu'on utilise (II.6) à la place de (II.50), les termes d'erreur de dissipation disparaissent. Seuls subsistent les termes d'erreur de dispersion, et nous avons vu que la valeur $\beta = \frac{1}{3}$ annulait le terme de dispersion d'ordre trois.

Pour les différencier des schémas d'ordre quatre obtenus ci-dessus avec $\beta = 0$ et avec matrice de masse, nous appellerons ces derniers schémas "schémas sans diffusion numérique".

II.4 Etude de stabilité des schémas

Etudions maintenant la stabilité des différents schémas vus précédemment. Nous utilisons une analyse de Fourier et posons:

$$u^n_{j,k} = \hat{u}^n e^{i(j\theta_1 + k\theta_2)} \ ,$$

avec $i^2 = -1$. On obtient alors la relation:

$$\hat{u}^{n+1} = G_{\theta_1,\theta_2} \ \hat{u}^n \ ,$$

où G_{θ_1,θ_2} est le facteur d'amplification dépendant de Δt, θ_1, θ_2, et la condition nécessaire et suffisante de stabilité s'écrit:

$$|G_{\theta_1,\theta_2}| \leq 1 \ \ \forall \ \ (\theta_1, \theta_2) \ \in [0, 2\Pi]^{\ 2} \ .$$

Nous allons dans cette section déterminer les limites de stabilité de tous les schémas présentés dans les sections précédentes. Ces valeurs limites du pas de temps assurant la stabilité du schéma seront déterminées numériquement, sauf dans certains cas où nous avons pu mener une étude analytique.

II.4.1 Schémas précis à l'ordre un

La figure II.3 représente les différents domaines de stabilité des schémas d'ordre un, avec ou sans matrice de masse: en faisant varier l'angle d'advection θ, on a tracé les valeurs maximales de $\alpha_1 = \frac{c_1 \Delta t}{\Delta x}$ et $\alpha_2 = \frac{c_2 \Delta t}{\Delta y}$ qui assurent la stabilité, avec α_1 en abscisse et α_2 en ordonnée. On observe que les domaines de stabilité des schémas **VF R** et **EF R** sont symétriques par rapport aux axes et à la première bissectrice, alors que le schéma **EF/VF T** privilégie la direction $\theta = \frac{\Pi}{4}$. On remarque aussi que la présence de la matrice de masse a tendance à réduire considérablement les domaines de stabilité pour les trois schémas, ce que nous expliquerons plus loin. Sans matrice de masse, le schéma **EF R** paraît le plus avantageux; avec matrice de masse, c'est le schéma **EF/VF T** qui a le plus grand domaine de stabilité.

Nous donnons quelques détails sur ces domaines de stabilité dans l'Annexe A. Le schéma **VF R** est le seul pour lequel il est possible de démontrer un résultat précis (la démonstration est donnée en Annexe) :

Lemme II.4.1 *Le schéma* **VF R** *d'ordre un sans matrice de masse est stable si et seulement si:*

$$\left|\frac{c_1 \Delta t}{\Delta x}\right| + \left|\frac{c_2 \Delta t}{\Delta y}\right| \leq 1 \ .$$

Le schéma **VF R** *d'ordre un avec matrice de masse est stable si et seulement si:*

$$\left|\frac{c_1 \Delta t}{\Delta x}\right| + \left|\frac{c_2 \Delta t}{\Delta y}\right| \leq \frac{1}{9} \ . \bullet$$

II.4.2 Schémas d'ordre plus élevé

Intéressons nous maintenant à la stabilité de schémas d'ordre plus élevé. Nous avons déjà utilisé à propos des équations équivalentes le fait que, pour une intégration temporelle de type Runge-Kutta d'ordre N, le coefficient d'amplification G_N peut s'écrire:

$$G_N = 1 + \sum_{p=1}^{N} \frac{G_1^p}{p!} \ .$$

On représente sur les figures II.4, II.5, et II.6 les domaines de stabilité des β-schémas (avec encore α_1 en abscisse et α_2 en ordonnée), précis à l'ordre deux, trois ou quatre en temps, pour différentes valeurs de β. On remarque que les domaines de stabilité avec ou sans matrice de masse s'ordonnent toujours de la même manière en fonction de β pour les trois schémas. Sans matrice de masse, les plus grands domaines de stabilité sont obtenus avec $\beta = \frac{1}{2}$ en RK2, avec $\beta = 0$ en RK3 et $\beta = 0$ en RK4. Les plus petits domaines de stabilité sont obtenus pour $\beta = 1$, quelle que soit l'approximation en temps effectuée. Avec matrice de masse, la taille des domaines de stabilité augmente toujours quand β diminue. On voit aussi que le schéma **VF R** avec matrice de masse est à écarter car les limites de stabilité obtenues sont trop restrictives (de plus l'étude des équations équivalentes avait montré que ce schéma n'est précis qu'à l'ordre deux en espace quelle que soit la valeur de β).

A propos de ces domaines de stabilité, nous pouvons énoncer deux résultats intéressants (et de portée assez générale). Le premier résultat (démontré en Annexe B) montre

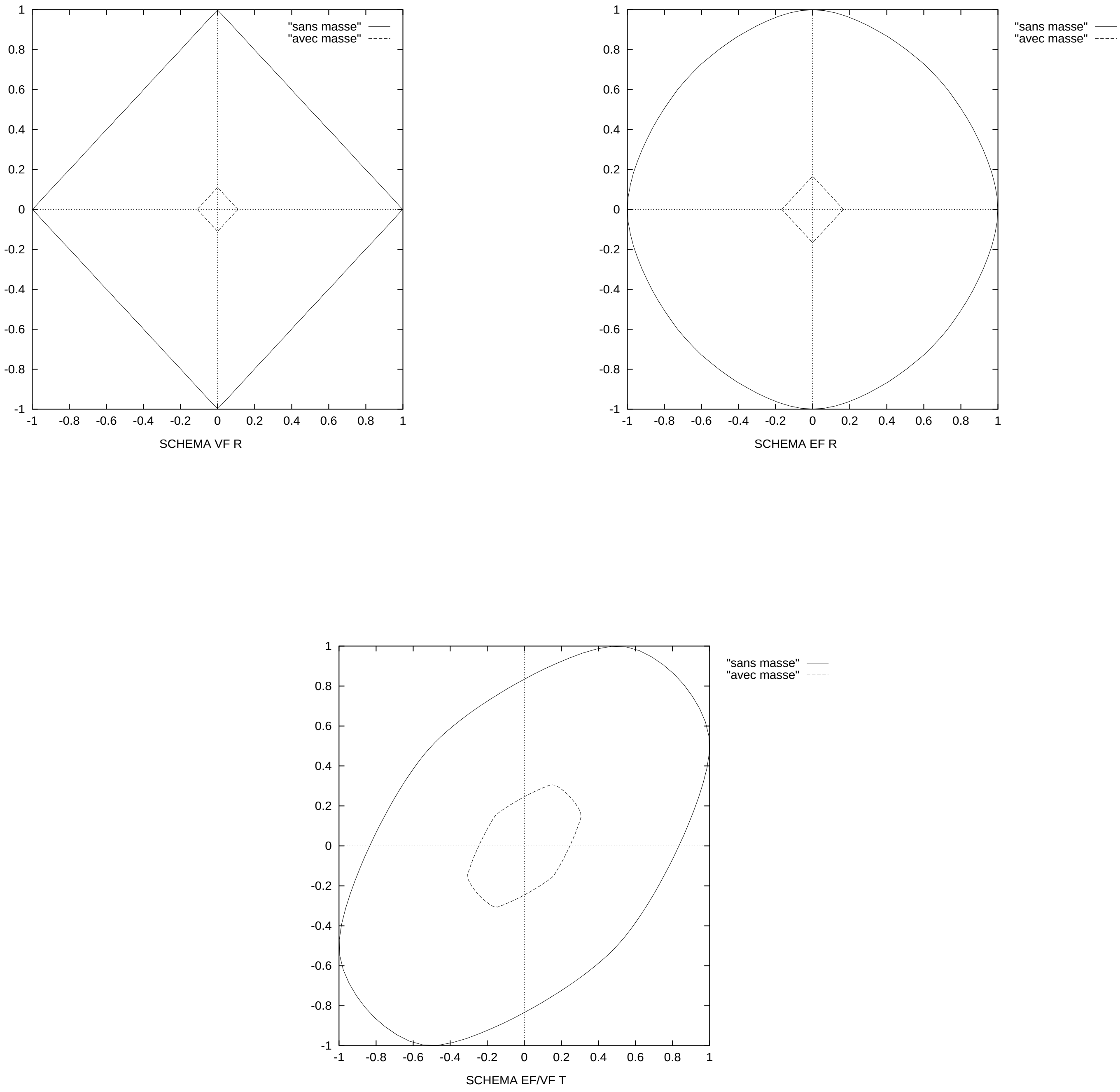


FIG. II.3 – *Domaines de stabilité des schémas avec et sans matrice de masse.*

que le domaine de stabilité du schéma **VF R** sans matrice de masse est toujours un carré, quelles que soient les valeurs des paramètres N et β :

Proposition II.4.1 *Soit $N \geq 2$ et $\beta \geq 0$ (si $N = 2$, on suppose que $\beta > 0$).*

Il existe une constante positive $\mathcal{L}_{\beta,N}$ telle que le β-schéma **VF R**, *avec intégration temporelle en Runge-Kutta d'ordre N et sans matrice de masse, est stable si et seulement si:*

$$\left|\frac{c_1\Delta t}{\Delta x}\right| + \left|\frac{c_2\Delta t}{\Delta y}\right| \leq \mathcal{L}_{\beta,N} \ . \bullet \tag{II.51}$$

En fait, la preuve de ce résultat montre que $\mathcal{L}_{\beta,N}$ est la limite de stabilité du schéma monodimensionnel correspondant (que l'on peut donc obtenir en faisant $c_2 = 0$). Cependant, obtenir explicitement la valeur de cette limite de stabilité $\mathcal{L}_{\beta,N}$ est difficile, sauf dans certains cas particuliers. Par exemple, pour $\beta = 0$, nous avons $a' = 0$ et $b' = -\alpha_1 \sin\theta_1$ (en prenant $\alpha_2 = 0$ d'après la remarque précédente). Il est aisé de voir que tout schéma antisymétrique (i.e. avec $a' = 0$) est instable en RK1 et en RK2, puisque d'une part $G_1 = 1 + ib'$ et $|G_1| > 1$ et que d'autre part $G_2 = 1 - \dfrac{b'^2}{2} + ib'$ et $|G_2| > 1$. De plus, des expressions $G_3 = G_2 + \dfrac{(ib')^3}{6}$ et $G_4 = G_3 + \dfrac{(ib')^4}{24}$, on tire aisément les valeurs des limites de stabilité lorsque $\beta = 0$ et $N = 3$ ou 4, données par (II.51) et:

$$\mathcal{L}_{0,3} = \sqrt{3}\ , \quad \mathcal{L}_{0,4} = 2\sqrt{2}\ .$$

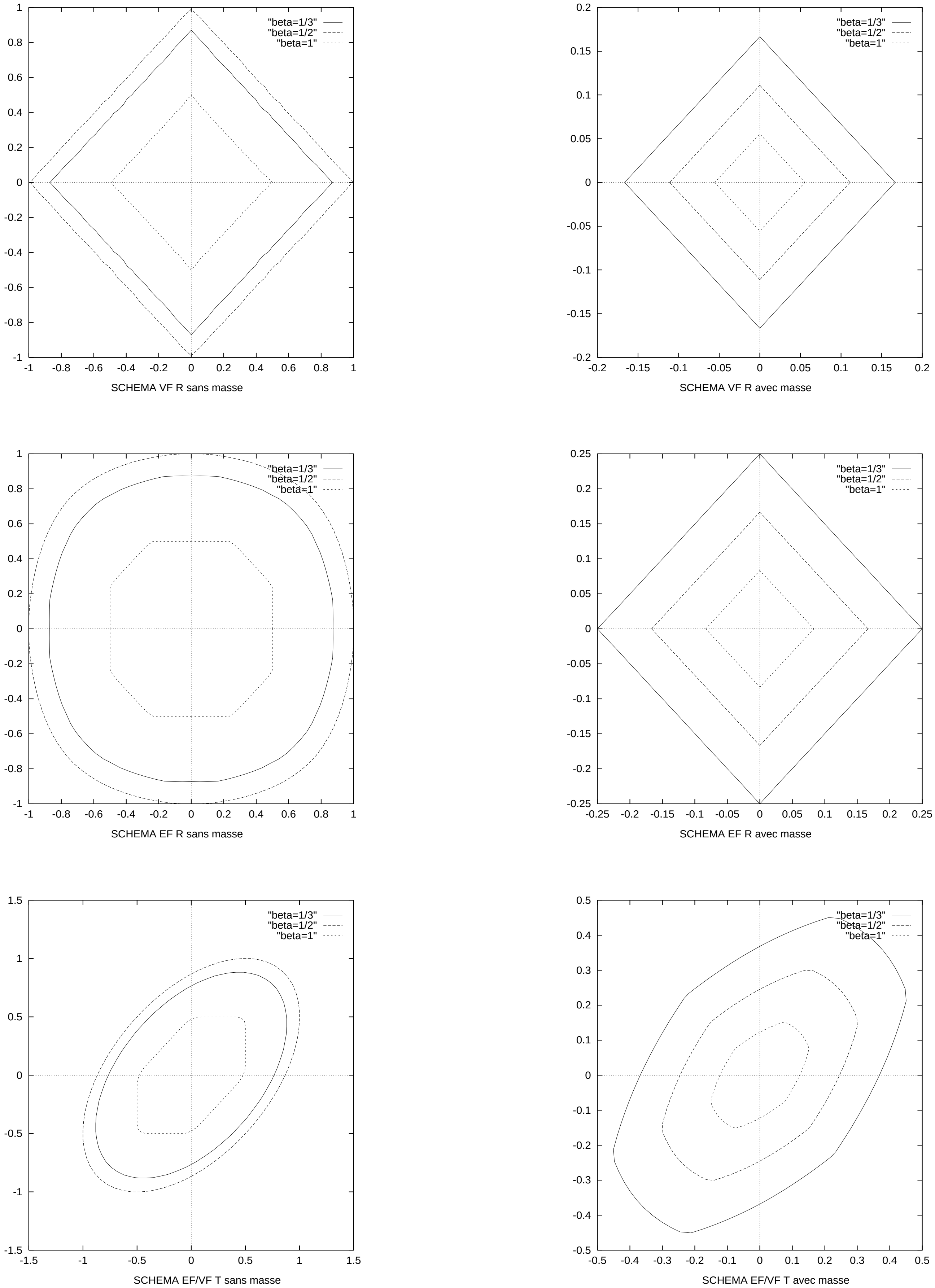


FIG. II.4 – *Domaines de stabilité des schémas précis à l'ordre deux en temps*

Le deuxième résultat montre que, sous des conditions très générales, l'utilisation de la matrice de masse réduit le domaine de stabilité :

Proposition II.4.2 *Si la matrice de masse est symétrique, inversible et a tous ses termes positifs, alors la condition sur le pas de temps Δt est plus restrictive pour un schéma avec matrice de masse que pour le schéma correspondant sans matrice de masse, quel que soit le schéma spatial et l'ordre en temps considérés.*

Preuve: Plaçons-nous pour simplifier dans le cas monodimensionnel. Nous voulons comparer le domaine de stabilité d'un schéma sans matrice de masse, que nous écrivons à l'ordre un en temps sous la forme:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = B_j^n \ ,$$

avec le domaine de stabilité du schéma correspondant avec matrice masse, qui s'écrit:

$$\sum_{k=-K}^{+K} a_k \frac{u_{j+k}^{n+1} - u_{j+k}^n}{\Delta t} = B_j^n \ ,$$

où les coefficients $a_k = \dfrac{1}{\Delta x} \displaystyle\int \Psi_j \Psi_{j+k}$ de la matrice de masse vérifient:

$$\sum_{k=-K}^{+K} a_k = 1 \ .$$

En développant u_j^n sous forme de séries de Fourier en ξ, on obtient le coefficient d'amplification, donné par:

$$\frac{G(\xi) - 1}{\Delta t} = F(\xi) \ \ \text{sans masse} \ , \quad S(\xi) \frac{G(\xi) - 1}{\Delta t} = F(\xi) \ \ \text{avec masse} \ ,$$

avec $S(\xi) = \displaystyle\sum_{k=-K}^{+K} a_k e^{ik\xi}$. Si la matrice de masse est symétrique et si tous ses coefficients sont positifs, on a alors:

$$S(\xi) = a_0 + \sum_{k>0} 2a_k \cos(k\xi) \leq a_0 + \sum_{k>0} 2a_k = 1 \ ;$$

si de plus la matrice de masse est inversible, on en déduit facilement que $0 < S(\xi) \leq 1$ pour tout ξ et la comparaison des expressions:

$$G(\xi)_{sans\ masse} = 1 + \Delta t F(\xi) \ , \quad G(\xi)_{avec\ masse} = 1 + \Delta t \frac{F(\xi)}{S(\xi)}$$

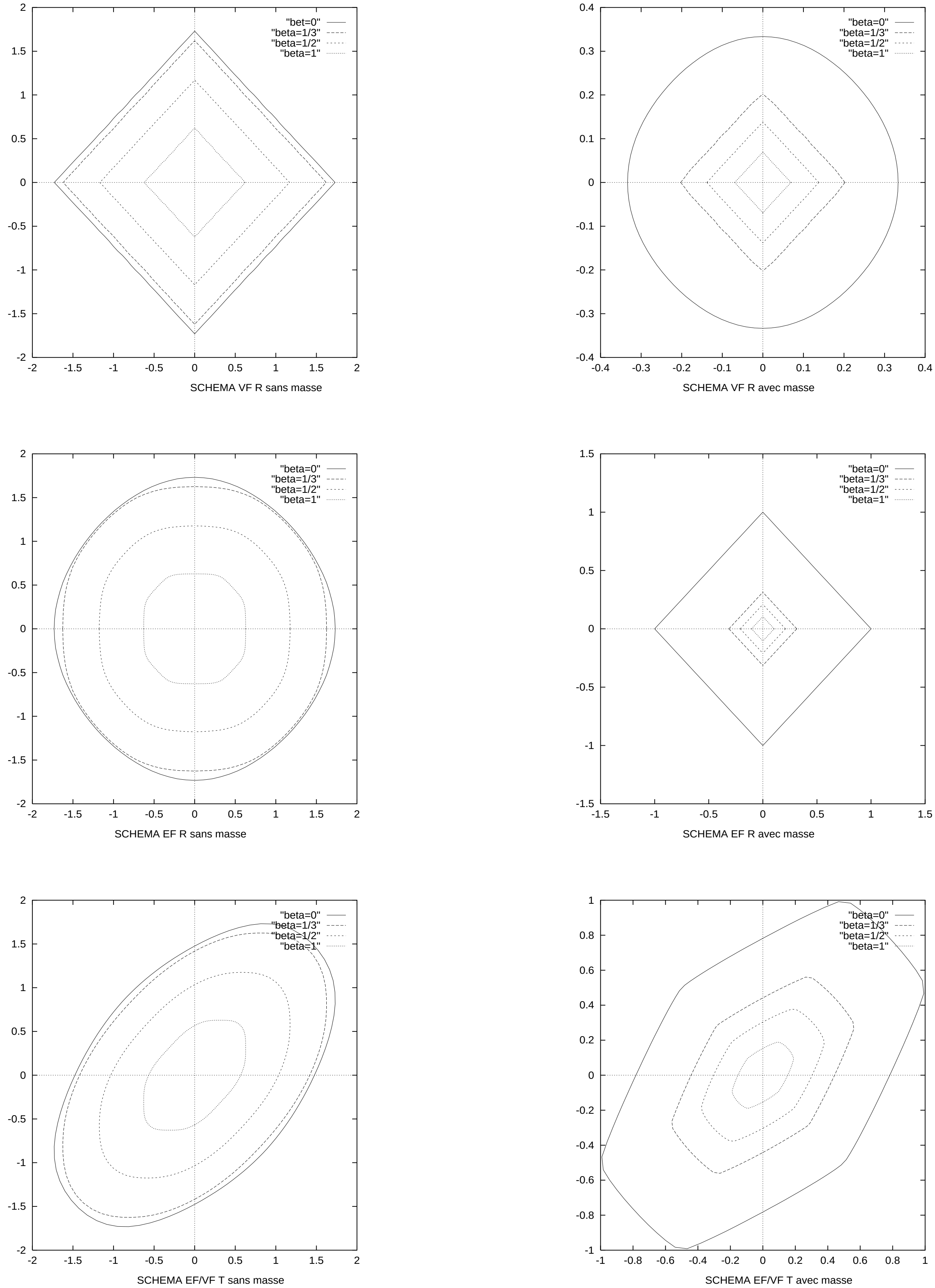


FIG. II.5 – *Domaines de stabilité des schémas précis à l'ordre trois en temps*

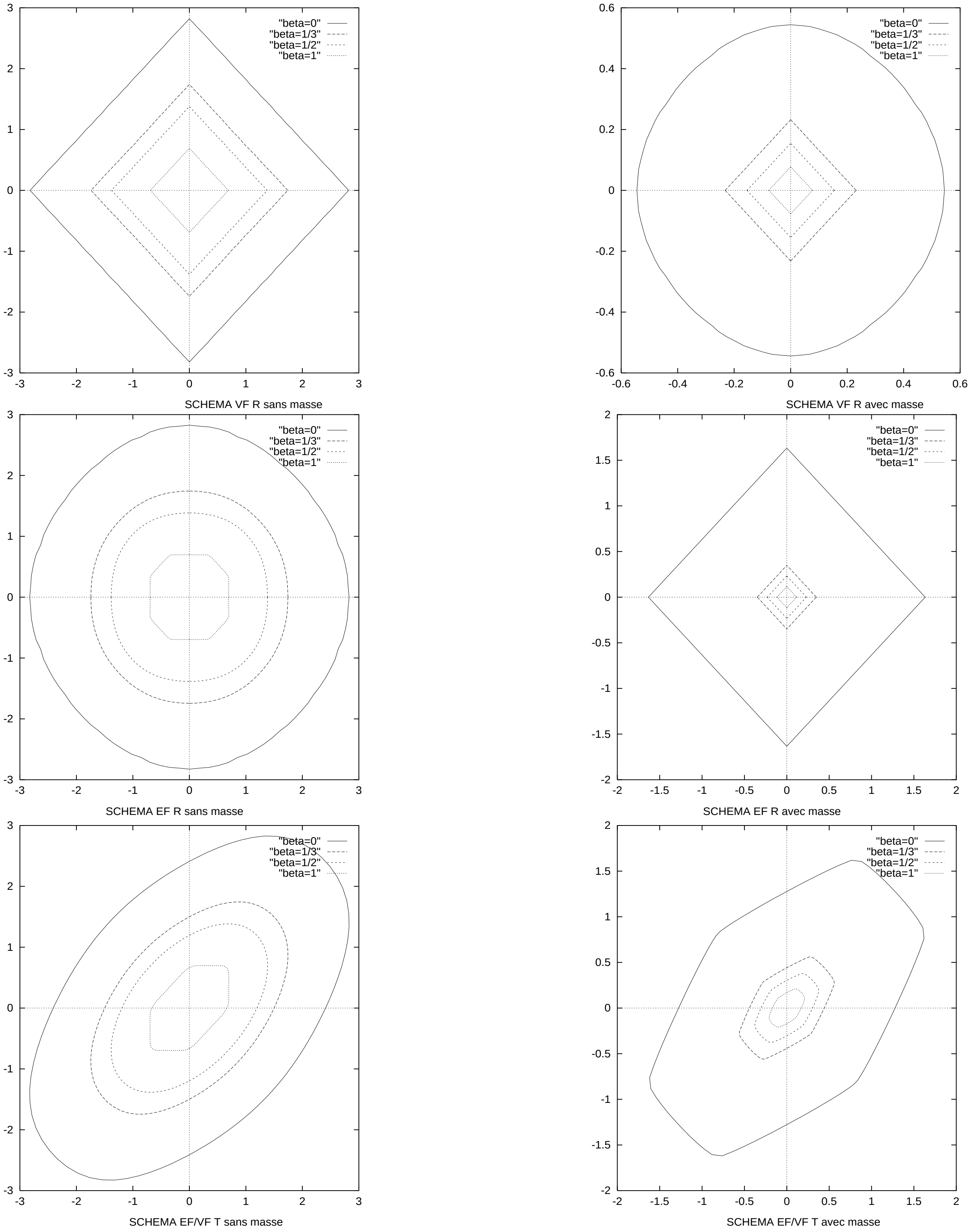


Fig. II.6 – *Domaines de stabilité des schémas précis à l'ordre quatre en temps* montre que l'utilisation de la matrice de masse conduit à une condition de stabilité plus sévère. Ceci reste vrai si l'on utilise le schéma de Runge-Kutta d'ordre N pour l'intégration en temps puisqu'alors:

$$G(\xi)_{sans\,masse} = 1 + \sum_{p=1}^{N} \Delta t^p F(\xi)^p \;, \quad G(\xi)_{avec\,masse} = 1 + \sum_{p=1}^{N} \Delta t^p \left(\frac{F(\xi)}{S(\xi)}\right)^p \;. \bullet$$

Nous complétons ces résultats généraux en donnant dans le tableau II.2 les limites de stabilité dans le cas où $\beta = 0$ et $\theta = 0$ (i.e. $c_2 = 0$), en RK3 et RK4 (lorsque $\beta = 0$, tous les schémas sont inconditionnellement instables en RK1 et RK2):

$\beta = 0$ $\theta = 0$	VF R sans masse	VF R avec masse	EF R sans masse	EF R avec masse	VF/EF T sans masse	VF/EF T avec masse
RK3	$\sqrt{3}$	$\frac{1}{3}$	$\sqrt{3}$	1	1.47	2.78
RK4	$2\sqrt{2}$	$\frac{2\sqrt{2}}{3\sqrt{3}}$	$2\sqrt{2}$	$\frac{2\sqrt{2}}{3}$	2.41	1.27

TAB. II.2 – *Limites de stabilité des schémas en RK3 et RK4 pour θ=0, β=0.*

On donne aussi dans le tableau II.3 les limites de stabilité obtenues numériquement en RK4 dans le cas monodimensionnel (θ=0) pour différentes valeurs de β.

RK4	VF R sans masse	VF R avec masse	EF R sans masse	EF R avec masse	VF/EF T sans masse	VF/EF T avec masse
$\beta = 0$	2.82	0.54	2.82	1.63	2.41	1.27
$\beta = \frac{1}{3}$	1.74	0.23	1.74	0.34	1.49	0.51
$\beta = \frac{1}{2}$	1.38	0.15	1.38	0.23	1.19	0.34
$\beta = 1$	0.69	0.07	0.69	0.11	0.66	0.17

TAB. II.3 – *Limites de stabilité des schémas pour $\theta = 0$ en RK4 pour différentes valeurs de β.*

II.4.3 Schémas d'ordre quatre

Intéressons nous de plus près aux schémas d'ordre quatre, que sont les schémas **EF R** et **EF/VF T** avec $\beta = 0$ et avec matrice de masse d'après l'étude des équations équivalentes. Ces schémas ont des limites de stabilité acceptables (avec $\theta = 0$, on obtient respectivement comme limite 1.63 et 1.27).

Les domaines de stabilité de ces deux schémas sont montrés ensemble sur la figure II.7. Pour $\theta = \frac{\Pi}{4}$, la limite de stabilité du schéma **EF/VF T** est nettement supérieure à celle obtenue pour le schéma **EF R**, alors que ces limites sont égales pour $\theta = \frac{3\Pi}{4}$; pour une vitesse d'advection parallèle aux axes, le schéma **EF R** permet d'utiliser un pas de temps plus grand.

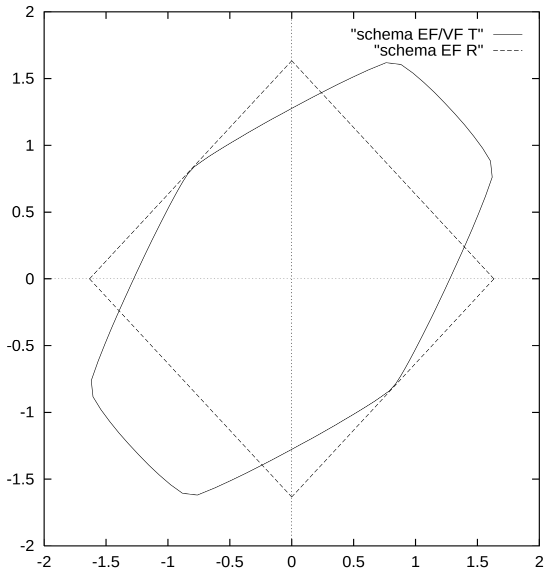


FIG. II.7 – *Schémas* **EF R** *et* **EF/VF T** *avec* $\beta = 0$ *et matrice de masse en RK4.*

D'autre part, nous disposons aussi d'autres schémas d'ordre quatre avec les schémas "sans diffusion numérique" de la section II.3.3 (avec RK4 et $\beta = \dfrac{1}{3}$, sans matrice de masse). Nous obtenons un résultat de stabilité grâce à la proposition II.4.1, que nous pouvons appliquer au schéma **VF R** :

Proposition II.4.3 *Soit* $\beta > 0$.

Il existe une constante positive $\mathcal{L}'_{\beta,4}$ *telle que le schéma* **VF R** *"sans diffusion numérique", avec intégration temporelle en Runge-Kutta d'ordre quatre et sans matrice de masse, est stable si et seulement si:*

$$\left|\frac{c_1 \Delta t}{\Delta x}\right| + \left|\frac{c_2 \Delta t}{\Delta y}\right| \leq \mathcal{L}'_{\beta,4} \ . \ \bullet \qquad \text{(II.52)}$$

Comme nous l'avons vu, le schéma "sans diffusion numérique" (II.50) est centré, il est donc antisymétrique et nous avons :

$$\begin{cases} x(\theta) = 0 \ , \\ y(\theta) = -(1+\beta) \ \sin\ \theta + \dfrac{\beta}{2} \sin\ 2\theta \ . \end{cases}$$

Par la suite, la démonstration est identique à celle effectuée pour un schéma décentré (voir Annexe B).

Nous représentons les domaines de stabilité des schémas **VF R**, **EF R** et **EF/VF T**, pour la valeur $\beta = \dfrac{1}{3}$. La figure II.8 montre que ces schémas ont des domaines de stabilité plus grands que ceux des schémas précédents; avec $\theta = 0$, la limite de stabilité est de 2.06 pour les schémas **VF R** et **EF R**, de 1.76 pour le schéma **EF/VF T**.

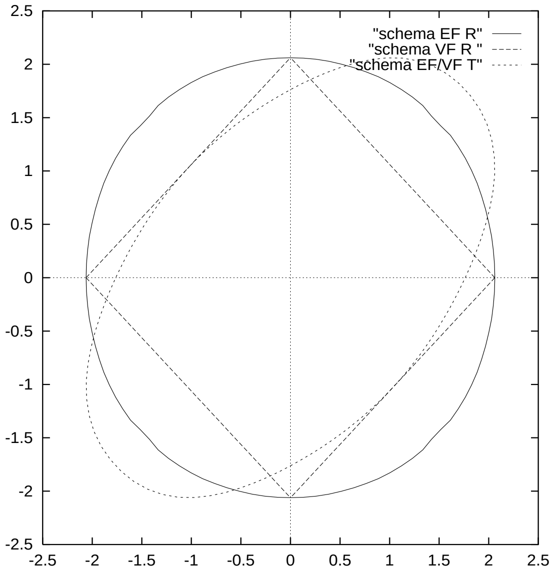


FIG. II.8 – *Domaines de stabilité des trois schémas "sans diffusion numérique" avec* $\beta = \dfrac{1}{3}$ *en RK4.*

II.5 Illustration numérique

Les études précédentes de stabilité et de précision nous permettent finalement de retenir cinq schémas précis à l'ordre quatre: d'une part, les schémas **VF R**, **EF R** et **EF/VF T** "sans diffusion numérique", pour $\beta = \dfrac{1}{3}$, avec RK4 et sans matrice de masse; d'autre part, les schémas **EF R**, **EF/VF T** avec matrice de masse, RK4 et $\beta = 0$. Dans ce paragraphe, nous effectuons quelques expériences numériques pour illustrer la précision de ces schémas.

Considérons le problème d'advection d'une "vague sinusoïdale", avec la donnée initiale $u(x, y, 0) = 2 + \sin(2\pi(x + y))$ pour $(x, y) \in [0, 1]^2$. La solution exacte à l'instant t s'écrit $u(x, y, t) = 2 + \sin(2\pi(x - c_1 t + y - c_2 t))$. On utilise des conditions aux limites périodiques et le domaine de calcul $[0, 1]^2$ est discrétisé en $N = 50$ mailles dans chaque direction:

$\Delta x = \Delta y = h = 2\ 10^{-2}$. Le pas de temps choisi vérifie la relation suivante :

$$\Delta t = CFL\,\frac{h}{c}$$

où CFL désigne le nombre de Courant, dont la valeur est déterminée par l'étude de stabilité.

Les figures II.9, II.10, et II.11 permettent de comparer les solutions obtenues au temps $T = 8$ avec $CFL = 1$, pour $\theta = \frac{\pi}{4}$ et pour différentes valeurs de β, dans le cas de schémas avec ou sans matrice de masse. Le schéma **VF R** sans matrice de masse est d'ordre deux en espace : pour la valeur $\beta = 0$ la solution admet un retard de phase important (figure II.9). Celui-ci disparaît lorsqu'on utilise le schéma **EF R** centré ($\beta = 0$) avec matrice de masse (figure II.10). De même, pour $\beta = \frac{1}{3}$, le schéma volumes finis classique est dissipatif au troisième ordre, ce qui n'est pas le cas du schéma **EF R** centré avec matrice de masse, globalement d'ordre quatre (figure II.11).

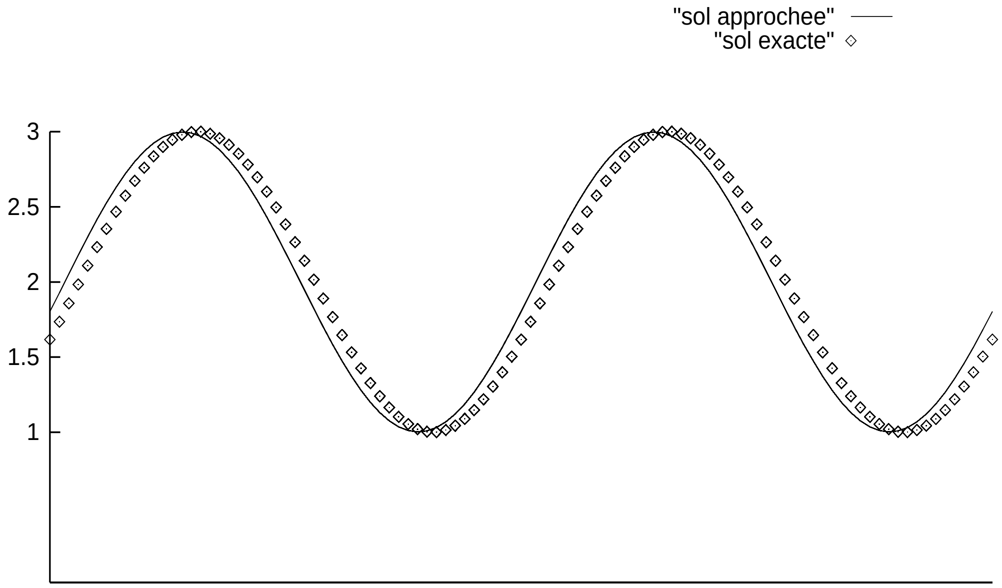


FIG. II.9 – *Solution obtenue avec le schéma VF R, sans masse, $\beta = 0$.*

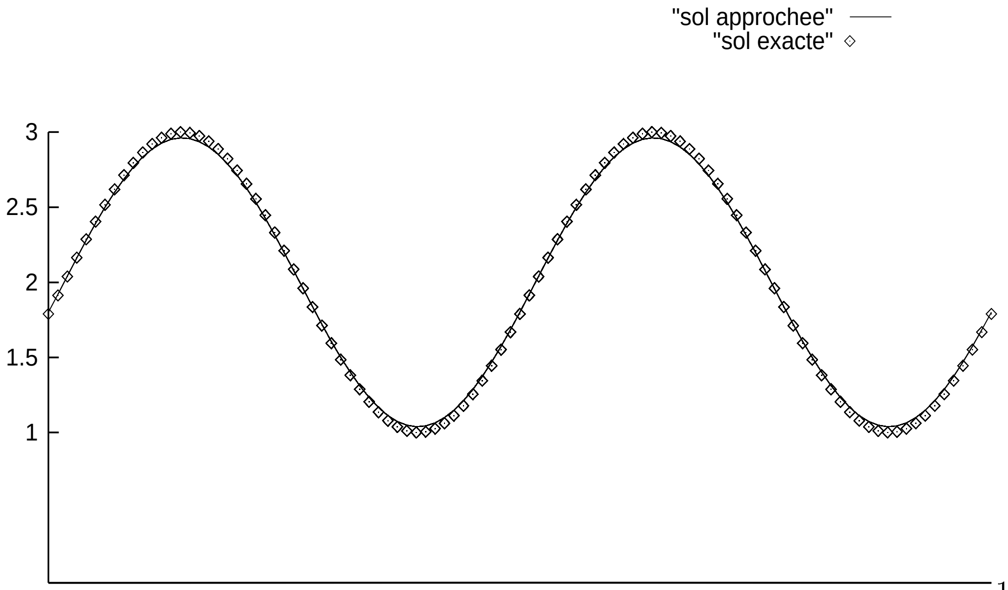


FIG. II.10 – *Solution obtenue avec le schéma VF R, sans masse,* $\beta = \frac{1}{3}$.

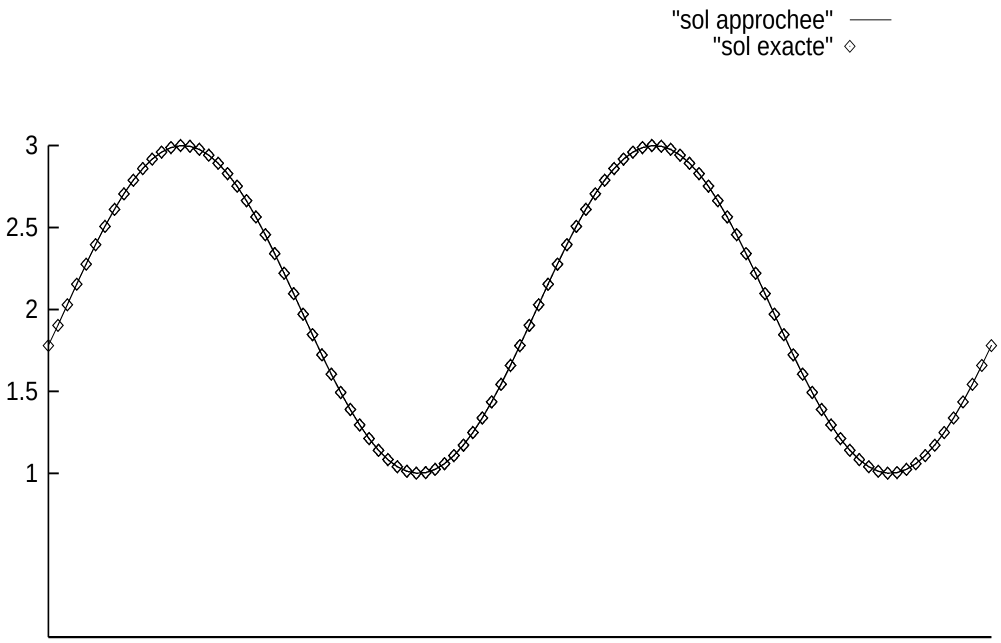


FIG. II.11 – *Solution obtenue avec le schéma EF R, avec masse,* $\beta = 0$.

Pour mieux évaluer la précision globale de chaque schéma, nous avons calculé l'erreur en norme l^2 (à un instant $T = n\Delta t$ fixé) :

$$E_2 = (\frac{1}{N}\left(\sum_{j=1}^{N}\sum_{k=1}^{N}(u_{jk}^n - u(j\Delta x, k\Delta y, n\Delta t))^2\right)^{\frac{1}{2}} \quad .$$

Les Tables IV,V,VI montrent les erreurs obtenues au temps $T = 1$. Pour chaque valeur

	VF R	EF R	VF/EF T
$\beta = 0$	$1.66\ 10^{-2}$	$3.30\ 10^{-2}$	$4.92\ 10^{-2}$
$\beta = \frac{1}{3}$	$1.05\ 10^{-3}$	$3.45\ 10^{-3}$	$5.39\ 10^{-3}$

TAB. II.4 – *Erreur l^2 pour les schémas sans matrice de masse.*

	EF R	VF/EF T
$\theta = 0$	$1.53\ 10^{-5}$	$4.62\ 10^{-5}$
$\theta = \frac{\Pi}{4}$	$6.08\ 10^{-5}$	$1.04\ 10^{-4}$
$\theta = \frac{3\Pi}{4}$	$5.29\ 10^{-14}$	$3.87\ 10^{-14}$

TAB. II.5 – *Erreur l^2 pour les schémas avec matrice de masse.*

	VF R	EF R	VF/EF T
$\theta = 0$	$5.50\ 10^{-5}$	$2.05\ 10^{-4}$	$3.55\ 10^{-4}$
$\theta = \frac{\Pi}{4}$	$1.57\ 10^{-4}$	$3.70\ 10^{-4}$	$5.81\ 10^{-4}$
$\theta = \frac{3\Pi}{4}$	$4.78\ 10^{-14}$	$3.44\ 10^{-14}$	$1.81\ 10^{-14}$

TAB. II.6 – *Erreur l^2 pour les schémas "sans diffusion numérique".*

de β, les calculs sont effectués en utilisant le même pas de temps pour tous les schémas (volumes ou éléments finis, rectangles ou triangles).

Dans le tableau III.4, on s'intéresse aux schémas décentrés **VF R**, **EF R**, **EF/VF T** d'ordre deux, sans matrice de masse, pour $\beta = 0$ et $\beta = \frac{1}{3}$. On choisit $\theta = \frac{\pi}{4}$ et $CFL = 1.5$ dans le cas $\beta = 0$, $CFL = 1$ pour $\beta = \frac{1}{3}$.

Dans le tableau II.5, on représente l'erreur E_2 des schémas **EF R** et **EF/VF T** avec matrice de masse, pour la valeur $\beta = 0$, en faisant varier l'angle de direction θ. Pour ces schémas on choisit $CFL = 1$.

On donne dans le tableau II.6 les erreurs obtenues pour les schémas **VF R**, **EF R**, **EF/VF T** "sans diffusion numérique", sans matrice de masse, pour la valeur $\beta = \frac{1}{3}$, en faisant varier l'angle de vitesse θ. On choisit un pas de temps commun pour les trois schémas en fixant le CFL à 1.2.

Les résultats des tableaux II.4, II.5 et II.6 illustrent bien l'étude des équations équivalentes : les schémas sans matrice de masse du tableau II.4 sont d'ordre deux pour $\beta = 0$ et d'ordre trois avec $\beta = \frac{1}{3}$. La prise en compte de la matrice de masse permet d'obtenir, pour la valeur $\beta = 0$, une précision d'ordre quatre pour les schémas **EF R** et **VF/EF T** (tableau II.5). Notons également la très bonne précision des trois schémas "sans diffusion numérique" (tableau II.6). En règle générale les schémas en rectangles s'avèrent plus précis que les schémas en triangles.

Les mesures d'erreur ont aussi été effectuées en choisissant un pas de temps propre à chaque schéma, proche de la limite de stabilité, et les remarques précédentes s'appliquent de la même manière.

Signalons aussi que la très bonne approximation de la solution par tous les schémas, pour la direction $\theta = \frac{3\pi}{4}$ ne doit pas surprendre : il est facile de voir en effet que, dans ce cas, la solution exacte du problème considéré est stationnaire et que tous les schémas sont alors exacts.

Pour conclure ce paragraphe, il nous a semblé intéressant de représenter l'évolution de l'erreur E_2 au cours du temps pour les schémas les plus performants ; sur la figure II.12, cette erreur est tracée en fonction du nombre d'itérations jusqu'à $T = 8$ (pour $\theta = \frac{\pi}{4}$).

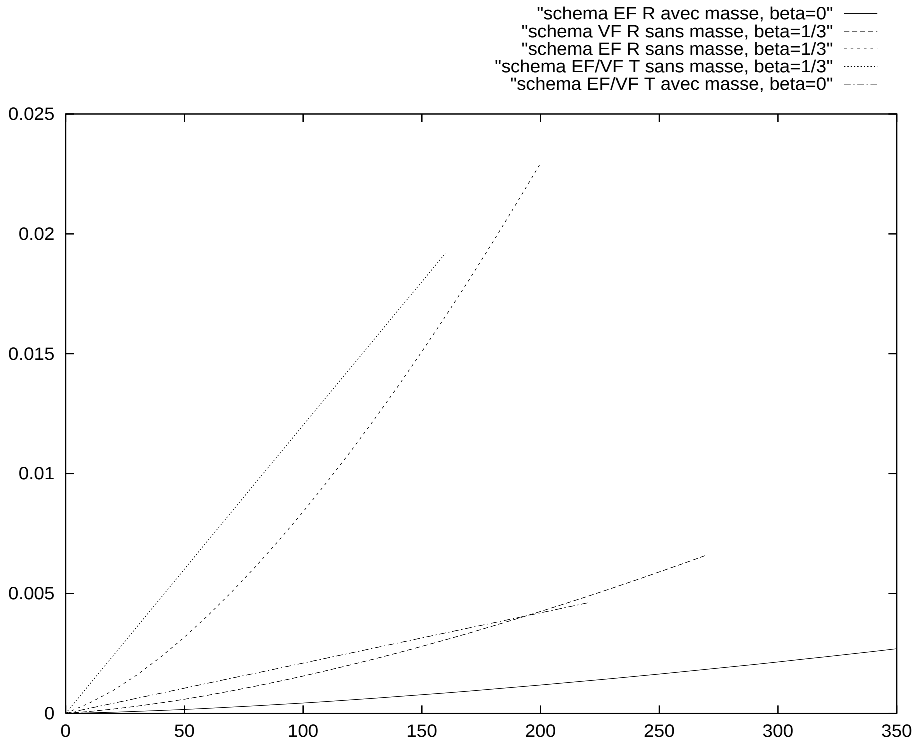


FIG. II.12 – *Evolution de l'erreur E_{l2} en fonction du nombre d'itérations pour les schémas d'ordre quatre.*

On observe ici que les schémas centrés ($\beta = 0$) avec matrice de masse sont les plus précis aussi bien en maillage rectangulaire qu'en maillage triangulaire : la prise en compte de la matrice de masse augmente donc bien la précision des schémas. Parmi les schémas "sans diffusion numérique" et sans matrice de masse, c'est le schéma **VF R** qui se comporte le mieux du point de vue de la précision .

II.6 Conclusion.

Nous avons présenté différents schémas numériques basés soit sur des formulations éléments finis et volumes finis classiques, soit sur des méthodes nouvelles mêlant les deux approches, appliqués à des maillages triangulaires et rectangulaires. Le but de notre analyse était d'obtenir des schémas d'ordre élevé : pour cela nous avons choisi une intégration temporelle d'ordre quatre et l'approximation spatiale utilise un flux centré ou décentré, caractérisé par un paramètre β. L'étude des équations équivalentes et de la stabilité nous a permis de comparer ces schémas et de retenir les plus précis.

Nous nous sommes aussi intéressés au rôle et aux propriétés de la matrice de masse dans l'intégration en temps: en effet dans les schémas explicites usuels la matrice de masse est généralement diagonalisée. Pour chacun des schémas étudiés nous avons mis en valeur les avantages et les inconvénients de l'utilisation de la matrice de masse en comparant les termes d'erreur des schémas avec et sans cette matrice. Celle-ci agit essentiellement sur les termes de dispersion : à l'ordre un, elle annule ces termes dans le cas des schémas **EF R** et **EF/VF T** ; à un ordre plus élevé, la valeur $\beta = 0$ annule à la fois les termes de dispersion et de dissipation, ce qui rend les schémas précédents d'ordre quatre en espace. Seul le schéma **VF R** avec matrice de masse reste globalement d'ordre deux, pour toute valeur de β. Pour ce qui concerne la stabilité, nous avons mis en évidence une propriété importante : l'utilisation de la matrice de masse impose *toujours* un choix de pas de temps plus restrictif que pour le schéma correspondant sans matrice de masse.

Dans le cas du schéma **VF R**, cette restriction apportée par la matrice de masse est très forte : pour une approximation spatiale de type volumes finis, la technique du "mass-lumping" sur les termes temporels s'avère donc tout à fait justifiée.

Finalement, notre étude a mis en valeur quatre schémas d'ordre quatre en temps et en espace : les schémas **EF R** et **EF/VF T** avec matrice de masse, pour la valeur $\beta = 0$, et les schémas **VF R**, **EF/VF T** sans "diffusion numérique", sans matrice de masse, avec $\beta = \dfrac{1}{3}$. La première catégorie de schémas offre la plus grande précision, la seconde a l'avantage d'être la moins coûteuse du point de vue temps de calcul et mise en oeuvre informatique.

Bien sûr, notre étude s'est limitée à la résolution de l'équation d'advection. Le but de cette étude était de déterminer les schémas les plus performants en vue de les appliquer à des problèmes non linéaires, comme les équations d'Euler, ou à des problèmes hyperboliques linéaires, par exemple les équations de Maxwell. Dans cette perspective, l'utilisation des schémas d'ordre quatre avec matrice de masse nous semble trop coûteuse par-rapport au gain de précision apporté. Ces méthodes nécessitent l'inversion de la matrice de masse et l'utilisation d'un pas de temps plus petit que pour les schémas "sans diffusion numérique". De plus, dans le cas des équations d'Euler ou de Maxwell, il est souvent impossible d'utiliser des schémas totalements centrés. On introduit alors un petit taux de diffusion ϵ pour assurer la stabilité. Dans le cas des schémas d'ordre quatre sans matrice de masse avec $\beta = \dfrac{1}{3}$, les limites de stabilité ne sont pas modifiées pour $\epsilon = 0.1$, par contre, pour les schémas avec matrice de masse et $\beta = 0$, avec $\epsilon = 0.1$, les limites de stabilité ne valent plus que 1.14 pour le schéma **EF R** et 1.13 pour le schéma **EF/VF T** (au lieu de 1.63 et 1.27 avec $\epsilon = 0$). Ces constatations nous incitent à conserver les schémas explicites d'ordre quatre "sans diffusion numérique", et nous nous proposons d'appliquer ces schémas aux équations de Maxwell, car il est important dans de nombreux cas de disposer de méthodes numériques de haute précision mais d'un coût raisonnable.

II.7 Annexe A.

Nous présentons dans cette partie une étude des limites de stabilité concernant les schémas d'ordre un. Pour chacun des schémas, nous donnons les parties réelle et imaginaire du coefficient d'amplication.
Posons donc $G_1 = G_{\theta_1,\theta_2} = a + ib$ avec a et b réels, et notons $\alpha_1 = c_1 \dfrac{\Delta t}{\Delta x} = \alpha \ \cos\ \theta$, $\alpha_2 = c_2 \dfrac{\Delta t}{\Delta y} = \alpha \ \sin\ \theta$.

II.7.1 Schémas sans matrice de masse

– Schéma **VF R**

On obtient:

$$\begin{cases} a = 1 - |\alpha_1|(1-\cos\theta_1) - |\alpha_2|(1-\cos\theta_2) \ , \\ b = -\alpha_1 \ \sin\ \theta_1 - \alpha_2 \ \sin\ \theta_2 \ . \end{cases}$$

Lemme II.7.1 *Le schéma* **VF R** *d'ordre un sans matrice de masse est stable si et seulement si:*

$$|\alpha_1| + |\alpha_2| \leq 1 \ . \bullet$$

Preuve: Dans cette démonstration ainsi que dans celles qui suivent, on limite l'étude au cas $0 \leq \alpha_1 \leq 1$, $0 \leq \alpha_2 \leq 1$, les autres cas s'obtenant par symétrie.

Notons d'abord que le schéma est instable si $\alpha_1 + \alpha_2 > 1$. En effet, en prenant $\theta_1 = \theta_2 = \Pi$, on obtient alors:

$$G_1^2 = 1 + 4(\alpha_1+\alpha_2)^2 - 4(\alpha_1+\alpha_2) > 1 \ .$$

Réciproquement, supposons que $\alpha_1 + \alpha_2 \leq 1$ et considérons les vecteurs $\vec{e_1} = \begin{pmatrix} \cos\ \theta_1 \\ -\sin\ \theta_1 \end{pmatrix}$ et $\vec{e_2} = \begin{pmatrix} \cos\ \theta_2 \\ -\sin\ \theta_2 \end{pmatrix}$. On peut alors écrire:

$$\vec{x} = \begin{pmatrix} a \\ b \end{pmatrix} = \alpha_1 \vec{e_1} + \alpha_2 \vec{e_2} + \begin{pmatrix} 1 - \alpha_1 - \alpha_2 \\ 0 \end{pmatrix} \ ,$$

si bien que:

$$|G_1| = ||\vec{x}|| \leq \alpha_1 + \alpha_2 + |1 - \alpha_1 - \alpha_2| = 1 \ . \bullet$$

– Schéma **EF R**:

On a ici:

$$\left\{\begin{array}{ll} a = & 1-\frac{2}{3}|\alpha_1|(1-\cos\theta_1)-\frac{2}{3}|\alpha_2|(1-\cos\theta_2) \\ & -\frac{1}{6}|\alpha_1+\alpha_2|(1-\cos(\theta_1+\theta_2))-\frac{1}{6}|\alpha_1-\alpha_2|(1-\cos(\theta_1-\theta_2)) \ , \\ b = & -\frac{2}{3}\alpha_1 \ \sin\ \theta_1-\frac{2}{3}\alpha_2\ \sin\ \theta_2-\frac{1}{3}\ \alpha_1 \sin\ \theta_1\ \cos\ \theta_2-\frac{1}{3}\ \alpha_2 \sin\ \theta_2\ \cos\ \theta_1 \ . \end{array}\right.$$

Il n'est plus possible d'obtenir une expression analytique exacte de la limite de stabilité. On peut cependant prouver l'encadrement suivant:

Lemme II.7.2 *Le schéma* **EF R** *d'ordre un sans matrice de masse est stable si* $|\alpha_1|+|\alpha_2|\leq 1$ *et instable si* $|\alpha_1|+|\alpha_2|>\frac{18}{13}$. •

Preuve: Si $\alpha_1+\alpha_2>\frac{18}{13}$ (avec $\alpha_1,\alpha_2\geq 0$), on voit que le schéma est instable en prenant $\theta_1=\theta_2=\frac{\Pi}{2}$ puisqu'alors:

$$G_1^2 = 1+\frac{13}{9}(\alpha_1+\alpha_2)^2-2(\alpha_1+\alpha_2)>1 \ .$$

Par ailleurs, supposons que $\alpha_1+\alpha_2\leq 1$ et considérons le cas $\alpha_1\geq\alpha_2$. Posons:

$$\vec{e_1}=\begin{pmatrix}\cos\ \theta_1\\ -\sin\ \theta_1\end{pmatrix},\ \vec{e_2}=\begin{pmatrix}\cos\ \theta_2\\ -\sin\ \theta_2\end{pmatrix},\ \vec{e_3}=\begin{pmatrix}\cos\ (\theta_1-\theta_2)\\ -\sin\ (\theta_1-\theta_2)\end{pmatrix},\ \vec{e_4}=\begin{pmatrix}\cos\ (\theta_1+\theta_2)\\ -\sin\ (\theta_1+\theta_2)\end{pmatrix},$$

nous obtenons:

$$\vec{x}=\begin{pmatrix}a\\ b\end{pmatrix}=\frac{2}{3}\alpha_1\vec{e_1}+\frac{2}{3}\alpha_2\vec{e_2}+\frac{1}{6}(\alpha_1-\alpha_2)\vec{e_3}+\frac{1}{6}(\alpha_1+\alpha_2)\vec{e_4}+\begin{pmatrix}1-\frac{2}{3}\alpha_2-\alpha_1\\ 0\end{pmatrix},$$

et:

$$|G_1|=||\vec{x}||\leq\frac{5}{6}(\alpha_1+\alpha_2)+\frac{1}{6}(\alpha_1-\alpha_2)+|1-\alpha_1-\frac{2}{3}\alpha_2|=1 \ ,$$

car $\alpha_1+\frac{2}{3}\alpha_2\leq 1$. On procède de la même manière dans le cas $\alpha_1\leq\alpha_2$. •

– Schéma **VF/EF T:**

On donne simplement les expressions des parties réelle et imaginaire du coefficient d'amplification G_1:

$$\left\{\begin{array}{l} a = 1-\frac{1}{3}|-2\alpha_1+\alpha_2|(1-\cos\ \theta_1)-\frac{1}{3}|-\alpha_1+2\alpha_2|(1-\cos\ \theta_2) \\ \quad -\frac{1}{3}|\alpha_1+\alpha_2|(1-\cos\ (\theta_1+\ \theta_2)) \ , \\ b = -\frac{1}{3}(-2\alpha_1+\alpha_2)\ \sin\ \theta_1-\frac{1}{3}(-\alpha_1+2\alpha_2)\ \sin\ \theta_2-\frac{1}{3}\ (\alpha_1+\alpha_2)\sin\ (\theta_1+\theta_2) \ . \end{array}\right.$$

II.7.2 Schémas avec matrice de masse

Avec matrice de masse, le coefficient d'amplification G_1 s'écrit:

$$G_1 = a' + ib' = 1 + \frac{1}{S(\theta_1, \theta_2)} \left[(a-1) + ib \right] ,$$

où $S(\theta_1, \theta_2)$ est le terme correspondant à la matrice de masse, après analyse de Fourier.

– <u>Schéma **VF R**</u>

Dans le cas d'un maillage rectangulaire, on a:

$$S(\theta_1, \theta_2) = \frac{4}{9} + \frac{2}{9} \cos\theta_1 + \frac{2}{9} \cos\theta_2 + \frac{1}{9} \cos\theta_1 \cos\theta_2 \ .$$

On observe notamment que:

$$\frac{1}{9} \leq S(\theta_1, \theta_2) \leq 1 \quad \forall (\theta_1, \theta_2) \in [0, 2\Pi]^2 \ . \tag{II.53}$$

Nous avons de nouveau un résultat précis pour le schéma **VF R**:

Lemme II.7.3 *Le schéma* **VF R** *d'ordre un avec matrice de masse est stable si et seulement si:*

$$|\alpha_1| + |\alpha_2| \leq \frac{1}{9} \ . \ \bullet$$

Preuve: Si $\alpha_1 + \alpha_2 > \frac{1}{9}$, on prend encore $\theta_1 = \theta_2 = \Pi$ et on voit que le schéma est instable car $S(\Pi, \Pi) = \frac{1}{9}$ et:

$$G_1^2 = 1 + \frac{4}{S(\Pi,\Pi)^2} (\alpha_1 + \alpha_2)^2 - \frac{4}{S(\Pi,\Pi)} (\alpha_1 + \alpha_2)^2 > 1 \ .$$

Réciproquement, si $\alpha_1 + \alpha_2 \leq \frac{1}{9}$, on a, avec les notations précédentes:

$$\vec{x} = \begin{pmatrix} a' \\ b' \end{pmatrix} = \frac{1}{S(\theta_1, \theta_2)} \left[\alpha_1 \vec{e_1} + \alpha_2 \vec{e_2} + \begin{pmatrix} S(\theta_1, \theta_2) - \alpha_1 - \alpha_2 \\ 0 \end{pmatrix} \right] ,$$

d'où:

$$|G_1| = ||\vec{x}|| \leq \frac{1}{S(\theta_1, \theta_2)} \left(\alpha_1 + \alpha_2 + |S(\theta_1, \theta_2) - \alpha_1 - \alpha_2| \right) \leq 1 ,$$

car $\alpha_1 + \alpha_2 \leq S(\theta_1, \theta_2)$ d'après (II.53). $\bullet$

– <u>Schéma **EF R**</u>

Nous avons ici l'encadrement suivant:

Lemme II.7.4 *Le schéma* **EF R** *d'ordre un avec matrice de masse est stable si* $|\alpha_1| + |\alpha_2| \leq \dfrac{1}{9}$ *et instable si* $|\alpha_1| + |\alpha_2| > \dfrac{1}{6}$. •

Preuve: En prenant $\theta_1 = \theta_2 = \Pi$, on voit facilement que le schéma est instable si $\alpha_1 + \alpha_2 > \dfrac{1}{6}$.

Si $\alpha_1 + \alpha_2 \leq \dfrac{1}{9}$ et $\alpha_1 \geq \alpha_2$, on procède comme pour le schéma sans matrice de masse ; il vient ici:

$$|G_1| \leq \frac{1}{S(\theta_1,\theta_2)}\left[\frac{5}{6}(\alpha_1+\alpha_2) + \frac{1}{6}(\alpha_1-\alpha_2) + |S(\theta_1,\theta_2) - \alpha_1 - \frac{2}{3}\alpha_2|\right] \leq 1 \ . \bullet$$

– Schéma **VF/EF T**

A partir de la matrice de masse P1 obtenue sur un maillage triangulaire, on déduit l'expression suivante de S pour ce schéma:

$$S(\theta_1,\theta_2) = \frac{1}{2} + \frac{1}{6}\cos\theta_1 + \frac{1}{6}\cos\theta_2 + \frac{1}{6}\cos(\theta_1+\theta_2) \ .$$

On a ici:

$$\frac{1}{4} \leq S(\theta_1,\theta_2) \leq 1 \ \ \forall(\theta_1,\theta_2) \in [0, 2\Pi]^2 \ .$$

II.8 Annexe B.

Nous nous proposons ici de démontrer la Proposition II.4.1 que nous rappelons :

Proposition II.4.1 : *Soit* $N \geq 2$ *et* $\beta \geq 0$ *(si* $N = 2$*, on suppose que* $\beta > 0$*).*

Il existe une constante positive $\mathcal{L}_{\beta,N}$ *telle que le* β*-schéma* **VF R***, avec intégration temporelle en Runge-Kutta d'ordre* N *et sans matrice de masse, est stable si et seulement si:*

$$\left|\frac{c_1\Delta t}{\Delta x}\right| + \left|\frac{c_2\Delta t}{\Delta y}\right| \leq \mathcal{L}_{\beta,N} \ . \bullet \tag{II.54}$$

Preuve: Posons de nouveau $\alpha_1 = \dfrac{c_1\Delta t}{\Delta x}$ et $\alpha_2 = \dfrac{c_2\Delta t}{\Delta y}$; nous limitons l'étude au cas $0 \leq \alpha_1$, $0 \leq \alpha_2$, les autres cas s'obtenant par symétrie.

Supposons $\beta \geq 0$ et $N \geq 2$ fixés (avec $\beta > 0$ si $N = 2$). Dans le cas du schéma **VF R** sans matrice de masse avec une intégration temporelle en Runge-Kutta d'ordre N, le coefficient d'amplification s'écrit :

$$G_N = \sum_{p=0}^{N} \frac{G_1^p}{p!} \ ,$$

où G_1 représente le coefficient d'amplication pour le schéma à l'ordre un et s'écrit :

$$G_1 = G_1(\theta_1, \theta_2) = \alpha_1 f(\theta_1) + \alpha_2 f(\theta_2) \ ,$$

θ_1 et θ_2 désignant les angles de Fourier et :

$$\left\{ \begin{array}{l} f(\theta) = x(\theta) + iy(\theta) \ , \\ x(\theta) = 2\beta \ \cos\ \theta - \dfrac{\beta}{2} \cos\ 2\theta - \dfrac{3\beta}{2} \ , \\ y(\theta) = -(1+\beta)\ \sin\ \theta + \dfrac{\beta}{2} \sin\ 2\theta \ . \end{array} \right.$$

Définissons le domaine de stabilité D_N de la méthode de Runge-Kutta d'ordre N comme le lieu dans le plan complexe des nombres z tels que $|\mathcal{P}_N(z)| \leq 1$, où $\mathcal{P}_N(z) = \sum_{p=0}^{N} \dfrac{z^p}{p!}$:

$$D_N = \{ z \in \mathbb{C} \ , \ \ |\mathcal{P}_N(z)| \leq 1 \} \ .$$

Ainsi, le schéma considéré est stable à l'ordre N si et seulement si, pour tout couple (θ_1, θ_2), $G_1(\theta_1, \theta_2) \in D_N$.

Soit alors Γ la courbe du plan complexe $\mathbb{C}$ définie par $\Gamma = \{ f(\theta) \ , \ \ -\pi \leq \theta \leq \pi \}$. Cette courbe délimite un domaine ouvert borné Ω (voir Figure II.13), à propos duquel nous démontrerons plus loin le lemme suivant :

Lemme II.8.1 *Le domaine Ω est un ouvert convexe appartenant à l'ensemble $\mathbb{C}^-$ des complexes à partie réelle négative et $0 \in \bar{\Omega}$.* •

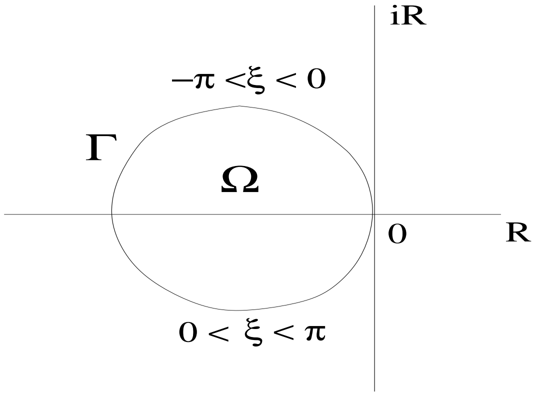


FIG. II.13 –

Pour $\lambda > 0$, notons alors $\lambda\Omega = \{ \lambda z \ , \ \ z \in \Omega \}$ le transformé de Ω par une homothétie de rapport λ et énonçons un deuxième lemme prouvé plus loin :

Lemme II.8.2 *On suppose que $\beta \geq 0$ et que $N \geq 2$ (avec $\beta > 0$ si $N = 2$). Alors l'ensemble $S_N = \{ \lambda \geq 0 \ , \ \ \lambda\bar{\Omega} \subset D_N \}$ est un intervalle fermé non vide $[0, \mathcal{L}_{\beta,N}]$.* •

En admettant les lemmes II.8.1 et II.8.2, il est ensuite facile d'achever la preuve de la Proposition II.4.1. En effet, en écrivant :

$$G_1(\theta_1,\theta_2) = (\alpha_1+\alpha_2)\left(\frac{\alpha_1}{\alpha_1+\alpha_2}f(\theta_1) + \frac{\alpha_2}{\alpha_1+\alpha_2}f(\theta_2)\right) ,$$

et en utilisant la convexité de Ω, nous voyons que $G_1(\theta_1,\theta_2) \in (\alpha_1+\alpha_2)\Omega$ pour tout couple (θ_1,θ_2). Alors, si $\alpha_1+\alpha_2 \le \mathcal{L}_{\beta,N}$, $G_1(\theta_1,\theta_2) \in D_N$, d'où $|G_N| = |\mathcal{P}_N(G_1(\theta_1,\theta_2))| \le 1$ par définition de D_N et le schéma est stable à l'ordre N. Inversement, si $\alpha_1+\alpha_2 > \mathcal{L}_{\beta,N}$, alors $(\alpha_1+\alpha_2)\Gamma \not\subset D_N$; nous pouvons alors choisir $\theta_1 = \theta_2 = \theta$ tel que $G_1(\theta,\theta) = (\alpha_1+\alpha_2)f(\theta) \notin D_N$, d'où $|G_N(\theta,\theta)| > 1$ et le schéma est instable.

Puisqu'on sait classiquement que $D_2 \subset D_3 \subset D_4$, on en déduit que, pour β fixé, on a les inégalités $\mathcal{L}_{\beta,2} \le \mathcal{L}_{\beta,3} \le \mathcal{L}_{\beta,4}$.

Il reste maintenant à démontrer les lemmes intermédiaires II.8.1 et II.8.2.

Preuve du lemme II.8.1 : Il est clair que x et y sont des fonctions de période 2π, et que $y(\theta) \le 0$ pour $\theta \in [0,\pi]$, $y(\theta) \ge 0$ pour $\theta \in [-\pi,0]$. Comme $x(\theta) = -\beta(\cos\theta - 1)^2 \le 0$, ceci établit que $\Omega \subset \mathbb{C}^-$ et que le domaine Ω est symétrique par-rapport à l'axe réel. De plus, comme $x(0) = y(0) = 0$, on voit bien que $0 \in \Gamma$.

Pour montrer que Ω est convexe, il suffit alors de montrer que la "moitié supérieure" de la courbe Γ (pour $\theta \in]-\pi,0[$) est concave. Or, on peut écrire le long de cette courbe :

$$\frac{dy}{dx} = \frac{y'(\theta)}{x'(\theta)} , \quad \frac{d^2y}{dx^2} = \frac{y''(\theta)x'(\theta) - x''(\theta)y'(\theta)}{(x'(\theta))^2} ,$$

si bien que, en posant $h(\theta) = y''(\theta)x'(\theta) - x''(\theta)y'(\theta)$, il s'agit de montrer que $h(\theta) \le 0$ pour tout $\theta \in]-\pi,0[$. En posant $\xi = \cos\theta$, on obtient $h(\theta) = 2\beta(1-\beta)\xi^3 + 6\beta^2\xi - 2\beta(1+2\beta)$, et il est aisé de voir que le second membre de cette égalité est une fonction croissante de ξ s'annulant en $\xi = 1$, ce qui achève la preuve du lemme II.8.1. •

Preuve du lemme II.8.2 : Remarquons d'abord que le domaine $\lambda\Omega$ est croissant avec λ (puisque Ω est convexe et que $0 \in \Omega$). Ceci montre que l'ensemble S_N est bien un intervalle. Comme S_N est nécessairement fermé et borné (puisque D_N l'est), il suffit de montrer que S_N est non vide, c'est-à-dire que $\varepsilon \in S_N$ pour $\varepsilon > 0$ petit. Nous traiterons d'abord le cas $N \ge 3$ puis ensuite la cas particulier $N = 2$.

a) Supposons que $N \ge 3$: nous allons montrer que le rectangle :

$$R_\varepsilon = \{z = x + iy \in \mathbb{C} , \quad -\varepsilon \le x \le 0 , \quad |y| \le \varepsilon\}$$

est inclus dans D_N pour $\varepsilon > 0$ suffisamment petit (voir Fig. 14), ce qui prouvera que l'on a aussi $\varepsilon\Omega \subset D_N$. Pour cela, posons $Q_N(x,y) = |\mathcal{P}_N(x+iy)|^2$. Nous obtenons, pour y petit, les développements limités suivants :

$$Q_3(0,y) = 1 - \frac{y^4}{12} + O(y^4) , \quad Q_N(0,y) = 1 - \frac{y^4}{24} + O(y^4) \quad \text{pour } N \le 4 ,$$

ce qui montre que le segment $S_\varepsilon = \{iy, |y| \leq \varepsilon\}$ est inclus dans D_N pour $\varepsilon > 0$ petit. Comme de plus, $Q_N(0,0) = 1$ et $\dfrac{\partial Q_N}{\partial x}(0,0) = 2$, on en déduit que $R_\varepsilon \subset D_N$.

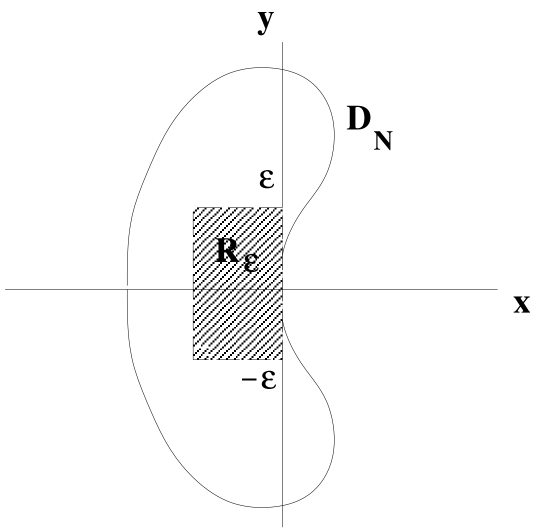


FIG. II.14 –

b) Supposons $N = 2$ et $\beta > 0$. Il n'est plus vrai que $S_\varepsilon \subset D_2$ (c'est pourquoi il faut exclure le cas $\beta = 0$). Si on résout l'équation $Q_2(x,y) = 1$ pour x et y petit, on trouve que la frontière du domaine D_2 a pour équation :

$$x = -\frac{y^4}{8} + O(y^4) \tag{II.55}$$

au voisinage de 0. Par ailleurs les fonctions $x(\theta)$ et $y(\theta)$ admettent les développements limités suivants au voisinage de 0 :

$$x(\theta) = -\frac{\beta\theta^4}{4} + O(\theta^4) \ , \quad y(\theta) = -\theta + O(\theta) \ ,$$

si bien que la courbe $\varepsilon\Gamma$, frontière du domaine $\varepsilon\Omega$ a pour équation au voisinage de 0 :

$$x = -\frac{\beta y^4}{4\varepsilon^3} \ . \tag{II.56}$$

La comparaison des équations (II.55) et (II.56) montre que, au voisinage de 0 et pour $\varepsilon > 0$ assez petit, la frontière du domaine $\varepsilon\Omega$ est située à gauche de la frontière du domaine D_2 (voir Figure II.15), d'où on déduit facilement que $\varepsilon\Omega \subset D_2$.•

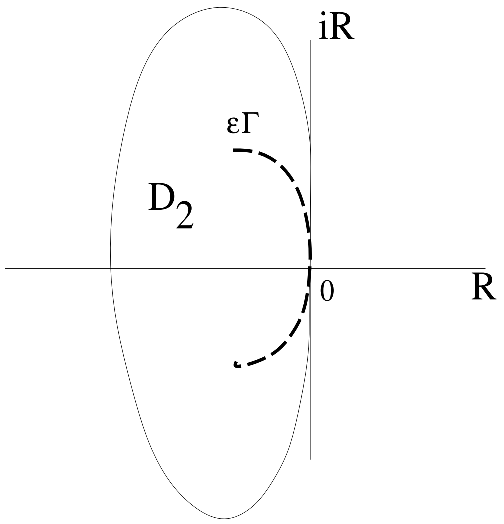


FIG. II.15 –

II.9 Annexe C

Nous donnons dans cette partie les expressions des termes d'erreur des schémas **VF R**, **EF R** et **EF/VF T** à l'ordre un en temps et en espace (tableaux II.7 à II.9), puis à l'ordre ≥ 2 (tableau II.10), pour toute valeur non nulle de Δx et Δy, et pour c_1, c_2 quelconques.

ORDRE 1	**erreur en temps**	**erreur en espace sans matrice de masse**	**erreur en espace avec matrice de masse**
u_{xx}	$-\dfrac{\Delta t}{2}c_1^2$	$\dfrac{\lvert c_1\rvert}{2}\Delta x$	$\dfrac{\lvert c_1\rvert}{2}\Delta x$
u_{yy}	$-\dfrac{\Delta t}{2}c_2^2$	$\dfrac{\lvert c_2\rvert}{2}\Delta y$	$\dfrac{\lvert c_2\rvert}{2}\Delta y$
u_{xxx}	$-\dfrac{\Delta t^2}{3}c_1^3+\dfrac{c_1^2}{2}\Delta t\Delta x$	$-c_1\dfrac{\Delta x^2}{6}$	0
u_{yyy}	$-\dfrac{\Delta t^2}{3}c_2^3+\dfrac{c_2^2}{2}\Delta t\Delta y$	$-c_2\dfrac{\Delta y^2}{6}$	0
u_{xy}	$-\Delta t c_1 c_2$	0	0
u_{xxy}	$-\Delta t^2 c_1^2 c_2+\Delta t\Delta x\dfrac{\lvert c_1\rvert}{2}c_2$	0	$\dfrac{c_2}{6}\Delta x^2$
u_{xyy}	$-\Delta t^2 c_1 c_2^2+\Delta t\Delta y\dfrac{c_1}{2}\lvert c_2\rvert$	0	$\dfrac{c_1}{6}\Delta y^2$

TAB. II.7 – *Schéma volumes finis en rectangles à l'ordre un.*

ORDRE 1	erreur en temps	erreur en espace sans matrice de masse	erreur en espace avec matrice de masse
u_{xx}	$-\frac{\Delta t}{2}c_1^2$	$\frac{\lvert c_1\rvert}{3}\Delta x+\frac{1}{12}\lvert c_1\Delta x+c_2\frac{\Delta x^2}{\Delta y}\rvert+\frac{1}{12}\lvert c_1\Delta x-c_2\frac{\Delta x^2}{\Delta y}\rvert$	$\frac{\lvert c_1\rvert}{3}\Delta x+\frac{1}{12}\lvert c_1\Delta x+c_2\frac{\Delta x^2}{\Delta y}\rvert+\frac{1}{12}\lvert c_1\Delta x-c_2\frac{\Delta x^2}{\Delta y}\rvert$
u_{yy}	$-\frac{\Delta t}{2}c_2^2$	$\frac{\lvert c_2\rvert}{3}\Delta y+\frac{1}{12}\lvert c_2\Delta y+c_1\frac{\Delta y^2}{\Delta x}\rvert+\frac{1}{12}\lvert c_2\Delta y-c_1\frac{\Delta y^2}{\Delta x}\rvert$	$\frac{\lvert c_2\rvert}{3}\Delta y+\frac{1}{12}\lvert c_2\Delta y+c_1\frac{\Delta y^2}{\Delta x}\rvert+\frac{1}{12}\lvert c_2\Delta y-c_1\frac{\Delta y^2}{\Delta x}\rvert$
u_{3x}	$-c_1^3\frac{\Delta t^2}{3}+\frac{\Delta t}{3}c_1\lvert c_1\rvert\Delta x$ $+\frac{c_1}{12}\Delta t\lvert c_1\Delta x+c_2\frac{\Delta x^2}{\Delta y}\rvert+\frac{c_1}{12}\Delta t\lvert c_1\Delta x-c_2\frac{\Delta x^2}{\Delta y}\rvert$	$-c_1\frac{\Delta x^2}{6}$	0
u_{3y}	$-c_2^3\frac{\Delta t^2}{3}+\frac{\Delta t}{3}c_2\lvert c_2\rvert\Delta y$ $+\frac{c_2}{12}\Delta t\lvert c_2\Delta y+c_1\frac{\Delta y^2}{\Delta x}\rvert+\frac{c_2}{12}\Delta t\lvert c_2\Delta y-c_1\frac{\Delta y^2}{\Delta x}\rvert$	$-c_2\frac{\Delta y^2}{6}$	0
u_{xy}	$-\Delta t c_1 c_2$	$\frac{1}{6}\lvert c_1\Delta y+c_2\Delta x\rvert-\frac{1}{6}\lvert c_1\Delta y-c_2\Delta x\rvert$	$\frac{1}{6}\lvert c_1\Delta y+c_2\Delta x\rvert-\frac{1}{6}\lvert c_1\Delta y-c_2\Delta x\rvert$
u_{xxy}	$-\Delta t^2c_1^2c_2+c_1\frac{\Delta t}{6}\lvert c_1\Delta y+c_2\Delta x\rvert+\lvert c_1\rvert c_2\frac{\Delta t}{3}\Delta x$ $-c_1\frac{\Delta t}{6}\lvert c_1\Delta y-c_2\Delta x\rvert+c_2\frac{\Delta t}{12}\lvert c_1\Delta x+c_2\frac{\Delta x^2}{\Delta y}\rvert+c_2\frac{\Delta t}{12}\lvert c_1\Delta x-c_2\frac{\Delta x^2}{\Delta y}\rvert$	$-c_2\frac{\Delta x^2}{6}$	0
u_{xyy}	$-\Delta t^2c_1c_2^2+c_2\frac{\Delta t}{6}\lvert c_1\Delta y+c_2\Delta x\rvert+c_1\lvert c_2\rvert\frac{\Delta t}{3}\Delta y$ $-c_2\frac{\Delta t}{6}\lvert c_1\Delta y-c_2\Delta x\rvert+c_1\frac{\Delta t}{12}\lvert c_2\Delta y+c_1\frac{\Delta y^2}{\Delta x}\rvert+c_1\frac{\Delta t}{12}\lvert c_2\Delta y-c_1\frac{\Delta y^2}{\Delta x}\rvert$	$-c_1\frac{\Delta y^2}{6}$	0

Table II: schéma éléments finis en quadrangles à l'ordre 1.

TAB. II.8 – *Schéma éléments finis en rectangles à l'ordre un.*

17

ORDRE 1	erreur en temps	erreur en espace sans matrice de masse	erreur en espace avec matrice de masse
u_{xx}	$-\dfrac{\Delta t}{2}c_1^2$	$\dfrac{1}{6}\lvert c_1\Delta x + c_2\dfrac{\Delta x^2}{\Delta y}\rvert + \dfrac{1}{6}\lvert -2c_1\Delta x + c_2\dfrac{\Delta x^2}{\Delta y}\rvert$	$\dfrac{1}{6}\lvert c_1\Delta x + c_2\dfrac{\Delta x^2}{\Delta y}\rvert + \dfrac{1}{6}\lvert -2c_1\Delta x + c_2\dfrac{\Delta x^2}{\Delta y}\rvert$
u_{yy}	$-\dfrac{\Delta t}{2}c_2^2$	$\dfrac{1}{6}\lvert c_1\dfrac{\Delta y^2}{\Delta x} + c_2\Delta y\rvert + \dfrac{1}{6}\lvert 2c_2\Delta y - c_1\dfrac{\Delta y^2}{\Delta x}\rvert$	$\dfrac{1}{6}\lvert c_1\dfrac{\Delta y^2}{\Delta x} + c_2\Delta y\rvert + \dfrac{1}{6}\lvert 2c_2\Delta y - c_1\dfrac{\Delta y^2}{\Delta x}\rvert$
u_{xxx}	$-c_1^3\dfrac{\Delta t^2}{3} + \dfrac{\Delta t}{6}c_1(\lvert c_1\Delta x + c_2\dfrac{\Delta x^2}{\Delta y}\rvert + \lvert -2c_1\Delta x + c_2\dfrac{\Delta x^2}{\Delta y}\rvert)$	$-c_1\dfrac{\Delta x^2}{6}$	0
u_{yyy}	$-c_2^3\dfrac{\Delta t^2}{3} + \dfrac{\Delta t}{6}c_2(\lvert c_2\Delta y + c_1\dfrac{\Delta y^2}{\Delta x}\rvert + \lvert 2c_2\Delta y - c_1\dfrac{\Delta y^2}{\Delta x}\rvert)$	$-c_2\dfrac{\Delta y^2}{6}$	0
u_{xy}	$-\Delta t c_1 c_2$	$\dfrac{1}{3}\lvert c_1\Delta y + c_2\Delta x\rvert$	$\dfrac{1}{3}\lvert c_1\Delta y + c_2\Delta x\rvert$
u_{xxy}	$-\Delta t^2 c_1^2 c_2 + c_1\dfrac{\Delta t}{3}\lvert c_1\Delta y + c_2\Delta x\rvert$ $+c_2\dfrac{\Delta t}{6}(\lvert c_1\Delta x + c_2\dfrac{\Delta x^2}{\Delta y}\rvert + \lvert -2c_1\Delta x + c_2\dfrac{\Delta x^2}{\Delta y}\rvert)$	$-\dfrac{1}{6}(c_1\Delta x\Delta y + c_2\Delta x^2)$	0
u_{xyy}	$-\Delta t^2 c_1 c_2^2 + c_2\dfrac{\Delta t}{3}\lvert c_1\Delta y + c_2\Delta x\rvert$ $+c_1\dfrac{\Delta t}{6}(\lvert c_2\Delta y + c_1\dfrac{\Delta y^2}{\Delta x}\rvert + \lvert 2c_2\Delta y - c_1\dfrac{\Delta y^2}{\Delta x}\rvert)$	$-\dfrac{1}{6}(c_1\Delta y^2 + c_2\Delta x\Delta y)$	0

TAB. II.9 – *Schéma volumes finis/éléments finis en triangles à l'ordre un.*

ORDRE 2	erreur en temps	schéma VF quadrangle		schéma EF quadrangle		schéma VF/EF triangle	
		erreur spatiale sans masse	erreur spatiale avec masse	erreur spatiale sans masse	erreur spatiale avec masse	erreur spatiale sans masse	erreur spatiale avec masse
u_{xxx}	RK2: $\frac{1}{6}c_1^3\Delta t^2$ RK3: 0 RK4: 0	$-\frac{c_1}{6}(1-3\beta)\Delta x^2$	$\frac{c_1}{2}\beta\Delta x^2$	$-\frac{c_1}{6}(1-3\beta)\Delta x^2$	$\frac{c_1}{2}\beta\Delta x^2$	$-\frac{c_1}{6}(1-3\beta)\Delta x^2$	$\frac{c_1}{2}\beta\Delta x^2$
u_{yyy}	RK2: $\frac{1}{6}c_2^3\Delta t^2$ RK3: 0 RK4: 0	$-\frac{c_2}{6}(1-3\beta)\Delta y^2$	$\frac{c_2}{2}\beta\Delta y^2$	$-\frac{c_2}{6}(1-3\beta)\Delta y^2$	$\frac{c_2}{2}\beta\Delta y^2$	$-\frac{c_2}{6}(1-3\beta)\Delta y^2$	$\frac{c_2}{2}\beta\Delta y^2$
u_{xxxx}	RK2: $\frac{1}{8}c_1^4\Delta t^3$ RK3: $-\frac{c_1^4}{24}\Delta t^3$ RK4: 0	$-\frac{\beta}{4}\lvert c_1\rvert\Delta x^3$	idem	$-\frac{c_1}{6}\beta\Delta x^3-\frac{\beta}{24}(\lvert c_1\Delta x^3+c_2\frac{\Delta x^4}{\Delta y}\rvert$ $+\lvert c_1\Delta x^3-c_2\frac{\Delta x^4}{\Delta y}\rvert)$	idem	$-\frac{\beta}{12}\lvert(\lvert c_1\Delta x^3+c_2\frac{\Delta x^4}{\Delta y}\rvert$ $+\lvert 2c_1\Delta x^3-c_2\frac{\Delta x^4}{\Delta y}\rvert)$	idem
u_{yyyy}	RK2: $\frac{1}{8}c_2^4\Delta t^3$ RK3: $-\frac{c_2^4}{24}\Delta t^3$ RK4: 0	$-\frac{\beta}{4}\lvert c_1\rvert\Delta x^3$	idem	$-\frac{c_2}{6}\beta\Delta y^3-\frac{\beta}{24}(\lvert c_2\Delta y^3+c_1\frac{\Delta y^4}{\Delta x}\rvert$ $+\lvert c_2\Delta y^3-c_1\frac{\Delta y^4}{\Delta x}\rvert)$	idem	$-\frac{\beta}{12}(\lvert c_2\Delta y^3+c_1\frac{\Delta y^4}{\Delta x}\rvert$ $+\lvert 2c_2\Delta y^3-c_1\frac{\Delta y^4}{\Delta x}\rvert)$	idem
u_{xxy}	RK2: $\frac{1}{2}c_1^2c_2\Delta t^2$ RK3: 0 RK4: 0	0	$\frac{c_2}{6}\Delta x\Delta y$	$-\frac{c_2}{6}(1-3\beta)\Delta x\Delta y$	$\frac{c_2}{2}\beta\Delta x\Delta y$	$-\frac{1}{6}(1-3\beta)$ $(c_1\Delta x\Delta y+c_2\Delta x^2)$	$\frac{\beta}{2}(c_1\Delta x\Delta y+c_2\Delta x^2)$
u_{xyy}	RK2: $\frac{1}{2}c_1c_2^2\Delta t^2$ RK3: 0 RK4: 0	0	$\frac{c_1}{6}\Delta x\Delta y$	$-\frac{c_1}{6}(1-3\beta)\Delta x\Delta y$	$\frac{c_1}{2}\beta\Delta x\Delta y$	$-\frac{1}{6}(1-3\beta)$ $(c_2\Delta x\Delta y+c_1\Delta y^2)$	$\frac{\beta}{2}(c_1\Delta y^2+c_2\Delta x\Delta y)$
u_{xxxy}	RK2: $\frac{1}{2}c_1^3c_2\Delta t^3$ RK3: $-\frac{c_1^3c_2}{6}\Delta t^3$ RK4: 0	0	0	$-\frac{\beta}{8}(\lvert c_1\Delta x^2\Delta y+c_2\Delta x^3\rvert$ $-\lvert c_1\Delta x^2\Delta y-c_2\Delta x^3\rvert)$	idem	$-\frac{\beta}{18}(\lvert -c_2\Delta x^3+2c_1\Delta x^2\Delta y\rvert$ $+5\lvert c_1\Delta x^2\Delta y+c_2\Delta x^3\rvert)$	idem
u_{xyyy}	RK2: $\frac{1}{2}c_1c_2^3\Delta t^3$ RK3: $-\frac{c_2^3c_1}{6}\Delta t^3$ RK4: 0	0	0	$-\frac{\beta}{8}(\lvert c_2\Delta y^2\Delta x+c_1\Delta y^3\rvert$ $-\lvert c_2\Delta x\Delta y^2-c_1\Delta y^3\rvert)$	idem	$-\frac{\beta}{18}(\lvert -c_1\Delta y^3+2c_1\Delta x\Delta y^2\rvert$ $+5\lvert c_2\Delta x\Delta y^2+c_1\Delta y^3\rvert)$	idem
u_{xxyy}	RK2: $\frac{3}{4}c_1^2c_2^2\Delta t^3$ RK3: $-\frac{c_1^2c_2^2}{24}\Delta t^3$ RK4: 0	0	0	$-\frac{\beta}{6}(\lvert c_1\rvert\Delta x\Delta y^2+\lvert c_1\Delta x\Delta y^2+c_2\Delta x^2\Delta y\rvert$ $+\lvert c_2\rvert\Delta x^2\Delta y+\lvert c_1\Delta x\Delta y^2-c_2\Delta x^2\Delta y\rvert)$	idem	$-\frac{\beta}{18}(\lvert 2c_1\Delta x\Delta y^2-c_2\Delta x^2\Delta y\rvert$ $+\lvert -c_1\Delta x\Delta y^2+2c_2\Delta x\Delta y\rvert$ $+14\lvert c_1\Delta x\Delta y^2+c_2\Delta x_2\Delta y\rvert$	idem

TAB. II.10 – *Termes d'erreur pour les trois β-schémas.*

Chapitre III

PRÉSENTATION DU SYSTÈME DE MAXWELL ET ÉTUDE DE STABILITÉ.

L'essentiel de ce chapitre est tiré du rapport Cermics (N^o 95-40) intitulé “Stability analysis for finite volume schemes on rectangular and triangular meshes applied to the two-dimensional Maxwell system”. Ce chapitre est principalement rédigé en anglais, sauf la dernière partie qui est écrite en français.

III.1 Introduction.

Nous nous intéressons ici à la modélisation des équations de Maxwell, qui régissent l' ensemble des phénomènes électromagnétiques. Au premier abord, les équations de Maxwell peuvent sembler simples car il s'agit d'un système linéaire; cependant leur résolution numérique fait apparaître certaines difficultés : par exemple, la précision de l'approximation numérique dépend du pas de discrétisation qui est une fraction de la longueur d'onde, les problèmes physiques doivent être résolus dans des géométries tridimensionnelles complexes comportant souvent des singularités, et la taille de ces problèmes est souvent très grande et demande des méthodes de résolution précises mais d'un coût non prohibitif.
Nous nous intéressons à la résolution des équations de Maxwell dans le domaine temporel. Plusieurs méthodes d'éléments finis ont été proposées : une méthode reposant sur des éléments finis conformes a été développée dans [21]. Néanmoins, celle-ci est basée sur l'introduction de multiplicateurs de Lagrange, ce qui alourdit la résolution numérique. D'autres méthodes utilisent des éléments finis $H(rot)$ [28, 3] parfaitement adaptés à la résolution des équations de Maxwell mais elles nécessitent la prise en compte de la matrice de masse, et la condensation de cette matrice soulève des difficultés [11, 12].
Nous avons choisi une méthode totalement explicite qui repose sur le caractère hyperbolique du système de Maxwell. Il s'agit de schémas de type volumes finis largement utilisés pour des problèmes non linéaires, comme les équations d'Euler. Ce type de méthodes, comme l'a montré J.P. Cioni dans sa thèse [7] est bien adapté à la modélisation de phénomènes électromagnétiques bidimensionnels ou tridimensionnels, comme le calcul de Surface Equivalente Radar, les problèmes de diffraction autour d'un objet de forme et de dimensions variées, les calculs de résonance dans des cavités. Ces schémas ne sont pas d'un coût excessif et l'extension des schémas d'ordre un en temps et en espace à des schémas d'ordre supérieur se fait très facilement.
Ici, on s'est principalement intéressé à l'étude de stabilité des schémas en volumes finis appliqués au système de Maxwell bidimensionnel, pour des maillages en rectangles et en triangles. On s'est restreint dans cette analyse au cas d'un milieu homogène, le vide par exemple.
Ce chapitre est divisé en quatre parties : dans les deux premières, on rappelle les équations de Maxwell et l'approximation numérique utilisée. La troisième partie décrit l'étude de stabilité pour des schémas d'ordre un en temps et en espace, ainsi que pour des schémas d'ordre élevé. Nous comparons les résultats obtenus avec ceux du schéma de Yee, qui est un schéma décalé très utilisé en électromagnétisme, ainsi qu'avec ceux obtenus pour l'équation d'advection. Nous terminons par une illustration numérique, il s'agit d'un calcul de propagation d'onde électromagnétique dans une cavité carrée.

III.2 Maxwell system.

III.2.1 Electromagnetic field equations.

The electric field $\mathbf{E} = \mathbf{E}(t, \mathbf{x})$ and the magnetic induction $\mathbf{B} = \mathbf{B}(t, \mathbf{x})$ are solutions in vacuum of the Maxwell equations:

$$\begin{cases} \dfrac{\partial \mathbf{E}}{\partial t} - c^2 rot(\mathbf{B}) = -\dfrac{\mathbf{j}}{\epsilon_0} \qquad (\mathbf{x} \in I\!R^3, t > 0) \\ \dfrac{\partial \mathbf{B}}{\partial t} + rot(\mathbf{E}) = 0 \\ \\ div(\mathbf{E}) = \dfrac{\rho}{\epsilon_0} \\ div(\mathbf{B}) = 0 \end{cases} \tag{III.1}$$

where c is the light velocity, ϵ_0 the vacuum electric permittivity and μ_0 the vacuum magnetic permeability. These values satisfy the relation: $\epsilon_0 \mu_0 c^2 = 1$.
We denote by $\mathbf{j} = \mathbf{j}(t, \mathbf{x})$ and $\rho = \rho(t, \mathbf{x})$ the given current and the given charge densities which are related by the conservation law:

$$\frac{\partial \rho}{\partial t} + div(\mathbf{j}) = 0 \tag{III.2}$$

We assume that the initial electric field $\mathbf{E}_0$ and the magnetic induction $\mathbf{B}_0$ are such that:

$$div\ \mathbf{E}_0 = \frac{\rho\,(t=0)}{\epsilon_0} = \frac{\rho_0}{\epsilon_0}, \qquad div\ \mathbf{B}_0 = 0 \tag{III.3}$$

In the Maxwell system, conditions (III.3) and the charge conservation law (III.2) imply that the divergence constraints $div\mathbf{E} = \dfrac{\rho}{\epsilon_0}$, $div\mathbf{B} = 0$ are satisfied for all $t > 0$. Hence, only the first two equations of III.1 will be considered in the numerical model since the divergence equations are redondant in the continuous one. This property of Maxwell system will be detailed precisely in chapter IV.

III.2.2 Conservative formulation and hyperbolic character.

System (III.1) can be written in the following conservative form:

$$\mathbf{Q}_t + \mathbf{F_1}(\mathbf{Q})_x + \mathbf{F_2}(\mathbf{Q})_y + \mathbf{F_3}(\mathbf{Q})_z = \mathbf{J} \tag{III.4}$$

where

$$\begin{cases} \mathbf{Q} = {}^t(E_1, E_2, E_3, B_1, B_2, B_3) \\ \mathbf{F_1}(\mathbf{Q}) = {}^t(0, c^2B_3, -c^2B_2, 0, -E_3, E_2) \\ \mathbf{F_2}(\mathbf{Q}) = {}^t(-c^2B_3, 0, c^2B_1, E_3, 0, -E_1) \\ \mathbf{F_3}(\mathbf{Q}) = {}^t(c^2B_2, -c^2B_1, 0, -E_2, E_1, 0) \\ \mathbf{J} = -\dfrac{1}{\epsilon_0}\,{}^t(j_1, j_2, j_3, 0, 0, 0) \end{cases}$$

or in condensed form:

$$\mathbf{Q}_t + \overrightarrow{\nabla}.\mathbb{F}(\mathbf{Q}) = \mathbf{J} \tag{III.5}$$

with $\mathbb{F}(\mathbf{Q}) = {}^t\left(\mathbf{F_1}(\mathbf{Q})\,;\,\mathbf{F_2}(\mathbf{Q})\,;\,\mathbf{F_3}(\mathbf{Q})\right)$.

One can easily check that system (III.5) is hyperbolic. Indeed, let us consider a linear combination of fluxes:

$$\boldsymbol{\mathcal{F}}(\mathbf{Q}, \boldsymbol{\eta}) = \boldsymbol{\eta}.\mathbb{F}(\mathbf{Q})$$

where $\boldsymbol{\eta} = {}^t(\eta_1, \eta_2, \eta_3)$ is any nonzero vector of $\mathbb{R}^3$.
The jacobian matrix $\mathcal{A}$ defined by:

$$\mathcal{A}(\mathbf{Q}, \boldsymbol{\eta}) = \boldsymbol{\eta}.\mathbb{F}'(\mathbf{Q}) = \eta_1\mathcal{A}_1 + \eta_2\mathcal{A}_2 + \eta_3\mathcal{A}_3 \quad , \quad (\mathcal{A}_i)_{i=1,\dots,3} = \frac{\partial}{\partial \mathbf{Q}}\mathbf{F_i}(\mathbf{Q})$$

is diagonalizable for any nonzero vector $\boldsymbol{\eta}$ of $\mathbb{R}^3$ and for any vector $\mathbf{Q}$ of $\mathbb{R}^6$.
Its three real eigenvalues of double multiplicity are given by:

$$\begin{cases} \lambda_1 = c||\boldsymbol{\eta}|| \\ \lambda_2 = -c||\boldsymbol{\eta}|| \\ \lambda_3 = 0 \end{cases}$$

In two dimensions, Maxwell equations can be split into two sets of systems associated to transverse electric polarizations noted **TE** ($\mathbf{E}.\mathbf{e_z}$=0) and transverse magnetic polarizations noted **TM** ($\mathbf{B}.\mathbf{e_z}$=0). From now, we shall restrict our study to the two-dimensional case, which writes :

$$\mathbf{Q}_t + \mathbf{F_1}(\mathbf{Q})_x + \mathbf{F_2}(\mathbf{Q})_y = \mathbf{J} \tag{III.6}$$

In the **TE** case one has :

$$\begin{cases} \mathbf{Q} = {}^t(E_1, E_2, B_3) \\ \mathbf{F_1}(\mathbf{Q}) = {}^t(0, c^2B_3, E_2) \\ \mathbf{F_2}(\mathbf{Q}) = {}^t(-c^2B_3, 0, -E_1) \\ \mathbf{J} = -\dfrac{1}{\epsilon_0}\,{}^t(j_1, j_2, 0) \end{cases}$$

ans the **TM** case:
$$\begin{cases} \mathbf{Q} = {}^t(B_1, B_2, E_3) \\ \mathbf{F_1}(\mathbf{Q}) = {}^t(0, -E_3, -c^2 B_2) \\ \mathbf{F_2}(\mathbf{Q}) = {}^t(E_3, 0, c^2 B_1) \\ \mathbf{J} = -\dfrac{1}{\epsilon_0}\, {}^t(0, 0, 0) \end{cases}$$

The conservative form as well as the hyperbolic character of the Maxwell system leads up naturally to the use of upwind schemes which are known to be well adapted to solve numerically hyperbolic conservative systems, see [17, 26, 31].

III.3 Numerical approximation.

The two-dimensional time domain solver presented here is based on a finite volume formulation using structured triangular or rectangular meshes. We focalise our study on high-order upwind schemes both in time and space. We describe briefly in the following section the finite volume method applied to the Maxwell equations. For more details on this method, one may refer to [18, 29].

III.3.1 Spatial formulation.

Let $\mathcal{T}_h$ be a standard finite element discretization of Ω_h, the polygonal approximation of a computational domain Ω :

$$\Omega_h = \bigcup_{j=1}^{nt} T_j$$

where T_j is a triangular or a rectangular element and nt is the number of elements. Another partition of Ω using finite volumes is then constructed as follows:

$$\Omega = \bigcup_{i=1}^{ns} C_i$$

where ns is the number of nodes and C_i is the control volume or cell whose construction has already been shown on Figures II.1 and II.2.

A weak formulation is then obtained by integrating system (III.5) on each control volume C_i taking the characteristic functions of the cells as test functions.

Assuming partial derivative $\mathbf{Q}_t$ to be constant in space on C_i and using a Green formula yields to the following equation written at each node of the mesh:

$$Area(C_i)\,(\mathbf{Q}_t)_i + \int_{\partial C_i} \mathbb{F}(\mathbf{Q}).\boldsymbol{\nu_i}\, d\sigma = \int_{C_i} \mathbf{J}\, d\mathbf{x}. \tag{III.7}$$

where $\boldsymbol{\nu_i}$ is the unit normal exterior to ∂C_i.

The integral term in equation (III.7) is splitted into a sum of internal fluxes and boundary terms. Since we are mainly interested in the study of stability conditions, we shall consider periodic boundary conditions, which makes the contribution of these boundary terms to be zero.

$$Area(C_i)\,(\mathbf{Q}_t)_i + \sum_{j=1}^{N_i} \mathbf{\Phi}_{ij} = Area(C_i)\mathbf{J_i} \qquad \text{(III.8)}$$

where N_i is the number of the neighbours of the node i and $\mathbf{\Phi}_{ij}$ is an approximation of the internal flux $\int_{\partial C_i \cap \partial C_j} \mathbb{F}(\mathbf{Q}).\boldsymbol{\nu_{ij}}\, d\sigma$ which will be discussed in the sequel.

III.3.2 First-order upwind scheme.

Since the Maxwell system is hyperbolic, we choose an upwind approximation for the evaluation of the numerical fluxes $\mathbf{\Phi}_{ij}$. Let us set :

$$\boldsymbol{\eta} = \int_{\partial C_i \cap \partial C_j} \boldsymbol{\nu_{ij}}\, d\sigma.$$

where $\partial C_i \cap \partial C_j$ represents the common interface between the two cells C_i and C_j. We recall that the Maxwell equations in vacuum form a linear system with constant coefficients. Thus all first-order upwind schemes reduce to the classical I.C.R (Isaac-Courant-Reeves) scheme [26] which writes :

$$\mathbf{\Phi}_{ij} = \mathbf{\Phi}(\mathbf{Q}_i, \mathbf{Q}_j, \boldsymbol{\eta}) = \frac{\mathcal{F}(\mathbf{Q}_i, \boldsymbol{\eta}) + \mathcal{F}(\mathbf{Q}_j, \boldsymbol{\eta})}{2} - \frac{1}{2}|\mathcal{A}(\boldsymbol{\eta})|\,(\mathbf{Q}_j - \mathbf{Q}_i) \qquad \text{(III.9)}$$

where $\mathbf{Q}_i$ denotes the value of $\mathbf{Q}$ at node i and $\mathcal{A}(\boldsymbol{\eta})$ is the jacobian matrix of $\mathcal{F}(\mathbf{Q}, \boldsymbol{\eta})$.

III.3.3 High order approximation.

The MUSCL (Monotonic Upwind Schemes for Conservation Laws) method [19] allows us to increase the precision of the schemes by defining new values $\mathbf{Q}_{ij}$ and $\mathbf{Q}_{ji}$ at the interface of the cells without altering the numerical fluxes fonction $\mathbf{\Phi}$. In the MUSCL method, these values are obtained by using a linear interpolation on each cell. We choose here a β-scheme formulation which writes :

$$\left\{ \begin{aligned} \boldsymbol{\Phi_{ij}} &= \boldsymbol{\Phi_{ij}}(\mathbf{Q}_{ij}, \mathbf{Q}_{ji}) \\ \mathbf{Q}_{ij} &= \mathbf{Q}_i + \frac{1}{2}\{\,(1-2\,\beta)(\mathbf{Q}_j - \mathbf{Q}_i) + 2\,\beta \overrightarrow{\nabla} \mathbf{Q}_i^H . \mathbf{S}_i\mathbf{S}_j\,\} \\ \mathbf{Q}_{ji} &= \mathbf{Q}_j - \frac{1}{2}\{\,(1-2\,\beta)(\mathbf{Q}_j - \mathbf{Q}_i) + 2\,\beta \overrightarrow{\nabla} \mathbf{Q}_j^H . \mathbf{S}_i\mathbf{S}_j\,\} \end{aligned} \right. \qquad \text{(III.10)}$$

where β is an upwinding parameter whose value determines the accuracy of the scheme. Choosing $\beta = \frac{1}{3}$ gives a third-order accurate scheme in space for structured schemes [18]. The formulation requires the evaluation of a nodal gradient $(\vec{\nabla}\mathbf{Q})_{i,j}^H$ which can be defined in several ways. We use here a finite element approach.

In the case of a rectangular mesh, it writes:

$$\begin{aligned} \vec{\nabla}\mathbf{Q}_i^H{}_R &= \frac{1}{Area(Supp(\varphi_i))}\int_{Supp(\varphi_i)} \vec{\nabla}\mathbf{Q}\,d\mathbf{x} \\ &= \frac{1}{Area(Supp(\varphi_i))}\sum_{R,i\in R}\sum_{k=1}^{4}\mathbf{Q}_{i^k}\int_R \vec{\nabla}\varphi_{i^k}d\mathbf{x} \end{aligned} \tag{III.11}$$

where the i^k $(k = 1, ..., 4)$ are the four vertices of the rectangle R and $\vec{\nabla}\varphi_{i^k}$ is the gradient of the bilinear Q^1 function at node i^k.

In the case of a triangular mesh, we use the following definition:

$$\begin{aligned} \vec{\nabla}\mathbf{Q}_i^H{}_T &= \frac{1}{Area(Supp(\varphi_i))}\int_{Supp(\varphi_i)} \vec{\nabla}\mathbf{Q}\,d\mathbf{x} \\ &= \frac{1}{Area(C_i)}\sum_{T,i\in T}\frac{Area(T)}{3}\sum_{k=1}^{4}\mathbf{Q}_{i^k}\vec{\nabla}\varphi_{i^k}(T) \end{aligned} \tag{III.12}$$

where the i^k $(k = 1, 2, 3)$ are the three vertices of the triangle T and $\vec{\nabla}\varphi_{i^k}(T)$ is the gradient of the linear P^1 function at node i^k, which is constant on T.

III.3.4 Time integration.

The time accuracy for unsteady problems is important that is why we choose explicit accurate time schemes. We use a Runge-Kutta multi-step explicit method; the step number for the accuracy of the scheme is fixed with regard to the value of β. The RKr algorithm is given below (in our case $r = 1, ..., 3$):

$$\begin{cases} \mathbf{Q}^0 = \mathbf{Q}^n \\ \mathbf{Q}^l = \mathbf{Q}^0 - \dfrac{\Delta t}{(r+1-l)}\mathbf{\Phi}(\mathbf{Q}^{l-1}) \quad l = 1, ..., r \\ \mathbf{Q}^{n+1} = \mathbf{Q}^r \end{cases}$$

where $t^n = n\Delta t$ and $\mathbf{\Phi}(\mathbf{Q}^{l-1})$ represent the fluxes calculated with fields $\mathbf{Q}^{l-1}$. For the values $\beta = \frac{1}{3}$ and $r = 3$, the scheme is third-order accurate in time and space since the Maxwell system is linear.

III.4 Stability analysis.

We study here the stability of the schemes presented above for both rectangular and triangular meshes. The Maxwell system is written dimensionless and we choose $c = 1$. We consider the first-order accurate scheme and we present a proof of the stability condition in the case of a rectangular grid. Then we study the β-scheme stability on both rectangular and triangular meshes by ajusting the parameter β. Stability study is based on Von Neumann analysis, but we first introduce some definitions before developping this analysis. We note :

$$Q^n_{j,k} = \hat{Q}^n e^{i(j\theta_1 + k\theta_2)} \qquad (p,q) \in \mathbb{Z}^2$$

where i^2=-1. Then we obtain the relation :

$$\hat{Q}^{n+1} = G_{\theta_1,\theta_2}\, \hat{Q}^n$$

where G_{θ_1,θ_2} is the 3x3 amplification matrix of the scheme which depends on the time increment Δt and the Fourier angles $\theta_1,\ \theta_2$.

III.4.1 First-order accurate schemes.

We recall that a necessary and sufficient stability condition (Von Neumann condition) writes:

$$\forall\ (\theta_1, \theta_2) \in [0, 2\pi]^{\,2}, \quad r(G_{\theta_1,\theta_2}) = \max_{l=1,2,3} |\mu^l_{\theta_1,\theta_2}| \leq 1 \tag{III.13}$$

where $\mu^l_{\theta_1,\theta_2}$ are the eigenvalues of G_{θ_1,θ_2} and r is the spectral radius of the matrix G_{θ_1,θ_2}.

Rectangular mesh.

In this part, we shall establish a necessary and sufficient stability condition for the first-order finite volume scheme on a rectangular mesh. The amplification matrix G_{θ_1,θ_2} writes in this case :

$$G_{\theta_1,\theta_2} = Id - \Delta t \begin{pmatrix} X_2 & 0 & \dfrac{i}{\Delta y}\sin(\theta_2) \\ 0 & X_1 & -\dfrac{i}{\Delta x}\sin(\theta_1) \\ \dfrac{i}{\Delta y}\sin(\theta_2) & -\dfrac{i}{\Delta x}\sin(\theta_1) & X_1 + X_2 \end{pmatrix} \tag{III.14}$$

where $X_1 = \dfrac{2}{\Delta x}\sin^2\dfrac{\theta_1}{2}$, $X_2 = \dfrac{2}{\Delta y}\sin^2\dfrac{\theta_2}{2}$.

We notice that G_{θ_1,θ_2} is a complex symmetric matrix.

Theorem III.4.1 *The first-order finite volume scheme applied to the Maxwell system using a rectangular mesh is stable if and only if* $\dfrac{\Delta t}{\Delta x} + \dfrac{\Delta t}{\Delta y} \leq 1$.

Proof: We prove first that the condition $\dfrac{\Delta t}{\Delta x} + \dfrac{\Delta t}{\Delta y} \leq 1$ is necessary and then that it is sufficient.

Proposition III.4.1 *If the scheme is stable then* $\dfrac{\Delta t}{\Delta x} + \dfrac{\Delta t}{\Delta y} \leq 1$.

Demonstrating this assertion is equivalent to show that if $\dfrac{\Delta t}{\Delta x} + \dfrac{\Delta t}{\Delta y} > 1$ there exits a couple (θ_1, θ_2) for which $\max\limits_{l=1,2,3} |\mu^l_{\theta_1,\theta_2}| > 1$. Taking $(\theta_1, \theta_2) = (\pi, \pi)$ leads up to:

$$G_{\theta_1,\theta_2} = Id - \Delta t \begin{pmatrix} \dfrac{2}{\Delta y} & 0 & 0 \\ 0 & \dfrac{2}{\Delta x} & 0 \\ 0 & 0 & \dfrac{2}{\Delta x} + \dfrac{2}{\Delta y} \end{pmatrix}$$

and $r(G_{\theta_1,\theta_2}) = |\ 1 - 2(\dfrac{\Delta t}{\Delta x} + \dfrac{\Delta t}{\Delta y})\ |$. Hence taking $\dfrac{\Delta t}{\Delta x} + \dfrac{\Delta t}{\Delta y} > 1$ leads clearly to an unstable scheme, which ends the proof of proposition (4.1).□

Proposition III.4.2 *If* $\dfrac{\Delta t}{\Delta x} + \dfrac{\Delta t}{\Delta y} \leq 1$, *then the scheme is stable.*

Proof: We first define a new matrix H_{θ_1,θ_2} by multiplying the third column of G_{θ_1,θ_2} by $-i$ and the third line by i. One can easily check that G_{θ_1,θ_2} and H_{θ_1,θ_2} are similar. Hence they have the same eigenvalues, and the stability condition (III.13) is identical when considering H_{θ_1,θ_2} as the scheme amplification matrix. H_{θ_1,θ_2} presents the advantage to be real and can be splitted into: $H_{\theta_1,\theta_2} = Id - \Delta t(D_{\theta_1,\theta_2} + A_{\theta_1,\theta_2})$ where D_{θ_1,θ_2} is a real diagonal matrix and A_{θ_1,θ_2} is a real antisymmetric one.

$$D_{\theta_1,\theta_2} = \begin{pmatrix} X_2 & 0 & 0 \\ 0 & X_1 & 0 \\ 0 & 0 & X_1 + X_2 \end{pmatrix} \quad , \quad A_{\theta_1,\theta_2} = \begin{pmatrix} 0 & 0 & -\dfrac{1}{\Delta y}\sin\theta_2 \\ 0 & 0 & \dfrac{1}{\Delta x}\sin\theta_1 \\ \dfrac{1}{\Delta y}\sin\theta_2 & -\dfrac{1}{\Delta x}\sin\theta_1 & 0 \end{pmatrix}$$

The matrix $(D_{\theta_1,\theta_2} + A_{\theta_1,\theta_2})$ has either three real eigenvalues or one real eigenvalue and two complex conjugate ones. Concerning the real eigenvalues we have the following result:

Lemma III.4.1 *The real eigenvalues $\lambda_{r,\theta_1,\theta_2}$ of $(D_{\theta_1,\theta_2} + A_{\theta_1,\theta_2})$ verify*

$$0 \leq \min(X_1, X_2) \leq \lambda_{r,\theta_1,\theta_2} \leq X_1 + X_2$$

We omit from now the subscripts θ_1, θ_2 in what follows. This lemma and all the following ones will be proved further.

Let μ be the eigenvalues of H and λ the eigenvalues of $(D + A)$. Then we have $\mu = 1 - \Delta t\lambda$.
We first consider the case of the real eigenvalues μ_r of H.
Equation (III.13) implies: $|\mu_r| \leq 1 \;\; \forall \; (\theta_1, \theta_2)$ i.e. $-1 \leq 1 - \Delta t\lambda_r \leq 1 \;\; \forall \; (\theta_1, \theta_2)$.
Using lemma III.4.1 one has $0 \leq \lambda_r \leq X_1 + X_2$, furthermore $X_1 + X_2 \leq \dfrac{2}{\Delta x} + \dfrac{2}{\Delta y}$.
Therefore, if the condition $\dfrac{\Delta t}{\Delta x} + \dfrac{\Delta t}{\Delta y} \leq 1$ is assumed one obtains $0 \leq \Delta t\lambda_r \leq 2$ and then $| \mu_r | \leq 1$.
We now consider the case of the complex eigenvalues μ_c of H. First we have:

$$| \mu_c |^2 = 1 - 2\Delta t Re(\lambda_c) + \Delta t^2 \mid \lambda_c \mid^2$$

The condition $| \mu_c | \leq 1$ writes $\Delta t^2 \mid \lambda_c \mid^2 - 2\Delta t Re(\lambda_c) \leq 0$
Assuming the real eigenvalue λ_r is strictly positive, we obtain by multiplying the previous inequality by λ_r:

$$\Delta t^2 \lambda_r \mid \lambda_c \mid^2 - 2\Delta t \lambda_r Re(\lambda_c) \leq 0 \tag{III.15}$$

Furthermore one has:
$\lambda_r \mid \lambda_c \mid^2 = det(D + A) = 2X_1X_2(\dfrac{1}{\Delta x} + \dfrac{1}{\Delta y})$, $\lambda_r + 2Re(\lambda_c) = Tr(D + A) = 2(X_1 + X_2)$.
Inequality (III.15) transforms into:

$$P(\lambda_r) = \lambda_r^2 - 2(X_1 + X_2)\lambda_r + 2X_1X_2(\frac{\Delta t}{\Delta x} + \frac{\Delta t}{\Delta y}) \leq 0 \tag{III.16}$$

Lemma III.4.2 *If the condition $\dfrac{\Delta t}{\Delta x} + \dfrac{\Delta t}{\Delta y} \leq 1$ is achieved, then $P(\lambda_r) \leq 0$.*

Using lemma III.4.2 allows us to conclude that $| \mu_c | \leq 1 \;\; \forall \; (\theta_1, \theta_2)$.

The following lemma treats the case $\lambda_r = 0$.

Lemma III.4.3 *If $\dfrac{\Delta t}{\Delta x} + \dfrac{\Delta t}{\Delta y} \leq 1$, the condition $| \mu_c | \leq 1 \;\; \forall \; (\theta_1, \theta_2)$ is ensured.*

We have finally proved that if $\dfrac{\Delta t}{\Delta x} + \dfrac{\Delta t}{\Delta y} \leq 1$, then the scheme is stable which concludes the proof of proposition (III.4.2) and the demonstration of theorem (III.4.1).□

We establish in the sequel the proof of all the intermediary lemma.

Proof of lemma III.4.1 :

Let $\mathbf{v}$ be the eigenvector of the real matrix $(D + A)$ associated to the eigenvalue λ_r. We have: $(D + A)\mathbf{v} = \lambda_r \mathbf{v}$ and ${}^t\mathbf{v}(D + A)\mathbf{v} = {}^t\mathbf{v}\lambda_r\mathbf{v}$. As A is an antisymmetric matrix, from ${}^t\mathbf{v}A\mathbf{v} = 0$ we deduce ${}^t\mathbf{v}D\mathbf{v} = {}^t\mathbf{v}\lambda_r\mathbf{v}$ which writes:

$$(X_2 - \lambda_r)v_1^2 + (X_1 - \lambda_r)v_2^2 + (X_1 + X_2 - \lambda_r)v_3^2 = 0$$

where v_i $(i = 1, 2, 3)$ are the components of the eigenvector $\mathbf{v}$.

We note that $X_1 + X_2 - \lambda_r \geq \max(X_1 - \lambda_r, X_2 - \lambda_r)$ as X_1 and X_2 are positive. Since the coefficients in front of v_i can not have all the same sign, one can deduce that $X_1 + X_2 - \lambda_r > 0$ and $(X_1 - \lambda_r)$ or $(X_2 - \lambda_r)$ are negative.
Thus one can conclude that $\lambda_r \leq X_1 + X_2$ and $\lambda_r \geq \min(X_1, X_2) \geq 0$. □

Proof of lemma III.4.2 :

We recall that:

$$P(\lambda_r) = \lambda_r^2 - 2(X_1 + X_2)\lambda_r + 2X_1X_2(\frac{\Delta t}{\Delta x} + \frac{\Delta t}{\Delta y}).$$

The discriminant of P writes:

$$\Delta = 4(X_1 + X_2)^2 - 8X_1X_2(\frac{\Delta t}{\Delta x} + \frac{\Delta t}{\Delta y}).$$

Assuming the condition $\dfrac{\Delta t}{\Delta x} + \dfrac{\Delta t}{\Delta y} \leq 1$ leads to $\Delta \geq 4(X_1^2 + X_2^2) \geq 0$.

The particular case $\Delta = 0$ corresponds to $X_1^2 + X_2^2 = 0$ which is equivalent to $\lambda_r = 0$ in view of lemma III.4.1. Since we consider the case $\lambda_r \neq 0$, the discriminant is strictly positive and the polynom $P(\lambda_r)$ has two distinct roots r_1, r_2 given by:

$$r_1 = \frac{2(X_1 + X_2) - \sqrt{\Delta}}{2} \quad , \quad r_2 = \frac{2(X_1 + X_2) + \sqrt{\Delta}}{2}$$

We obtain that $P(\lambda_r) \leq 0$ ever since $\lambda_r \in [r_1, r_2]$. Lemma (III.4.1) establishes that $0 < \lambda_r \leq X_1 + X_2 \leq r_2$, then we still have to show that $\lambda_r \geq r_1$.

From $\Delta \geq 4(X_1^2 + X_2^2)$ we have $r_1 \leq X_1 + X_2 - \sqrt{X_1^2 + X_2^2}$. Furthermore $\sqrt{X_1^2 + X_2^2} \geq \max(X_1, X_2)$ and $X_1 + X_2 - \sqrt{X_1^2 + X_2^2} \leq X_1 + X_2 - \max(X_1, X_2) = \min(X_1, X_2) \leq \lambda_r$ thanks to lemma (III.4.1).
Finally $r_1 \leq \lambda_r \leq r_2$ and $P(\lambda_r) \leq 0$ which ends the proof of lemma III.4.2. □

Proof of lemma III.4.3 :

We consider here the case of a zero eigenvalue λ_r. From lemma III.4.1, if $\lambda_r = 0$ then $\min(X_1, X_2) = 0$ that is to say $\theta_1 = 2k\pi$ or $\theta_2 = 2k\pi$, ($k \in \mathbb{Z}$). Conversely, if θ_1 or $\theta_2 = 2k\pi$ then $\lambda_r = 0$. Hence $\lambda_r = 0$ is equivalent to $\theta_1 = 2k\pi$ or $\theta_2 = 2k\pi$.
Assuming $\theta_1 = 2k\pi$, the matrix $(D + A)$ writes:

$$D + A = \begin{pmatrix} X_2 & 0 & -\dfrac{1}{\Delta y}\sin\theta_2 \\ 0 & 0 & 0 \\ \dfrac{1}{\Delta y}\sin\theta_2 & 0 & X_2 \end{pmatrix}$$

In this case, the eigenvalues of the matrix are: $0, X_2 \pm \dfrac{i}{\Delta y}\sin\theta_2$. The condition $|\mu_c|^2 = |1 - \Delta t \lambda_c|^2 \leq 1$ writes: $\Delta t^2 X_2^2 + \dfrac{\Delta t^2}{\Delta y^2}\sin^2\theta_2 - 2\Delta t X_2 \leq 0$.

One can easily check that this condition is achieved since $\dfrac{\Delta t}{\Delta y} \leq 1$. In the same way if we consider the case $\theta_2 = 2k\pi$, the stability condition is satisfied if $\dfrac{\Delta t}{\Delta x} \leq 1$.
To sum up in the case of a zero eigenvalue, the condition $\dfrac{\Delta t}{\Delta x} + \dfrac{\Delta t}{\Delta y} \leq 1$ implies $|\ \mu_c\ | \leq 1$, which ends the proof of lemma III.4.3. □

We have proved that a necessary and sufficient stability condition for the first-order scheme on a rectangular mesh was $\dfrac{\Delta t}{\Delta x} + \dfrac{\Delta t}{\Delta y} \leq 1$.
A way to represent the stability domain is to obtain numerically the maximum values of the couple $(\dfrac{\Delta t}{\Delta x}, \dfrac{\Delta t}{\Delta y})$ such that the condition (III.13) may be verified. To represent this domain, we choose the variables $\dfrac{\Delta t}{\Delta x}$ and $\dfrac{\Delta t}{\Delta y}$ as coordinates in the plane.

Remarks:

- First we can notice that if we consider one direction infinite, for instance Δy, we obtain the one-dimensional stability condition $\dfrac{\Delta t}{\Delta x} \leq 1$. We note also that the stability domain represented on Figure III.1 is the same as the one obtained when

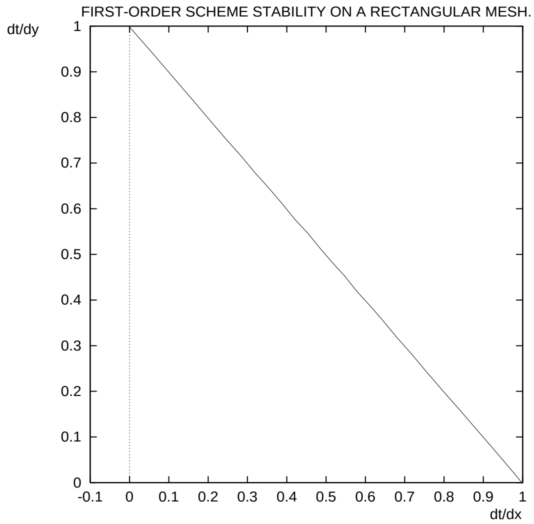


FIG. III.1 – *Maxwell system and convection equation.*

considering the first-order scheme applied to the two-dimensional scalar convection equation $u_t + u_x + u_y = 0$ on a rectangular mesh. For more details on the stability analysis concerning the convection equation, one may refer to [6, 18].

– Maxwell system (III.1) can be written into a non conservative form:

$$\mathbf{Q}_t + A\mathbf{Q}_x + B\mathbf{Q}_y = 0$$

where A and B are the jacobians of the fluxes $\mathbf{F_1}(\mathbf{Q})$ and $\mathbf{F_2}(\mathbf{Q})$.
In the one-dimensional case, one can diagonalize the jacobian matrix which leads up to a splitted system: each component is solution of the convection equation with speeds (c,-c,0).
Unfortunately, in the two-dimensional case, the matrixes A and B are not diagonalizable in the same basis for the two space coordinates (x, y). Thus it is not possible to transform the Maxwell system in order to obtain a system which each component may verify the convection equation, as it is in one dimension. However we find the same stability condition for the first-order scheme applied either to the Maxwell system or to the convection equation. As we shall see later, we do not observe the same concerning the β-schemes and the schemes using a triangular mesh.

Triangular mesh.

In this part we study the stability in the case of a first-order scheme applied to the Maxwell system. The mesh used here is a structured triangular mesh obtained by cutting rectangles diagonally. We use again a Fourier analysis but in this case the matrix H writes: $H = Id - \Delta t(D + A + S)$ where S is a symmetric matrix, which prevents us to apply the same demonstration as for the rectangular mesh.

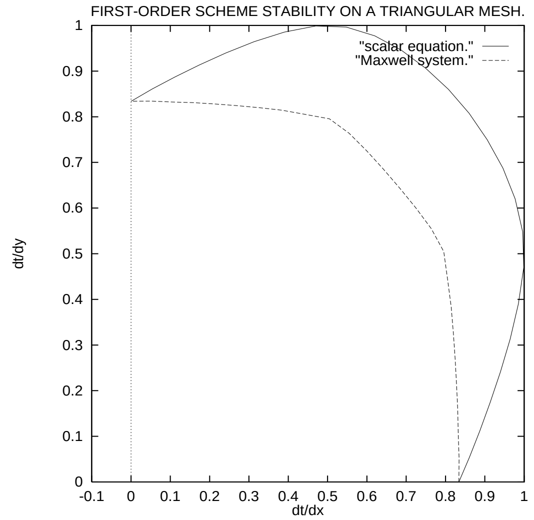


FIG. III.2 – *Maxwell system and convection equation.*

The eigenvalues of the amplification matrix are calculated numerically in order to obtain numerically a sufficient stability condition verifying (III.13). As done in the rectangular case, we choose the variables $(\dfrac{\Delta t}{\Delta x}, \dfrac{\Delta t}{\Delta y})$ to represent the stability domain. A comparison between the convection equation and the Maxwell system is shown on Figure III.2. We notice that the stability domain obtained when considering the convection equation is wider than the one obtained with the Maxwell system.
If we choose $\Delta x = \Delta y$, we can take CFL=1.17 in the case of the convection equation, and CFL=0.93 for the Maxwell system, where CFL is the Courant-Friedricks-Levy number. The stability limit is generally higher for the triangular mesh (see Figures III.1 and III.2). However, when a direction Δx or Δy is almost infinite, then the rectangular mesh gives a higher stability limit.

Comparison with the Yee scheme.

One recalls that Yee introduced a set of finite-difference equations to discretize Maxwell equations. Yee algorithm consists in using finite difference expressions for the space and time derivatives, and in positioning the components of $\mathbf{E}$ and $\mathbf{B}$ orthogonally to each other. In order to achieve a second-order accurate scheme, in time and space, $\mathbf{E}$ and $\mathbf{B}$ are evaluated at half-time and half-space steps. The stability criterion for the two-dimensional

Yee scheme writes as:

$$\Delta t\sqrt{\frac{1}{\Delta x^2}+\frac{1}{\Delta y^2}}\leq 1 \tag{III.17}$$

and a proof of the above result can be found in [30].

Choosing the variables $\frac{\Delta t}{\Delta x}$ and $\frac{\Delta t}{\Delta y}$ to represent the stability domain leads up to a quarter-circular unit domain. Therefore, this stability condition is less restrictive than the one obtained for the first-order upwind scheme on both rectangular and triangular meshes (see Figure III.3). It is mainly due to the second-order accuracy of the Yee scheme.

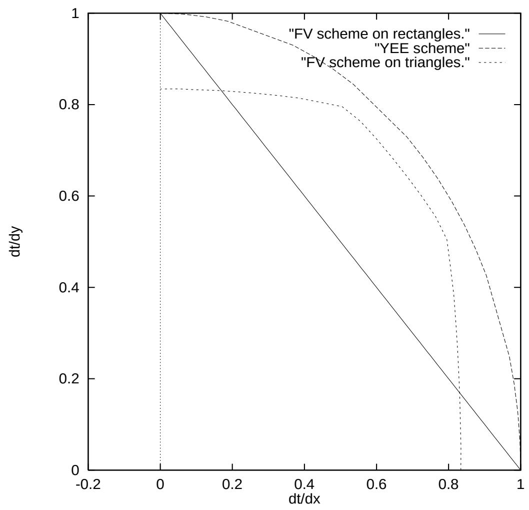


FIG. III.3 – *Stability domains for Yee and first-order finite volume schemes.*

III.4.2 Higher order schemes.

In the case of a three-step Runge-Kutta time integration we introduce the characteristic polynom

$$g(z)=1+z+\frac{z^2}{2}+\frac{z^3}{6}.$$

For $z=A\Delta t$, we recall that the polynom $\mathcal{G}(A\ \Delta t)$ represents the amplification matrix of the Runge-Kutta method applied to the differential system $Q_t=A\,Q$ where A is the 3x3 scheme matrix. We obtain the following relation using a Fourier analysis :

$$\hat{Q}^{n+1}=\mathcal{G}_{\theta_1,\theta_2}(A\ \Delta t)\ \hat{Q}^n$$

and Von Neumann theorem (III.13) still applies to $\mathcal{G}_{\theta_1,\theta_2}$.

Rectangular mesh.

In this section we plot some stability domains computed with different values of the upwinding parameter β.
We recall that for $\beta = 0$ we obtain a centered scheme, for $\beta = \frac{1}{2}$ the scheme is half-centered, $\beta = 1$ gives an upwind scheme.
Figure III.4 shows that the closer to 1 β is, the smaller the stability limit is, which means that using a centered scheme allows us to take a higher time step.

As a comparison, we represent on Figure III.5 the stability domain in the case of the convection equation. Although they vary in the same way with β, the stability domains are different except for $\beta = 1$ where we obtain in both cases the numerical stability limit: $\frac{\Delta t}{\Delta x} + \frac{\Delta t}{\Delta y} = 0.62$. If we choose Δx or Δy infinite we find in both cases the one-dimensional stability limit. For Δx and Δy finite and for a fixed value of β the stability domain is wider for the Maxwell system than for the advection equation, especially for $\beta = \frac{1}{3}$ or $\beta = 0$.

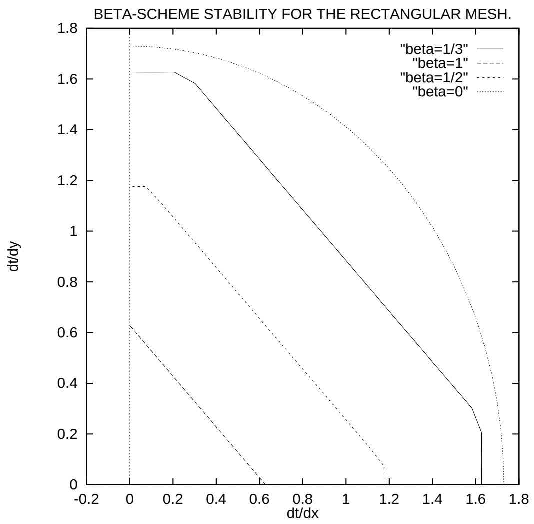


FIG. III.4 – *Maxwell system.*

Triangular mesh.

The stability domains obtained for the β-schemes on a triangular mesh are still different concerning the Maxwell system and the convection equation (see Figures III.6 and III.7). As for the rectangular case, the stability limit decreases when β is closer to 1. However, the stability domains are wider in the case of the convection equation for all value of the upwinding parameter β.

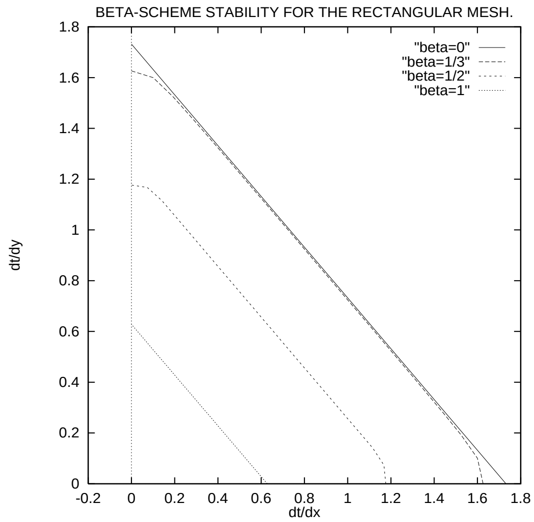


FIG. III.5 – *Convection equation.*

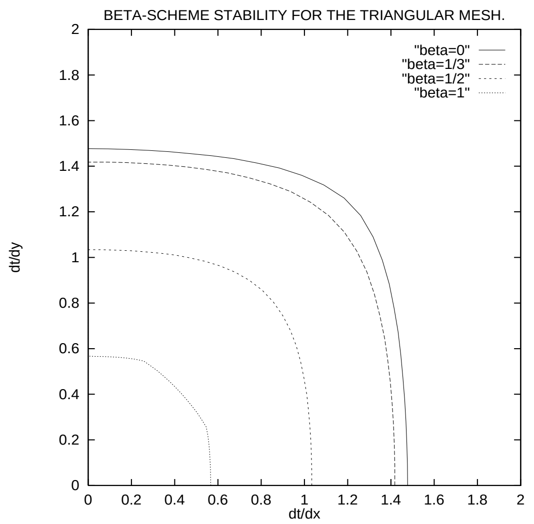


FIG. III.6 – *Maxwell system.*

If we take Δx or Δy infinite we find again the same one-dimensional stability limit for the Maxwell system and for the convection equation.
We can notice that using a triangular mesh gives the privilege to the direction $\Delta x = \Delta y$ concerning the stability: it is the direction where we can choose the highest time step, on the contrary to the rectangular mesh where imposing $\Delta x = \Delta y$ is the most restrictive choice. However, if a direction Δx or Δy is close to infinity, the use of rectangular meshes is more interesting concerning the stability limit. The two last remarks concern the Maxwell system as well as the scalar convection equation.

Comparison with the Yee scheme.

We now compare the stability domains obtained for the Yee scheme and for third-order finite volume schemes applied to the Maxwell system. Figure III.8 shows that stability domains are wider for β-schemes ($\beta = \dfrac{1}{3}$), on both rectangular and triangular meshes.
On the contrary to the Yee scheme, the finite volume approach has the advantage to extend easily first-order upwind schemes to β-schemes, and then to achieve high-order accuracy in time and space. This method is also more flexible, as it can be applied to many sorts of meshes. However, β-schemes require an additionnal CPU time cost compared to the use of the Yee scheme.

In the case of third-order finite volume schemes, stability domains are wider, on both rectangular and triangular meshes.

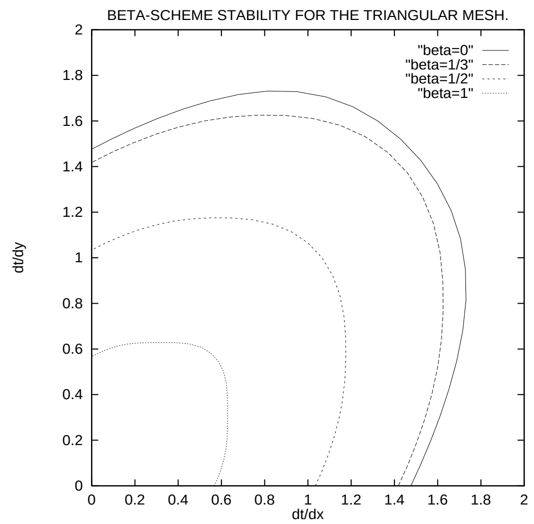


FIG. III.7 – *Convection equation.*

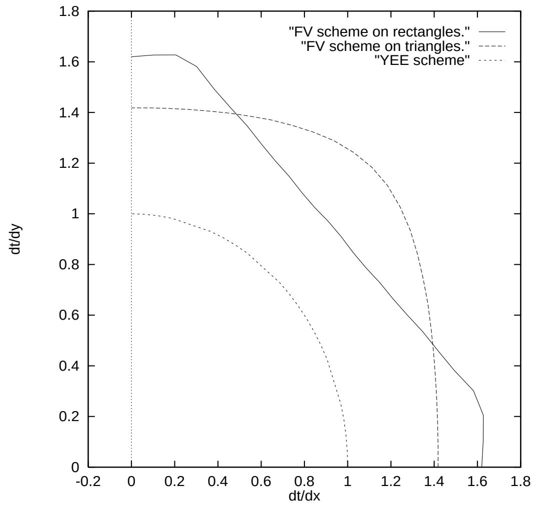


FIG. III.8 – *Stability domains for Yee and third-order finite volume schemes.*

III.4.3 Domaines de stabilité pour les schémas centrés.

Comme dans le cas scalaire linéaire (voir section II.3.3), on s'intéresse aussi aux schémas "sans diffusion numérique" appliqués au système de Maxwell. Ces schémas sont obtenus en remplaçant l'expression du flux numérique (III.9) par le flux centré :

$$\mathbf{\Phi}_{ij} = \mathbf{\Phi}(\mathbf{Q}_i, \mathbf{Q}_j, \boldsymbol{\eta}) = \frac{\mathcal{F}(\mathbf{Q}_i, \boldsymbol{\eta}) + \mathcal{F}(\mathbf{Q}_j, \boldsymbol{\eta})}{2} \tag{III.18}$$

Utilisé avec la valeur $\beta = \dfrac{1}{3}$ et avec une intégration d'ordre quatre en temps de type Runge-Kutta, le schéma est d'ordre quatre en temps et en espace, dans le cas d'un maillage structuré en rectangles ou en triangles. Comme dans le cas scalaire, cela se voit aisément à partir des équations équivalentes des schémas, que nous établissons dans le chapitre IV.

Nous comparons ici les domaines de stabilité obtenus avec ces schémas dans le cas de l'équation de convection et du système de Maxwell.

Dans le cas d'un maillage rectangulaire, la Figure III.9 montre que le domaine de stabilité obtenu pour le système de Maxwell est bien plus grand que celui obtenu pour l'équation de convection. Au contraire, pour un maillage en triangles, c'est pour l'équation de convection que le domaine de stabilité est le plus large (voir Figure III.10). Comme précédemment, on voit que dans le cas monodimensionnel, on retrouve la même limite de stabilité dans le cas scalaire et dans le cas du système de Maxwell. Comparés au schéma de Yee, les domaines de stabilité obtenus pour les schémas (III.18) permettent un choix de pas temps très grand (le nombre de Courant vaut 2.06 pour le schéma en rectangles et 1.76 pour celui en triangles); néanmoins ils nécessitent une approximation d'ordre quatre en temps, ce qui est relativement coûteux.

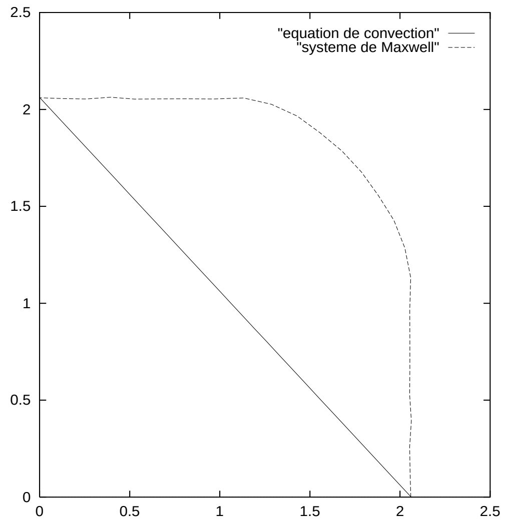


FIG. III.9 – *Domaines de stabilité pour un maillage rectangulaire.*

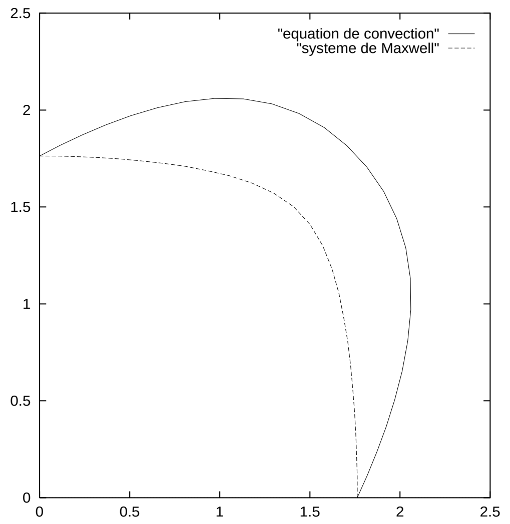


FIG. III.10 – *Domaines de stabilité pour un maillage triangulaire.*

III.5 Illustration numérique.

Nous souhaitons comparer le schéma décentré d'ordre trois en temps et en espace avec le schéma “sans diffusion numérique” d'ordre quatre en temps et en espace. Pour cela, nous nous intéressons à la propagation d'une onde électromagnétique dans une cavité carrée, en l'absence de charge et de courant. Le schéma d'ordre quatre peut dans certains cas, poser des difficultés pour obtenir la solution numérique, cela provient de son caractère centré. Un moyen de le rendre plus robuste consiste, soit à lui adjoindre des limiteurs (nous verrons cela au chapitre VII), soit à considérer le schéma décentré d'ordre trois en espace et quatre en temps, pour lequel le terme de diffusion est multiplié par un paramètre γ, choisi le plus petit possible (dans notre cas, $\gamma = 0.1$). Nous baptisons ce schéma “schéma modifié d'ordre quatre”.

On considère une onde transverse magnétique (**TM**) sur le domaine $\Omega =]0,1[\text{x}]0,1[$ avec des conditions aux limites périodiques. Nous allons comparer les solutions lorsque l'on discrétise le domaine avec des rectangles ou avec des triangles structurés, comportant 21 points par longueur d'onde. Nous prenons pour tous les schémas le même nombre de Courant : $CFL = 1$. On initialise les composantes du champ électromagnétique par :

$$\begin{cases} B_x = 0 \\ B_y = 0 \\ E_z = \sin(\pi x)\sin(\pi y) \end{cases}$$

Dans ce cas, la solution exacte s'écrit :

$$\begin{cases} B_x = -\frac{1}{\sqrt{2}}\pi\sin(\pi x)\cos(\pi y)\sin t \\ B_y = \frac{1}{\sqrt{2}}\pi\cos(\pi x)\sin(\pi y)\sin t \\ E_z = \sin(\pi x)\sin(\pi y)\cos t \end{cases}$$

On représente tout d'abord l'erreur L^2 commise par les schémas sur la composante E_z en fonction du temps, pour le schéma décentré d'ordre trois, le schéma centré d'ordre quatre et le schéma modifié d'ordre quatre.

Les figures III.11 et III.12 montrent la supériorité en gain de précision du schéma modifié d'ordre quatre. L'erreur augmente très rapidement pour le schéma décentré d'ordre trois, qui a tendance à dissiper la solution. L'erreur obtenue avec le schéma centré est moins importante et elle s'avère périodique au cours du temps. Avec la figure III.13, nous pouvons comparer l'erreur du schéma modifié en fonction du choix de la discrétisation : pour un maillage en triangles nous voyons que l'erreur a doublé au temps $t = 20$ secondes par-rapport au maillage en rectangles.

Maintenant, nous considérons un maillage en triangles comportant 31 points par longueur d'onde. Nous présentons sur les figures III.14 et III.15, les valeurs des composantes

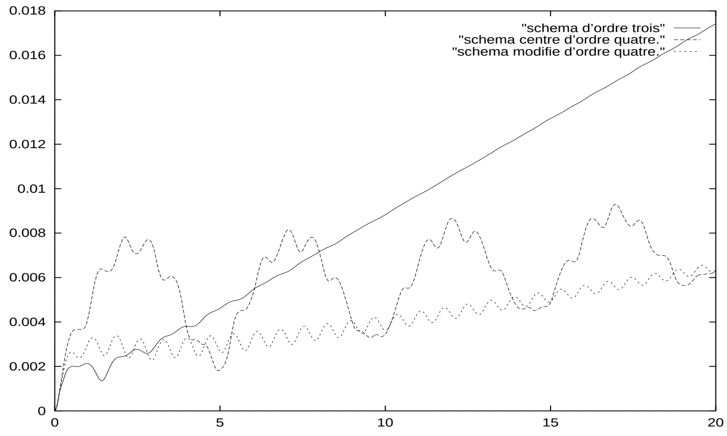


FIG. III.11 – *Comparaison des schémas pour un maillage en rectangles.*

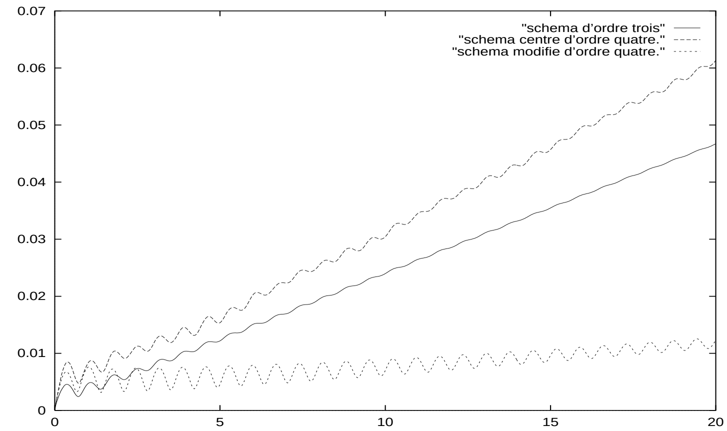


FIG. III.12 – *Comparaison des schémas pour un maillage en triangles.*

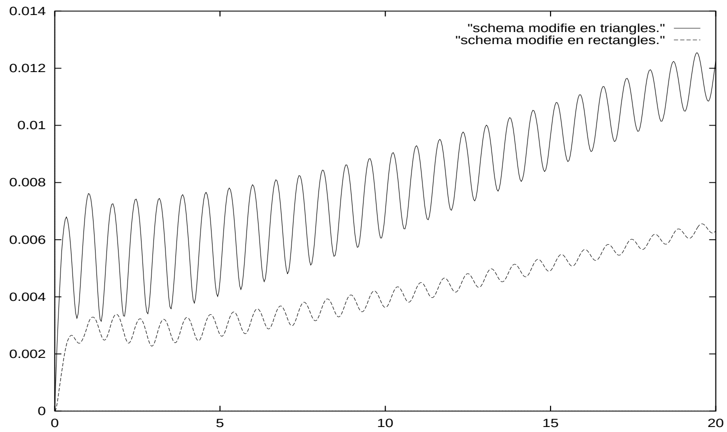


FIG. III.13 – *Comparaison entre le maillage en rectangles et celui en triangles .*

du champ électromagnétique sur la diagonale de la cavité (pour $x = y$), prises au temps $t = 1$ seconde. Nous comparons le schéma décentré d'ordre trois avec le schéma modifié d'ordre quatre.

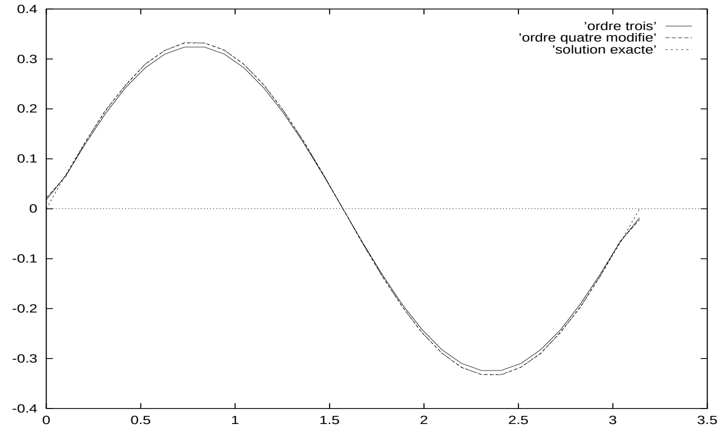


FIG. III.14 – *Comparaison entre l'ordre trois et l'ordre quatre pour* B_x.

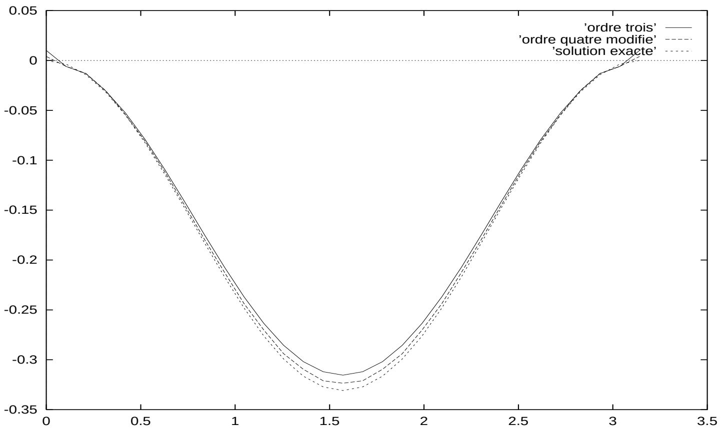


FIG. III.15 – *Comparaison entre l'ordre trois et l'ordre quatre pour* E_z.

Là encore, le schéma modifié d'ordre quatre augmente bien la précision de la solution par-rapport au schéma décentré d'ordre trois.

Sur les figures III.16, III.17, III.18, est représentée l'erreur commise par les schémas sur les relations de divergence $div\mathbf{E} = 0$, $div\mathbf{B} = 0$. Nous verrons dans le chapitre suivant pourquoi ces équations ne sont pas vérifiées numériquement. Nous constatons sur les figures III.16 et III.17 que l'erreur sur la divergence dépend du pas de maillage : quand on augmente le nombre de points par longueur d'onde, l'erreur diminue sensiblement.

On compare sur la figure III.18 l'erreur sur la divergence pour le schéma décentré d'ordre trois et le schéma modifié d'ordre quatre : les erreurs obtenues sont faibles et on

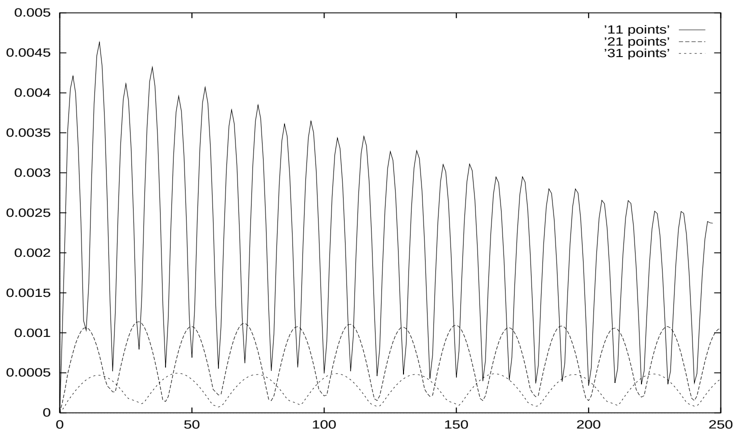


FIG. III.16 – *Influence du pas de maillage sur la divergence pour les schémas d'ordre trois.*

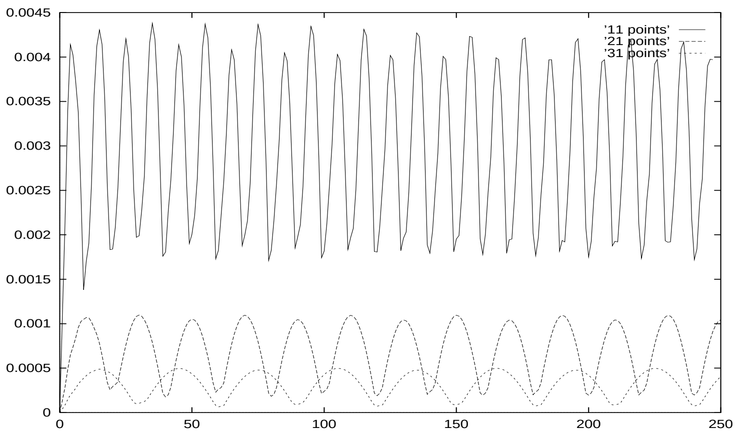


FIG. III.17 – *Influence du pas de maillage sur la divergence pour les schémas modifiés d'ordre quatre.*

n'observe pas de différence notoire entre les deux schémas. Cependant, il a été montré dans [8] que la précision en espace des schémas a une grande influence sur la diminution de l'erreur de la divergence.

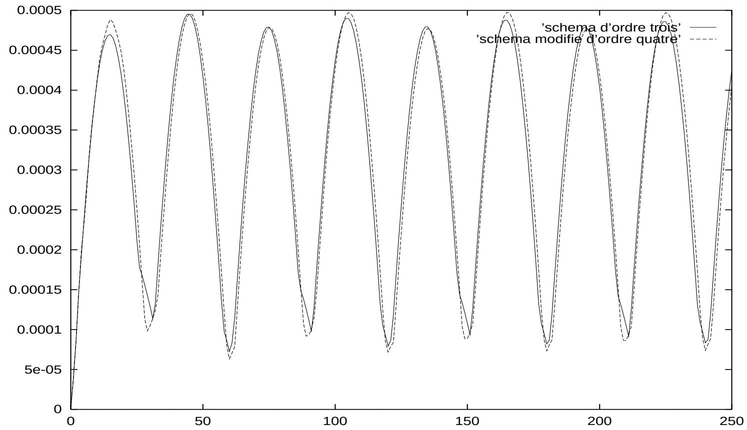


FIG. III.18 – *Comparaison entre l'ordre trois et l'ordre quatre pour la divergence.*

Dans le chapitre suivant, nous nous intéressons plus en détail à ce problème de la non-conservation des propriétés de divergence par nos schémas de type volumes finis/éléments finis, et nous proposons une nouvelle méthode afin de mieux vérifier ces relations au niveau discret.

Chapitre IV

UNE NOUVELLE FORMULATION DU SYSTÈME DE MAXWELL.

Réalisé avec Didier Issautier*

*CERMICS-INRIA, 06902 Sophia-Antipolis Cedex, France

Ce chapitre est tiré du rapport Cermics (N^o 95-39) intitulé “Application aux schémas volumes finis d’une méthode de pénalisation des contraintes pour le système de Maxwell”. Il a été accepté pour publication dans la revue *Mathematical Modelling and Numerical analysis* sous le titre “A new constrained formulation of the Maxwell system”.

IV.1 Introduction.

On s'intéresse ici au problème de la conservation des propriétés de divergence des champs électrique et magnétique dans les équations de Maxwell :

$$\begin{cases} div(\mathbf{E}) = \dfrac{\rho}{\epsilon_0} \\ div(\mathbf{B}) = 0 \end{cases} \tag{IV.1}$$

Nous rappelons le résultat suivant. Si à l'instant initial, on se donne un champ électromagnétique $(\mathbf{E_0}, \mathbf{B_0})$ vérifiant les contraintes (IV.1), alors ces contraintes sont satisfaites pour tout $t > 0$.

En effet, en prenant la divergence des équations (III.1) du système de Maxwell et en utilisant les propriétés $div(rot\mathbf{E}) = 0$, $div(rot\mathbf{B}) = 0$, on obtient :

$$\begin{cases} \dfrac{\partial(div\ \mathbf{E})}{\partial t} = -\dfrac{div\ \mathbf{j}}{\epsilon_0} \\ \dfrac{\partial(div\ \mathbf{B})}{\partial t} = 0 \end{cases}$$

Si les densités de charge et de courant ρ et $\mathbf{j}$ vérifient l'équation de conservation de la charge (III.2), on obtient le résultat annoncé :

$$\begin{cases} \dfrac{\partial}{\partial t}(div\mathbf{E} - \dfrac{\rho}{\epsilon_0}) = 0 \\ \dfrac{\partial(div\ \mathbf{B})}{\partial t} = 0 \end{cases}$$

Si l'équation de continuité (III.2) est vérifiée, les conditions de divergence (IV.1) sont redondantes, et au niveau continu, on considère seulement les deux premières équations du système de Maxwell.

Cependant, au niveau discret, ces conditions ne sont pas toujours préservées au cours du temps par le schéma. Ainsi, ρ et $\mathbf{j}$ ne vérifient pas nécessairement (III.2), ou alors les opérateurs discrets de divergence et rotationnel ne satisfont pas toujours : $div_h(rot_h\ (.)) = 0$.

Il existe diverses méthodes pour pallier à ces problèmes. Lorsque la loi de conservation de la charge n'est pas vérifiée numériquement, la méthode classique consiste à introduire des multiplicateurs de Lagrange associés aux contraintes (IV.1), on dispose alors d'inconnues supplémentaires pour satisfaire (IV.1), voir [1, 22]. Dans [1], on résout un problème du second ordre où le champ électromagnétique est solution de l'équation des ondes. Cependant, cette méthode conduit à résoudre une équation de Laplace à chaque pas de temps.
En l'absence de charges ($\rho = 0$, $\mathbf{j} = 0$), les relations de divergence s'écrivent :

$$\begin{cases} div\,(\mathbf{E}) = 0 \\ div\,(\mathbf{B}) = 0 \end{cases} \tag{IV.2}$$

Les méthodes usuelles pour résoudre les équations de Maxwell sont des méthodes de différences finies, d'éléments finis ou de volumes finis. Certains schémas aux différences finies comme les schémas de Yee [33], préservent les relations (IV.2) mais sont traditionnellement limités au cas de grilles orthogonales. La plupart des problèmes physiques doivent être résolus pour des géométries complexes et l'utilisation de maillages non structurés s'avère nécessaire.
Dans le cas des méthodes de volumes finis, les contraintes (IV.2) sont vérifiées lorsque l'on utilise des maillages duaux [22]. Cependant, la contruction de ces maillages qui utilise des triangulations de Voronoï-Delaunay est encore mal maîtrisée, en particulier au voisinage de la frontière du domaine de calcul en trois dimensions d'espace.
Les méthodes d'éléments finis ne vérifient pas à priori les contraintes (IV.2). Des méthodes de correction de la divergence ont été développées dans [1, 34]. On peut aussi considérer l'équation des ondes qui satisfait naturellement les contraintes [27], ou alors utiliser des éléments finis de type $H(rot)$ ou $H(div)$ [28].

La méthode que nous proposons est une méthode mixte Eléments finis/Volumes finis qui consiste à introduire un terme de diffusion dans les équations de Maxwell, de manière à satisfaire les contraintes de divergence (IV.1). De plus, elle présente l'avantage de garder le caractère de type loi de conservation du système de Maxwell.

IV.2 Nouvelle formulation du système de Maxwell.

IV.2.1 Présentation de la méthode.

Notre approche consiste à introduire un terme de viscosité dans les équations de Maxwell classiques. Soient α et γ deux constantes positives. On considère le nouveau problème:

$$\begin{cases} \dfrac{\partial \mathbf{E}}{\partial t} - c^2 rot\,(\mathbf{B}) - \dfrac{1}{\alpha}\nabla(div\,\mathbf{E} - \dfrac{\rho}{\epsilon_0}) = -\dfrac{1}{\epsilon_0}\mathbf{j} \qquad (\mathbf{x} \in I\!R^3, t > 0) \\ \\ \dfrac{\partial \mathbf{B}}{\partial t} + rot\,(\mathbf{E}) - \dfrac{1}{\gamma}\nabla(div\,\mathbf{B}) = 0 \end{cases} \tag{IV.3}$$

avec comme conditions initiales $\mathbf{E}(0,\mathbf{x}) = \mathbf{E}_0(\mathbf{x})$ et $\mathbf{B}(0,\mathbf{x}) = \mathbf{B}_0(\mathbf{x})$ qui vérifient comme pour le système de Maxwell classique :

$$\begin{cases} div(\mathbf{E}_0) = \dfrac{\rho(0,\mathbf{x})}{\epsilon_0} \\ div(\mathbf{B}_0) = 0 \end{cases} \tag{IV.4}$$

Le système IV.3 est complété par des conditions aux limites. On utilise en général deux sortes de conditions aux limites sur la frontière $\Gamma = \delta\Omega = \Gamma_b \cap \Gamma_\infty$: une condition de conducteur parfait sur Γ_b et des conditions aux limites absorbantes sur Γ_∞.

Remarque IV.2.1 *Les paramètres α et γ ont pour dimension des s/m^2 qui est la dimension inverse d'un coefficient de viscosité. Lorsque α et γ tendent vers $+\infty$, on retrouve le système de Maxwell classique.*

Concernant les équations (IV.3), nous avons les résultats suivants:

Proposition IV.2.1 *Les formulations (III.1) et (IV.3) du système de Maxwell sont équivalentes si:*

- *les données initiales $\mathbf{E}_0$, $\mathbf{B}_0$ vérifient (IV.4).*
- *les densités de charge et de courant ρ et $\mathbf{j}$ vérifient l'équation de conservation de la charge (III.2).*

Proposition IV.2.2 *La nouvelle formulation (IV.3) des équations de Maxwell préserve les estimations d'energie.*

La preuve de ces deux propositions peut être trouvée dans [24, 23] nous rappelons succintement celle de la proposition (IV.2.1).
De manière évidente, une solution des équations de Maxwell classiques est aussi solution de (IV.3). Réciproquement, en prenant la divergence des deux équations du nouveau système, et en tenant compte de l'équation de conservation de la charge (III.2), on obtient :

$$\left\{ \begin{array}{ll} \dfrac{\partial(div\,\mathbf{E} - \dfrac{\rho}{\epsilon_0})}{\partial t} - \dfrac{1}{\alpha}\Delta(div\mathbf{E} - \dfrac{\rho}{\epsilon_0}) = 0 & \qquad (\mathbf{x} \in I\!\!R^3, t > 0) \\ \\ \dfrac{\partial(div\,\mathbf{B})}{\partial t} - \dfrac{1}{\gamma}\Delta(div\,\mathbf{B}) = 0 & \end{array} \right.$$

Ces deux équations de la chaleur pour $(div\,\mathbf{E} - \dfrac{\rho}{\epsilon_0})$ et $div\,\mathbf{B}$ impliquent (IV.1) si les conditions intiales vérifient (IV.4). Une solution de (IV.3) est donc aussi solution du système de Maxwell classique.

IV.2.2 Adimensionnement.

Pour résoudre le nouveau système (IV.3), la méthode pratique consiste à adimensionner les équations, ce qui permet de mettre en évidence les différentes échelles de grandeur. On introduit les grandeurs de référence suivantes :

- $t_\star$: temps en secondes s.
- $x_\star$: longueur en mètres m.
- $B_\star$: induction magnétique en Tesla T.
- $E_\star$: champ électrique en volt par mètre V/m.
- $\rho_\star$: densité de charge en coulomb par mètre cube C/m^3.
- $j_\star$: densité de courant en ampère par mètre carré A/m^2.

On obtient les quantités adimensionnées $t = \dfrac{\bar{t}}{t_\star}$,... que l'on reporte dans (IV.3). On obtient ainsi le système:

$$\begin{cases} \dfrac{\partial \bar{\mathbf{E}}}{\partial \bar{t}} - c^2 \dfrac{t_\star B_\star}{x_\star E_\star} rot\,(\bar{\mathbf{B}}) - \dfrac{t_\star}{\alpha x_\star^2} \nabla (div\,\bar{\mathbf{E}} - \dfrac{x_\star \rho_\star}{\epsilon_0 E_\star} \bar{\rho}) = -\dfrac{t_\star j_\star}{\epsilon_0 E_\star} \bar{\mathbf{j}} \\[2ex] \dfrac{\partial \bar{\mathbf{B}}}{\partial \bar{t}} + \dfrac{t_\star E_\star}{x_\star B_\star} rot\,(\bar{\mathbf{E}}) - \dfrac{t_\star}{\gamma x_\star^2} \nabla (div\,\bar{\mathbf{B}}) = 0 \end{cases}$$

Si on choisit $x_\star = c\,t_\star$ et $E_\star = c\,B_\star$, il vient:

$$\begin{cases} \dfrac{\partial \bar{\mathbf{E}}}{\partial \bar{t}} - rot\,(\bar{\mathbf{B}}) - \dfrac{t_\star}{\alpha x_\star^2} \nabla (div\,\bar{\mathbf{E}} - \dfrac{x_\star \rho_\star}{\epsilon_0 E_\star} \bar{\rho}) = -\dfrac{t_\star j_\star}{\epsilon_0 E_\star} \bar{\mathbf{j}} \\[2ex] \dfrac{\partial \bar{\mathbf{B}}}{\partial \bar{t}} + rot\,(\bar{\mathbf{E}}) - \dfrac{t_\star}{\gamma x_\star^2} \nabla (div\,\bar{\mathbf{B}}) = 0 \end{cases}$$

Les densités de charge et de courant sont déduites de l'équation de conservation de la charge (III.2) qui s'écrit:

$$\frac{\partial \bar{\rho}}{\partial \bar{t}} + \frac{t_\star j_\star}{\rho_\star x_\star} div \bar{j} = 0$$

On choisit $j_\star$ tel que $\dfrac{t_\star j_\star}{\rho_\star x_\star} = 1$, d'où $j_\star = c\,\rho_\star$. On détermine maintenant $\rho_\star$ et $j_\star$ tels que:

$$\frac{\rho_\star x_\star}{\epsilon_0 E_\star} = 1, \quad \text{donc} \quad \frac{j_\star t_\star}{\epsilon_0 E_\star} = 1$$

Finalement, en posant $\bar{\alpha} = \dfrac{\alpha x_\star^2}{t_\star}$ et $\bar{\gamma} = \dfrac{\gamma x_\star^2}{t_\star}$, la nouvelle formulation des équations de Maxwell s'écrit sous forme adimensionnée (on supprime les notations -) :

$$\begin{cases} \dfrac{\partial \mathbf{E}}{\partial t} - rot(\mathbf{B}) - \dfrac{1}{\alpha}\nabla(div\mathbf{E} - \rho) = -\mathbf{j} \\[2ex] \dfrac{\partial \mathbf{B}}{\partial t} + rot(\mathbf{E}) - \dfrac{1}{\gamma}\nabla(div\mathbf{B}) = 0 \end{cases} \tag{IV.5}$$

IV.2.3 Formulation faible.

Les équations (IV.5) peuvent s'écrire sous la forme:

$$\mathbf{Q}_t + \mathbf{F_1}(\mathbf{Q})_x + \mathbf{F_2}(\mathbf{Q})_y + \mathbf{F_3}(\mathbf{Q})_z = \mathbf{J} + \mathbf{G_1}(\mathbf{Q})_x + \mathbf{G_2}(\mathbf{Q})_y + \mathbf{G_3}(\mathbf{Q})_z \tag{IV.6}$$

où :

$$\begin{cases} \mathbf{Q} = {}^t(E_1, E_2, E_3, B_1, B_2, B_3) \\ \mathbf{F_1}(\mathbf{Q}) = {}^t(0, B_3, -B_2, 0, -E_3, E_2) \\ \mathbf{F_2}(\mathbf{Q}) = {}^t(-B_3, 0, B_1, E_3, 0, -E_1) \\ \mathbf{F_3}(\mathbf{Q}) = {}^t(B_2, -B_1, 0, -E_2, E_1, 0) \\ \mathbf{J} = -{}^t(j_1, j_2, j_3, 0, 0, 0) \\ \mathbf{G_1}(\mathbf{Q}) = {}^t(\dfrac{1}{\alpha}(div\mathbf{E} - \rho), 0, 0, \dfrac{1}{\gamma}div\mathbf{B}, 0, 0) \\ \mathbf{G_2}(\mathbf{Q}) = {}^t(0, \dfrac{1}{\alpha}(div\mathbf{E} - \rho), 0, 0, \dfrac{1}{\gamma}div\mathbf{B}, 0) \\ \mathbf{G_3}(\mathbf{Q}) = {}^t(0, 0, \dfrac{1}{\alpha}(div\mathbf{E} - \rho), 0, 0, \dfrac{1}{\gamma}div\mathbf{B}) \end{cases}$$

En utilisant les notations introduites précédemment, la formulation faible de (IV.6) s'écrit:

$$\begin{aligned} \int_{S_i} (\mathbf{Q}_t)\psi_i \, d\mathbf{x} + \int_{S_i} (\mathbf{F_1}(\mathbf{Q})_x + \mathbf{F_2}(\mathbf{Q})_y + \mathbf{F_3}(\mathbf{Q})_z)\psi_i \, d\mathbf{x} &= \int_{S_i} \mathbf{J}\psi_i \, d\mathbf{x} \\ &+ \int_{S_i} (\mathbf{G_1}(\mathbf{Q})_x + \mathbf{G_2}(\mathbf{Q})_y + \mathbf{G_3}(\mathbf{Q})_z)\psi_i \, d\mathbf{x} \end{aligned} \tag{IV.7}$$

où S_i désigne le support de la fonction test ψ_i. La formulation mixte volumes finis-éléments finis apparaît ici avec le choix des fonctions tests ψ_i suivant les flux intégrés. Pour les flux convectifs et l'intégrale de courant nous utilisons comme fonction test la fonction caractéristique de la cellule C_i; pour les flux diffusifs, nous prenons la fonction de base φ_i, associée au noeud i. On obtient ainsi:

$$Vol(C_i)\,(\mathbf{Q}_t)_i + \int_{C_i} \vec{\nabla}.\mathbb{F}(\mathbf{Q})d\mathbf{x} = \int_{C_i} \mathbf{J}d\mathbf{x} + \int_{Supp(\varphi_i)} (\mathbf{G_1}(\mathbf{Q})_x + \mathbf{G_2}(\mathbf{Q})_y + \mathbf{G_3}(\mathbf{Q})_z)\varphi_i \, d\mathbf{x} \tag{IV.8}$$

En appliquant à (IV.8) la formule de Green, il vient:

$$Vol(C_i)\,(\mathbf{Q}_t)_i + \sum_{j\in K(i)} \int_{\partial C_{ij}} \mathbb{F}(\mathbf{Q}).\boldsymbol{\nu_{ij}}\,d\sigma + \int_{\partial C_i\cap(\Gamma_b\cup\Gamma_\infty)} \mathbb{F}(\mathbf{Q}).\boldsymbol{n}\,d\sigma = \int_{C_i} \mathbf{J}\,d\mathbf{x}$$
$$- \int_{Supp(\varphi_i)} (\mathbf{G_1}(\mathbf{Q})\frac{\partial \varphi_i}{\partial x} + \mathbf{G_2}(\mathbf{Q})\frac{\partial \varphi_i}{\partial y} + \mathbf{G_3}(\mathbf{Q})\frac{\partial \varphi_i}{\partial z})d\mathbf{x}$$
$$+ \int_{\partial C_i\cap(\Gamma_b\cup\Gamma_\infty)} (\mathbf{G_1}(\mathbf{Q}) + \mathbf{G_2}(\mathbf{Q}) + \mathbf{G_3}(\mathbf{Q})).\mathbf{n}d\sigma \tag{IV.9}$$

où $\boldsymbol{\nu_{ij}}$ est la normale extérieure de l'interface ∂C_{ij} entre deux cellules C_i et C_j et $K(i)$ l'ensemble des noeuds voisins d'un sommet i. L'approximation des termes de dérivée temporelle, de densité de courant et de convection est identique à celle du système de Maxwell classique.

Le terme de viscosité dans le membre de droite est évalué d'après la méthode des éléments finis et s'écrit pour un maillage en tétraèdres:

$$\sum_{T,S_i\in T} Vol(T)(\mathbf{G_1}(\mathbf{Q})\mid_T \frac{\partial \varphi_i^T}{\partial x} + \mathbf{G_2}(\mathbf{Q})\mid_T \frac{\partial \varphi_i^T}{\partial y} + \mathbf{G_3}(\mathbf{Q})\mid_T \frac{\partial \varphi_i^T}{\partial z})$$

où $\mathbf{G_1}(\mathbf{Q})\mid_T$ est la valeur de $\mathbf{G_1}(\mathbf{Q})$ sur T (élément structuré ou non structuré). Plus précisément:

$$\mathbf{G_1}(\mathbf{Q})\mid_T = {}^t(\frac{1}{\gamma}(div\mathbf{B})\mid_T, 0, 0, \frac{1}{\alpha}((div\mathbf{E})\mid_T - \rho\mid_T), 0, 0)$$

avec:

$$(div\mathbf{E})\mid_T = \sum_{k=1}^{nt}(E_1^k\frac{\partial \phi_k^T}{\partial x} + E_2^k\frac{\partial \phi_k^T}{\partial y} + E_3^k\frac{\partial \phi_k^T}{\partial z}) \qquad , \qquad \rho\mid_T = \frac{1}{nt}\sum_{k=1}^{nt}\rho_k$$

où nt représente le nombre de sommets de l'élément T.

Dans le cas bidimensionnel, qui sera notre cadre d'étude par la suite, les équations de Maxwell peuvent être découplées en deux sous-systèmes associés aux ondes transverse électrique **T.E** et transverse magnétique **T.M**. Ces deux types d'ondes vérifient respectivement $\mathbf{E}.\mathbf{e_z} = 0$ et $\mathbf{B}.\mathbf{e_z} = 0$. Dans le cas d'une onde **T.E**, le système (IV.6) s'écrit donc:

$$\mathbf{Q}_t + \mathbf{F_1}(\mathbf{Q})_x + \mathbf{F_2}(\mathbf{Q})_y = \mathbf{J} + \mathbf{G_1}(\mathbf{Q})_x + \mathbf{G_2}(\mathbf{Q})_y \tag{IV.10}$$

avec :

$$\begin{cases} \mathbf{Q} = {}^t(E_1, E_2, B_3) \\ \mathbf{F_1}(\mathbf{Q}) = {}^t(0, B_3, E_2) \\ \mathbf{F_2}(\mathbf{Q}) = {}^t(-B_3, 0, -E_1) \\ \mathbf{J} = -{}^t(j_1, j_2, 0) \\ \mathbf{G_1}(\mathbf{Q}) = {}^t(\frac{1}{\alpha}(div\mathbf{E} - \rho), 0, 0) \\ \mathbf{G_2}(\mathbf{Q}) = {}^t(0, \frac{1}{\alpha}(div\mathbf{E} - \rho), 0) \end{cases}$$

IV.2.4 Traitement des conditions aux limites.

Dans de nombreux problèmes électromagnétiques, comme les problèmes de diffraction, on distingue la frontière de l'obstacle Γ_b et la frontière artificielle Γ_∞ délimitant le domaine de calcul. Dans le cas d'objets métalliques parfaitement conducteurs, on impose sur Γ_b la condition: $\mathbf{Ex}\nu = 0$. Pour cela, nous précisons l'approximation du terme:

$$\int_{\partial C_i \cap \Gamma_b} \mathbb{F}(\mathbf{Q}).\boldsymbol{n}\, d\sigma = \int_{\partial C_i \cap \Gamma_b} (\mathbf{F_1} n_{ix} + \mathbf{F_2} n_{iy} + \mathbf{F_3} n_{iz})\, d\sigma.$$

Ce dernier terme s'écrit en utilisant la définition des $\mathbf{F_i}$, puis la relation $\mathbf{Ex}\nu = 0$:

$$\int_{\partial C_i \cap \Gamma_b} \begin{pmatrix} n_{iy}E_3 - n_{iz}E_2 \\ -n_{ix}E_3 + n_{iz}E_1 \\ n_{ix}E_2 - n_{iy}E_1 \\ -n_{iy}B_3 + n_{iz}B_2 \\ n_{ix}B_3 - n_{iz}B_1 \\ -n_{ix}B_2 + n_{iz}B_1 \end{pmatrix} d\sigma = \int_{\partial C_i \cap \Gamma_b} \begin{pmatrix} 0 \\ 0 \\ 0 \\ -n_{iy}B_3 + n_{iz}B_2 \\ n_{ix}B_3 - n_{iz}B_1 \\ -n_{ix}B_2 + n_{iz}B_1 \end{pmatrix} d\sigma$$

On l'évalue de la manière suivante:

$$\int_{\partial C_i \cap \Gamma_b} \begin{pmatrix} 0 \\ 0 \\ 0 \\ -n_{iy}B_3 + n_{iz}B_2 \\ n_{ix}B_3 - n_{iz}B_1 \\ -n_{ix}B_2 + n_{iz}B_1 \end{pmatrix} d\sigma \simeq B_{i1} \int_{\partial C_i \cap \Gamma_b} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -n_{iz} \\ n_{iy} \end{pmatrix} d\sigma + B_{i2} \int_{\partial C_i \cap \Gamma_b} \begin{pmatrix} 0 \\ 0 \\ 0 \\ n_{iz} \\ 0 \\ -n_{ix} \end{pmatrix} d\sigma$$

$$+ B_{i3} \int_{\partial C_i \cap \Gamma_b} \begin{pmatrix} 0 \\ 0 \\ 0 \\ -n_{iy} \\ n_{ix} \\ 0 \end{pmatrix} d\sigma$$

où B_{ik} (k=1,2,3) représente la kième composante du champ $\mathbf{B}$ au sommet i. La condition $\mathbf{Ex}\nu = 0$ est prise en compte faiblement.

De la même façon, nous avons à évaluer l'intégrale suivante:

$$\int_{\partial C_i \cap \Gamma_b} (\mathbf{G_1} n_{ix} + \mathbf{G_2} n_{iy} + \mathbf{G_3} n_{iz})\, d\sigma \qquad \text{(IV.11)}$$

De par la définition de $\mathbf{G_1}$, $\mathbf{G_2}$, $\mathbf{G_3}$, cela revient à calculer les quantités suivantes:

$$\int_{\partial C_i \cap \Gamma_b} div\mathbf{B}\, \mathbf{n}\, d\sigma, \qquad \int_{\partial C_i \cap \Gamma_b} (div\mathbf{E} - \rho)\mathbf{n}\, d\sigma$$

Or nous imposons sur la frontière Γ_b les conditions suivantes : $div\mathbf{E} - \dfrac{\rho}{\epsilon_0} = 0$ et $div\mathbf{B} = 0$, ce qui met à zéro la contribution (IV.11).

En ce qui concerne la frontière Γ_∞, on utilise en général un décentrage d'ordre un, qui permet de garder le même type d'approximation que pour les points internes. Pour plus de précisions sur la prise en compte des conditions aux limites, on pourra se référer à [7, 9].

IV.2.5 Intégration en temps.

Comme pour le système de Maxwell classique, nous utilisons un schéma de Runge-Kutta explicite pour l'avancement en temps de la solution numérique.

IV.3 Etude de stabilité des schémas.

Avec la nouvelle formulation (IV.3), nous souhaitons mieux vérifier numériquement, en l'absence de termes source, l'équation :

$$div\,\mathbf{Q} = 0 \tag{IV.12}$$

Le but de notre étude est de déterminer une stratégie pour le choix du paramètre α. Si on prend la divergence de (IV.10), on obtient une équation de la chaleur pour $div\,\mathbf{Q}$:

$$\frac{(div\,\mathbf{Q})}{\partial t} - \frac{1}{\alpha}\Delta(div\,\mathbf{Q}) = 0 \tag{IV.13}$$

On voit que $div\,\mathbf{Q} \longrightarrow 0$ quand $t \longrightarrow +\infty$ d'autant plus vite que le paramètre α est petit. Cependant, l'utilisation d'un α trop petit entraînerait une contrainte très forte sur le pas de temps, ce que nous voulons éviter. On sait, en effet, que le terme de viscosité a une grande influence sur la stabilité du schéma. En particulier, lorsque le coefficient de diffusion $(1/\alpha)$ augmente, le domaine de stabilité devient très réduit [20].

Nous introduisons la définition suivante qui précise le choix du paramètre α :

Définition IV.3.1 *On appellera α_{opt} la plus petite valeur du paramètre α, pour laquelle la condition de stabilité est la même que pour le système de Maxwell classique. Pour une telle valeur de α, on pourra donc utiliser le même pas de temps Δt que pour la résolution des équations de Maxwell classiques.*

Dans la suite on effectue l'étude de stabilité linéaire des schémas volumes finis présentés précédemment pour des maillages réguliers (de pas Δx,Δy), en rectangles et en triangles. Comme pour l'étude du système de Maxwell classique (chapitre III), on considère tout d'abord les schémas d'ordre un en temps et en espace, puis les schémas d'ordre supérieur pour lesquels on étudiera l'influence du paramètre de décentrage β sur la stabilité. Cette étude repose sur une analyse de Fourier analogue à celle effectuée dans le chapitre III pour le système de Maxwell.

IV.3.1 Schémas précis à l'ordre un.

– Cas d'un maillage rectangulaire.

On représente sur la figure IV.1 les domaines de stabilité en fonction de Δt et de α pour différentes valeurs de $h = \Delta x = \Delta y$. La valeur optimale α_{opt} est celle à partir de laquelle la valeur de Δt est constante.

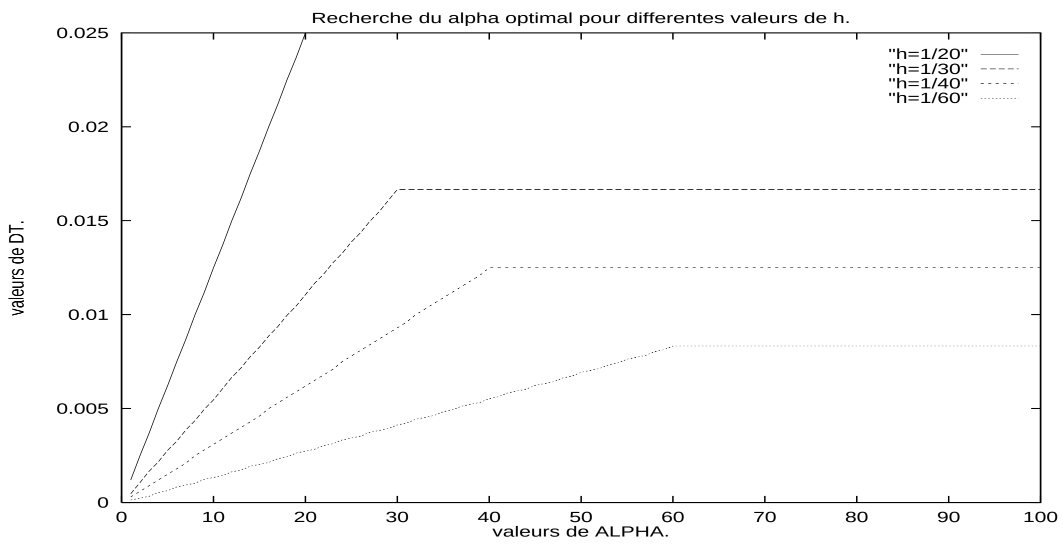


FIG. IV.1 – *Maillage en rectangles.*

On voit sur ces courbes qu'il existe une valeur de α optimale α_{opt} au sens défini précédemment. Ainsi pour des valeurs de $\alpha \geq \alpha_{opt}$ le terme de viscosité n'impose pas de contraintes sur le pas de temps puisqu'on peut utiliser le même Δt que dans le cas du système de Maxwell classique ($\alpha \longrightarrow +\infty$).
On remarque que lorsque h devient grand, le pas de temps Δt augmente et le α_{opt} diminue: autrement dit pour des pas d'espace de plus en plus grands, on donne une importance de plus en plus grande au terme de viscosité.
On résume dans le tableau IV.1 les valeurs α_{opt} obtenues pour différentes valeurs de h.

On a vu qu'à la valeur $\alpha = \alpha_{opt}$ correspond un pas de temps maximal, il s'agit du pas de temps vérifiant la condition: $\alpha_1 + \alpha_2 = 1$ (avec les notations du chapitre III: $\alpha_1 = \dfrac{\Delta t}{\Delta x}$

h	Δt_{max}	α_{opt}
1/20	$2.5\ 10^{-2}$	20
1/30	$1.66\ 10^{-2}$	30
1/40	$1.25\ 10^{-2}$	40
1/60	$8.33\ 10^{-3}$	60
1/250	$2\ 10^{-3}$	250

TAB. IV.1 – *Valeurs de α_{opt} en fonction de h.*

h	Δt_{max}	α_{opt}
1/15	$4.4\ 10^{-2}$	45
1/20	$3.3\ 10^{-2}$	60
1/30	$2.2\ 10^{-2}$	90
1/40	$1.6\ 10^{-2}$	120
1/60	$9\ 10^{-3}$	145

TAB. IV.2 – *Valeurs de α_{opt} en fonction de h.*

et $\alpha_2 = \dfrac{\Delta t}{\Delta y}$) , ce qui peut encore s'écrire, pour $\Delta x = \Delta y = h$:

$$\Delta t_{max} = \frac{h}{2}.$$

On observe numériquement (voir le tableau IV.1) que le coefficient de diffusion ($1/\alpha_{opt}$) suit une loi linéaire en h.

$$\boxed{\textit{pour } h \leq \frac{1}{20},\ \alpha_{opt} = \frac{1}{h} = \frac{1}{2\Delta t_{max}}}$$

– Cas d'un maillage triangulaire.

On représente sur la figure IV.2 les domaines de stabilité en fonction de Δt et de α pour différentes valeurs de $h = \Delta x = \Delta y$.

On remarque sur la figure IV.2 que pour h fixé, le pas de temps maximal obtenu est plus grand que celui obtenu dans le cas rectangulaire. Cela s'explique par le fait que la limite de stabilité du schéma en triangles est nettement supérieure à celle obtenue dans le cas rectangulaire, dans le cas Δx=Δy=h. D'autre part, à h fixé ou à Δt fixé, on remarque que la valeur de α_{opt} est plus grande que pour le schéma en rectangles.

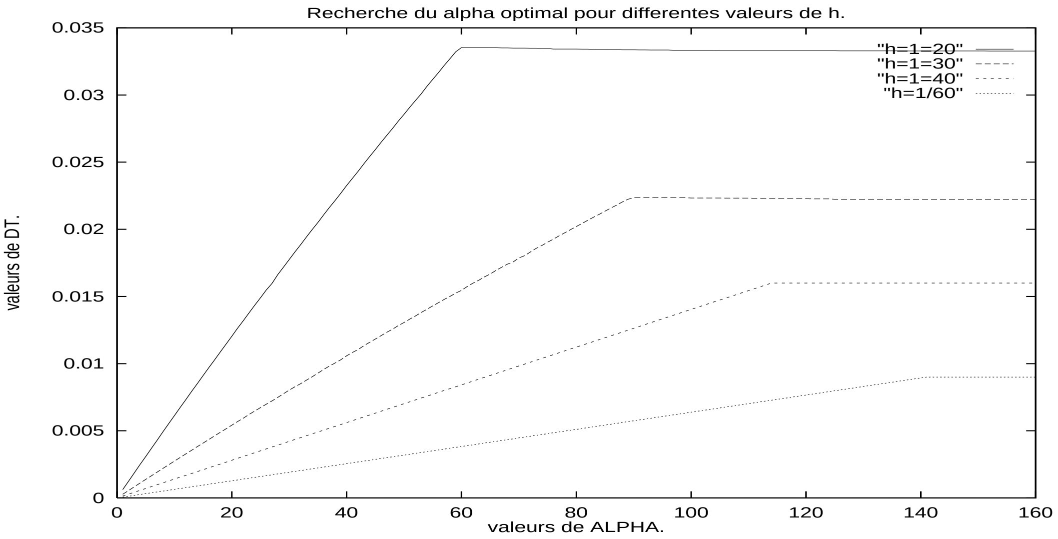


FIG. IV.2 – *Maillage en triangles.*

Le tableau IV.2 nous montre que, de même que pour un maillage rectangulaire, le coefficient de diffusion $(1/\alpha_{opt})$ suit une loi linéaire en h.

$$\boxed{pour\ h \leq \frac{1}{15},\ \alpha_{opt} = \frac{3}{h}}$$

IV.3.2 Schémas précis d'ordre supérieur.

Ils sont obtenus à l'aide de β-schémas utilisant une approximation en temps de type Runge-Kutta à trois pas. L'étude de stabilité se fait comme dans le cas du système de Maxwell classique (voir section III.4.2).

De même que précédemment, on cherche à déterminer le paramètre α_{opt} nous permettant d'utiliser le pas de temps le plus grand possible, dans le cas d'un schéma volumes finis d'ordre deux au moins en espace et d'ordre trois en temps.

– Cas d'un maillage rectangulaire.

On fixe $\Delta x = \Delta y = h = \dfrac{1}{20}$. On représente sur la figure IV.3 les domaines de stabilité en fonction de Δt et α pour différentes valeurs du paramètre β.

On voit sur la figure IV.3 que plus β diminue (ie plus on se rapproche d'un schéma totalement centré), plus le pas de temps optimal augmente; ce qui est en accord avec les limites de stabilité obtenues pour le système de Maxwell classique (III.1). De même, plus β diminue, plus le paramètre α_{opt} augmente, c'est à dire que pour un schéma centré

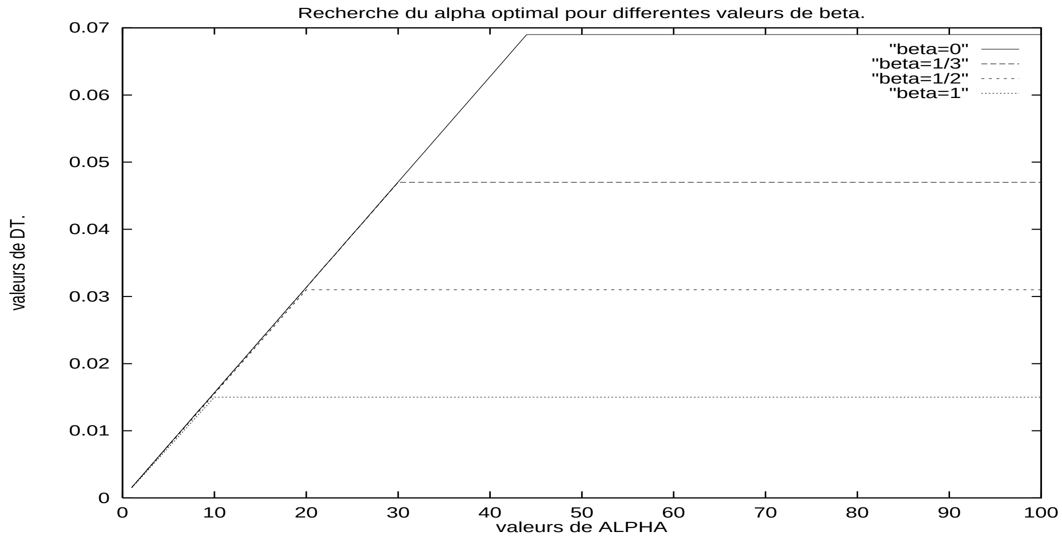


FIG. IV.3 – *Maillage en rectangles.*

β	Δt_{max}	α_{opt}
0	$6.9\ 10^{-2}$	45
$1/3$	$4.7\ 10^{-2}$	30
$1/2$	$3.1\ 10^{-2}$	20
1	$1.5\ 10^{-2}$	10

TAB. IV.3 – *Valeurs de α_{opt} en fonction de β.*

l'influence du terme de diffusion est moins forte que dans le cas d'un schéma décentré. On note que le pas de temps maximal varie linéairement avec α_{opt} quand on fait varier β.

On résume dans le tableau IV.3 les valeurs de α_{opt} pour différentes valeurs de β.

On considère maintenant le cas $\beta = \dfrac{1}{3}$: on a alors un schéma exactement d'ordre trois en espace et en temps (en l'absence du terme de diffusion). On représente sur la figure IV.4 les domaines de stabilité en fonction de Δt et de α pour différentes valeurs de $h = \Delta x = \Delta y$.

On remarque que les courbes varient de la même manière en fonction de h à l'ordre trois et à l'ordre un: à Δt fixé, plus h augmente, plus α_{opt} diminue c'est à dire plus grande est l'importance donnée au terme de diffusion.

Les valeurs du tableau IV.4 montrent que le coefficient de diffusion $(1/\alpha_{opt})$ suit une

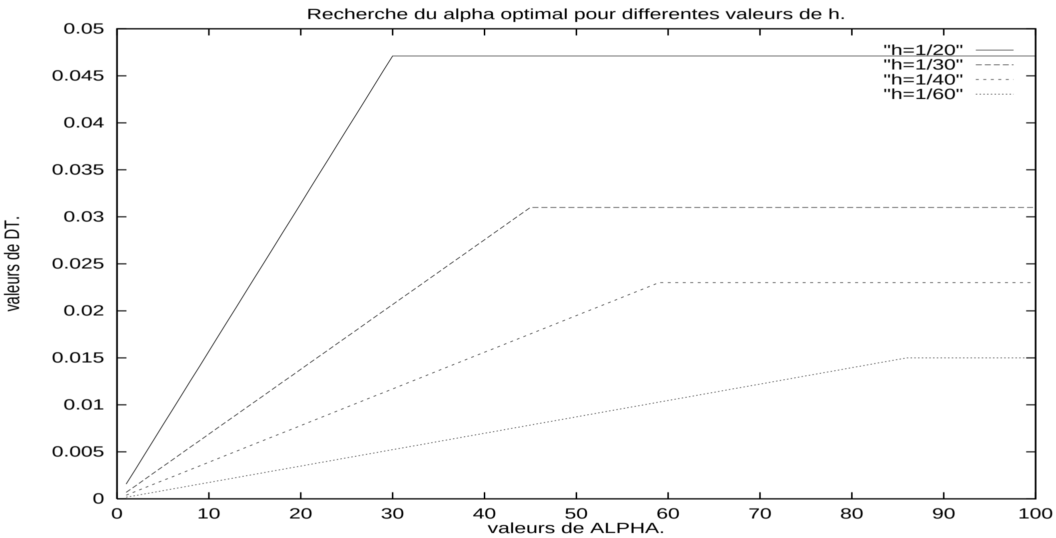


FIG. IV.4 – *Maillage en rectangles.*

h	Δt_{max}	α_{opt}
$1/20$	$4.7\ 10^{-2}$	30
$1/30$	$3.1\ 10^{-2}$	45
$1/40$	$2.3\ 10^{-2}$	60
$1/50$	$1.8\ 10^{-2}$	75
$1/60$	$1.5\ 10^{-2}$	90

TAB. IV.4 – *Valeurs de α_{opt} en fonction de h.*

loi linéaire en h.

$$\boxed{\textit{pour } h \leq \frac{1}{20},\ \alpha_{opt} = \frac{3}{2h}}$$

– Cas d'un maillage triangulaire.

De même que dans le cas rectangulaire, nous fixons le pas d'espace $h = \dfrac{1}{20}$ et on représente sur la figure IV.5 les domaines de stabilité en fonction de Δt et α pour différentes valeurs du paramètre β.

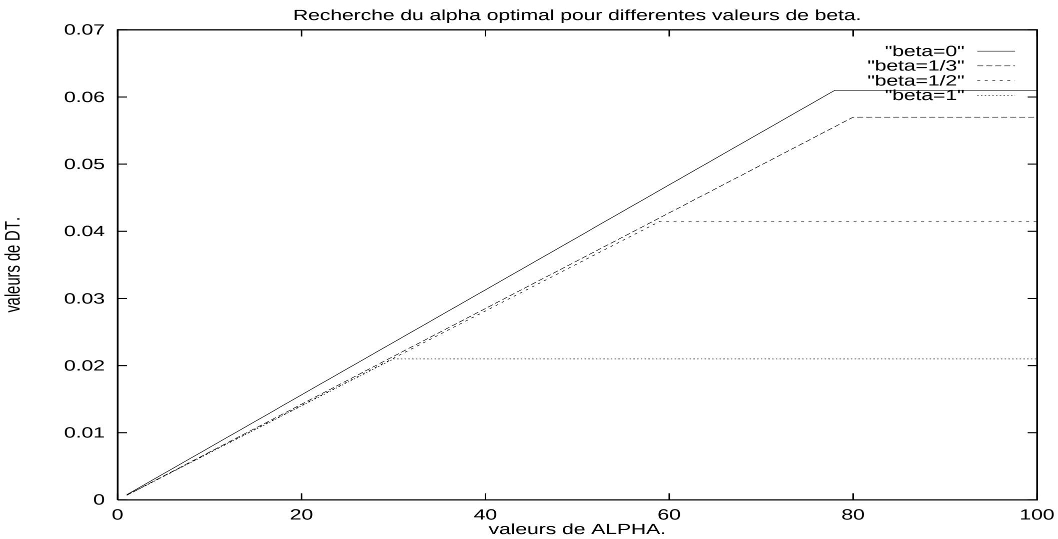


FIG. IV.5 – *Maillage en triangles.*

Les courbes varient de la même manière pour un maillage rectangulaire ou triangulaire. A β fixé, on remarque que le pas de temps maximal est plus grand dans le cas d'un maillage triangulaire que pour un maillage en rectangles, sauf dans le cas d'un schéma centré (β=0). De plus, à β fixé, et pour un même pas de temps Δt, on voit que la valeur de α_{opt} est plus petite en rectangles qu'en triangles (d'où une plus grande influence donnée au terme de diffusion pour un maillage en rectangles).
Contrairement au schéma d'ordre trois en maillage rectangulaire, ici Δt ne dépend pas linéairement de α_{opt} quand on fait varier β.

On considère à nouveau le cas $\beta = \dfrac{1}{3}$. On représente sur la figure IV.6 les domaines de stabilité en fonction de Δt et de α pour différentes valeurs de $h = \Delta x = \Delta y$.

A h fixé, le pas de temps optimal est plus grand pour un maillage triangulaire que pour un maillage rectangulaire, ce quelle que soit la valeur de h. Toutefois, utiliser un

β	Δt_{max}	α_{opt}
0	$6.1\ 10^{-2}$	78
$1/3$	$5.7\ 10^{-2}$	80
$1/2$	$4.1\ 10^{-2}$	58
1	$2.1\ 10^{-2}$	30

TAB. IV.5 – *Valeurs de α_{opt} en fonction de β.*

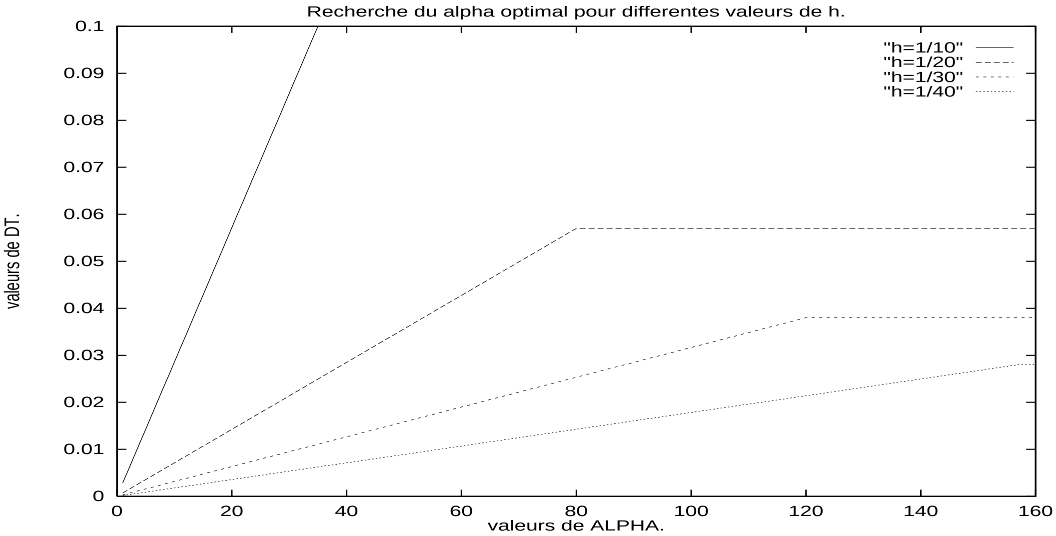


FIG. IV.6 – *Maillage en triangles.*

h	Δt_{max}	α_{opt}
$1/20$	$5.7\ 10^{-2}$	80
$1/30$	$3.8\ 10^{-2}$	120
$1/40$	$2.8\ 10^{-2}$	160
$1/50$	$2.2\ 10^{-2}$	200

TAB. IV.6 – *Valeurs de α_{opt} en fonction de h.*

schéma précis en temps s'avère plus intéressant en rectangles. En effet,

$$\frac{\Delta t_{RK3}}{\Delta t_{RK1}} = 1.8 \text{ en rectangles}$$

$$\frac{\Delta t_{RK3}}{\Delta t_{RK1}} = 1.7 \text{ en triangles}$$

On remarque également que le coefficient de diffusion $(1/\alpha_{opt})$ suit une loi linéaire en h:

$$\boxed{pour\ h \leq \frac{1}{10},\ \alpha_{opt} = \frac{4}{h}}$$

Cette étude de stabilité nous montre que les schémas en triangles permettent, à $\Delta x{=}\Delta y{=}h$ fixé, un choix de pas de temps plus grand qu'en rectangles. Elle nous a permis de déterminer une valeur optimale de α (α_{opt}) en fonction des schémas et du maillage utilisés. On a vu également, que, quel que soit l'ordre en temps et l'approximation spatiale choisis, le coefficient de diffusion $(1/\alpha_{opt})$ suit une loi linéaire en h. La figure IV.7 nous montre l'influence de ce paramètre sur les différents schémas en fonction de h. On voit que la valeur de α_{opt} est plus petite pour les schémas en rectangles que pour les schémas en triangles. Ceci montre que les schémas en rectangles permettent de donner plus d'importance au terme de viscosité et par conséquent de mieux vérifier numériquement IV.1.

De même la valeur de α_{opt} devient plus grande quand on augmente la précision du schéma, aussi bien en triangles qu'en rectangles.

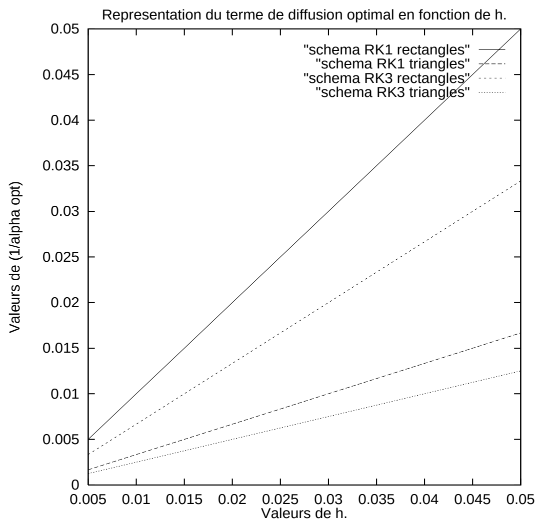


FIG. IV.7 – *Valeurs de α_{opt} en fonction de h.*

IV.4 Equations équivalentes.

Nous souhaitons mettre en évidence les termes d'erreur de troncature de nos schémas, en particulier les erreurs de dispersion et de dissipation. Pour cela, nous cherchons à établir les équations équivalentes des schémas vus précédemment. Cette technique, introduite par Warming and Hyett [32] donne l'équation effective résolue par le schéma. Comme dans le cas de l'équation d'advection (voir section II.3), nous utilisons la méthode décrite dans [4]; celle-ci s'étend facilement au cas d'un système linéaire et nous l'appliquons au système de Maxwell classique (III.1) et au nouveau système (IV.3). Nous nous intéressons plus particulièrement à l'équation que vérifie $div\mathbf{Q}$ pour justifier l'introduction de la nouvelle formulation IV.3.
On rappelle que dans le cas bidimensionnel, pour une onde **TE** en l'absence de charges, la formulation IV.3 du système de Maxwell s'écrit :

$$\left\{\begin{array}{l}
\dfrac{\partial E_1}{\partial t} - \dfrac{\partial B_3}{\partial y} - \dfrac{1}{\alpha}\dfrac{\partial(div\mathbf{E})}{\partial x} = 0 \\
\dfrac{\partial E_2}{\partial t} + \dfrac{\partial B_3}{\partial x} - \dfrac{1}{\alpha}\dfrac{\partial(div\mathbf{E})}{\partial y} = 0 \\
\dfrac{\partial B_3}{\partial t} + \dfrac{\partial E_2}{\partial x} - \dfrac{\partial E_1}{\partial y} = 0 \\
div\mathbf{E} = \dfrac{\partial E_1}{\partial x} + \dfrac{\partial E_2}{\partial y} = 0
\end{array}\right. \tag{IV.14}$$

IV.4.1 Analyse et comparaison des termes d'erreur des schémas d'ordre un.

Par la suite nous noterons $E_x^l = \dfrac{\partial E_l}{\partial x}, \;\; l = 1, 2, 3.$

Maillage rectangulaire.

On donne ici les termes d'erreur pour la première composante E^1, les équations équivalentes des autres composantes s'obtenant par symétrie par-rapport à Δx, Δy.
Pour la valeur $\alpha = \infty$, on obtient l'équation équivalente du système de Maxwell classique :

$$\begin{aligned}E_t^1 + B_y^3 = {} & \frac{1}{2}(\Delta y - \Delta t)\, E_{yy}^1 + (\frac{\Delta t}{2} + \frac{1}{\alpha})\, E_{xy}^2 + \frac{E_{xx}^1}{\alpha} - \frac{\Delta t}{2\alpha^2}(E_{xxxx}^1 + E_{xxyy}^1 + E_{xyyy}^2 + E_{xxxy}^2) \\ & + O(\Delta t, \Delta x, \Delta y)^2\end{aligned}$$

On voit ici clairement que la nouvelle formulation consiste à ajouter des termes de diffusion d'ordre deux $(\dfrac{E_{xx}^1}{\alpha}, \dfrac{E_{xy}^2}{\alpha})$ et d'ordre quatre $(-\dfrac{\Delta t}{2\alpha^2}\,(E_{xxxx}^1 + E_{xyyy}^1 + E_{xyyy}^2 + E_{xxxy}^2))$.
Nous établissons l'équation équivalente sur la divergence que nous obtenons aisément à partir des équations équivalentes du champ **E**.

En effet:

$$\begin{aligned}(div\mathbf{E})_t &= E^1_{tx} + E^2_{ty} \\ &= \frac{\Delta y}{2} E^1_{xyy} + \frac{\Delta x}{2} E^2_{xxy} + \frac{E^2_{yyy}}{\alpha} + \frac{E^1_{xyy}}{\alpha} + \frac{E^1_{xxx}}{\alpha} + \frac{E^2_{xxy}}{\alpha} \\ &-\frac{\Delta t}{2\alpha^2}\,(E^1_{xxxxx} + 2E^1_{xxxyy} + E^1_{xyyyy} + E^2_{yyyyy} + E^2_{xxyyy} + E^2_{xxxxy}) \\ &+O(\Delta t, \Delta x, \Delta y)^2\end{aligned}$$

ce qui peut encore s'écrire:

$$(div\,\mathbf{E})_t - \frac{1}{\alpha}\,\Delta\; div\mathbf{E} = \frac{\Delta y}{2} E^1_{xyy} + \frac{\Delta x}{2} E^2_{xxy} - \frac{\Delta t}{2\alpha^2}\,\Delta^2\; div\mathbf{E} + O(\Delta t, \Delta x, \Delta y)^2 \quad \text{(IV.15)}$$

où $\Delta^2\; div\mathbf{E}$ représente le bilaplacien de $div\mathbf{E}$. Là encore on voit clairement que la nouvelle formulation consiste pour Δx et Δy fixés à rajouter des termes de diffusion sur la divergence du champ électrique $\mathbf{E}$, et donc permet de mieux vérifier numériquement $div\mathbf{Q} = 0$.

Maillage triangulaire.

On proçède de la même manière que précédemment. Pour ne pas alourdir les calculs, on pose Δx=Δy=h.

On obtient comme équation équivalente pour la première composante du champ électrique :

$$\begin{aligned}E^1_t + B^3_y &= (\frac{ah}{6} + \frac{1}{\alpha})\,E^1_{xx} + (\frac{ah}{6} - \frac{\Delta t}{2})\,E^1_{yy} + \frac{h}{3\sqrt{2}}\,E^1_{xy} + \frac{bh}{6}\,(E^2_{xx} + E^2_{yy}) \\ &+(-\frac{h}{3\sqrt{2}} + \frac{\Delta t}{2} + \frac{1}{\alpha})\,E^2_{xy} - \frac{\Delta t}{2\alpha^2}(E^1_{xxxx} + E^1_{xxyy} + E^2_{xyyy} + E^2_{xxxy}) + O(\Delta t, h)^2\end{aligned}$$

où on a posé $a = \dfrac{1}{\sqrt{2}} + \dfrac{1}{\sqrt{5}}$ et $b = \dfrac{2}{\sqrt{5}} - \dfrac{1}{\sqrt{2}}$.
On remarque que les termes en temps sont les mêmes pour le schéma en triangles que pour le schéma en rectangles. Seuls les termes liés à l'approximation spatiale diffèrent: le schéma est beaucoup plus dissipatif dans le cas d'un maillage triangulaire; en effet parmi les termes d'erreur spatiale interviennent tous les termes de dissipation d'ordre deux pour les composantes E^1 et E^2 du champ électrique.
L'équation équivalente sur la divergence du champ électrique $\mathbf{E}$ s'écrit:

$$\begin{aligned}(div\mathbf{E})_t - \frac{1}{\alpha}\,\Delta\; div\mathbf{E} &= \frac{h}{6\sqrt{2}}(E^1_{xxx} + E^1_{xxy} + E^1_{yyy} - E^1_{xyy} - E^2_{xxx} + E^2_{yyy} + E^2_{xyy} - E^2_{xxy}) \\ &+\frac{h}{6\sqrt{5}}(E^1_{xxx} + 2E^1_{xxy} + 2E^1_{yyy} + E^1_{xyy} + 2E^2_{xxx} + E^2_{yyy} + 2E^2_{xyy} + E^2_{xxy}) \\ &-\frac{\Delta t}{2\alpha^2}\,\Delta^2\; div\mathbf{E} + O(\Delta t, h)^2\end{aligned} \quad \text{(IV.16)}$$

Comme pour le maillage rectangulaire, on obtient dans le membre de gauche une équation de la chaleur pour la divergence, avec dans le membre de droite les termes d'erreur d'ordre un en temps et en espace inhérents au schéma, ainsi que le terme $-\dfrac{\Delta t}{2\alpha^2}\,\Delta^2\, div\mathbf{E}$. Les termes d'erreur en espace comprennent toutes les dérivées troisièmes de $\mathbf{E}$: ces termes d'erreur proviennent des termes de diffusion spatiale du schéma.

IV.4.2 Analyse et comparaison des termes d'erreur des β-schémas.

On se propose dans cette partie d'écrire les équations équivalentes dans le cas d'une approximation spatiale au moins d'ordre deux et d'un schéma en temps RK3.

Maillage rectangulaire.

On donne les termes d'erreur du schéma pour la première composante du champ électrique $\mathbf{E}$:

$$\begin{aligned}
E^1_t + B^3_y = & -\frac{\Delta y^2}{6}(1-3\beta)\,B^3_{yyy} - \frac{\beta}{4}\,(E^1_{yyyy}\Delta y^3 + E^1_{xxyy}\Delta x^2\Delta y)\\
& -\frac{\Delta t^3}{24}\,(E^1_{yyyy} + E^1_{xxyy} - E^2_{xyyy} - E^2_{xxxy}) + \frac{E^1_{xx}}{\alpha} + \frac{E^2_{xy}}{\alpha}\\
& +\frac{1}{12\alpha}\,(E^1_{xxxx}\Delta x^2 + 2\,E^1_{xxyy}\Delta y^2 + 2\,E^2_{xxxy}\Delta x^2 + 2\,E^2_{xyyy}\Delta y^2)\\
& -\frac{\Delta t^3}{24\alpha^4}\,(E^1_{xxxxxxxx} + 3\,E^1_{xxxxxxyy} + 3\,E^1_{xxxxyyyy} + E^1_{xxyyyyyy} + E^2_{xxxxxxxy}\\
& +3\,E^2_{xxxxxyyy} + 3\,E^2_{xxxxyyyy} + E^2_{xxxxxxyy}) + O(\Delta t, \Delta x, \Delta y)^4
\end{aligned}$$

Cette équation équivalente nous montre que deux valeurs de β sont intéressantes quand on considère le système de Maxwell classique ($\alpha = \infty$): la valeur $\beta = \dfrac{1}{3}$ annule les termes de dispersion spatiale et permet ainsi d'avoir un schéma d'ordre trois en temps et en espace; avec la valeur β=0 on a toujours un schéma d'ordre deux, mais cette valeur permet d'annuler les termes de dissipation en espace.
En ce qui concerne la nouvelle formulation (IV.3) du système de Maxwell, on remarque que l'on obtient des termes d'erreur en $\dfrac{\Delta x^2}{\alpha}$,$\dfrac{\Delta y^2}{\alpha}$ et également en $\dfrac{\Delta t^3}{24\alpha^4}$, faisant intervenir respectivement les dérivées quatrièmes et huitièmes de $\mathbf{E}$, c'est à dire des termes de dissipation.

Comme précédemment, on recherche les termes d'erreur sur la divergence du champ électrique. Pour plus de simplicité on donne son équation équivalente dans le cas $\Delta x = \Delta y = h$.

$$(div\mathbf{E})_t - \frac{1}{\alpha}\,\Delta\, div\mathbf{E} = -\frac{h^2}{6}(1-3\beta)\,(B^3_{xyyy} - B^3_{xxxy}) - \frac{\beta h^3}{4}\,(E^1_{xyyyy} + E^2_{xxxxy} + E^1_{xxxyy} + E^2_{xxyyy}) + \frac{h^2}{6\alpha}\,\Delta^2 div\mathbf{E} - \frac{h^2}{12\alpha}\,(E^1_{xxxxx} + E^2_{yyyyy}) - \frac{\Delta t^3}{24\alpha^4}\,\Delta^4 div\mathbf{E} + O(\Delta t, h)^4 \tag{IV.17}$$

avec $\Delta^4 div\mathbf{E} = \Delta\circ\Delta\circ\Delta\circ\Delta(div\mathbf{E})$. Dans le cas du système de Maxwell classique ($\alpha = \infty$), on remarque que l'erreur sur la divergence est a priori d'ordre deux en espace, sauf pour $\beta = \dfrac{1}{3}$, où elle est d'ordre trois. Dans ce cas, les termes d'erreur du second membre proviennent des termes de dissipation spatiale du schéma.
Dans le cas de la formulation (IV.3) du système de Maxwell, on voit que les termes d'erreur liés au terme de viscosité sont des termes de dissipation en temps et en espace. Là aussi la méthode proposée rajoute des termes de dissipation sur la divergence de E.

Maillage triangulaire.

On donne l'équation équivalente du β-schéma triangulaire pour la première composante E^1. On pose Δx=Δy=h.

$$\begin{aligned}
E^1_t + B^3_y = &-\frac{h^2}{6}(1-3\beta)\,(B^3_{yyy} + B^3_{xxy} + B^3_{xyy}) - \beta h^3\, g(.,.)\\
&-\frac{\Delta t^3}{24}\,(E^1_{yyyy} + E^1_{xxyy} - E^2_{xyyy} - E^2_{xxxy}) + \frac{E^1_{xx}}{\alpha} + \frac{E^2_{xy}}{\alpha}\\
&+\frac{1}{12\alpha}\,(E^1_{xxxx}\Delta x^2 + 2\,E^2_{xxxy}\Delta x^2 + 2\,E^2_{xyyy}\Delta y^2)\\
&-\frac{\Delta t^3}{24\alpha^4}\,(E^1_{xxxxxxxx} + 3\,E^1_{xxxxxxyy} + 3\,E^1_{xxxxyyyy} + E^1_{xxyyyyyy}\\
&+E^2_{xxxxxxxy} + 3\,E^2_{xxxxxyyy} + 3\,E^2_{xxxxyyyy} + E^2_{xxxxxxyy}) + O(\Delta t, \Delta h)^4
\end{aligned}$$

où g dépend linéairement des dérivées quatrièmes en espace de E^1 et E^2.
Dans le cas du système de Maxwell classique ($\alpha = \infty$), comme pour le schéma en rectangles, on voit que la valeur $\beta = \dfrac{1}{3}$ permet d'annuler les termes de dispersion spatiale et que les termes de dissipation spatiale sont fonction de β.
Comme à l'ordre un, les termes d'erreur de dispersion et de dissipation sont plus nombreux en triangles qu'en rectangles. Ici interviennent des termes de dispersion croisés B^3_{xxy}, B^3_{xyy}, ainsi que tous les termes de dissipation spatiale des deux premières composantes E^1, E^2 dans g. On remarque que les termes en temps sont identiques à ceux obtenus pour le β-schéma en rectangles.
Le β-schéma en maillage triangulaire est donc plus dissipatif que celui en maillage rectangulaire; en effet, l'approximation des flux numériques pour le maillage triangulaire choisi

fait intervenir 19 noeuds de calcul alors que le schéma en rectangles est un schéma à 9 noeuds.
Dans le cas de la formulation (IV.3) du système de Maxwell, les termes liés à α sont des termes de dissipation en temps et en espace, comme pour le schéma en rectangles.
Nous écrivons l'équation équivalente de la divergence de $\mathbf{E}$ dans le cas $\Delta x{=}\Delta y{=}h$:

$$\begin{aligned}(div\mathbf{E})_t - \frac{1}{\alpha}\,\Delta\, div\mathbf{E} = -\beta h^3\, k(.,.) + \frac{h^2}{6\alpha}\,\Delta^2 div\mathbf{E} - \frac{h^2}{12\alpha}\,(E^1_{xxxxx} + E^2_{yyyyy} + 2\,E^1_{xxxyy} \\ +2\,E^2_{xxyyy}) - \frac{\Delta t^3}{24\alpha^4}\,\Delta^4 div\mathbf{E} + O(\Delta t, h)^4\end{aligned} \tag{IV.18}$$

où k dépend linéairement des dérivées cinquième en espace de E^1 et E^2. Dans le cas particulier $\Delta x{=}\Delta y{=}h$, les termes d'erreur provenant de la dispersion du schéma s'annihilent. Dans ce cas-là, l'erreur sur la divergence dans le cas du système de Maxwell classique est d'ordre trois. Les termes d'erreur proviennent alors des termes de dissipation du schéma. Dans le cas de la nouvelle formulation du système de Maxwell, on voit comme précédemment que les termes de viscosité sont des termes de dissipation sur la divergence.

Pour résumer, les équations équivalentes montrent que l'erreur sur la divergence est très sensible à la précision spatiale du schéma : en effet dans le cas du système de Maxwell classique ($\alpha = \infty$), les termes d'erreur sur la divergence proviennent des termes de dispersion et de dissipation du schéma. La méthode de viscosité proposée ($\alpha < \infty$) consiste à augmenter la dissipation des schémas et ainsi obtenir une équation de la chaleur pour la divergence.

IV.5 Résultats numériques

Dans cette partie on s'intéresse plus particulièrement à la vérification numérique des relations de divergence (IV.1).
Nous constaterons que l'introduction d'un terme de viscosité dans les équations de Maxwell permet de mieux satisfaire numériquement les contraintes sur la divergence du champ électromagnétique.

En l'absence de charges

On considère la formulation (IV.3) en deux dimensions d'espace pour une onde transverse électrique (**T.E**) sur $\Omega =]0,1[\times]0,1[$ avec des conditions aux limites périodiques. On initialise les composantes du champ électromagnétique par une combinaison linéaire d'ondes sinusoïdales de diverses fréquences.

On représente sur la figure IV.8 la norme infinie de la divergence du champ électrique $||div\,\mathbf{E}||_\infty$ en fonction du temps à $\Delta x = \Delta y$ fixé. Ceci est obtenu pour différentes valeurs de α sur un maillage en rectangles pour les schémas d'ordre un et trois. On a choisi la norme infinie pour observer plus précisément les différences entre les divers schémas.

On représente ensuite la même chose sur la figure IV.9 pour un maillage triangulaire.

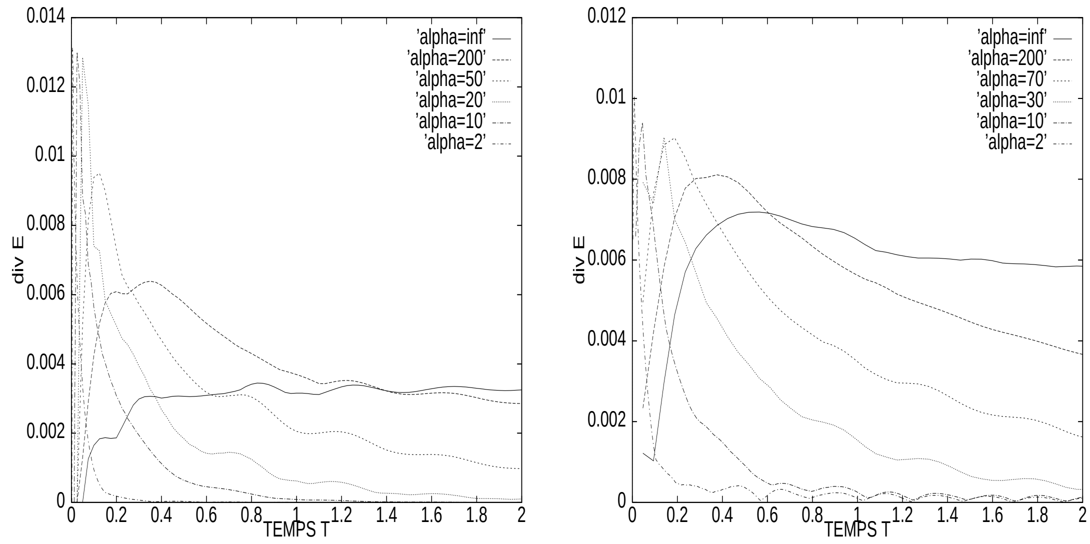


Schéma à l'ordre un   Schéma à l'ordre trois

FIG. IV.8 – *Maillage en rectangles.*

On remarque que plus le paramètre de correction est petit, ce qui correspond à une influence plus importante des termes de viscosité introduits dans les équations de Maxwell, plus la divergence du champ électrique $\mathbf{E}$ est petite. En contrepartie, le fait d'utiliser un paramètre de correction très petit va entraîner une contrainte sur le pas de temps Δt pour des raisons de stabilité. Néanmoins, nous avons montré précédemment l'existence d'une valeur optimale du paramètre de correction qui n'introduit aucune contrainte supplémentaire sur le pas de temps Δt. On remarque également que $div_h\,\mathbf{E}_h \to 0$ quand $t \to +\infty$, ceci est dû au caractére dissipatif de la nouvelle formulation (IV.3).

On compare maintenant l'influence du maillage (triangulaire ou rectangulaire) sur la divergence du champ $\mathbf{E}$.

On représente sur la figure IV.10, $||div\,\mathbf{E}||_\infty$ en fonction du temps à $\Delta x = \Delta y$ fixé avec la valeur optimale du paramètre de correction pour les schémas d'ordre un et trois.

On remarque, pour les schémas précis à l'ordre un et à l'ordre trois, que la divergence est plus petite lorsque l'on considère un maillage en rectangles. Ceci est dû au fait que la valeur optimale du paramètre de correction est plus petite pour le schéma en rectangles, par conséquent l'influence des termes de viscosité est plus importante. Cependant il faut

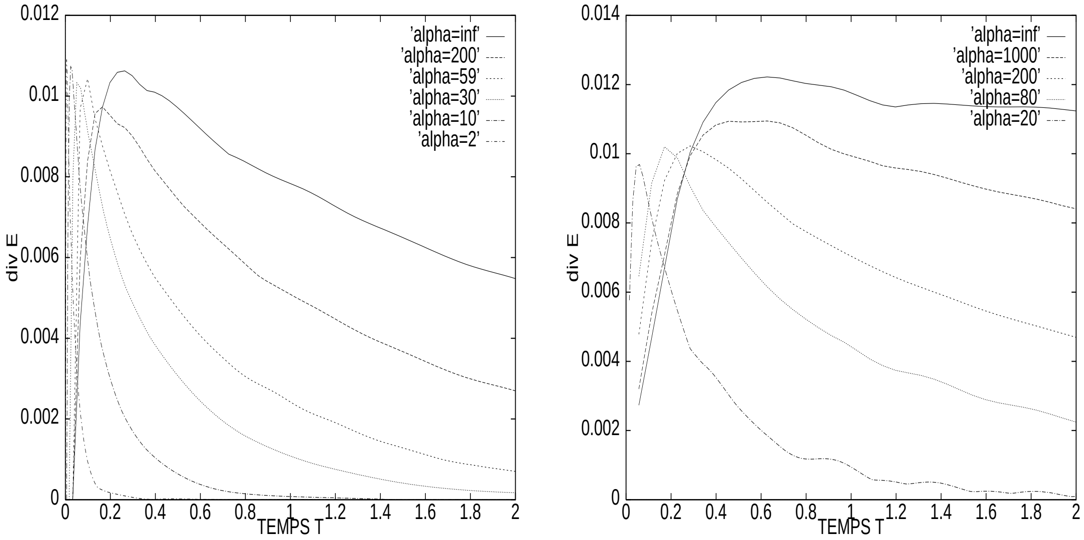
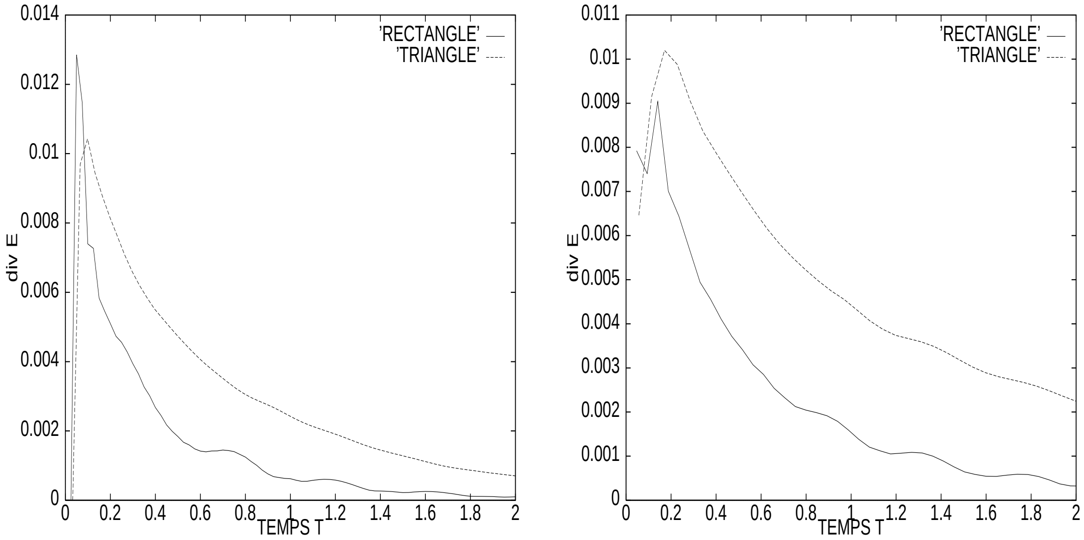


Schéma à l'ordre un Schéma à l'ordre trois

FIG. IV.9 – *Maillage en triangles.*



Schémas à l'ordre un Schémas à l'ordre trois

FIG. IV.10 – *Comparaison triangles-rectangles.*

préciser que l'on peut utiliser des pas de temps plus grands pour le schéma en triangles et donc diminuer le temps de calcul.
On va maintenant étudier l'influence du paramètre de décentrage β sur la divergence du champ **E**.
Plus précisément on représente sur la figure IV.11 la norme infinie de la divergence du champ électrique **E** en fonction du temps à $\Delta x = \Delta y$ fixé. Ceci est obtenu en considérant la valeur optimale du paramètre de correction pour chaque valeur de β dans le cas des schémas en rectangles et en triangles. On rappelle que lorque $\beta = 0$ le schéma est centré

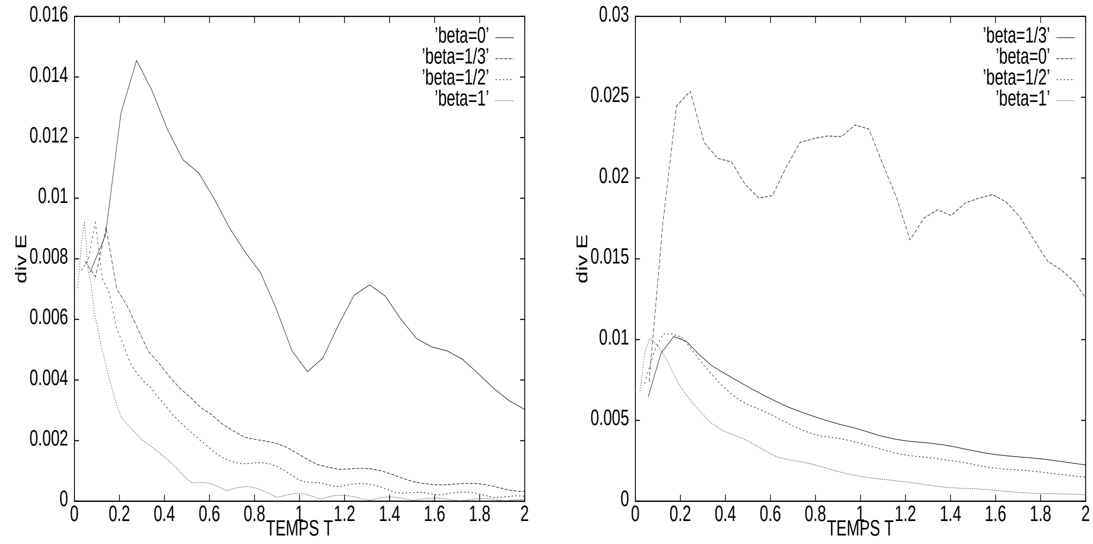


Schéma en rectangles Schéma en triangles

FIG. IV.11 – *Influence du paramètre de décentrage β.*

et lorsque $\beta = 1$ il est totalement décentré.
On remarque, aussi bien pour le schéma en triangles que pour le schéma en rectangles, que la divergence diminue lorsque β augmente. Ceci s'explique par le fait que lorsque β augmente, la valeur optimale du paramètre de correction (α_{opt}) diminue. On note cependant que la contrainte sur le pas de temps augmente avec β. Il semble au vu des résultats qu'il vaut mieux prendre $\beta = 1/3$, valeur pour laquelle la divergence est petite, la contrainte sur le pas de temps n'est pas importante, et surtout le schéma est d'ordre trois en espace.
On s'intéresse maintenant à l'influence des termes de viscosité sur le champ électromagnétique. Pour ce faire on considère une onde plane (**T.E**) solution des équations de Maxwell du type :

$$\begin{cases} E_x(t,x,y) = -cos(x+y-\sqrt{2}t) \\ E_y(t,x,y) = cos(x+y-\sqrt{2}t) \\ B_z(t,x,y) = \sqrt{2}cos(x+y-\sqrt{2}t) \end{cases}$$

On représente sur la figure IV.12 la norme infinie de l'erreur sur le champ magnétique B_z en fonction du temps pour différentes valeurs du paramètre du correction et de $h = \Delta x = \Delta y$, pour les schémas précis à l'ordre un et trois en triangles. On remarque, notamment sur le

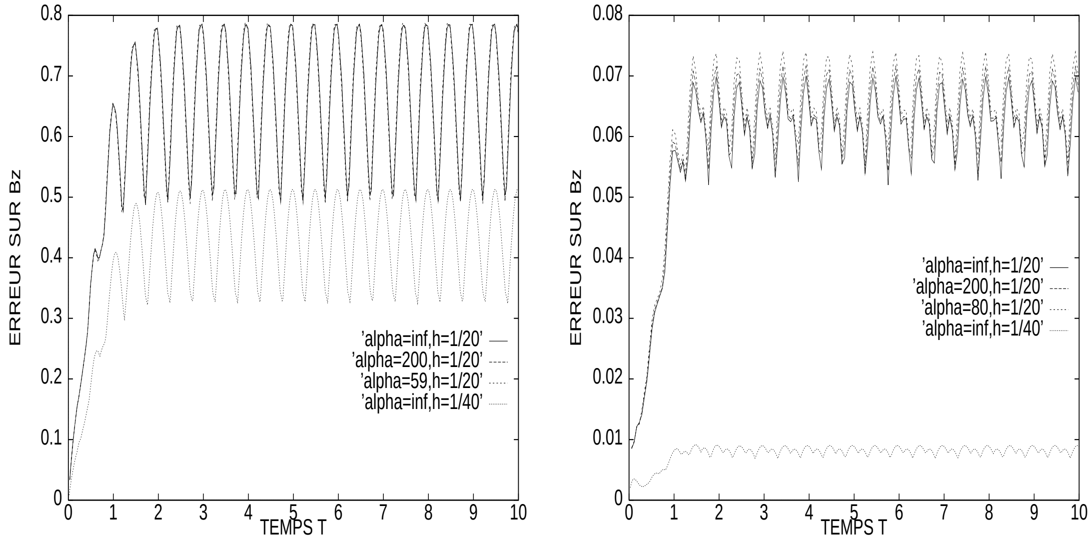


Schéma à l'ordre un — Schéma à l'ordre trois

FIG. IV.12 – *Erreur sur le champ magnétique pour un maillage triangulaire.*

schéma à l'ordre trois, que l'introduction d'un paramètre de correction n'induit pas une erreur supplémentaire importante sur le champ électromagnétique. Rajouter un terme de viscosité dans les équations de Maxwell revient numériquement à rajouter de la diffusion dans nos schémas, ce qui explique cette petite erreur supplémentaire. D'autre part on remarque l'influence de la finesse du maillage sur la précision du champ électromagnétique. Lorsque l'on divise le pas de discrétisation par deux, on divise l'erreur d'un facteur deux pour le schéma à l'ordre un et d'un facteur huit pour le schéma à l'ordre trois.
Il nous reste à étudier l'influence de la discrétisation en espace sur la divergence. On représente sur la figure IV.13 $||div\, \mathbf{E}||_\infty$ en fonction du temps pour différentes valeurs de $h = \Delta x = \Delta y$ en fixant le paramètre de correction, dans le cas du schéma d'ordre trois sur des maillages rectangulaires et triangulaires. On remarque bien que lorsque l'on divise le pas de discrétisation par deux on diminue sensiblement l'erreur sur la divergence du champ $\mathbf{E}$, aussi bien pour un maillage en rectangles que pour un maillage en triangles. On peut ainsi conclure que la finesse du maillage a une forte influence sur la divergence.

En présence de charges

On considère la formulation (IV.3) du système de Maxwell sur un domaine $\Omega =]0,1[^2$. On suppose qu'à l'instant initial le champ électromagnétique est nul et que la frontière

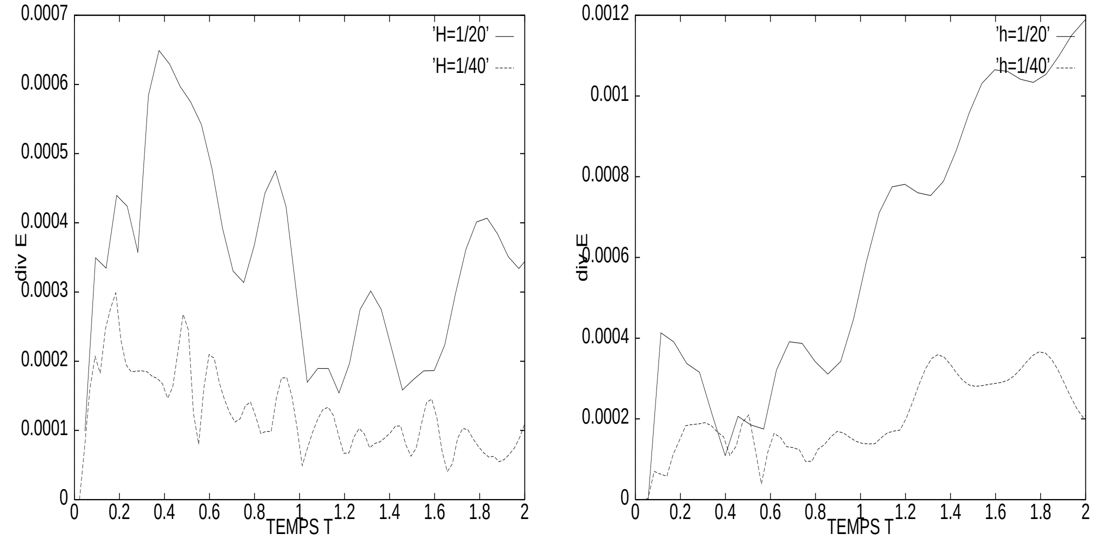


Schéma en rectangles Schéma en triangles

FIG. IV.13 – *Erreur sur la divergence pour des schémas d'ordre trois.*

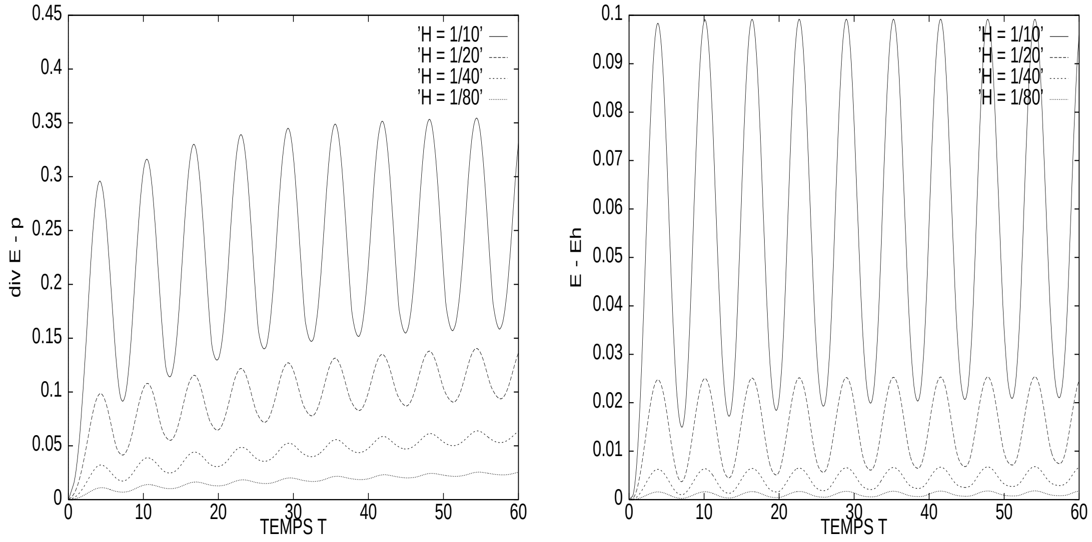
$\Gamma = \partial\Omega$ est parfaitement conductrice: $n \times \mathbf{E} = 0$ sur Γ.
On se donne une densité de charges et de courants:

$$\begin{cases} \rho(t,x,y) = \sin(t)(\sin(\pi y) + \sin(\pi x)) \\ j_x(t,x,y) = (\cos(t) - 1)(\pi\cos(\pi x) + \pi^2 x \sin(\pi y)) - x\cos(t)\sin(\pi y) \\ j_y(t,x,y) = (\cos(t) - 1)(\pi\cos(\pi y) + \pi^2 y \sin(\pi x)) - y\cos(t)\sin(\pi x) \end{cases}$$

Pour de telles données, la solution exacte de (IV.3) est donnée par:

$$\mathbf{E} = \sin(t)\begin{pmatrix} x\sin(\pi y) \\ y\sin(\pi x) \\ 0 \end{pmatrix} \qquad \mathbf{B} = (cos(t) - 1)\begin{pmatrix} 0 \\ 0 \\ \pi y\cos(\pi x) - \pi x\cos(\pi y) \end{pmatrix}$$

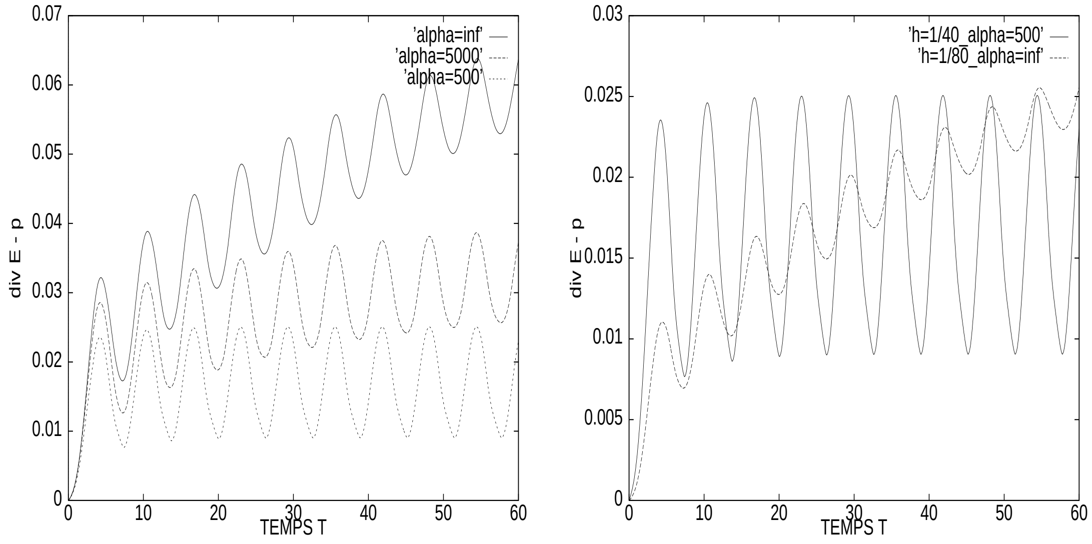
On considère ici le schéma à l'ordre trois pour un maillage en triangles du domaine de calcul Ω. On représente sur la figure IV.14, $\parallel div\mathbf{E_h} - \rho \parallel_{L^1(\Omega)}$ et $\parallel \mathbf{E} - \mathbf{E_h} \parallel_{L^1(\Omega)}$ en fonction du temps, pour différentes valeurs du pas de discrétisation $h = \Delta x = \Delta y$ et pour une valeur très grande fixée du paramètre de correction α. On remarque que l'utilisation de maillages fins diminue notablement l'erreur sur le champ électromagnétique, de même la relation de divergence $div\mathbf{E} - \rho = 0$ est mieux vérifiée.
On représente maintenant, dans le cas d'un maillage en triangles pour le schéma à l'ordre trois, sur la figure IV.15 $\parallel div\mathbf{E_h} - \rho \parallel_{L^1(\Omega)}$ en fonction du temps à $h = \Delta x = \Delta y$ fixé pour différentes valeurs du paramètre de correction α. Sur la figure IV.15, on représente



$div\mathbf{E_h} - \rho$ en fonction de h $\qquad$ $\mathbf{E} - \mathbf{E_h}$ en fonction de h

FIG. IV.14 – *Influence du pas de maillage h.*

également $\| \; div\mathbf{E_h} - \rho \; \|_{L^1(\Omega)}$ en fonction du temps pour, $h = \dfrac{1}{40}$ avec la valeur optimale du paramètre de correction, et pour $h = \dfrac{1}{80}$ sans correction ($\alpha = +\infty$). Comme en



$div\mathbf{E_h} - \rho$ en fonction de α $\qquad$ $div\mathbf{E_h} - \rho$ en fonction de α et h

FIG. IV.15 – *Influence de α et de h sur la divergence.*

l'absence de charges, on note que lorsque la valeur du paramètre de correction diminue, ce qui correspond à une influence plus importante des termes de viscosité, la relation de

divergence $div\mathbf{E} - \rho = 0$ est mieux vérifiée. On remarque aussi que l'erreur sur $div\mathbf{E_h} - \rho$ est du même ordre pour un maillage donné en utilisant la nouvelle méthode (IV.3) que pour un maillage deux fois plus fin en résolvant le système de Maxwell classique.
On compare maintenant sur la figure IV.16 $\| div\mathbf{E_h} - \rho \|_{L^1(\Omega)}$ pour les schémas à l'ordre un et trois, à $h = \Delta x = \Delta y$ fixé et pour une valeur fixée du paramètre α. On remarque que

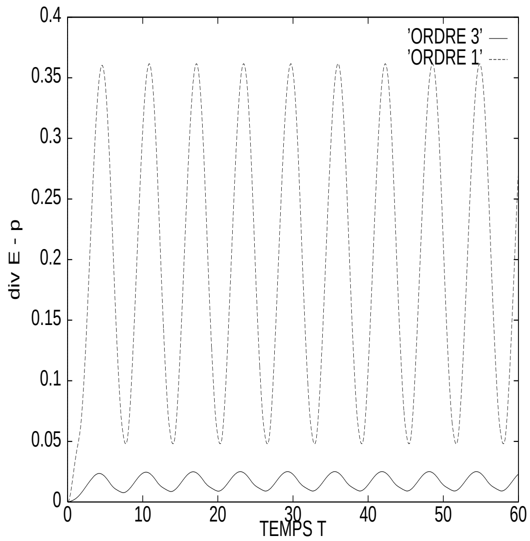


FIG. IV.16 – *Comparaison entre les schémas d'ordre un et trois.*

la précision du schéma a une très forte influence sur la relation de divergence $div\mathbf{E} - \rho = 0$. Il est donc très intéressant d'utiliser un schéma précis à l'ordre trois.

Temps de calcul

Il est intéressant d'évaluer le coût, en terme de temps de calcul et de stockage, qu'entraîne l'introduction d'un terme de viscosité dans les équations de Maxwell.
La résolution numérique de la formulation (IV.3) nécessite environ 7% de temps de calcul supplémentaire par rapport à la résolution du système de Maxwell classique. Le surcoût en terme d'occupation mémoire est d'environ 4%. On peut donc, au vu de ces chiffres, conclure que la méthode de viscosité proposée permet de prendre en compte les contraintes de divergence pour le champ électromagnétique à un coût raisonnable.

IV.6 Conclusion

Cette étude nous montre que la méthode de pénalisation proposée permet que les relations de divergence (IV.1) soient mieux vérifiées numériquement.

On a vu dans l'étude de stabilité que l'on peut trouver un paramètre de diffusion optimal qui n'introduit aucune contrainte supplémentaire sur le pas de temps, quel que soit le maillage utilisé, rectangulaire ou triangulaire.
L'étude des équations équivalentes nous a permis de mettre en évidence les termes d'erreur sur les schémas eux-mêmes et également sur la variation de la divergence au cours du temps. La valeur $\beta = \dfrac{1}{3}$ permet d'obtenir des schémas d'ordre trois en temps et en espace, ainsi qu'une erreur sur la divergence d'ordre trois; aussi bien pour les maillages rectangulaires que triangulaires.
Les schémas en triangles semblent introduire plus de dissipation numérique qu'en rectangles, et les relations de divergence sont mieux vérifiées dans le cas d'un schéma en rectangles. Cela peut provenir du fait que l'influence du terme de diffusion introduit est plus importante en rectangles qu'en triangles. Toutefois, dans notre étude nous avons considéré des maillages triangulaires structurés, ce qui pénalise dans ce cas les schémas par-rapport aux maillages rectangulaires.
On a aussi remarqué l'influence du pas d'espace h sur la divergence. A ce sujet, cette méthode peut s'avérer intéressante dans le cas tridimensionnel, en permettant par exemple d'utiliser un pas d'espace plus grand (moins de points) par longueur d'onde.

Chapitre V

CONCLUSION.

L'objectif de cette première partie était de construire des schémas précis en temps et en espace par des méthodes volumes finis/éléments finis dans le but de les appliquer aux équations de Maxwell.

Tout d'abord, nous avons construit et étudié une classe de schémas, en maillage triangulaire et rectangulaire, basés sur des volumes finis, sur des éléments finis, ou bien sur des méthodes mélant les deux approches. L'étude de stabilité et les équations équivalentes nous ont permis de comparer ces schémas entre eux et de retenir les plus précis. La formulation éléments finis fait intervenir une matrice de masse dans l'intégration en temps : nous avons montré que les schémas avec matrice de masse ont toujours des limites de stabilité inférieures à celles des schémas correspondants sans matrice de masse. Pour simuler des phénomènes électromagnétiques, nous recherchons des schémas précis mais d'un coût raisonnable afin de minimiser le nombre de points du maillage. En effet, certaines applications industrielles tridimensionnelles peuvent nécessiter des maillages comportant jusqu'à dix millions de mailles. Pour cela, nous préférons les schémas explicites où la matrice de masse est diagonalisée.
En maillage rectangulaire, nous avons comparé l'approche volumes finis classique avec celle utilisant des éléments finis Q1. Le schéma **EF R** a un domaine de stabilité plus large que le schéma **VF R**, mais le gain de précision est faible par-rapport aux volumes finis classiques. Par contre, le schéma **EF R** s'avère plus coûteux en termes de temps de calcul et d'occupation mémoire puisqu'il nécessite un calcul de flux avec chaque voisin d'un noeud.
Les schémas les plus précis et les moins coûteux que nous avons étudié sont les schémas "sans diffusion numérique" d'ordre quatre en temps et en espace (avec $\beta = \dfrac{1}{3}$ et sans matrice de masse).

Par la suite, nous avons repris les schémas en volumes finis vus précédemment et nous les avons appliqué au système de Maxwell, dans un cadre bidimensionnel. Une étude

de stabilité concernant cette classe de schémas a été présentée, pour des maillages en rectangles et en triangles, ce qui n'a pas été fait à notre connaissance pour le système de Maxwell, excepté dans des cas particuliers. Une condition nécessaire et suffisante de stabilité a été démontrée dans le cas du schéma décentré d'ordre un, pour un maillage en rectangles. Dans ce cas, on retrouve la condition de stabilité obtenue pour l'équation d'advection : $u_t + u_x + u_y = 0$.
Pour les schémas d'ordre supérieur (β-schémas), on constate que les domaines de stabilité obtenus pour l'équation d'advection sont plus grands avec un maillage triangulaire. Au contraire, dans le cas d'un maillage en rectangles, les limites de stabilité sont plus élevées pour le système de Maxwell. Néanmoins, les limites de stabilité varient de la même manière avec β pour les deux modèles : les plus grandes limites de stabilité sont obtenues avec les schémas centrés, en particulier, les schémas "sans diffusion numérique". On a aussi comparé les schémas décalés (schémas de Yee) avec l'approximation en volumes finis. Dans le cas de schémas au moins d'ordre trois en temps et en espace, les domaines de stabilité sont beaucoup plus grands, aussi bien pour des maillages en rectangles qu'en triangles. Contrairement au schéma de Yee, l'extension des schémas décentrés d'ordre un à un ordre supérieur se fait facilement avec l'approche volumes finis. De plus, cette méthode peut s'appliquer à toutes sortes de maillages.
Nous effectuons ensuite un calcul de propagation d'onde dans une cavité carrée. Nous comparons le schéma décentré d'ordre trois, le schéma "sans diffusion numérique" d'ordre quatre et le schéma "modifié d'ordre quatre", pour lequel on a rajouté un peu de diffusion numérique. Ce dernier schéma améliore beaucoup la précision de la solution, lorsqu'on regarde les champs B_x et E_z. Par contre, nous voyons que les relations de divergence du système de Maxwell ne sont pas vérifiées par nos schémas. En particulier, en l'absence de charges et de courant, la divergence du champ électrique et la divergence de l'induction magnétique ne tendent pas vers 0.

Pour remédier à ce problème, nous avons proposé une nouvelle formulation des équations de Maxwell, qui consiste à rajouter un terme de viscosité dans les équations de Maxwell classiques. La résolution numérique de ce nouveau système engendre un très faible surcoût du point de vue temps de calcul, par-rapport au système de Maxwell classique. Nous avons établi les équations équivalentes des schémas pour le système de Maxwell classique et pour la nouvelle formulation, ce qui nous a permis de montrer pourquoi notre nouvelle méthode vérifie mieux numériquement les relations de divergence. Nous avons aussi montré l'influence du pas de maillage et de la précision du schéma sur la divergence. Nous avons étudié la stabilité des schémas avec la nouvelle formulation, dans le but de préciser le choix du paramètre de viscosité qui n'introduit aucune contrainte supplémentaire sur le pas de temps par-rapport au système de Maxwell classique. Nous constatons que les relations de divergence sont mieux vérifiées avec un maillage en rectangles plutôt qu'en triangles : d'une part, les schémas en triangles sont plus dissipatifs, d'autre part, l'étude de stabilité montre que l'influence du terme de viscosité introduit est plus grande pour un schéma en rectangles que pour celui en triangles.

Cette nouvelle méthode a été également validée pour résoudre le système couplé de Vlasov-Maxwell [23], où les densités de charge ρ et de courant $\mathbf{j}$ sont déterminées à partir d'une fonction de distribution f, solution de l'équation de Vlasov relativiste. Cela a été fait dans un cadre bidimensionnel où on a considéré comme cas d'école, un tube électronique comprenant une cavité. Les résultats numériques obtenus sont en accord avec les résultats attendus [23].

Une des perspectives de ce travail serait de valider la nouvelle formulation proposée dans le cadre tridimensionnel, puis de paralléliser cette méthode. En effet, la résolution effective de systèmes, comme celui de Vlasov-Maxwell, est très coûteuse en termes de temps de calcul, et l'utilisation de machines parallèles pourrait alors permettre de considérer des dispositifs plus complexes.

Une autre perspective serait maintenant de construire des schémas volumes finis/éléments finis utilisant des maillages hybrides (par exemple, en mélangeant les mailles triangulaires et rectangulaires). Pour l'instant, les schémas en volumes finis appliqués aux problèmes électromagnétiques ont été utilisé sur des maillages en triangles non structurés, afin de pouvoir mailler des géométries complexes. Cependant, cela s'avère coûteux en termes de temps de calcul et d'occupation mémoire, en particulier dans les zones où l'on n'est pas intéressé par la solution. Ainsi, certains problèmes comme la diffraction autour d'un objet requièrent des domaines de calcul non bornés : en pratique, on place une frontière artificielle assez loin de l'objet afin que d'éventuelles réflexions parasites n'intéragissent avec l'objet diffractant. Dans ce cas, l'utilisation de maillages hybrides permettrait de mailler finement autour de l'objet (par exemple, avec des triangles) et d'utiliser un maillage plus grossier (grilles orthogonales, maillage en quadrangles,...) loin de l'objet, jusqu'à la frontière artificielle. L'étude et l'implémentation de tels schémas sont actuellement en cours par F. Bonnet [2].

Bibliographie

[1] ASSOUS F., DEGOND P., SEGRE J., *A particle-tracking method for 3D electromagnetic PIC codes on unstructured meshes*, Comp. Phys. Comm., Vol 72, pp. 105-114, (1992).

[2] BONNET F., *Une méthode de volumes finis hybride pour la résolution du système de Maxwell instationnaire.*, Rapport de recherche CERMICS à paraître.

[3] BOSSAVIT A., *Electromagnétisme, en vue de la modélisation*, Springer-Verlag, Paris (1993).

[4] CARPENTIER R., de la BOURDONNAYE A., LARROUTUROU B., *On the derivation of the modified equation for the analysis of linear numerical methods*, Rapport de recherche CERMICS no.26 (1994).

[5] CARPENTIER R., *Approximation et analyse numérique d'écoulements instationnaires, application à des instabilités tourbillonnaires*, Thèse de Doctorat de l'Université de Nice-Sophia-Antipolis (1995).

[6] CHANG S. C., *A critical analysis of the modified equation technique of Warming and Hyett*, J. Comp. Phys., Vol 86, pp. 107-126, (1990).

[7] CIONI J. P., *Résolution numérique des équations de Maxwell instationnaires par une méthode de volumes finis*, Thèse de Doctorat de l'Université de Nice-Sophia-Antipolis (1995).

[8] CIONI J.P., FÉZOUI L., STÈVE H., *Approximation des équations de Maxwell par des schémas décentrés en éléments finis*, Rapport de recherche INRIA no.1601 (1992).

[9] CIONI J.P., FÉZOUI L., STÈVE H., *A parallel time-domain Maxwell solver using upwind schemes and triangular meshes*, IMPACT in computing in science and engeenering No 165 (1993)

[10] CIONI J.P., FÉZOUI L., ISSAUTIER D., *High-order upwind schemes for solving time-domain Maxwell equation*, La Recherche Aérospatiale, numéro spécial électromagnétisme.

[11] COHEN G., MONK P., *Efficient edge finite element schemes in computational electromagnetism*, actes Third international conference on mathematical and numerical aspects of wave propagation, pp. 250-259, Mandelieu la Napoule (1995).

[12] COHEN G., JOLY P., TORDJMAN N., *Higher order triangular finite elements with mass lumping for the wave equation*, actes Third international conference on mathematical and numerical aspects of wave propagation, pp. 270-279, Mandelieu la Napoule (1995).

[13] DAUTRAY R., LIONS J.L. *Analyse mathématique et calcul numérique*, Vol. 1, pp 68-127, Masson (1987).

[14] DEPEYRE S., ISSAUTIER D., *Application aux schémas volumes finis d'une méthode de pénalisation des contraintes pour le système de Maxwell*, Rapport CERMICS no 39 (1995) et accepté pour publication dans "Mathematical Modelling and Numerical analysis".

[15] DEPEYRE S., *Stability analysis for the finite volume schemes on rectangular and triangular meshes applied to the 2D Maxwell system*, Rapport CERMICS no 40 (1995)

[16] DEPEYRE S., LARROUTUROU B., CARPENTIER R., *Méthodes numériques décentrées d'ordre élevé en deux dimensions d'espace*, Rapport CERMICS no 41 (1995).

[17] DERVIEUX A., FÉZOUI L., LORIOT F., *On high resolution variants of Lagrange-Galerkin finite-element schemes*, Rapport de recherche INRIA no.1703.

[18] DÉSIDÉRI J.A, GOUDJO A., SELMIN V., *Third-order numerical schemes for hyperbolic problems*, Rapport de recherche INRIA no.607 (1987).

[19] FÉZOUI F., *Résolution des équations d'Euler par un schéma de Van Leer en éléments finis*, Rapport de recherche INRIA no.358 (1985).

[20] GLINSKY N., *Simulation numérique d'écoulements hypersoniques réactifs hors-équilibre chimique*, Thèse de l'Université de Nice-Sophia-Antipolis (1990).

[21] HEINTZÉ E., *Résolution des équations de Maxwell tridimensionnelles instationnaire par une méthode d'éléments finis conformes*, Thèse de Doctorat de L'Université Paris VI, (1992).

[22] HERMELINE F., *Two coupled particle-finite volume methods using Delaunay-Voronoï meshes for the approximation of Vlasov-Poisson and Vlasov-Maxwell equations*, Jour. Comp. Phy., Vol 106, pp. 1-18, (1993).

[23] ISSAUTIER D., *Méthodes particulaires en cinétique des gaz et transport de particules chargées*, Thèse de Doctorat de l'Université de Nice-Sophia-Antipolis (1995).

[24] ISSAUTIER D. - CIONI J.P. - POUPAUD F. - FÉZOUI L., *A 2-D Vlasov-Maxwell solver on unstructured meshes*, Third international conference on mathematical and numerical aspects of wave propagation phenomena, Mandelieu, Avril 1995.

[25] LANTÉRI S., *Simulation d'écoulements aérodynamiques instationnaires sur une architecture massivement parallèle*, Thèse de Doctorat de l'Université de Nice-Sophia-Antipolis (1991).

[26] LAX P.D., HARTEN A., VAN LEER B., *On upstream differencing and Godunov type schemes for hyperbolic conservation laws*, SIAM Revue, Vol. 25, no. 1 (1983).

[27] LYNCH D. R., PAULSEN K.D., *Time-domain integration of the Maxwell equations on finite elements*, IEEE Trans. Ant. Prop., Vol 38, pp. 1933-1942 (1990).

[28] NÉDÉLEC J.C., *Mixed finite elements in* $I\!R^3$, Num. Math., Vol 35, pp. 315-341 (1980). Ecoles CEA/EDF/INRIA (1994).

[29] SELMIN V., *Finite element solution of hyperbolic equations; I: one dimensional case, II: two dimensional case*, Rapport de recherche INRIA no. 655 (1987).

[30] TAFLOVE A., BRODWIN M.E., *Numerical Solution of Steady-State Electromagnetism Scattering Problems Using the Time-Dependent Maxwell's Equations*, IEEE Trans. Microwave Theory and Techniques 23, Vol 8, 623-630 (1975).

[31] VAN LEER B., *Flux vector splitting for the Euler equations*, Lecture Notes in Physics, Vol 170, pp 405-512 (1982)

[32] WARMING R.F.& HYETT F., *The modified equation approach to the stability and accuracy analysis of finite-difference methods*, J. Comp. Phys.,14, (2), p.159, (1974).

[33] YEE K.S., *Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media*, IEEE Trans. Ant. Prop., Vol 14, pp. 302-307, (1993).

[34] YOUNG J. L., BRUECKNER F.P., *A time-domain, weighted residual formulation of Maxwell's equations*, AIAA Paper, 93-0462, (1993).

DEUXIÈME PARTIE :

CAS NON LINÉAIRE.

Chapitre VI

ÉTUDE DE MODÈLES NON LINÉAIRES ET APPROXIMATION NUMÉRIQUE.

VI.1 Introduction.

Ce chapitre est consacré à l'étude et à la résolution numérique de deux modèles conservatifs non linéaires. Nous considérons tout d'abord le cas d'une équation scalaire, l'équation de Burgers, dans un cadre monodimensionnel, et nous décrivons brièvement la méthode de Godunov. Par la suite, nous nous intéressons au système des équations d'Euler, qui régit l'écoulement d'un fluide parfait compressible. Nous nous plaçons alors dans un cadre bidimensionnel, et nous rappellons l'approche mixte éléments finis/volumes finis décrite dans [7] pour des maillages en triangles non structurés. Ce type de maillages prend en compte de façon naturelle les géométries complexes rencontrées dans les écoulements industriels, et permet, d'autre part, des raffinements locaux du maillage, lorsque cela est nécessaire. Nous utilisons des fonctions de flux numériques centrées et décentrées. Comme dans le cas linéaire (pour l'équation d'advection ou pour le système de Maxwell), l'ordre de précision spatiale des schémas est élevé à l'aide de l'approche MUSCL (Monotonic Upstream Scheme for Conservation Laws) introduite par Van Leer [21].
De nombreuses études ont été faites [18, 19], dans le cadre des équations d'Euler ou de Navier-Stokes, afin de s'affranchir de la condition de stabilité CFL (condition de Courant-Friedrichs-Lewy) inhérente aux schémas explicites. Pour cela, nous nous tournons vers des méthodes implicites, qui permettent d'utiliser de plus grands pas de temps et d'atteindre l'état stationnaire (lorsqu'il est recherché) plus rapidement. Nous nous intéressons plus particulièrement aux schémas implicites linéarisés décentrés, avec une méthode de Jacobi pour résoudre le système linéaire. L'efficacité et la robustesse de cette méthode

ont été montrées numériquement pour des calculs complexes des équations d'Euler [18]. L'avantage de la méthode de Jacobi par-rapport à d'autres méthodes de relaxation comme celle de Gauss-Seidel, est le faible encombrement mémoire puisqu'on évite le stockage des termes extra-diagonaux de la matrice du système linéaire.

VI.2 L'équation de Burgers.

On considère le problème de Cauchy suivant :

$$\begin{cases} u_t + f(u)_x = 0 \quad \text{pour} \quad (x,t) \ \in \mathbb{R}\text{x}\mathbb{R}^+, \\ u(x,0) = u_0(x) \quad \text{pour} \quad x \ \in \mathbb{R}. \end{cases} \tag{VI.1}$$

où $f(u) = \dfrac{u^2}{2}$ est une fonction non linéaire, strictement convexe.

Contrairement au cas linéaire, même si la fonction f et la donnée initiale u_0 sont très régulières, la solution u du problème de Cauchy n'est pas régulière, et des discontinuités peuvent apparaître en temps fini. On introduit alors la notion de solutions faibles, ce sont des solutions qui peuvent présenter des discontinuités appelées chocs par analogie à la mécanique des fluides. Il n'y a pas unicité des solutions faibles, on impose alors une condition supplémentaire, appelée inégalité d'entropie, qui permet de sélectionner la solution physique du problème. Cette inégalité permet, dans le cas scalaire, d'énoncer un résultat d'existence et d'unicité : le problème (VI.1) a, pout toute donnée initiale u_0, une unique solution faible entropique [17, 10, 11].

En particulier, si on considère le problème de Riemann pour l'équation de Burgers :

$$\begin{cases} u_t + (\dfrac{u^2}{2})_x = 0, \\ u(x,0) = \begin{cases} u_g & \text{si } x < 0 \\ u_d & \text{si } x > 0 \end{cases} \end{cases} \tag{VI.2}$$

L'unique solution entropique de (VI.2) est :

– une détente si $u_g \leq u_d$ et la solution est continue:

$$u(x,t) = \begin{cases} u_g & \text{si } \dfrac{x}{t} \leq u_g \\ \dfrac{x}{t} & \text{si } u_g < \dfrac{x}{t} < u_d \\ u_d & \text{si } \dfrac{x}{t} \geq u_d \end{cases}$$

– un choc si $u_g > u_d$ et la solution est discontinue:

$$u(x,t) = \begin{cases} u_g & \text{si } \dfrac{x}{t} < \dfrac{1}{2}(u_g + u_d) \\ u_d & \text{si } \dfrac{x}{t} > \dfrac{1}{2}(u_g + u_d) \end{cases}$$

VI.2.1 Approximation d'ordre un.

Pour résoudre numériquement (VI.1), nous choisissons une formulation en volumes finis. Nous rappelons que l'idée de base est de diviser le domaine spatial considéré en cellules et de former les équations discrètes à partir de la forme intégrale de l'équation de conservation écrite pour chaque cellule. Ainsi, les inconnues u_i^n sont des approximations au temps t^n de la moyenne de u sur la cellule C_i. On choisit un maillage régulier tel que :

$$\mathcal{C}_i = [(i-1/2)\Delta x, (i+1/2)\Delta x].$$

Nous utilisons un schéma conservatif, qui s'écrit :

$$\frac{u_i^{n+1}-u_i^n}{\Delta t} + \frac{\phi_{i+\frac{1}{2}} - \phi_{i-\frac{1}{2}}}{\Delta x} = 0, \tag{VI.3}$$

où le flux numérique $\phi_{i+\frac{1}{2}}$ entre les cellules $\mathcal{C}_i$ et $\mathcal{C}_{i+1}$ est donné par $\phi_{i+\frac{1}{2}} = \phi(u_i^n, u_{i+1}^n)$.

La bonne façon de généraliser le schéma décentré pour les problèmes non linéaires repose sur la méthode de Godunov. En effet, dans le cas de l'équation de Burgers, les caractéristiques sont des droites, elles ont donc des pentes constantes mais dont le signe peut être quelconque. Un schéma conservatif décentré en volumes finis doit donc pouvoir calculer avec précision les flux aux interfaces entre les cellules en tenant compte du fait que la solution discrète u^n est constante par cellule. La méthode de Godunov est un schéma de volumes finis dont la construction utilise la résolution exacte de problèmes de Riemann aux interfaces entre deux cellules. Si on prend pour condition initiale celle de (VI.2), le flux de Godunov s'écrit :

$$\phi_{\text{Godunov}}(u_G, u_D) = \begin{cases} f(u_G) & \text{if } 0 \le u_G \le u_D \text{ or } u_G \ge |u_D| \\ f(u_D) & \text{if } u_G \le u_D \le 0 \text{ or } |u_G| \le -u_D \\ 0 & \text{if } u_G \le 0 \le u_D \end{cases}, \tag{VI.4}$$

et le schéma de Godunov pour l'équation de Burgers s'écrit :

$$\frac{u_i^{n+1}-u_i^n}{\Delta t^n} + \frac{\phi_{i+\frac{1}{2}} - \phi_{i-\frac{1}{2}}}{\Delta x} = 0, \text{ avec } \phi_{i+\frac{1}{2}} = \phi_{\text{Godunov}}(u_i^n, u_{i+1}^n). \tag{VI.5}$$

où le pas de temps Δt^n varie au cours du temps.

Théorème VI.2.1 *Ce schéma est monotone sous la condition :*

$$\frac{\Delta t^n}{\Delta x}\|u^n\|_{L^\infty} \le 1. \tag{VI.6}$$

La démonstration est assez simple et résulte des propriétés des solutions faibles entropiques, en particulier, la variation totale de u ne croît pas au cours du temps [11, 22]. Ce résultat permet d'assurer que le schéma de Godunov est T.V.D (Total Variation Diminishing), d'ordre un et L^∞ décroissant sous la condition de type CFL non linéaire (VI.6). Il est donc convergent dans L^1 [22]. Nous reviendrons dans le chapitre VII sur la notion de schémas T.V.D.

VI.2.2 Extension à un ordre supérieur.

L'extension à un ordre supérieur en espace s'effectue, comme dans le cas linéaire au moyen d'un β-schéma. On utilise la même fonction de flux qu'à l'ordre un, seuls les arguments changent, ils s'obtiennent par interpolation affine dans chaque cellule. On obtient ainsi :

$$\phi_{i+\frac{1}{2}} = \phi_{\text{Godunov}}(u^n_{i+\frac{1}{2}^-}, u^n_{i+\frac{1}{2}^+}). \tag{VI.7}$$

avec :

$$\begin{aligned} u_{i+\frac{1}{2}^-} &= u_i + \frac{1}{2}\left[(1-\beta)(u_{i+1}-u_i) + \beta(u_i - u_{i-1})\right], \\ u_{i+\frac{1}{2}^+} &= u_{i+1} - \frac{1}{2}\left[(1-\beta)(u_{i+1}-u_i) + \beta(u_i - u_{i-1})\right], \end{aligned} \tag{VI.8}$$

Nous avons vu que dans le cas linéaire le schéma obtenu est au moins d'ordre deux en espace, et qu'il est d'ordre trois pour la valeur $\beta = \dfrac{1}{3}$. En conséquence, nous recherchons des intégrations temporelles d'ordre deux ou trois. Nous utilisons des schémas de type “Runge-Kutta non linéaires”. L'algorithme RKk est donné ci-dessous (pour $k = 1, ..., 3$) :

$$\begin{cases} u^0 = u^n \\ u^l = u^0 - \displaystyle\sum_{j=1}^{k} \Delta t \ \alpha_{jl} \ \Delta(u^{j-1}) \quad l = 1, ..., k \\ u^{n+1} = u^k \end{cases} \tag{VI.9}$$

où $t^n = n\Delta t$ et $D(u^{j-1}) = \phi_{i+\frac{1}{2}} - \phi_{i-\frac{1}{2}}$ représente le flux total calculé avec la valeur u^{j-1}. Les matrices des coefficients $(\alpha_{jl})_{j,l=1,...,k}$ sont données ci-après :

$$(\alpha_{jl})_{k=2} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \end{pmatrix}, \quad (\alpha_{jl})_{k=3} = \begin{pmatrix} 1/3 & 0 & 0 \\ 0 & 2/3 & 0 \\ 1/4 & 0 & 3/4 \end{pmatrix}$$

Nous nous intéressons maintenant à l'étude de systèmes hyperboliques non linéaires, comme celui des équations d'Euler.

VI.3 Le système des équations d'Euler.

VI.3.1 Généralités.

Les équations d'Euler décrivant l'état d'un fluide compressible sont obtenues à partir de principes de conservation de la mécanique, comme la conservation de la masse, la conservation de la quantité de mouvement, et la conservation de l'énergie.

Le système des équations d'Euler en deux dimensions d'espace s'écrit :

$$\left\{ \begin{array}{l} \partial_t \mathbf{W} + \nabla . \mathbb{F}(\mathbf{W}) = 0 \;, \mathbf{W} \in \Omega \text{ x } \mathbb{R}^{+\star} \\ \\ \text{avec conditions initiales et conditions aux limites} \end{array} \right. \tag{VI.10}$$

où Ω est un ouvert borné de $\mathbb{R}^2$.
La fonction $\mathbf{W}$ est appelée variable d'état et s'écrit :

$$\mathbf{W} = \begin{pmatrix} \rho \\ \rho \mathbf{V} \\ E \end{pmatrix} \tag{VI.11}$$

où ρ est la masse volumique, $\mathbf{V} = (u, v)$ est le vecteur champ de vitesse du fluide et E représente l'énergie totale par unité de volume.

Le flux convectif $\mathbb{F} = (\mathbf{F}, \mathbf{G})$ est donné par :

$$\mathbf{F}(\mathbf{W}) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho u v \\ u(E + p) \end{pmatrix} \;, \quad \mathbf{G}(\mathbf{W}) = \begin{pmatrix} \rho v \\ \rho u v \\ \rho v^2 + p \\ v(E + p) \end{pmatrix} \;. \tag{VI.12}$$

où p est la pression qui vérifie la loi d'état des gaz parfaits :

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

γ est le rapport des chaleurs spécifiques supposé constant et égal à 1.4 pour un gaz parfait diatomique.

VI.3.2 Hyperbolicité.

Considérons une combinaison linéaire des flux :

$$\mathcal{F}(\mathbf{W}, \boldsymbol{\eta}) = \eta_x \mathbf{F}(\mathbf{W}) + \eta_y \mathbf{G}(\mathbf{W}), \tag{VI.13}$$

où $\boldsymbol{\eta} = (\eta_x, \eta_y)$ est un vecteur quelconque et non nul de $\mathbb{R}^2$.
La matrice jacobienne $\mathcal{A}$ définie par :

$$\mathcal{A}(\mathbf{W}, \boldsymbol{\eta}) = \eta_x A(\mathbf{W}) + \eta_y B(\mathbf{W}), \tag{VI.14}$$

avec

$$A = \frac{\partial \mathbf{F}}{\partial \mathbf{W}}(\mathbf{W}), \quad B = \frac{\partial \mathbf{G}}{\partial \mathbf{W}}(\mathbf{W}) \tag{VI.15}$$

est diagonalisable et toutes les valeurs propres sont réelles pour tout vecteur non nul $\boldsymbol{\eta} \in \mathbb{R}^2$ et $\mathbf{W} \in \mathbb{R}^4$. Les valeurs propres $(\lambda_i)_{i=1,4}$ valent :

$$\left\{ \begin{array}{l} \lambda_1 = \lambda_2 = \eta_x u + \eta_y v \\ \\ \lambda_3 = \lambda_1 - c||\boldsymbol{\eta}|| \\ \\ \lambda_4 = \lambda_1 + c||\boldsymbol{\eta}|| \end{array} \right. \tag{VI.16}$$

où la vitesse du son dans le gaz c vérifie : $c = \sqrt{\dfrac{\gamma p}{\rho}}$

La matrice $\mathcal{A}$ s'écrit : $\mathcal{A} = T^{-1}\Lambda T$ avec $\Lambda = \text{diag}(\lambda_k)$ matrice des valeurs propres λ_k de $\mathcal{A}$ et T matrice de passage inversible. Les expressions de $\mathcal{A}$ et de T sont données dans l'Annexe D.

VI.4 Approximation spatiale.

Là encore, nous utilisons une méthode de volumes finis à partir d'une discrétisation en triangles du domaine de calcul. La formulation variationnelle a déjà été décrite dans la section III.3.1 du chapitre III. La figure (VI.1) montre la construction d'une cellule en volumes finis pour un maillage triangulaire non structuré.

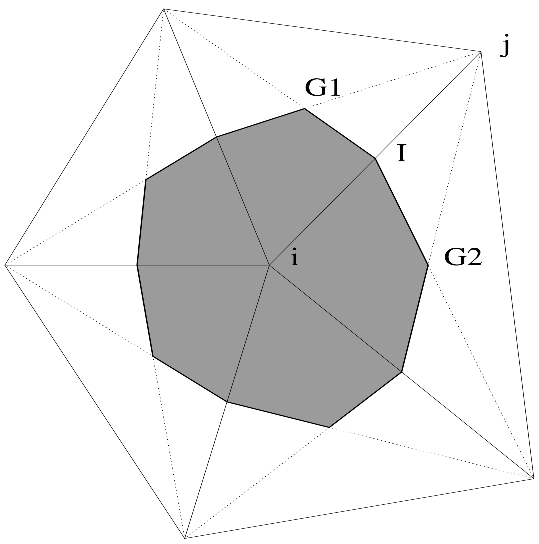


FIG. VI.1 – *Construction d'une cellule pour un maillage triangulaire.*

VI.4.1 Calcul des flux.

Nous notons $\boldsymbol{\nu_{ij}}$ la normale extérieure à l'interface ∂C_{ij} entre deux cellules C_i et C_j.

L'approximation du terme $\displaystyle\int_{\partial C_{ij}} \mathbb{F}(\mathbf{W}).\boldsymbol{\nu_{ij}}\ d\sigma$ se fait à l'aide d'une fonction de flux numérique:

$$\boldsymbol{\Phi}_{ij} = \Phi(\mathbf{W}_i, \mathbf{W}_j, \boldsymbol{\eta}) \tag{VI.17}$$

où $\boldsymbol{\eta} = (\eta_1, \eta_2) = \displaystyle\int_{\partial C_{ij}} \boldsymbol{\nu_{ij}}\ d\sigma$.

Pour évaluer ce flux, on peut, comme dans le cas scalaire, utiliser une méthode de Godunov. Cependant, l'extension de cette méthode au cas d'un système, en deux dimension d'espace, s'avère trop coûteuse. C'est pourquoi des schémas du même type mais résolvant des problèmes de Riemann approchés ont été développés : on pense en particulier au schéma de Roe, explicité ci-dessous, ou à ceux de Van Leer ou d'Osher, dont on ne parlera pas ici.

Le flux de Roe, basé sur une linéarisation du système (VI.10), se construit naturellement comme un solveur de Riemann local dans la direction de la normale $\boldsymbol{\eta}$:

$$\left\{ \frac{\partial \mathbf{W}}{\partial t} + \frac{\partial}{\partial \boldsymbol{\eta}} \left(\mathbb{F}(\mathbf{W}).\boldsymbol{\eta} \right) = 0 \right.$$

$$\mathbf{W}(\mathbf{x}, t) = \begin{cases} W_i & \text{si } \ \mathbf{x} \in C_i \\ W_j & \text{si } \ \mathbf{x} \in C_j \end{cases}$$

Il s'écrit :

$$\boldsymbol{\Phi}_{ij} = \boldsymbol{\Phi}(\mathbf{W}_i, \mathbf{W}_j, \boldsymbol{\eta}) = \frac{\mathcal{F}\left(\mathbf{W}_i, \boldsymbol{\eta}\right) + \mathcal{F}\left(\mathbf{W}_j, \boldsymbol{\eta}\right)}{2} - d^R\left(\mathbf{W}_i, \mathbf{W}_j, \boldsymbol{\eta}\right)\ , \tag{VI.18}$$

$$d^R\left(\mathbf{W}_i, \mathbf{W}_j, \boldsymbol{\eta}\right) = \gamma_c \mid \mathcal{R}\left(\mathbf{W}_i, \mathbf{W}_j, \boldsymbol{\eta}\right) \mid \frac{(\mathbf{W}_j - \mathbf{W}_i)}{2}\ , \tag{VI.19}$$

où $\mathcal{R}\left(\mathbf{W}_i, \mathbf{W}_j, \boldsymbol{\eta}\right)$ est appelée matrice de Roe.
Le paramètre γ_c est un paramètre de décentrage : lorsque $\gamma_c = 1$, le flux $\boldsymbol{\Phi}_{ij}$ est décentré; quand $\gamma_c = 0$, il est centré. A l'ordre un, nous utilisons des fonctions de flux décentrées pour des raisons de stabilité. A un ordre supérieur, nous pourrons utiliser des schémas centrés, qui ont l'avantage de ne pas introduire de diffusion numérique.

La matrice $\mathcal{R}\left(\mathbf{W}_i, \mathbf{W}_j, \boldsymbol{\eta}\right)$ possède certaines propriétés :

$$\bullet\ \mathcal{R}\left(\mathbf{W}_i, \mathbf{W}_j, \boldsymbol{\eta}\right)\left(\mathbf{W}_j - \mathbf{W}_i\right) = \mathcal{F}\left(\mathbf{W}_j, \boldsymbol{\eta}\right) - \mathcal{F}\left(\mathbf{W}_i, \boldsymbol{\eta}\right)\ , \tag{VI.20}$$

qui est en fait une linéarisation des relations de saut,

$$\bullet\ \mathcal{R}(\mathbf{W}_i, \mathbf{W}_j, \boldsymbol{\eta}) \longrightarrow \mathcal{F}'(\mathbf{W}_i, \boldsymbol{\eta}) \ \text{ lorsque } \ \mathbf{W}_j \longrightarrow \mathbf{W}_i \tag{VI.21}$$

assurant la consistance du schéma numérique, et enfin :

- $\mathcal{R}(\mathbf{W}_i, \mathbf{W}_j, \boldsymbol{\eta})$ doit être diagonalisable à valeurs propres réelles.

La propriété (VI.20) a plusieurs conséquences intéressantes; elle conduit à d'autres expressions plus simples du flux numérique :

$$\Phi^R\left(\mathbf{W}_i,\mathbf{W}_j,\boldsymbol{\eta}\right)=\mathcal{F}\left(\mathbf{W}_j,\boldsymbol{\eta}\right)-\frac{1}{2}\left[\mathcal{R}\left(\mathbf{W}_i,\mathbf{W}_j,\boldsymbol{\eta}\right)+\gamma_c\mid\mathcal{R}\left(\mathbf{W}_i,\mathbf{W}_j,\boldsymbol{\eta}\right)\mid\right]\left(\mathbf{W}_j-\mathbf{W}_i\right)\ , \tag{VI.22}$$

ou encore

$$\Phi^R\left(\mathbf{W}_i,\mathbf{W}_j,\boldsymbol{\eta}\right)=\mathcal{F}\left(\mathbf{W}_i,\boldsymbol{\eta}\right)+\frac{1}{2}\left[\mathcal{R}\left(\mathbf{W}_i,\mathbf{W}_j,\boldsymbol{\eta}\right)-\gamma_c\mid\mathcal{R}\left(\mathbf{W}_i,\mathbf{W}_j,\boldsymbol{\eta}\right)\mid\right]\left(\mathbf{W}_j-\mathbf{W}_i\right). \tag{VI.23}$$

C'est l'expression (VI.22) que nous avons choisie et que nous détaillerons dans la suite.

La matrice $\mathcal{R}$ n'est pas unique, toutefois Roe a proposé de définir $\mathcal{R}$ comme suit :

$$\mathcal{R}\left(\mathbf{W}_i,\mathbf{W}_j,\boldsymbol{\eta}\right)=\mathcal{A}\left(\tilde{\mathbf{W}},\boldsymbol{\eta}\right)\ ,$$

où $\mathcal{A}$ est la matrice jacobienne du flux évaluée pour l'état $\tilde{\mathbf{W}}$ qui est en fait la "moyenne de Roe" des deux états $\mathbf{W}_i$ et $\mathbf{W}_j$.
C'est cette matrice de Roe que nous prendrons pour le calcul du flux numérique de la phase gazeuse.

On définit l'état $\tilde{\mathbf{W}}$ par les relations suivantes :

$$\mathbf{W}_i=\begin{pmatrix}\rho_i\\ \rho_i u_i\\ \rho_i v_i\\ E_i\end{pmatrix}\ ,\quad \mathbf{W}_j=\begin{pmatrix}\rho_j\\ \rho_j u_j\\ \rho_j v_j\\ E_j\end{pmatrix}\ ,\quad \tilde{\mathbf{W}}=\begin{pmatrix}\tilde{\rho}\\ \tilde{\rho u}\\ \tilde{\rho v}\\ \tilde{E}\end{pmatrix}\ , \tag{VI.24}$$

avec:

$$\tilde{\rho}=\frac{\sqrt{\rho_i}\rho_i+\sqrt{\rho_j}\rho_j}{\sqrt{\rho_i}+\sqrt{\rho_j}}\ ,$$

$$\tilde{u}=\frac{\sqrt{\rho_i}u_i+\sqrt{\rho_j}u_j}{\sqrt{\rho_i}+\sqrt{\rho_j}}\ ,$$

$$\tilde{v}=\frac{\sqrt{\rho_i}v_i+\sqrt{\rho_j}v_j}{\sqrt{\rho_i}+\sqrt{\rho_j}}\ ,$$

$$\tilde{H}=\frac{\sqrt{\rho_i}H_i+\sqrt{\rho_j}H_j}{\sqrt{\rho_i}+\sqrt{\rho_j}}\ ,$$

où $H=\dfrac{E+p}{\rho}$ représente l'enthalpie par unité de masse.

Avec cette formulation où $\mathbf{W}_i$ et $\mathbf{W}_j$ sont constants sur leur cellule, le schéma (VI.18),(VI.19) est précis au premier ordre en espace.

VI.4.2 Traitement des conditions aux limites.

On utilise en général deux sortes de conditions aux limites sur $\Gamma = \partial\Omega = \Gamma_b \cup \Gamma_\infty$: une condition de glissement sur la paroi Γ_b et des conditions sur le bord infini Γ_∞. L'équation (VI.10) s'écrit en incluant les termes de bord :

$$\begin{aligned}
\mathrm{Aire}\, C_i\, (\mathbf{W}_t)_i \quad = \quad & - \sum_{j \in K(i)} \int_{\partial C_{ij}} \mathbb{F}(\mathbf{W}).\boldsymbol{\nu_{ij}}\, d\sigma \quad < 1 > \\
& - \int_{\partial C_i \cap \Gamma_b} \mathbb{F}(\mathbf{W}).\boldsymbol{n}\, d\sigma \qquad < 2 > \\
& - \int_{\partial C_i \cap \Gamma_\infty} \mathbb{F}(\mathbf{W}).\boldsymbol{n}\, d\sigma \qquad < 3 >
\end{aligned} \tag{VI.25}$$

où $\mathbf{n}$ représente la normale unitaire sortante sur le segment de bord.

Traitement de la paroi.

On introduit une condition de glissement sur la paroi : $\mathbf{V}.\mathbf{n} = 0$. L'expression du flux sur le bord devient :

$$\mathcal{F}(\mathbf{W}_i, \mathbf{n}) = \begin{pmatrix} 0 \\ p_i n_x \\ p_i n_y \\ 0 \end{pmatrix};$$

où p_i est la pression au noeud i, considérée comme constante sur tout le segment de bord. L'intégrale de pression s'écrit alors :

$$\boldsymbol{\Phi}_{ib}(\mathbf{W}_i, \boldsymbol{\eta}) = \int_{\partial C_i \cap \Gamma b} \mathcal{F}\,(\mathbf{W}, \mathbf{n})\, d\sigma \;=\; \begin{pmatrix} 0 \\ p_i \eta_x \\ p_i \eta_y \\ 0 \end{pmatrix}; \tag{VI.26}$$

avec

$$\boldsymbol{\eta} = \int_{\partial C_i \cap \Gamma b} \mathbf{n} d\sigma$$

La condition de glissement est ainsi vérifiée faiblement seulement.

Conditions à l'infini.

Il reste à évaluer l'intégrale des flux convectifs sur les bords correspondant aux frontières infinies amont et aval.

On se donne un champ $\mathbf{W}_\infty$ représentatif de l'écoulement à l'extérieur du domaine Ω_h. Le problème de Riemann défini par les valeurs $\mathbf{W}_i$ à l'intérieur du domaine Ω_h et $\mathbf{W}_\infty$, est résolu en utilisant, comme pour les flux internes au domaine de calcul, un solveur de Riemann approché:

$$\int_{\partial C_i \cap \Gamma_\infty} \mathcal{F}(\mathbf{W},\mathbf{n})\,d\sigma = \Phi(\mathbf{W}_i,\mathbf{W}_\infty,\boldsymbol{\eta}_{i\infty}) = \mathbf{\Phi}_{i\infty} \ , \tag{VI.27}$$

avec:

$$\boldsymbol{\eta}_{i\infty} = \int_{\partial C_i \cap \Gamma_\infty} \mathbf{n} d\sigma \ . \tag{VI.28}$$

On utilise une décomposition de flux de type Steger-Warming : les échanges avec le milieu extérieur sont calculés à partir de l'expression suivante:

$$\mathbf{\Phi}^{SW}(\mathbf{W}_i,\mathbf{W}_\infty,\boldsymbol{\eta}_{i\infty}) = \mathcal{A}^+(\mathbf{W}_i,\boldsymbol{\eta}_{i\infty})\,\mathbf{W}_i + \mathcal{A}^-(\mathbf{W}_i,\boldsymbol{\eta}_{i\infty})\,\mathbf{W}_\infty \ . \tag{VI.29}$$

où $\mathcal{A}^+(\mathbf{W}_i,\boldsymbol{\eta}_{i\infty})$ et $\mathcal{A}^-(\mathbf{W}_i,\boldsymbol{\eta}_{i\infty})$ sont respectivement la partie positive et la partie négative de la matrice jacobienne du flux gazeux $\mathcal{A}(\mathbf{W}_i,\boldsymbol{\eta}_{i\infty})$.
Cett décomposition est en fait fondée sur l'homogénéité d'ordre un de la fonction flux :

$$\mathcal{F}(\mathbf{W},\mathbf{n}) = \mathcal{A}(\mathbf{W},\mathbf{n})\mathbf{W}$$

Comme $\mathcal{A}(\mathbf{W},\boldsymbol{\eta}) = \mathcal{A}^+(\mathbf{W},\boldsymbol{\eta}) + \mathcal{A}^-(\mathbf{W},\boldsymbol{\eta})$, il s'ensuit que le flux de Steger-Warming est bien consistant avec les équations d'Euler : $\Phi(\mathbf{W},\mathbf{W},\mathbf{n}) = \mathcal{F}(\mathbf{W},\mathbf{n})$.

VI.5 Approximation d'ordre supérieur en temps et en espace.

Comme dans le cas linéaire, pour l'équation d'advection et pour le système de Maxwell, l'extension à l'ordre deux en espace se fait à l'aide des β-schémas (III.10), qui permettent de définir de nouvelles valeurs interpolées $\mathbf{W}_{ij}$ et $\mathbf{W}_{ji}$ sans modifier la fonction de flux $\mathbf{\Phi}$. Comme nous l'avons déjà vu, dans le cas linéaire avec un maillage structuré, le schéma est d'ordre trois avec $\beta = \dfrac{1}{3}$ et $\gamma_c = 1$, et d'ordre quatre avec $\beta = \dfrac{1}{3}$ et $\gamma_c = 0$ [6, 5]. Comme il serait dommage d'utiliser des approximations en temps d'ordre inférieur, on choisit donc des schémas "Runge-Kutta non linéaires". Les schémas d'ordre deux et trois sont donnés dans la section VI.2.2, on donne simplement la matrice des coefficients $(\alpha_{jl})_{j,l=1,\dots,4}$ dans le cas d'un schéma d'ordre quatre.

$$(\alpha_{jl})_{k=4} = \begin{pmatrix} 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1/6 & 1/3 & 1/3 & 1/6 \end{pmatrix} \tag{VI.30}$$

Nous avons vu dans les chapitres II et III que dans le cas linéaire monodimensionnel, le schéma d'ordre trois est stable en norme L^2 pour un nombre de Courant $CFL \leq 1.63$ et que celui d'ordre quatre l'est pour un $CFL \leq 2.06$.

VI.6 Schémas implicites.

Nous sommes par la suite intéressés par le calcul de solutions stationnaires obtenues par une approche instationnaire. Dans ce cas, les schémas implicites s'avèrent particulièrement intéressants car ils n'ont pas en général de restriction sur le pas de temps. Nous utilisons des schémas implicites linéarisés d'ordre un en temps et d'ordre supérieur en espace, avec une méthode de Jacobi comme méthode de relaxation.

Le schéma implicite d'ordre un en temps s'écrit :

$$\frac{\mathbf{W}^{n+1}-\mathbf{W}^{n}}{\Delta t}+\mathbf{H}\left(\mathbf{W}^{n+1}\right)=0 \tag{VI.31}$$

où $\mathbf{H}(\mathbf{W})$ approche le terme $\nabla.\mathbb{F}(\mathbf{W})$. La linéarisation du flux à l'aide d'un développement de Taylor à l'ordre un en temps, lorsque $\mathbf{H}(\mathbf{W})$ est différentiable, s'écrit :

$$\mathbf{H}\left(\mathbf{W}^{n+1}\right)=\mathbf{H}\left(\mathbf{W}^{n}\right)+\mathbf{H}'\left(\mathbf{W}^{n}\right)\left(\mathbf{W}^{n+1}-\mathbf{W}^{n}\right)+O\left(\Delta t^{2}\right)\ , \tag{VI.32}$$

où $\mathbf{H}'(\mathbf{W}^n)$ désigne la matrice Jacobienne de $\mathbf{H}(\mathbf{W}^n)$.
On obtient alors la version linéarisée du schéma implicite précédent sous la forme d'un δ-schéma (on note $\delta\mathbf{W}=\mathbf{W}^{n+1}-\mathbf{W}^{n}$).

$$\left(\frac{Id}{\Delta t}+\mathbf{H}'\left(\mathbf{W}^{n}\right)\right)\delta\mathbf{W}=-\mathbf{H}\left(\mathbf{W}^{n}\right)\ , \tag{VI.33}$$

La précision des solutions *stationnaires* obtenues avec le schéma (VI.33) est déterminée par celle du second membre. Pour obtenir des solutions stationnaires précises à l'ordre deux, il suffit d'utiliser (III.10) comme expression du flux dans le second membre.

Toutefois lorsque le flux $\mathbf{H}(\mathbf{W})$ n'est pas différentiable (c'est le cas du flux de Roe), on introduit un opérateur linéaire que nous notons $\mathbf{P}^n$ qui approche la matrice Jacobienne $\mathbf{H}'(\mathbf{W}^n)$.
Le système (VI.33) s'écrit alors:

$$\left(\frac{Id}{\Delta t}+\mathbf{P}^{n}\right)\delta\mathbf{W}=-\mathbf{H}\left(\mathbf{W}^{n}\right)\ . \tag{VI.34}$$

Lorsque le flux est différentiable et que $\Delta t\rightarrow+\infty$, le schéma (VI.33) est une méthode de Newton pour le calcul de solutions stationnaires. La convergence est théoriquement quadratique au voisinage de la solution. Lorsque $\mathbf{P}^n$ est différent du vrai jacobien $\mathbf{H}'(\mathbf{W}^n)$, le schéma (VI.33) devient une méthode de Newton modifiée : la matrice du système linéaire joue en fait le rôle de préconditionneur, et la convergence est seulement linéaire.

VI.6.1 Linéarisation des flux convectifs:

D'après l'expression du flux numérique (VI.18),(VI.19), on a :

$$\Phi_{ij}^{n+1} = \Phi\left(\mathbf{W}_i^n, \mathbf{W}_j^n, \mathbf{W}_i^{n+1}, \mathbf{W}_j^{n+1}, \boldsymbol{\eta}\right)$$

On linéarise le flux de la manière suivante :

$$\Phi_{ij}^{n+1} = \Phi_{ij}^n + \mathcal{A}(\mathbf{W}_i^n, \boldsymbol{\eta})\delta\mathbf{W}_i - \tilde{\mathcal{A}}^+(\mathbf{W}_i^n, \mathbf{W}_j^n, \boldsymbol{\eta})(\delta\mathbf{W}_i - \delta\mathbf{W}_j)$$

avec $\mathcal{A}(\mathbf{W}_i^n, \boldsymbol{\eta})$ la matrice jacobienne du flux, et $\tilde{\mathcal{A}}^+(\mathbf{W}_i^n, \mathbf{W}_j^n, \boldsymbol{\eta})$ la partie positive de la matrice de Roe entre les états $\mathbf{W}_i$ et $\mathbf{W}_j$ prise à l'instant t^n.

VI.6.2 Linéarisation sur le bord $\delta\Omega$:

Comme dans le cas explicite, le calcul sur les bords se décompose en deux parties :

– <u>Sur Γ_∞ :</u>

Nous calculons un flux décentré de Steger-Warming entre la valeur $\mathbf{W}_i^n$ au noeud i et sa valeur imposée au bord, $\mathbf{W}_\infty$.
Ce flux se linéarise donc uniquement sur la valeur $\mathbf{W}_i$, il s'écrit :

$$\Phi_{i\infty}^{n+1} = \Phi_{i\infty}^n + \mathcal{A}^+(\mathbf{W}_i^n, \boldsymbol{\eta})\delta\mathbf{W}_i$$

où $\mathcal{A}^+(\mathbf{W}_i^n, \boldsymbol{\eta})$ désigne la partie positive de la matrice jacobienne du flux $\Phi_{i\infty}$.

– <u>Sur Γ_b :</u>

Sur la paroi Γ_b (bord glissant), on différencie le flux $\Phi_{ib}(\mathbf{W}, \boldsymbol{\eta})$ construit à partir de la condition de glissement $\mathbf{V}.\mathbf{n} = 0$, on obtient son jacobien $\mathcal{A}_b(\mathbf{W}, \boldsymbol{\eta})$.

$$\mathcal{A}_b(\mathbf{W}, \boldsymbol{\eta}) = (\gamma - 1)\begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2}(u^2+v^2)\eta_x & -u\eta_x & -v\eta_x & \eta_x \\ \frac{1}{2}(u^2+v^2)\eta_y & -u\eta_y & -v\eta_y & \eta_y \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

VI.6.3 Méthode de résolution.

L'algorithme implicite s'écrit sous la forme d'un schéma à deux phases : une phase explicite qui prend en compte les données physiques du problème à résoudre,

$$\mathrm{RHS} = -\left(\sum_{j\in K(i)} \Phi_{ij}^n + \Phi_{i\infty}^n + \Phi_{ib}^n\right) \tag{VI.35}$$

puis une phase implicite qui résout le système suivant :

$$\left\{\begin{array}{l} \dfrac{\text{aire } \mathrm{C_i}}{\Delta t}\delta \mathbf{W}_i + \displaystyle\sum_{j\in K(i)} \Big(\mathbf{A}(\mathbf{W}_i^n, \boldsymbol{\eta})\delta\mathbf{W}_i - \tilde{\mathbf{A}}^+(\mathbf{W}_i^n, \mathbf{W}_j^n, \boldsymbol{\eta})(\delta\mathbf{W}_i - \delta\mathbf{W}_j)\Big) \\ \qquad\qquad + \mathbf{A}^+(\mathbf{W}_i^n, \boldsymbol{\eta}_{i\infty})\delta\mathbf{W}_i + \ \mathbf{A}_b(\mathbf{W}_i^n, \boldsymbol{\eta})\delta\mathbf{W}_i = \ \mathrm{RHS} \\ \mathbf{W}_i^{n+1} = \mathbf{W}_i^n + \delta\mathbf{W}_i \end{array}\right. \tag{VI.36}$$

Le calcul de $\mathbf{W}^{n+1}$ à partir de $\mathbf{W}^n$ nécessite la résolution d'un système linéaire à chaque pas de temps qui s'écrit de manière simplifiée :

$$(\frac{\text{aire } \mathrm{C_i}}{\Delta t} Id + \mathcal{B})\delta\mathbf{W} = RHS \tag{VI.37}$$

Nous cherchons à éviter un stockage prohibitif de la matrice $\mathcal{B}$, c'est pourquoi nous utilisons comme méthode de résolution la méthode de Jacobi,

$$\left\{\begin{array}{l} \delta\mathbf{W}_i^{\alpha+1} = \left(\dfrac{\text{aire } \mathrm{C_i}}{\Delta t} + \mathcal{B}_{ii}\right)^{-1} RHS - \displaystyle\sum_{j=1}^{n} \left(\dfrac{\text{aire } \mathrm{C_i}}{\Delta t} + \mathcal{B}_{ii}\right)^{-1} \mathcal{B}_{ij}\ \delta\mathbf{W}_j^\alpha \\ \delta\mathbf{W}_i^0 = 0 \end{array}\right. \tag{VI.38}$$

et nous obtenons la solution au temps $\mathrm{n}+1$:

$$\mathbf{W}_i^{n+1} = \mathbf{W}_i^n + \delta\mathbf{W}_i \tag{VI.39}$$

Dans cette méthode, les blocs diagonaux $\mathcal{B}_{ii}$ sont inversés puis stockés. Les termes extra-diagonaux $\mathcal{B}_{ij}$ sont recalculés à chaque itération de Jacobi mais ne sont pas stockés.

VI.6.4 Convergence vers la solution stationnaire.

On peut montrer que le schéma implicite linéarisé décentré d'ordre un est linéairement inconditionnellement stable [18, 19].

Pour accélérer la convergence vers l'état stationnaire, nous utilisons un pas de temps local par cellules. Pour chaque cellule C_i, nous avons :

$$\Delta t_i^n = CFL\ \min_T\left(\frac{h_i}{\max\limits_{k=1,3}\lambda_{\max}^k}\right)$$

où h_i la hauteur minimale issue du triangle T et $\lambda^k_{\max}$ est la vitesse d'onde maximale à l'instant n du kième noeud de T.

On ne peut pas calculer les solutions stationnaires des équations d'Euler en commençant le processus itératif avec de grands pas de temps.
On note RES^n le résidu du processus itératif; il constitue un test d'arrêt pour la convergence vers la solution stationnaire. On choisit :

$$\mathrm{RES}^n = \frac{||\mathbf{G}(\rho^n)||_1}{||\mathbf{G}(\rho^0)||}$$

où $||.||_1$ désigne la norme l_1.

Nous verrons dans le chapitre suivant que la formulation en β-schémas ne préserve pas la monotonie de l'interpolation : elle peut créer ou amplifier des oscillations parasites, néfastes pour la stabilité et la convergence dans le cas stationnaire. Pour remédier à ce problème, la méthode classique consiste à introduire des limiteurs de flux afin d'obtenir des schémas T.V.D. Dans le chapitre VII, nous étudions différents limiteurs et nous proposons une nouvelle méthode de limitation que nous appliquons aux équations d'Euler, dans le cas stationnaire et instationnaire.

Chapitre VII

UN NOUVEAU LIMITEUR DE FLUX POUR LES ÉQUATIONS D'EULER.

Réalisé avec Serge Piperno*.

*CERMICS-INRIA, 06902 Sophia-Antipolis Cedex, France

Ce chapitre, qui est le fruit d'une collaboration avec Serge Piperno, est rédigé en anglais, dans le but d'en soumettre la majeure partie à publication.

VII.1 Introduction

The aim of this paper is the construction of a new limiter for the resolution of hyperbolic conservation laws. Some limiters have been proposed in the past years [H, I] in order to obtain TVD (Total Variation Diminishing) schemes, a notion introduced by Harten [B]. Our first intention was a comparison of existing limiters, as had already been done [G] in the main frame of conservative finite-volume methods using the MUSCL (Monotone Upwind Schemes for Conservation Laws) extension. However, some third-order TVD schemes were proposed later [F, E], which do not show an optimal behaviour. As a result, we tried to explain the performances of these limiters, aiming for the construction of a new one.

Following the analysis of spurious oscillations presented by Sweby [G], one can see high accuracy schemes as a perturbation of the first-order upwind scheme. This perturbation can be seen as an antidiffusive term which can provoke oscillations near discontinuities. In this case, the anti-diffusion term is too important and should be "limited" by multiplication by a limiter function φ, which depends on a the ratio r of consecutive slopes.

Introducing an upwinding parameter β in the MUSCL method, as done in [F], we notice that all unlimited β-schemes do not have the same behaviour near discontinuities. Indeed, when considering the advection of a square wave, spurious oscillations in discontinuities are produced for the fully-centered ($\beta = 0$)-scheme when $r = 0$, and for the fully-upwind ($\beta = 1$)-scheme when $r = \pm\infty$. We choose to build a limiter which behaves like the unlimited ($\beta = 0$)-scheme when $r = \pm\infty$, and like the ($\beta = 1$)-scheme when $r = 0$, in order to produce no spurious oscillations.

The new limiter is first tested on the one-dimensional linear advection and Burgers equations. Its properties are evaluated on unsteady test cases, where an exact solution is known. We then make the extension to the hyperbolic system of inviscid Euler equations in two dimensions and perform some numerical tests on both unsteady and steady bidimensional solutions.

The paper is organized as follows. We first set the frame of this work, made of conservative finite-volume methods using high-order TVD schemes. In the first paragraph, we recall the need of flux limiters and their construction. We then introduce new limiters, which are shown to be more efficient than previous ones for linear and non-linear monodimensional hyperbolic equations. Finally, these methods using new flux limiters are extended to the two-dimensional non-linear hyperbolic system of inviscid Euler equations. This extension is not straightforward, but somewhat classical. The limiter is tested on unsteady shock tube test cases and a steady transonic flow around a NACA0012 airfoil.

VII.2 Existing high-order TVD schemes

In this section, we review existing high-order TVD schemes for monodimensional hyperbolic conservation equations, which write

$$\begin{cases} u_t + (f(u))_x = 0 \\ u(x,0) = u_0(x) \end{cases}, \tag{VII.1}$$

where f is a regular function called flux and subscripts stand for derivation. The linear advection equation is obtained when $f(u) = cu$ and Burgers equation appears when $f(u) = u^2/2$. These simple equations play an important role because Euler equations solutions take the same aspects.

We use a finite-volume formulation which is well-adapted to conservation equations. The method has been described in section VI.2.1, and we shall use the conservative scheme given in (VI.3).

It is well-known that accurate numerical fluxes produce spurious oscillations when applied to solutions with sharp discontinuities. Several monotonicity notions have been considered for numerical schemes. Harten [B] introduced TVD schemes:

Definition VII.2.1 *A numerical scheme is TVD (Total Variation Diminishing) if and only if*

$$TV(u^{n+1}) \leq TV(u^n),$$

where the discrete total variation $TV(u^n)$ is given by:

$$TV(u^n) = \sum_{i=-\infty}^{+\infty} |u_i^n - u_{i-1}^n|.$$

It is well known that such a scheme preserves monotonicity (*i.e.* the property $\forall i$, $\min(u_{i-1}, u_{i+1}) \leq u_i \leq \max(u_{i-1}, u_{i+1})$ is conserved by the scheme).

The basic tool for the construction of TVD schemes is also due to Harten [B].

Theorem VII.2.1 *A numerical scheme used for (VII.1) written using differences $\Delta u_{i+\frac{1}{2}}^n = u_{i+1}^n - u_i^n$ in the following way*

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x}\left(A_{i+\frac{1}{2}}^n \Delta u_{i+\frac{1}{2}}^n + B_{i-\frac{1}{2}}^n \Delta u_{i-\frac{1}{2}}^n\right), \tag{VII.2}$$

is TVD if

$$\forall i,\ \forall n, \quad A_{i+\frac{1}{2}} \leq 0 \leq B_{i+\frac{1}{2}}, \quad B_{i+\frac{1}{2}} - A_{i+\frac{1}{2}} \leq \frac{\Delta x}{\Delta t}. \tag{VII.3}$$

The proof is elementary and can be found in [B].

For example, the first-order upwind flux applied to the linear advection equation $u_t + cu_x = 0$, defined by

$$\phi_{i+\frac{1}{2}} = \phi(u_i, u_{i+1}) = c^+ u_i + c^- u_{i+1}, \text{ with } \begin{cases} c^+ &= \max(c, 0) \\ c^- &= \min(c, 0) \end{cases}, \tag{VII.4}$$

is TVD under the CFL condition $\nu \equiv |c|\Delta t/\Delta x \leq 1$.

For Burgers equation, the Godunov's scheme (VI.4) is shown to be TVD under the non-linear CFL-type condition (VI.6).

However, no linear TVD schemes can be more than first-order accurate [B]. Thus, non-linear flux limiters have been introduced to transform high-order accurate - non TVD - linear numerical fluxes into high-order accurate TVD non-linear numerical fluxes.

VII.2.1 The linear advection equation.

The extension to higher accuracy is made through the use of the MUSCL method [H]. The spatial part of the equation is handled with the same upwind numerical flux function as previously, but taken on interpolated states at the interface between cells on both sides. The linear interpolation increases the global accuracy of the spatial scheme. The numerical flux now writes:

$$\phi_{i+\frac{1}{2}} = c^+ u_{i+\frac{1}{2}^-} + c^- u_{i+\frac{1}{2}^+} \equiv \phi\left(u_{i+\frac{1}{2}^-}, u_{i+\frac{1}{2}^+}\right), \tag{VII.5}$$

where ϕ is the upwind flux function of (VII.4) and interpolated states are

$$u_{i+\frac{1}{2}^-} = u_i + \frac{1}{2}\left[(1-\beta)\Delta u_{i+\frac{1}{2}} + \beta \Delta u_{i-\frac{1}{2}}\right], \tag{VII.6}$$

$$u_{i+\frac{1}{2}^+} = u_{i+1} - \frac{1}{2}\left[(1-\beta)\Delta u_{i+\frac{1}{2}} + \beta \Delta u_{i+\frac{3}{2}}\right], \tag{VII.7}$$

where β is an upwinding parameter. The difference $\Delta u_{i+\frac{1}{2}}$ is given by

$$\Delta u_{i+\frac{1}{2}} = \Delta_-\{u_{i+1}\} = u_{i+1} - u_i. \tag{VII.8}$$

The previous interpolated states are more commonly defined in function of local slopes $p_{i+\frac{1}{2}} = \dfrac{\Delta u_{i+\frac{1}{2}}}{\Delta x}$ (Δx is assumed to be constant). They also write

$$u_{i+\frac{1}{2}^-} = u_i + \frac{\Delta x}{2}\left[(1-\beta)p_{i+\frac{1}{2}} + \beta p_{i-\frac{1}{2}}\right],$$

$$u_{i+\frac{1}{2}^+} = u_{i+1} - \frac{\Delta x}{2}\left[(1-\beta)p_{i+\frac{1}{2}} + \beta p_{i+\frac{3}{2}}\right].$$

We get centered slopes when $\beta = 0$, and upwind slopes when $\beta = 1$. This scheme is spatially at least second-order accurate, and third-order accurate if $\beta = 1/3$. The troncature error analysis is valid only if the solution u is sufficiently regular. In the case of the advection equation, the troncature error writes:

$$\epsilon^x = (1-3\beta)\frac{|c|\Delta x}{6}u_{xxx} + \beta\frac{|c|\Delta x}{4}u_{xxxx} + O(\Delta x^4). \tag{VII.9}$$

Used with linear Runge-Kutta time schemes, these schemes show out to be L^2-stable under CFL-like conditions. However, being linear and more than first order accurate, they can not be TVD [B].

The analysis of spurious oscillations was first presented by Sweby [G] and extended to MUSCL β-schemes by Spekreijse [F]. The basic idea is to see the previous high accuracy scheme as a perturbation of the first order upwind scheme. If we consider the advection equation with $c > 0$, the β-scheme used with the explicit forward-Euler time scheme writes:

$$u_i^{n+1} = u_i^n - \nu\Delta u_{i-\frac{1}{2}} - \nu\Delta_- \left\{\frac{1-\beta}{2}\Delta u_{i+\frac{1}{2}} + \frac{\beta}{2}\Delta u_{i-\frac{1}{2}}\right\}. \tag{VII.10}$$

The last term of the previous equation can be seen as an anti-diffusion term, which provokes spurious oscillations near sharp discontinuities. In order to avoid this, Sweby introduced a limiter function φ_i, and using notations of (VII.8), the limited scheme writes:

$$u_i^{n+1} = u_i^n - \nu \quad \Delta_- \left\{u_i + \varphi_i\left[\frac{1-\beta}{2}\Delta u_{i+\frac{1}{2}} + \frac{\beta}{2}\Delta u_{i-\frac{1}{2}}\right]\right\}. \tag{VII.11}$$

φ_i is chosen such that the accuracy is not lost, and the obtained scheme is TVD for a Courant number ν as high as possible. Classically, the limiters are constructed as functions of consecutive slopes fractions:

$$\varphi_i = \varphi(r_i) \quad \text{with} \quad r_i = \frac{\Delta u_{i-\frac{1}{2}}}{\Delta u_{i+\frac{1}{2}}}.$$

The order of the truncation error of the limited scheme can be estimated with no difficulty in areas where the derivative u_x does not vanish. Indeed, the argument of the function φ is close to 1 when $u_x \neq 0$. It is easy to verify that, if the function φ is regular enough, and if $\varphi(1+h) = 1 + O(h^p)$ for small h, the truncation error introduced by the addition of the limiter is a $O(h^{p+1})$. Thus, if the scheme is pth-order accurate, the limited scheme is also pth-order accurate (extrema excepted, i.e. when $u_x \neq 0$).

Harten's theorem VII.2.1 gives us a sufficient condition on φ and ν to have a TVD scheme.

Proposition VII.2.1 *If there exist three constants m_1, m_2 and M such that*

$$\left\{\begin{array}{lclclcl} \forall r \in I\!R & , & m_1 & \leq & \left(\frac{\beta}{2}+\frac{1-\beta}{2r}\right)\varphi(r) & \leq & M \\ \forall r \in I\!R & , & m_2 & \leq & \left(\frac{\beta}{2}r+\frac{1-\beta}{2}\right)\varphi(r) & \leq & 1+m_1 \end{array}\right. , \tag{VII.12}$$

then the scheme (VII.11) is TVD if

$$\nu \leq \frac{1}{1+M-m_2}. \tag{VII.13}$$

For the general case of advection ($c > 0$ or $c < 0$), the β-scheme can be limited in the same way. It is TVD under the same condition on φ and ν as in (VII.12-VII.13) if it writes:

$$\phi_{i+\frac{1}{2}} = c^+ u_{i+\frac{1}{2}^-} + c^- u_{i+\frac{1}{2}^+} \equiv \phi\left(u_{i+\frac{1}{2}^-}, u_{i+\frac{1}{2}^+}\right), \tag{VII.14}$$

with

$$\begin{array}{lcl} u_{i+\frac{1}{2}^-} & = & u_i \quad + \frac{\varphi_i^+}{2}\left[(1-\beta)\Delta u_{i+\frac{1}{2}} + \beta\Delta u_{i-\frac{1}{2}}\right], \\ u_{i+\frac{1}{2}^+} & = & u_{i+1} - \frac{\varphi_i^-}{2}\left[(1-\beta)\Delta u_{i+\frac{1}{2}} + \beta\Delta u_{i+\frac{3}{2}}\right], \end{array} \tag{VII.15}$$

where the limiters ϕ_i^+ et ϕ_i^- are given by

$$\begin{array}{lcl} \varphi_i^+ = \varphi(r_i) & \text{and} & r_i = \dfrac{\Delta u_{i-\frac{1}{2}}}{\Delta u_{i+\frac{1}{2}}}, \\ \varphi_i^- = \varphi(t_i) & \text{and} & t_i = \dfrac{\Delta u_{i+\frac{3}{2}}}{\Delta u_{i+\frac{1}{2}}}. \end{array} \tag{VII.16}$$

We present here some classical examples. First, the average of van Albada writes:

$$\varphi(r) = \frac{2r}{(r^2+1)} \quad (\text{if } r > 0, \text{ else } 0).$$

When used along with $\beta = 1/2$, it makes a second-accurate TVD scheme for $\nu \leq 0.6$ according to (VII.13). It was numerically observed that the use of the second order Runge-Kutta explicit time-scheme pushes this limit to $\nu \leq 1$.

Secondly, the limiter introduced by Spekreijse with $\beta = 1/3$, uses the following function (sketched on Figure VII.1 with Van Albada's):

$$\varphi(r) = \frac{3r^3 - 2r^2 + 3r}{2r^4 + 2} \quad (\text{if } r > 0, \text{ else } 0). \tag{VII.17}$$

This gives a third-order accurate TVD scheme for $\nu \leq 0.52$ according to (VII.13), and it was numerically observed that it is still TVD only for $\nu \leq 0.96$ when used with the third order-accurate explicit Runge-Kutta scheme (VI.9).

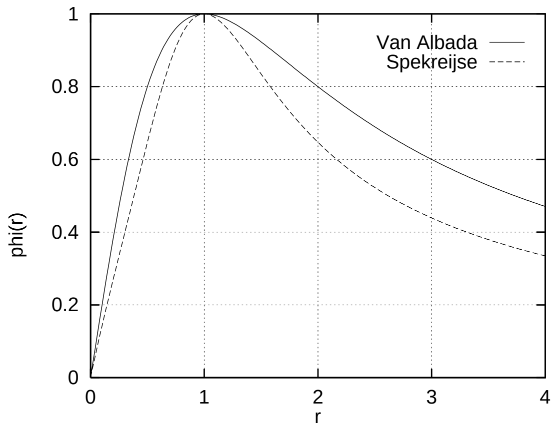


FIG. VII.1 – *Limiter functions : Van Albada's average and Spekreijse's.*

VII.2.2 The Burgers equation.

The preceding conservative schemes can be easily applied to Burgers equation, this time using Godunov's flux function (VI.4) instead of the upwind flux function (VII.14). The numerical flux writes now

$$\phi_{i+\frac{1}{2}} = \phi_{\text{Godunov}}(u_{i+\frac{1}{2}^-}, u_{i+\frac{1}{2}^+}), \tag{VII.18}$$

where limited interpolated states are given by (VII.15) and (VII.16).

In conclusion, limiters for hyperbolic scalar equations exist. However, the two most commonly used are not perfect. The limiter based on van Albada's average is only TVD for the second-order accurate $\beta = 1/2$ MUSCL scheme. And the limiter introduced by Spekreijse produces a third-order accurate TVD scheme which seems to lack some robustness, as will be shown on two-dimensional numerical results.

VII.3 New limiters for hyperbolic scalar equations

In this section, we construct two new limiters. The first one is constructed for an upwind flux function and a β-scheme with $\beta = 1/3$, like the limiter of Spekreijse. The second one is built for centered fluxes, in order to obtain a fourth-order accurate TVD scheme.

VII.3.1 Limiters for upwind fluxes

As it has already been said in the introduction, the construction of the new limiter is based on this remark : all unlimited β-schemes do not have the same behaviour near discontinuities. For linear advection, the MUSCL interpolation produces spurious oscillations in discontinuities where $r = 0$ for $\beta = 0$ and where $r = \pm\infty$ for $\beta = 1$ (see Figure VII.2).

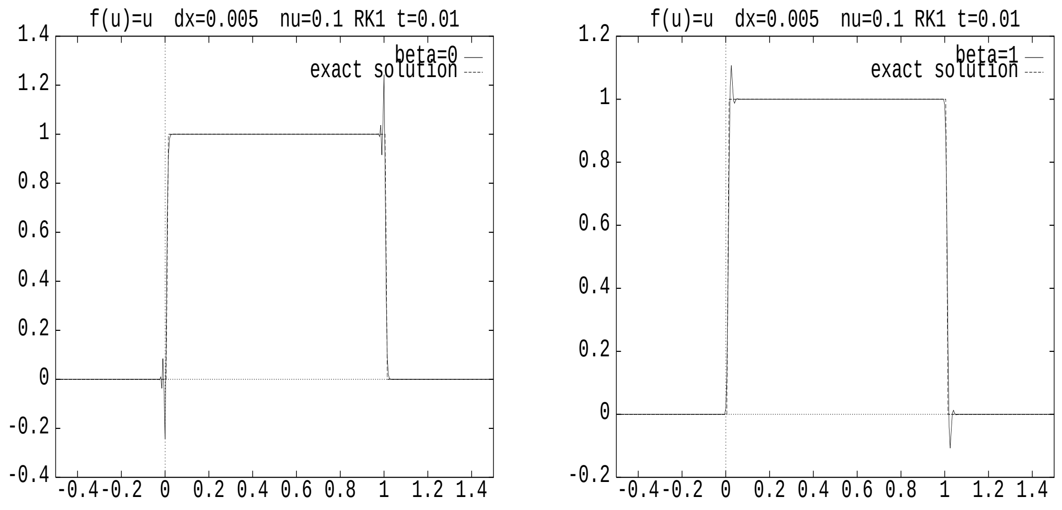


FIG. VII.2 – *Spurious oscillations for $\beta = 0$ and $\beta = 1$.*

The limited flux in (VII.11) can be rewritten as

$$\phi_{i+\frac{1}{2}} = u_i + \frac{1}{2}\Delta u_{i-\frac{1}{2}}\theta(r_i), \tag{VII.19}$$

with

$$\theta(r) = \varphi(r)\left[\beta + \frac{(1-\beta)}{r}\right]. \tag{VII.20}$$

We intend to build a limiter which behaves like the unlimited ($\beta = 0$)-scheme when $r = \pm\infty$ and like the ($\beta = 1$)-scheme when $r = 0$ (in order to produce no spurious oscillations on both sides of the square wave).

On the one hand, when $r \to \pm\infty$, we should have $\theta(r) \sim 1/(2r)$, which is not the case of Spekreijse's limiter with $\beta = 1/3$. For the value $\beta = 1/3$, this is achieved if φ is such that $\varphi(r) \sim 3/r$, when $r \to \pm\infty$.

On the other hand, when $r = 0$, the unlimited ($\beta = 1$)-scheme and the ($\beta = 1/3$)-scheme with Spekreijse's limiter have the same behaviour, i.e. $\theta(r) \to 1/2$ when $r \to 0$. For $\beta = 1/3$, φ must be chosen such that $\varphi(r) \sim 3r/2$, when $r \to 0$.

We also require that $\varphi(1+h) = 1 + O(h^3)$ when $h \to 0$ in order to have even less active limiters when the solution is regular.

We summarize below the three conditions which are required to obtain the function $\varphi(r)$:

- $\varphi(1+h) = 1 + O(h^3)$ when $h \to 0$.
- for $\beta = \dfrac{1}{3}$, $\varphi(r) = \dfrac{3r}{2}$ when $r \to 0$.

– for $\beta = \dfrac{1}{3}$, $\varphi(r) = \dfrac{3}{r}$ when $r \to \pm\infty$.

– we also ask for $\varphi(r) = 0$ if $r < 0$ (this restriction has no influence on Spekreijse limiter).

We construct the function φ such that φ is a polynomial in $[0, 1]$. Constraints are $\varphi(0) = 0$, $\varphi'(0) = 3/2$, $\varphi(1) = 1$, $\varphi'(1) = 0$, $\varphi''(1) = 0$. We select the unique polynom of degree four satisfaying these constraints. In $[1, +\infty]$, φ is a rational function of the form:

$$\varphi(r) = \frac{3r^2 + ar + b}{(r-1)^3 + 3r^2 + ar + b}.$$

All criteria are automatically satisfied if the denominator has no zero in $[1, +\infty]$ (including $r = 1$). Parameters are chosen in order to minimize the bound M and to maximize the bound m_2. We select $a = -6$ and $b = 19$.

The definition of the function is summed up as

$$\varphi(r) = \begin{cases} 0 & \text{if} \quad r < 0 \\ (3r^4 - 7r^3 + 3r^2 + 3r)/2 & \text{if} \quad r \leq 1 \\ (3r^2 - 6r + 19)/(r^3 - 3r + 18) & \text{if} \quad 1 \leq r \end{cases} \tag{VII.21}$$

and it is plotted on Figure VII.3 with the Spekreijse function.

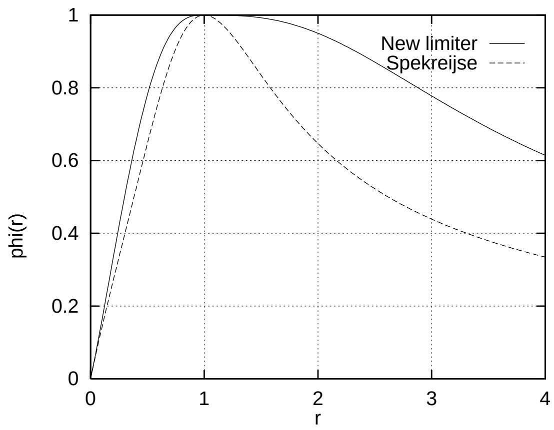


FIG. VII.3 – *The new limiter function compared with Spekreijse's.*

This new limiter gives a third-order accurate TVD scheme for $\nu \leq 0.6$ according to (VII.13), and it was numerically observed that it is still TVD for $\nu \leq 1.05$ when used with (VI.9). This is quite better than with the limiter function of Spekreijse (0.52 and 0.96).

VII.3.2 Limiters for centered fluxes

We now try to apply this method to centered numerical fluxes.

The advection equation.

Indeed, for the linear case, a centered flux combined with the MUSCL β-scheme with $\beta = 1/3$ gives a fourth order accurate linear (non TVD) scheme. We would like to build non-linear limiters which make this scheme TVD. The unlimited centered flux writes:

$$\phi^c_{i+\frac{1}{2}} = c \ \frac{u_{i+\frac{1}{2}^-} + u_{i+\frac{1}{2}^+}}{2}, \tag{VII.22}$$

where interpolated states are those of (VII.6-VII.7). When $\beta = 1/3$, the spatial truncation error is of fourth order, and no even term (diffusion term) appears in the infinite development of this truncation error. Indeed, it writes :

$$\epsilon^x = \frac{-|c|\Delta x^5}{30} u_{xxxxx} + O(\Delta x^6). \tag{VII.23}$$

Multiple limiting factors φ_i and ψ_i, are added to the numerical flux. The limited flux now writes (linear advection with $c = 1$):

$$\phi_{i+\frac{1}{2}} = u_i + \frac{\varphi_i}{2}\Delta u_{i+\frac{1}{2}} + \frac{\beta\psi_i}{4}\left(\Delta u_{i-\frac{1}{2}} - \Delta u_{i+\frac{3}{2}}\right). \tag{VII.24}$$

Limiters are chosen as functions of r_i and r_{i+1} of the following form:

$$\left\{\begin{array}{lcl} \varphi_i &=& \varphi(r_i), \\ \psi_i &=& \psi_1(r_i)\psi_2(r_{i+1}). \end{array}\right. \tag{VII.25}$$

Again, if the three functions φ, ψ_1 and ψ_2 are regular enough and almost equal to 1 near $r = 1$ (first and second derivatives equal to zero), then the limited scheme is fourth-order accurate. Harten's theorem helps showing that the spatial scheme, combined with the explicit forward Euler time scheme, results in a TVD scheme. We give the following proposition, easily obtained with Harten's theorem VII.2.1.

Proposition VII.3.1 *If there exist two constants M_1 and M_2 such that*

$$\left\{\begin{array}{rl} \forall r \in I\!R, & 0 \le \varphi(r) \le M_1, \\ \forall r \in I\!R, & \varphi(r) \le M_2\ |r|, \\ \forall (r,s) \in I\!R^2, & \beta\,|\psi_1(r)\psi_2(s)\,(1 - 1/(rs))|\max(1,|r|) \le 2 - M_1, \end{array}\right. \tag{VII.26}$$

then the scheme (VII.22,VII.24) using a first-order time accuracy is TVD if

$$\frac{1}{\nu} \ge 2 + \frac{M_2 - M_1}{2}. \tag{VII.27}$$

A possible choice for the limiter functions with $\beta = 1/3$ is the following (see Figure VII.4):

$$\left\{\begin{array}{ll} r \le 0 & \varphi(r) = \psi_1(r) = \psi_2(r) = 0 \\ 0 \le r \le 1 & \varphi(r) = 1 - (3r+1)(r-1)^4,\ \psi_1(r) = \psi_2(r) = \frac{r^3}{r^3+(1-r)^3} \\ 1 \le r & \varphi(r) = \psi_2(r) = 1,\ \psi_1(r) = \frac{1}{1+(r-1)^3} \end{array}\right. \tag{VII.28}$$

These functions verify criteria similar to those required for the limiters applied to the upwind scheme. For these limiter functions, the explicit forward Euler scheme is TVD if

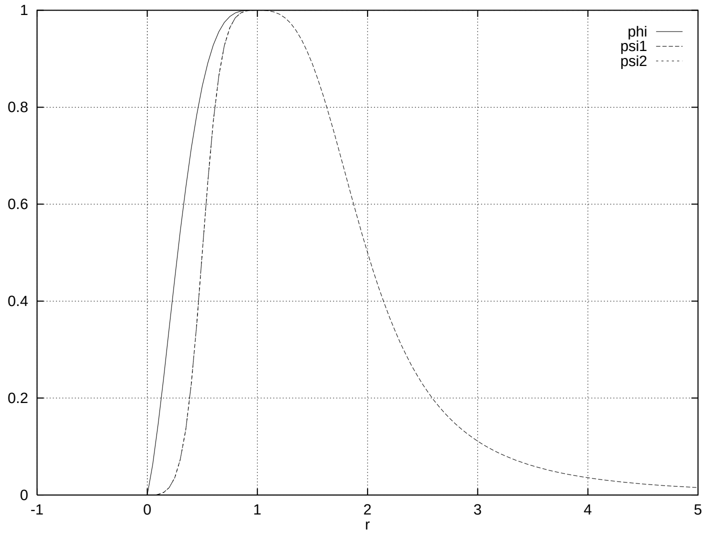


FIG. VII.4 – *Limiter fonctions φ, ψ_1 et ψ_2 for $\beta = 1/3$ (VII.28).*

$\nu \leq 0.47$. The global scheme is TVD for $\nu \leq 1.28$ if the time scheme is the fourth-order Runge-Kutta scheme (VI.9,VI.30).

We then have found a fourth order TVD scheme for the linear advection. However, this limiter is not valid for the general case of advection (c not necessarily positive) or even for Burgers equation. As a matter of fact, the expression (VII.25), the unique form we found, is directly related to the (positive) sign of c. The numerical meaning of these facts is simple. A centered flux can make no discrimination on the sign of the diffusive flux added by limitation. This discrimination can only be made with upwinding. For linear advection, we propose the use of the numerical flux of (VII.14) with the new interpolated states:

$$\begin{aligned} u_{i+\frac{1}{2}^-} &= u_i + \tfrac{\varphi(r_i)}{2}\Delta u_{i+\frac{1}{2}} + \tfrac{\beta}{4}\psi_1(r_i)\psi_2(r_{i+1})\left(\Delta u_{i-\frac{1}{2}} - \Delta u_{i+\frac{3}{2}}\right), \\ u_{i+\frac{1}{2}^+} &= u_{i+1} - \tfrac{\varphi(t_i)}{2}\Delta u_{i+\frac{1}{2}} - \tfrac{\beta}{4}\psi_1(t_i)\psi_2(t_{i-1})\left(\Delta u_{i-\frac{1}{2}} - \Delta u_{i+\frac{3}{2}}\right). \end{aligned} \tag{VII.29}$$

With no limitation, both previous states are equal, and the upwind flux between these two identical states is equal to the centered unlimited flux. The global scheme used with a forward Euler time scheme is again TVD for $\nu \leq 0.47$. However, this flux has at least the same computational cost as an upwind flux.

The Burgers equation.

The method is straightforwardly applicable to Burgers equation with the previous interpolated states (VII.29) and Godunov's flux function (VI.4).

VII.3.3 Numerical comparison of the limited schemes.

We now give a short comparison of performances of different limited TVD schemes. For the linear advection of a very smooth wave :

$$u_0(x) = 1024x^5(1-x)^5, \text{for } 0 \leq x \leq 1,$$

limited numerical fluxes were used along with Runge-Kutta time schemes of the same accuracy, with the maximal Courant number ν for which the global scheme is TVD. The spatial step Δx was adjusted to have the same computational cost. Data are presented on Table VII.1 (lines **A***), and L^1-norm of the error is shown on Figure VII.5.

	Time scheme	flux limiters	ν	Δx
A1	RK3 (VI.9)	Spekreijse (VII.17)	0.96	0.00125
A2	RK3 (VI.9)	new upwind (VII.21)	1.05	0.00116
A3	RK4 (VI.30)	new centered (VII.28)	1.28	0.00131
B1	RK2	Van albada	1.00	0.00185
B2	RK3 (VI.9)	Spekreijse (VII.17)	0.96	0.00225
B3	RK3 (VI.9)	new upwind (VII.21)	1.05	0.00213
B4	RK4 (VI.9)	new centered (VII.28)	1.28	0.00280

TAB. VII.1 – *Simulations parameters for linear advection and Burgers equation.*

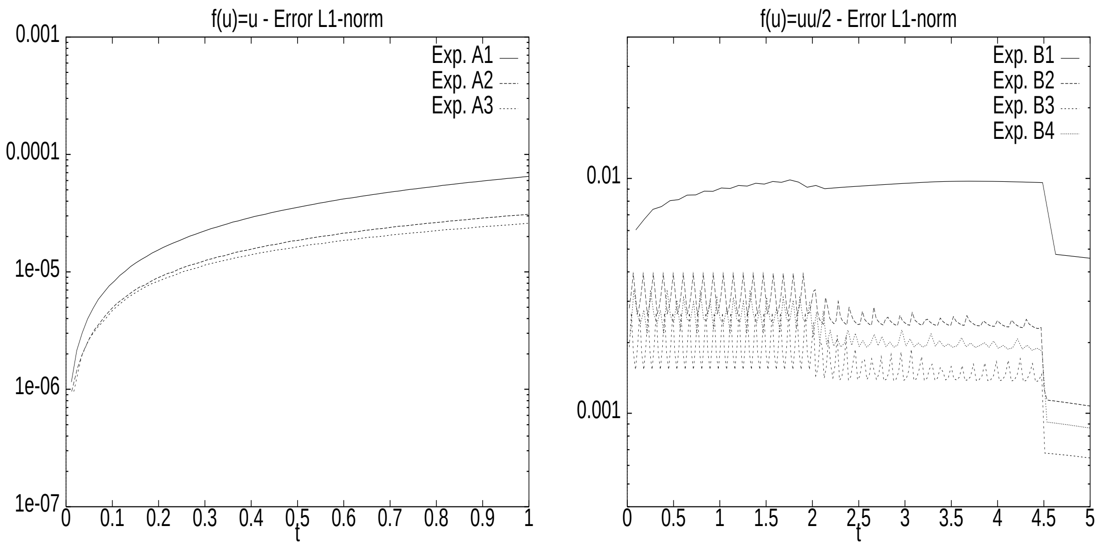


FIG. VII.5 – *Error L_1-norm for hyperbolic equations simulations.*

We see that new schemes are a little more efficient that the one using the limiter of

Spekreijse. Also, because of the high computational cost of each time step, the fourth-order scheme is not more efficient.

For Burgers equations, initial data was chosen as zero, except on $[-3; -2]$ and $[2; 3]$, where $u_0(x) = -sign(x)$. Simulation parameters are reported on Table VII.1 (lines **B***). The space step is adjusted to have the same computational cost. Again, the L^1-norm of the error is shown on Figure VII.5. The second order scheme is clearly the less efficient (in the sense that it is the less accurate for a given computational cost). The same conclusions as before hold: the upwind fluxes limited by Spekreijse are less efficient than new centered or upwind limited fluxes. And the fourth-order accurate limited centered fluxes has the drawback to need a four step fourth-order Runge-Kutta time scheme.

VII.4 Extension to two-dimensional Euler equations

The extension to two-dimensional Euler equations is not really complex. The third order limited flux presented in this paper is still based on an upwind flux function. We do not dispose of a cheap exact Riemann solver for Euler equations, like for linear advection (VII.4) or Burgers equation (VI.4). We will use the approximate Riemann solver of Roe [C], which can be seen as a two-dimensional upwind flux. However, two points must be considered. First, the solution state is now a vector (of conservative variables). Limitations will have to be active on all components of the vectors. Second, finite difference (VII.8) used for one-dimensional hyperbolic equations will have to be constructed in two dimensions. They will be replaced by gradients with a multiple choice for their calculation.

The vector of conservative variables W equal to ${}^t(\rho, \rho u, \rho v, E)$, where ρ, u, v and E respectively denote the density, the velocity along the x-axis and the y-axis, and the volumic total energy, is solution of inviscid Euler equations:

$$\left\{\begin{array}{c} \rho_t + (\rho u)_x + (\rho u)_y = 0, \\ (\rho u)_t + (\rho u^2 + p)_x + (\rho uv)_y = 0, \\ (\rho v)_t + (\rho uv)_x + (\rho v^2 + p)_y = 0, \\ E_t + [u(E+p)]_x + [v(E+p)]_y = 0 \end{array}\right.$$

The system is closed by the law of perfect gas giving the pressure P:

$$P = (\gamma - 1)\left[E - \frac{\rho}{2}(u^2 + v^2)\right]$$

with the value $\gamma = 1.4$ for a diatomic perfect gas. We shall not get into details on Roe's approximate Riemann solver, which has been detailed more than once. The flux between two state vectors W_i and W_j through a cell boundary of normal $\vec{\eta}$ will be denoted by $\Phi_{Roe}(W_i, W_j, \vec{\eta})$.

VII.4.1 MUSCL scheme for unstructured triangular meshes

The MUSCL scheme was first introduced by Van Leer [I] and then adapted to unstructured triangular meshes [A]. It produces a second-order accurate numerical flux based on Roe's approximate Riemann solver. The principle is quite the same as in one dimension. Assume W_i and W_j are average values of W in neighbouring finite volume cells C_i and C_j, respectively surrounding vertices S_i and S_j. We have to compute interpolated states W_{ij} and W_{ji} on both sides of the cells interface I_{ij} as described on Figure VII.6. These

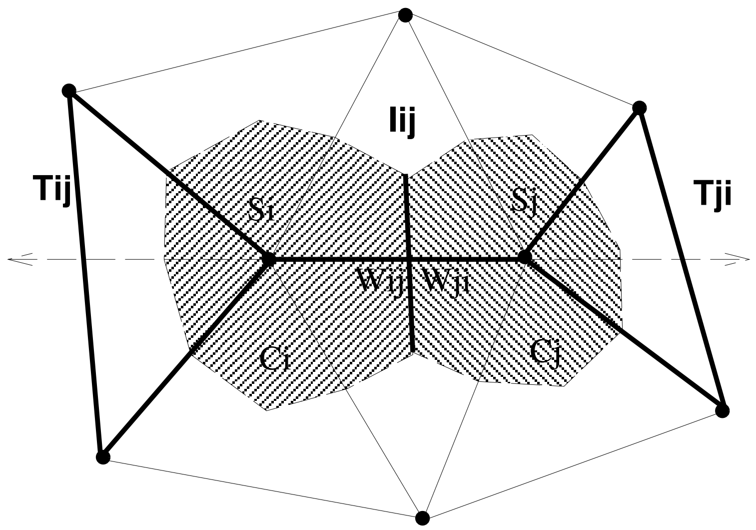


FIG. VII.6 – *MUSCL extension in two dimensions, and upwind and downwind triangles.*

interpolated states are directly corresponding to those of (VII.6-VII.7). They now write

$$\begin{aligned} W_{ij} &= W_i + \tfrac{1}{2}\vec{\nabla} W_i.\vec{ij}, \\ W_{ji} &= W_j - \tfrac{1}{2}\vec{\nabla} W_j.\vec{ij}, \end{aligned} \tag{VII.30}$$

where $\vec{ij} = \vec{S_iS_j}$ and the gradients $\vec{\nabla} W_i$ and $\vec{\nabla} W_j$ correspond to one dimensional slopes. The centered part of these gradients is simply taken such that

$$\vec{\nabla} W_i^{cent}.\vec{ij} = W_j - W_i. \tag{VII.31}$$

The upwind/downwind part of the gradients can be built on two possible choices. First, nodal averaged gradients $\vec{\nabla} W_i^{ave}$ are computed as the average on cell C_i of gradients of W on each triangle included in C_i. If N_k^T denotes the linear function equal to one on vertex k and equal to zero on both other vertices of triangle T, the nodal averaged gradients are given by:

$$\vec{\nabla} W_i^{ave} = \frac{1}{mes(C_i)} \sum_{T \subset C_i} \frac{mes(T)}{3} \sum_{k \in T} W_k \vec{\nabla} N_k^T. \tag{VII.32}$$

This nodal averaged gradients correspond in one dimension to the term $\frac{1}{2}(\Delta u_{i-1/2} + \Delta u_{i+1/2})$. The general MUSCL β-scheme based on nodal averaged gradients writes:

$$\vec{\nabla} W_i.\vec{ij} = (1-2\beta)\vec{\nabla} W_i^{cent}.\vec{ij} + 2\beta\vec{\nabla} W_i^{ave}.\vec{ij}. \tag{VII.33}$$

A second choice for gradients correspond to the mono-dimensional upwind $\Delta u_{i-1/2}$ or downwind $\Delta u_{i+3/2}$. They are based on corresponding upwind and downwind triangles T_{ij} and T_{ji} determined as shown on Figure VII.6. They are defined as

$$\left\{ \begin{array}{lcl} \vec{\nabla} W_{ij}^{up/down} & = & \vec{\nabla} W(T_{ij}) \\ \vec{\nabla} W_{ji}^{up/down} & = & \vec{\nabla} W(T_{ji}) \end{array} \right. \tag{VII.34}$$

In that case, the general MUSCL β-scheme writes:

$$\vec{\nabla} W_i.\vec{ij} = (1-\beta)\vec{\nabla} W_i^{cent}.\vec{ij} + \beta\vec{\nabla} W_{ij}^{up/down}.\vec{ij}. \tag{VII.35}$$

It is easy to check that the expressions (VII.30) for W_{ij} and W_{ji} with upwinding based either on averaged gradients (VII.33) or upwind/downwind gradients (VII.35) gives back interpolated states $u_{i+\frac{1}{2}^-}$ and $u_{i+\frac{1}{2}^+}$ in (VII.6-VII.7) (with $j = i+1$ and W instead of u).

VII.4.2 Effective limitation for the vector W

The scalar limitation procedures described in Section VII.3 are applied successively to each component of W. However, the limitation applied directly to conservative variables might not be optimal. We shall see in the following that a limitation applied to the so-called physical variables, *i.e.* $W^\star = (\rho, u, v, P)$, often gives significatively better results (probably because it avoids oscillations on the density and the pressure, two variables which should always be positive). In that case, physical variables are computed from conservative variables W. Then limitation is applied to each component of $W^\star$, and limited conservative variables are computed from limited physical variables.

VII.4.3 Preliminary results on a shock tube

We first test the different schemes on the classical shock tube problem of Sod [D].

We compare numerical results with the exact solution at $t = 0.16$. Figure VII.7 shows that the MUSCL scheme is not TVD when $\beta = 1/2$ or $\beta = 1/3$.

Limitation is needed to avoid spurious oscillations. These oscillations are deleted with the addition of limiters. We compare on Figure VII.8 the results for the third order accurate MUSCL $\beta = 1/3$ scheme with the limiter of Spekreijse (VII.17) and the new limiter (VII.21). The new limiter adds less numerical diffusion and seems to avoid oscillations.

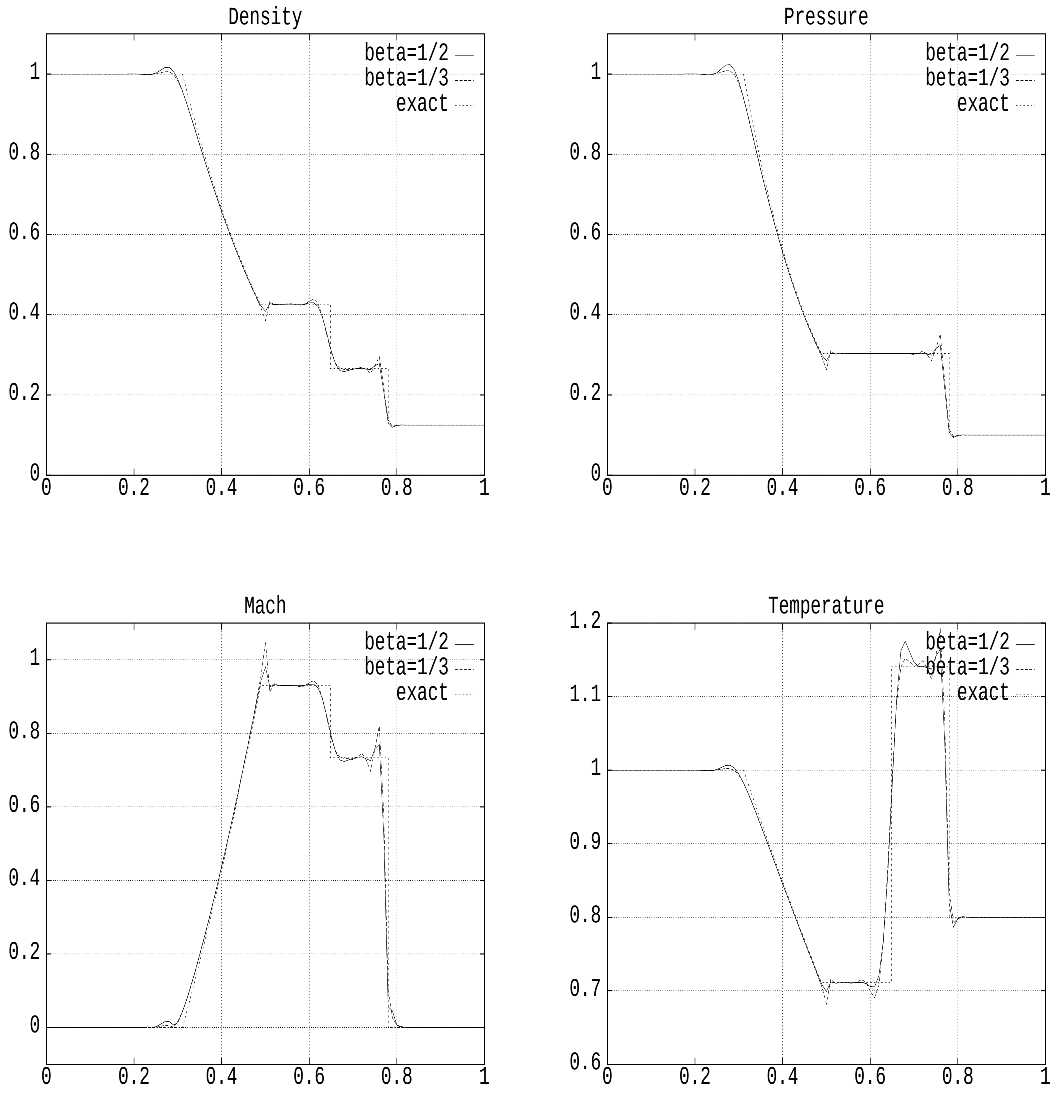


Fig. VII.7 – *Shock tube solutions with no limitation.*

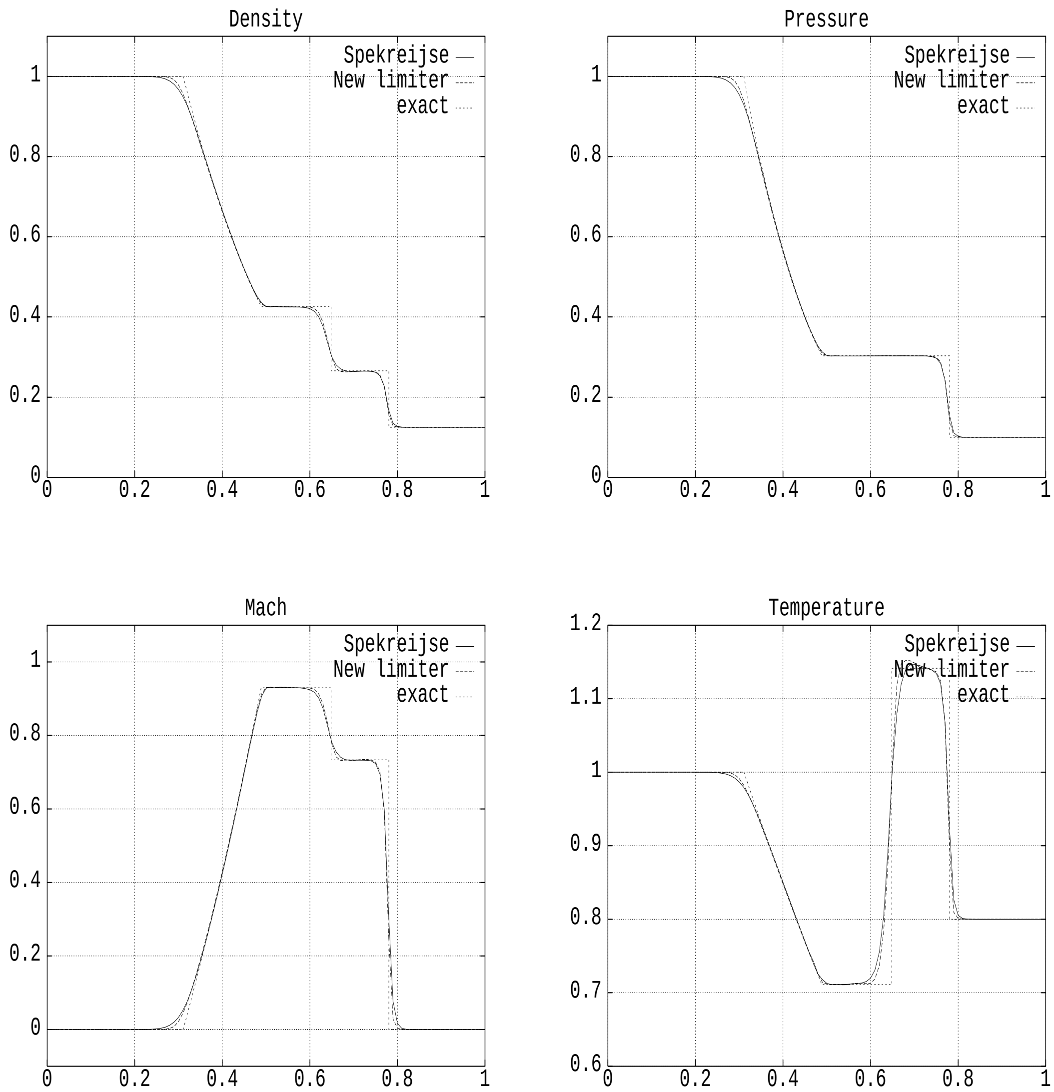


Fig. VII.8 – *Third-order shock tube solutions.*

On Figure VII.9 and VII.10 we compare the third-order accurate scheme and the fourth-order centered scheme. This scheme is computed with a fourth-order Runge-Kutta time scheme. On Figure VII.9 we use no limitation : the solutions obtained for both schemes are almost identical. We plot the solutions on Figure VII.10 when using the new limiters (VII.21),(VII.28). As before, we note that the fourth-order centered scheme is hardly more efficient that the third-order upwind scheme.

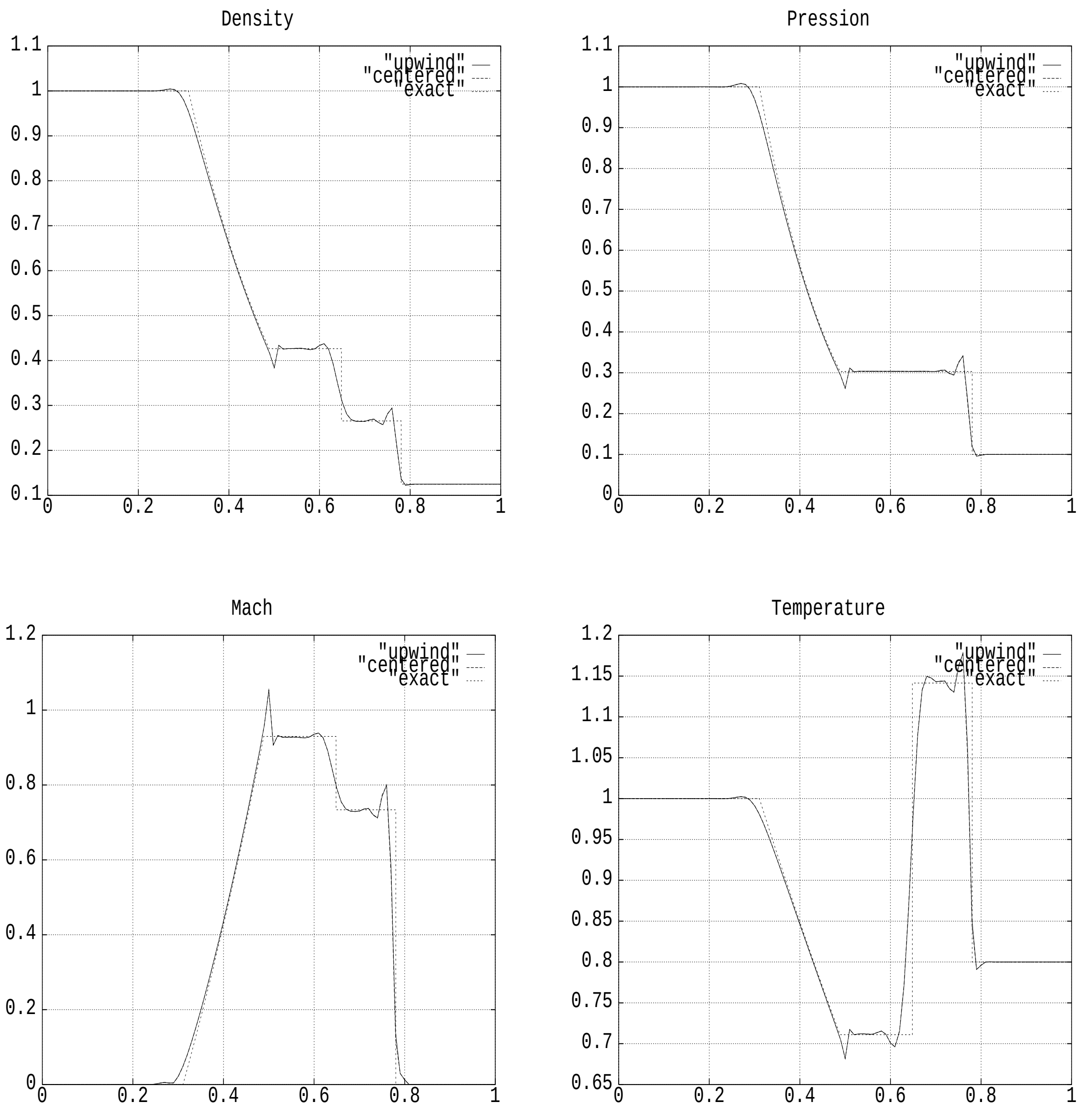


FIG. VII.9 – *Shock tube solutions with no limitation.*

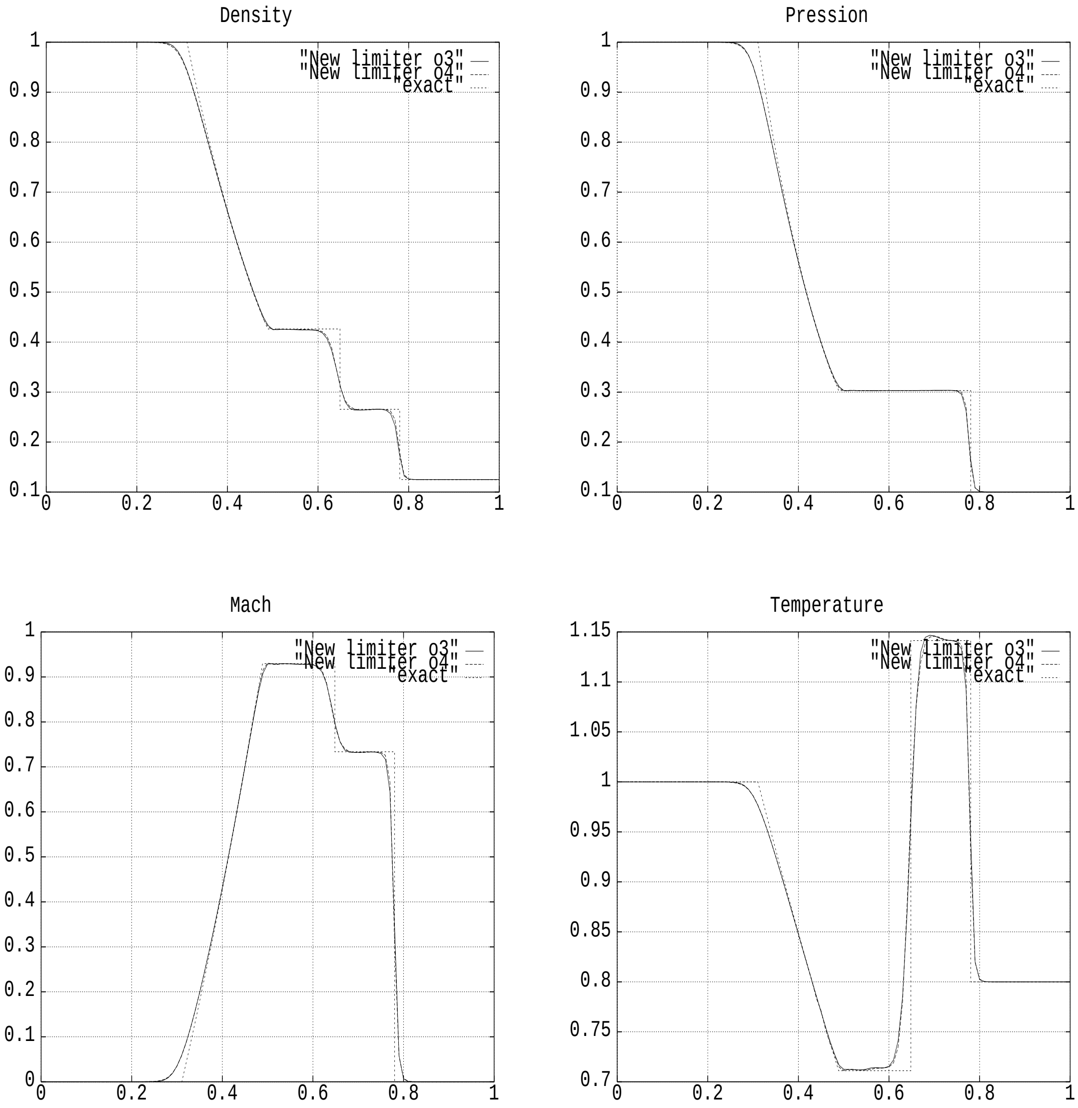


FIG. VII.10 – *Third-order and fourth-order shock tube solutions.*

These tests have been performed with a two-dimensional formulation of this 1D problem. We have observed that the use of nodal average gradients (VII.32) adds numerical diffusion compared to upwind/downwind gradients (VII.34). We have also observed that, using a third-order or a fourth-order accurate explicit Runge-Kutta time scheme, the recomputation of gradients at each Runge-Kutta iteration is less necessary with upwind/downwind gradients than with averaged gradients.

However, since the exact solution has discontinuities, the scheme is not of high accuracy in zones of shock and contact discontinuities. Thus, the difference between fourth-order, third-order or even second-order accurate schemes is not dramatic. In this shock tube case, we found the same accuracy behaviour of all schemes as in the similar unsteady and one-dimensional Burgers equation of Section VII.3.

Altogether, the limiter on fourth order accurate centered fluxes produces an accurate, TVD spatial scheme which has two drawbacks. First, it requires at least a fourth-order accurate time scheme, to preserve its accuracy. Moreover, since the formulation cannot be actually centered, but upwind between near (if not equal) states, it still has the cost of an upwind scheme, which is really important, compared to the cost of a true centered flux, for multidimensional hyperbolic systems. Thus, we will not lead further investigations on this method in the sequel.

VII.4.4 Two-dimensional numerical results

We now test the above schemes on the computation of the steady inviscid flow around a NACA0012 airfoil. We consider the classical transonic test case with $M_\infty = 0.85$ and no incidence, which is a much harder test for numerical methods than subsonic or totally supersonic configurations. The general method we used to obtain steady flows around the airfoil was a pseudo-unsteady implicit (Backward Euler) time scheme. The numerical fluxes based on Roe's approximate Riemann solver were linearized at each time step. The solution algorithm is based on a Jacobi iteration process, which fits well with the unstructured nature of the fluid mesh. To accelerate convergence, we have used local time-stepping, but no multigrid approach or defect correction for second-order accuracy as in [F]. Indeed, we have found that both second-order and third-order unlimited schemes converge easily on this test case with our Jacobi relaxation method, based on the linearization of first-order terms only. The implicit equation before linearization writes:

$$W_i^{n+1} = W_i^n - \Delta t \sum_{j/I_{ij}\neq\emptyset} \|I_{ij}\| F_{Roe}\left(W_{ij}^{n+1}, W_{ji}^{n+1}, \vec{\eta}_{ij}\right) \ , \qquad \text{(VII.36)}$$

where I_{ij} is the interface between neighbouring cells, $\|I_{ij}\|$ and $\vec{\eta}_{ij}$ denote respectively the length and the normal to the cell interface. First order part of Roe's numerical fluxes are linearized. We use the approximation:

$$\begin{aligned} F_{Roe}\left(W_{ij}^{n+1}, W_{ji}^{n+1}, \vec{\eta}_{ij}\right) &= F_{Roe}\left(W_{ij}^{n}, W_{ji}^{n}, \vec{\eta}_{ij}\right) \\ &+ [A_{Roe}^{+}]_{ij}^{n}\left(W_i^{n+1} - W_i^n\right) \\ &+ [A_{Roe}^{-}]_{ij}^{n}\left(W_j^{n+1} - W_j^n\right), \qquad \text{(VII.37)} \end{aligned}$$

where $[A_{Roe}^{\pm}]_{ij}^{n}$ denote the positive and negative parts of the Euler flux Jacobian based on $\vec{\eta}_{ij}$ taken at Roe's average of both states W_{ij}^n and W_{ji}^n. The linearization is far from exact. However, this scheme works perfectly well for unlimited schemes, and we would like to obtain the same behaviour for the limited schemes, which assumes a certain robustness of the limiter functions.

We first try to converge on a relatively coarse grid of 800 vertices and 1514 triangles with 34 points on the profile. Best solutions are obtained with the third order $\beta = 1/3$ MUSCL scheme with the new limiter (VII.21). Contours for the adimensionalized density and pressure, the Mach number and the entropy (given by $s = log((P/P_\infty)/(\rho/\rho_\infty)^\gamma)$) are shown on Figure VII.11.

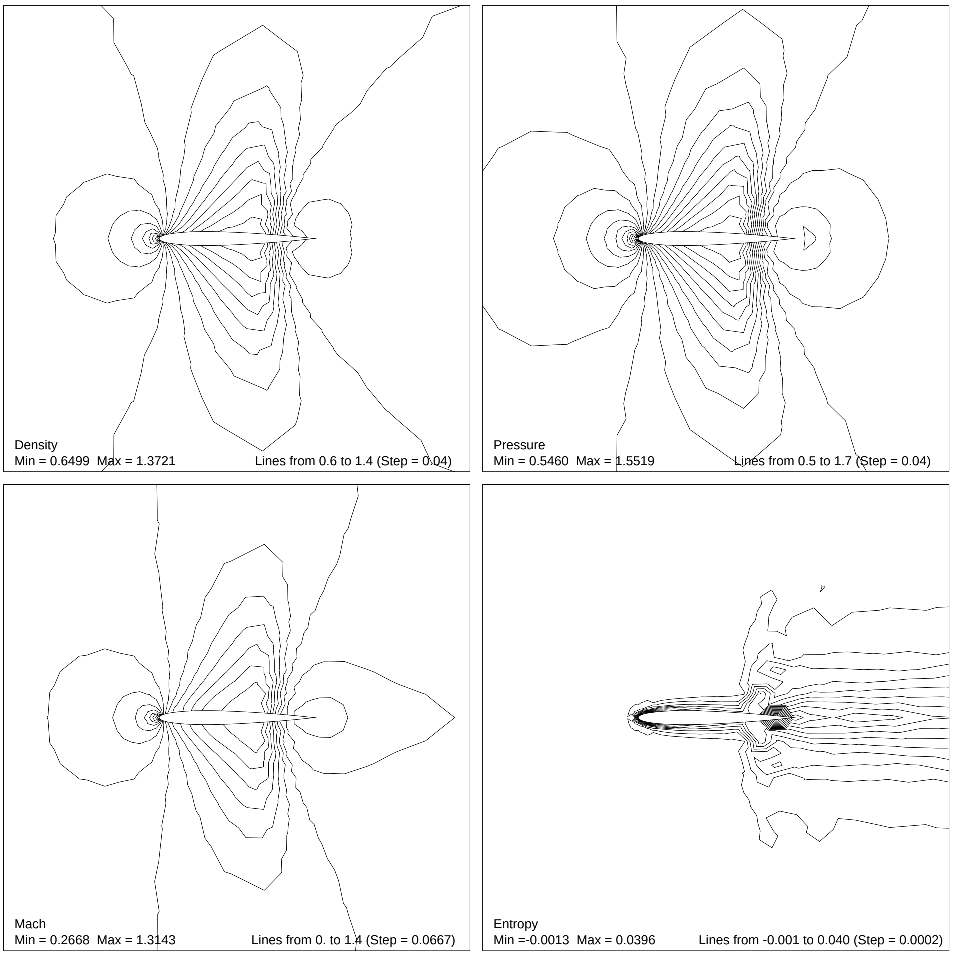


FIG. VII.11 – *Density, pressure, Mach and entropy contours for the steady solution with* $\beta = 1/3$ *and the new limiter (800 vertices).*

The supposed third-order accurate solution is not very smooth, because of the coarseness of the mesh used. We compare on Figure VII.12 the entropy of different converged solutions.

The observation of these curves confirms that the second-order accurate $\beta = 1/2$ MUSCL scheme with Van Albada average produces more diffusion than the third-order

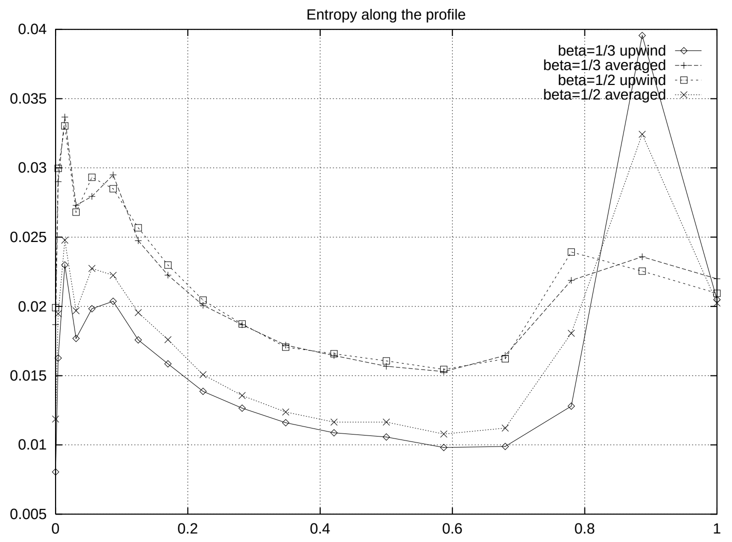


FIG. VII.12 – *Entropy along the profile for different gradients and MUSCL limited schemes (800 vertices).*

$\beta = 1/3$ MUSCL scheme with the new limiter (VII.21). It is also confirmed that averaged gradients (VII.33) produce more numerical diffusion than upwind/downwind gradients (VII.34).

For this coarse grid, we were not able to converge to a solution with the $\beta = 1/3$ MUSCL scheme limited with Spekreijse's limiter (VII.17). This was possibly due to the coarseness of the grid. It is clear that even for the new limiter, the entropy of the steady solution along the profile shows an important oscillation near the trailing edge.

We then test all schemes on a finer grid with 2280 vertices and 4320 triangles (and 120 points along the airfoil profile). Entropy curves along the profile are given on Figure VII.13. The use of the new limiter reduces oscillations of the MUSCL $\beta = 1/3$ scheme. Averaged gradients produce more diffusion than upwind and downwind gradients. The solution with new limiters is now more accurate. Convergence to a steady solution could not be obtained with Spekreijse's limiters. Oscillations in time near the shocks but far from the profile were observed. Anyway, the almost steady entropy profile along the airfoil is rather inaccurate. For comparison, we have plotted on Figure VII.14 the L^2-norm residual of the density through iterations (normalized at the first iteration). The residual vanishes quickly for the $\beta = 1/3$ MUSCL scheme with no limiter or with the new limiter. However, no steady flow can be reached with Spekreijse's limiter (time oscillations appear near shocks but away from the profile).

With this finer mesh, the solution is rather nice. Contours for the density, the pressure, the Mach number and the entropy are given on Figure VII.15.

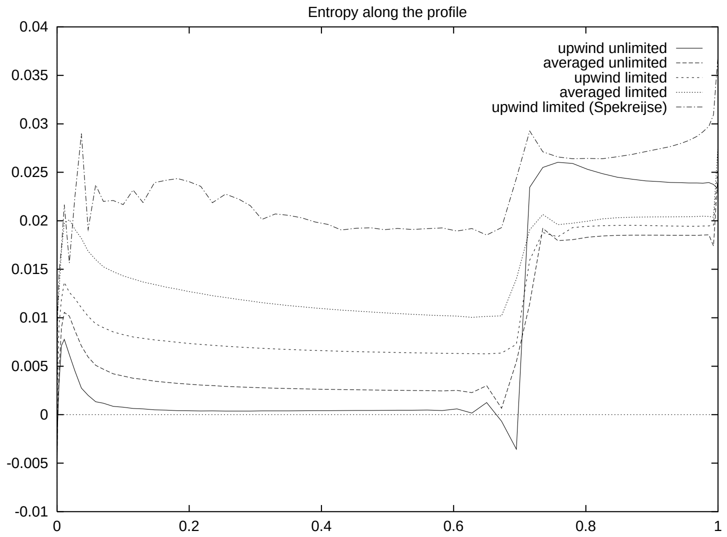


FIG. VII.13 – *Profile entropy for different gradients and limiters (2280 vertices).*

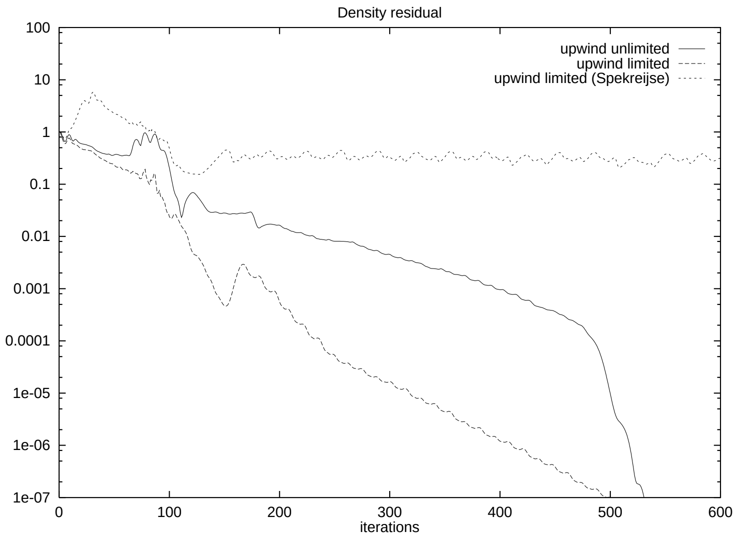


FIG. VII.14 – L^2*-norm residual of the density through iterations (2280 vertices).*

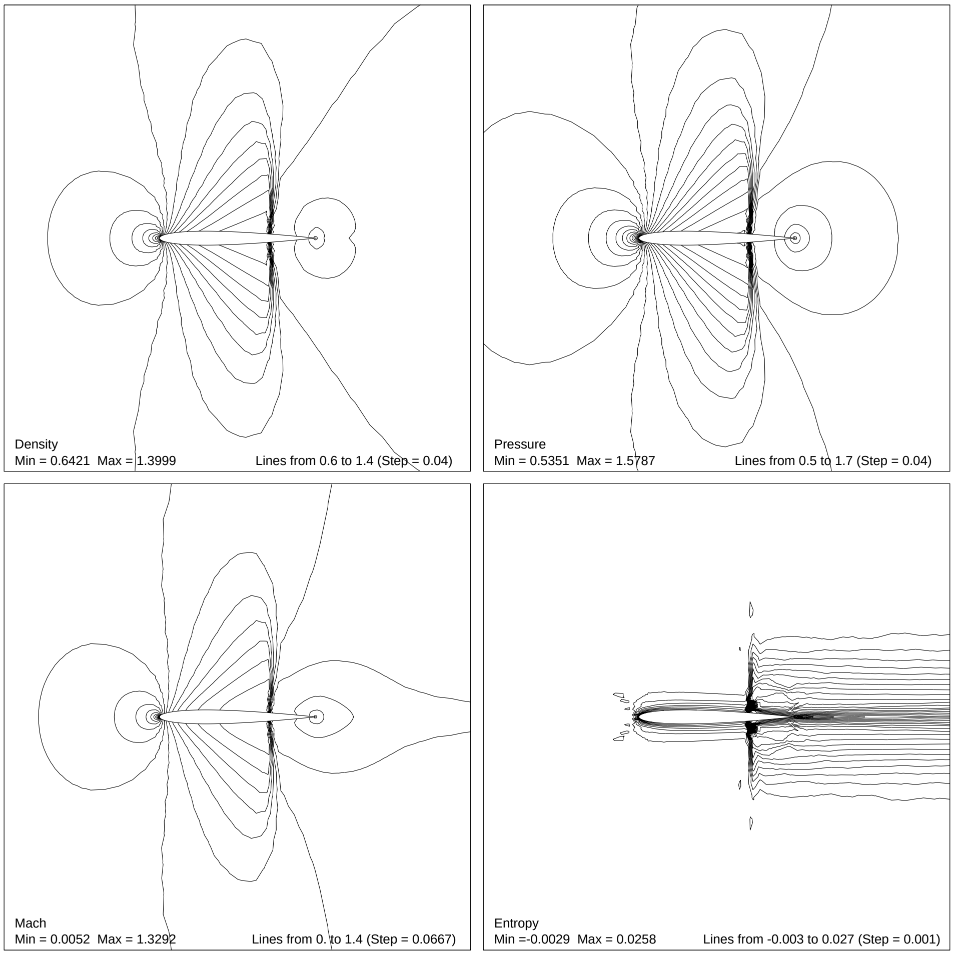


FIG. VII.15 – *Density, pressure, Mach and entropy contours for the steady solution with* $\beta = 1/3$ *and the new limiter (2280 vertices).*

Finally, we test the new limiter on a very fine grid (12284 vertices and 24224 triangles, 136 vertices on the profile) where numerical diffusion will be very small. We test the robustness of the limiter. We find that the direct algorithm with linearization of only first-order fluxes, as described in (VII.37) is not sufficient. Even the new limiter has difficulties to converge to a steady solution. We have modified the linearization in order to have a more accurate Jacobian of the numerical fluxes. However, we conserved a diagonally dominant matrix to have a good convergence of Jacobi iterations at each time step. We got convergence to a steady solution with the following linearization:

$$\begin{aligned} F_{Roe}\left(W_{ij}^{n+1}, W_{ji}^{n+1}, \vec{\eta}_{ij}\right) &= F_{Roe}\left(W_{ij}^{n}, W_{ji}^{n}, \vec{\eta}_{ij}\right) \\ &+ \left|[A_{Roe}]_{ij}^{n}\right| \left(\left|\frac{\partial W_{ij}}{\partial W_i}\right| + \left|\frac{\partial W_{ji}}{\partial W_i}\right|\right)\left(W_i^{n+1} - W_i^n\right) \\ &+ \left([A_{Roe}^{+}]_{ij}^{n}\left|\frac{\partial W_{ij}}{\partial W_j}\right| + [A_{Roe}^{-}]_{ij}^{n}\left|\frac{\partial W_{ji}}{\partial W_j}\right|\right)\left(W_j^{n+1} - W_j^n\right) \end{aligned} \tag{VII.38}$$

Contours for the very accurate steady solution are given on Figure VII.16.

VII.5 Conclusion

In this paper, we have reviewed existing limiters for MUSCL-type schemes based on upwind numerical fluxes. The non linear limitation makes these schemes accurate and TVD. However, we have given a sufficient condition on the limitation and the Courant number used to produce a TVD scheme. This criterion allowed us to construct new limiters for both upwind and centered numerical fluxes.

The fourth-order accurate centered scheme cannot be limited with no upwind information. We have presented a limited upwind scheme, which reduces to the centered scheme when the limitation is omitted. The scheme is TVD, but accuracy is necessarily obtained with high computational costs.

On the contrary, we have been able to produce an efficient new limiter for the upwind $\beta = 1/3$ third-order accurate MUSCL scheme. This new limiter appears to be more accurate, stable and robust than existing limiters on one-dimensional test cases of linear advection and Burgers equation. All conclusions were confirmed and extended to two-dimensional cases with numerical tests made of unsteady shock tube problems and steady state computations around a NACA airfoil. In this last case, the new limiter allowed us to use a very simple algorithm for the difficult accurate calculation of the transonic steady flow around the airfoil with an implicit time scheme, which was not possible with existing limiters, like the limiter proposed by Spekreijse.

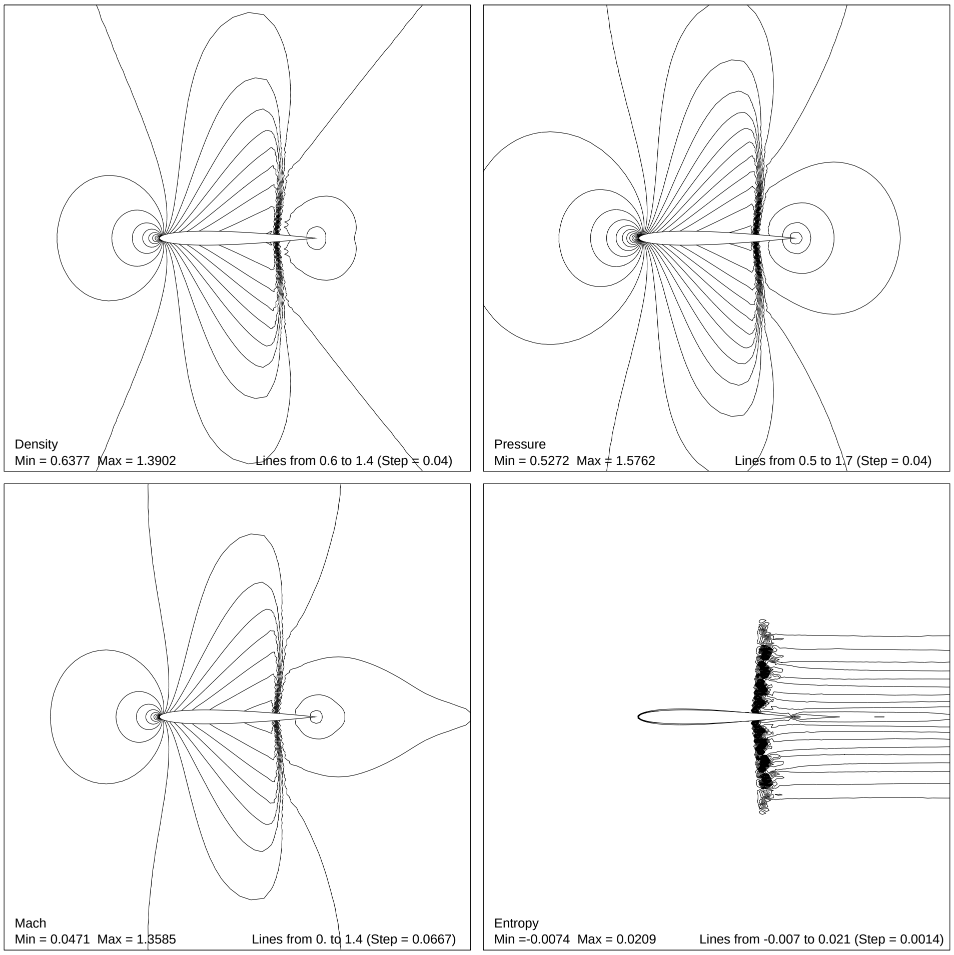


FIG. VII.16 – *Density, pressure, Mach and entropy contours for the steady solution with* $\beta = 1/3$ *and the new limiter (12214 vertices).*

Bibliographie

[A] FEZOUI L., *Résolution des équations d'Euler par un schéma de Van Leer en éléments finis*, Rapport de recherche INRIA no.358 (1985).

[B] HARTEN A., *High Resolution Schemes for Hyperbolic Conservation Laws*, Journal of Computational Physics, Vol 49, pp 357-393, (1983).

[C] ROE P. L., *Approximate riemann solvers, parameters vectors, and difference schemes*, Journal of Computational Physics, Vol 4, pp 357-371, (1981).

[D] SOD G. A., *A survey of several finite difference methods for systems of non-linear hyperbolic conservation laws.*, Journal of Computational Physics, Vol 27, pp 1-31, (1978).

[E] SPEKREIJSE S. P., *Multigrid solution of monotone second-order discretizations of hyperbolic conservation laws*, Math. Comp., Vol 49, pp 135-155, (1987).

[F] SPEKREIJSE S. P., *Multigrid Solution of the Steady Euler Equations*, PhD thesis, Centrum voor Wiskunde en Informatica, Amsterdam, Nov. (1987).

[G] SWEBY P. K., *High resolution schemes using flux limiters for hyperbolic conservation laws*, SIAM Numer. Anal., Vol 21, pp 995-1011, (1984).

[H] VAN LEER B., *Towards the Ultimate Conservative Difference Scheme II: Monotonicity and conservation combined in a second order scheme*, Journal of Computational Physics, Vol 14, pp 361-370, (1974).

[I] VAN LEER B., *Towards the Ultimate Conservative Difference Scheme V: a Second-Order Sequel to Godunov's Method*, Journal of Computational Physics, Vol 32, pp 361-370, (1979).

Chapitre VIII

UNE MÉTHODE COUPLÉE POUR LA SIMULATION D'ÉCOULEMENTS DIPHASIQUES.

VIII.1 Introduction.

La modélisation d'écoulements diphasiques composés d'un nuage de gouttes liquides ou de particules solides est l'objet de nombreuses études depuis quelques années. De tels écoulements gouvernent le fonctionnement des moteurs diesel, des moteurs à injection directe ou encore des moteurs cryogéniques des fusées. Dans ce cas, la phase gazeuse est composée d'hydrogène et la phase liquide d'oxygène. Un autre type d'écoulement diphasique a lieu dans les fusées à poudre. En effet, dans les fusées modernes, on introduit dans le propergol solide de petites particules métalliques (par exemple, des particules d'aluminium) pour augmenter la poussée de la fusée. La présence de ces particules modifie la vitesse du son dans la fusée et atténue les instabilités de combustion; de plus, elles produisent des particules de fumée de très petite taille (1μm), et des agglomérats de grosse taille ($100\,\mu$m) par coalescence.

Pour modéliser ces types de mélange, nous choisissons une approche Eulérienne, où chacune des phases (gaz, particules ou gouttes) est traitée comme un fluide continu décrit par des quantités macroscopiques. Cette approche est beaucoup moins coûteuse d'un point de vue numérique, mais aussi moins riche qu'une approche cinétique, dans la mesure où les phénomènes de coalescence et de turbulence ne sont pas pris en compte. On suppose en particulier que toutes les particules ont localement le même rayon r.

Le modèle utilisé est obtenu par un processus de moyennes à partir des équations de Navier-Stokes qui régissent l'écoulement au niveau microscopique. Soit α la fraction volumique du gaz. Le gaz est supposé parfait, les variables liées au gaz sont : sa densité $\alpha\rho_g$, son impulsion $\alpha\rho_g\mathbf{u}_g$ et son énergie spécifique $\alpha\rho_g e_g$. Les variables décrivant la phase dispersée sont: sa densité $(1-\alpha)\rho_l$, son impulsion $(1-\alpha)\rho_l\mathbf{u}_l$ et son énergie spécifique $(1-\alpha)\rho_l e_l$. On suppose que la densité des particules est constante et très supérieure à celle du gaz. L'intéraction entre les deux phases se fait par un échange d'impulsion et d'énergie, au moyen de la force de traînée $\mathbf{f}_d$ et de son travail. Si on suppose l'écoulement laminaire autour des particules, on voit que cette force est proportionnelle à l'expression $(\mathbf{u}_g-\mathbf{u}_l)/\epsilon$, où ϵ représente un temps de relaxation proportionnel à r^2. La force de traînée a donc tendance à rendre égales les vitesses du gaz et des particules, après un temps de l'ordre de ϵ.

En pratique, dans une fusée, en dehors d'une zone située autour des injecteurs, α est très proche de 1 (typiquement $1-\alpha \leq 10^{-3}$). On ne traite donc pas l'accumulation de particules. De plus, dans les écoulements que nous considérons les effets de la traînée sont dominants devant les effets diffusifs de la viscosité du gaz; on utilisera donc les équations d'Euler plutôt que celles de Navier-Stokes.

Le système que nous obtenons, en l'absence de termes source est un système hyperbolique non conservatif. L'étude de ce système a été faite dans [13, 14] et a mis en évidence le faible couplage entre la phase gazeuse et la phase dispersée lorsque α est proche de 1. Cette analyse a permis de mettre au point un modèle simplifié hyperbolique conservatif. L'avantage de ce modèle est d'obtenir deux sous-modèles entièrement découplés, dont l'un est celui de la dynamique des gaz pour les variables $(\alpha\rho_g, \alpha\rho_g\mathbf{u}_g, \alpha\rho_g e_g)$.

Ici, nous nous intéressons à la simulation numérique de ce nouveau modèle. Du fait de la présence de termes source raides losque ϵ devient petit, nous avons, dans un premier temps utilisé une méthode de pas fractionnaires en temps. Cependant, cette méthode peut conduire à des solutions très peu précises, en particulier à un amortissement numérique important des ondes sonores. Une méthode alternative a été proposée dans [2], l'expansion de Chapman-Enskog, qui consiste à obtenir un système dont le membre de droite est un terme de diffusion proportionnel à ϵ. F. Béreux [2] a étudié l'atténuation d'une onde acoustique de pulsation ω dans un tube, pour une grande échelle de valeurs de ϵ. Il a montré que pour des valeurs ϵw petites devant 1, la méthode de "splitting" donne des solutions imprécises alors que l'expansion de Chapman-Enskog donne de très bons résultats. Inversement, pour des valeurs ϵw de l'ordre de 1 ou plus grandes que 1, la méthode de "splitting" permet d'obtenir de très bonnes solutions alors que l'autre méthode ne représente pas correctement la solution physique du problème.
Nous proposons ici une méthode couplée, qui consiste à prendre en compte en même temps la partie convective et le terme source. Pour des grandes valeurs de ϵw, l'approximation numérique se fera au moyen de schémas volumes finis explicites en temps et en espace, d'ordre élevé (deux, trois, ou quatre en temps et en espace). Lorsque ϵw devient petit devant 1, pour éviter l'utilisation de pas de temps trop petits, nous construisons des schémas implicites précis en temps et en espace.
Pour valider cette méthode, nous reprenons l'étude de la propagation d'une onde sonore à l'intérieur d'un tube, et nous comparons les solutions obtenues par cette nouvelle méthode, avec celles obtenues par la méthode de "splitting". Nous montrons que la nouvelle méthode donne de très bons résultats pour une grande échelle de valeurs de ϵ.
Nous effectuons aussi une simulation d'écoulement diphasique dans une tuyère, et nous comparons la solution stationnaire obtenue avec la méthode de "splitting" à celle obtenue avec la méthode couplée.

VIII.2 Présentation du modèle.

Nous présentons les hypothèses physiques permettant d'écrire le modèle dans un cadre monodimensionnel.
On considère un écoulement constitué d'un brouillard de gouttes liquides suspendues dans une atmosphère gazeuse. On suppose l'écoulement compressible autour des gouttes et

l'écoulement liquide incompressible à l'intérieur des gouttes. Nous rappelons l'ensemble des variables qui décrivent le gaz: sa densité $\rho_g(x,t)$, sa vitesse $u_g(x,t)$, et son énergie spécifique totale $e_g(x,t)$. Celle-ci vérifie :

$$e_g = \frac{u_g^2}{2} + \varepsilon_g \tag{VIII.1}$$

où ε_g représente l'énergie spécifique interne.
On modélise le nuage de gouttes comme un fluide. Nous supposons que toutes les gouttes ont localement le même rayon r. On définit la fraction volumique du gaz α comme suit:

$$1 - \alpha = \frac{\text{masse du nuage de gouttes par unité de volume}}{\rho_l} \tag{VIII.2}$$

où ρ_l est la densité massique de la phase liquide, supposée constante.
Par construction, α est compris entre 0 et 1. Nous introduisons la définition suivante de l'énergie spécifique totale de la phase liquide:

$$e_l = \frac{u_l^2}{2} + \varepsilon_l \tag{VIII.3}$$

où ε_l représente l'énergie spécifique interne des gouttes.

On omet les échanges de masse et le transfert de chaleur entre les deux phases. Nous prenons en compte seulement la force de traînée f_d:

$$f_d = \frac{4\pi r^3}{3(1-\alpha)}\Gamma \tag{VIII.4}$$

où Γ représente l'échange d'impulsion entre les deux phases. A partir de la formule de Stokes, ce terme s'écrit:

$$\Gamma = \frac{(1-\alpha)\rho_l(u_g - u_l)}{\epsilon} \tag{VIII.5}$$

et le temps de relaxation ϵ est donné par:

$$\epsilon = \frac{16 r^2 \rho_l}{81 \mu_g} \tag{VIII.6}$$

où μ_g représente la viscosité du gaz supposée petite.
Pour les applications que nous considérons ici, ϵ varie entre 10^{-6} et 10^{-2} secondes et peut être très petit par-rapport à la période des ondes de pression se propageant dans le milieu diphasique. Le terme d'échange d'énergie entre les deux phases s'écrit:

$$I = \Gamma . u_l \tag{VIII.7}$$

VIII.2.1 Le modèle original.

D'après L. Sainsaulieu [13, 14], en tenant compte des hypothèses précédentes, le modèle diphasique s'écrit (en une dimension d'espace):

$$(VIII.2.1.i) \qquad \partial_t(\alpha\rho_g) + \partial_x(\alpha\rho_g u_g) = 0 \ ,$$

$$(VIII.2.1.ii) \qquad \partial_t(\alpha\rho_g u_g) + \partial_x(\alpha\rho_g u_g^2) + \alpha\partial_x p_g = -\Gamma \ ,$$

$$(VIII.2.1.iii) \qquad \partial_t(\alpha\rho_g e_g) + \partial_x(\alpha\rho_g e_g u_g + \alpha p_g u_g) + p_g\partial_x((1-\alpha)u_l) = -I \ ,$$

$$(VIII.2.1.iv) \qquad \partial_t((1-\alpha)\rho_l) + \partial_x((1-\alpha)\rho_l u_l) = 0 \ ,$$

$$(VIII.2.1.v) \qquad \partial_t((1-\alpha)\rho_l u_l) + \partial_x((1-\alpha)\rho_l u_l^2) + (1-\alpha)\partial_x p_g + \partial_x\theta = \Gamma \ ,$$

$$(VIII.2.1.vi) \quad \partial_t((1-\alpha)\rho_l e_l) + \partial_x((1-\alpha)\rho_l e_l u_l) + (1-\alpha)u_l\partial_x p_g + \partial_x(\theta u_l) = I \ .$$

où la pression p_g vérifie la loi des gaz parfaits, et la pression θ est définie comme suit :

$$p_g = (\gamma - 1)\rho_g\varepsilon_g, \qquad \theta = \theta_0(1-\alpha)^\delta \tag{VIII.8}$$

γ représente le coefficient adiabatique du gaz, θ_0 caractérise la pression d'arrêt du gaz sur les gouttes, et δ est une constante positive. Nous avons en général $\gamma = 1.4$ pour un gaz parfait diatomique et $\delta = 4/3$ pour un écoulement diphasique [14]. L'équation de conservation de l'impulsion totale s'écrit :

$$\partial_t(\alpha\rho_g u_g + (1-\alpha)\rho_l u_l) + \nabla.(\alpha\rho_g u_g^2 + (1-\alpha)\rho_l u_l^2) + \nabla p_{\text{eff}} = 0, \tag{VIII.9}$$

où la pression effective de l'écoulement est:

$$p_{\text{eff}} = p_g + \theta \tag{VIII.10}$$

Physiquement, la pression θ est petite devant la pression du gaz p_g, si bien que la prise en compte de ce terme est peu importante d'un point de vue quantitatif. Mais du point de vue mathématique, c'est ce terme qui rend le système (VIII.2.1) hyperbolique. En effet, le système écrit ci-dessus, en l'absence des termes source, est un système hyperbolique non conservatif, au moins dans la limite $\alpha \to 1$: il possède lorsque l'écart de vitesses entre les deux phases est subsonique, c'est-à-dire si $|\mathbf{u}_g - \mathbf{u}_l| < c_g$ où c_g est la vitesse du son dans le gaz, six vitesses caractérisques réelles dont on peut calculer un développement limité en $1-\alpha$ [14] :

$$\left\{\begin{aligned} &\lambda_1 = u_g - c_g + o\Big((1-\alpha)^{\delta-1}\Big) \ , \\ &\lambda_2 = u_g + c_g + o\Big((1-\alpha)^{\delta-1}\Big) \ , \\ &\lambda_3 = u_l - \Big(\tfrac{-\theta'(\alpha)}{\rho_l}\Big)^{1/2} + o\Big((1-\alpha)^{\frac{\delta-1}{2}}\Big) \ , \\ &\lambda_4 = u_l + \Big(\tfrac{-\theta'(\alpha)}{\rho_l}\Big)^{1/2} + o\Big((1-\alpha)^{\frac{\delta-1}{2}}\Big) \ , \\ &\lambda_5 = u_g \ , \\ &\lambda_6 = u_l \ . \end{aligned}\right. \tag{VIII.11}$$

VIII.2.2 Un modèle simplifié.

Dans sa thèse [13], L. Sainsaulieu a introduit un solveur numérique de type Roe pour résoudre le système **non conservatif** (VIII.2.1) ci-dessus.
Pour un système conservatif, la matrice de Roe doit satisfaire certaines relations, calquées sur celles de Rankine-Hugoniot. Le problème pour un système non conservatif est de pouvoir définir des relations de Rankine-Hugoniot approchées.
Les solutions onde de choc de ce système non-conservatif sont définies comme les limites de profils visqueux solutions de (VIII.2.1) lorsque la viscosité tend vers 0 [13, 14]. L'analyse des solutions obtenues montre que ces ondes de choc sont de deux types : ou bien elles portent essentiellement sur les variables décrivant la phase gazeuse (ρ_g, u_g et e_g) et les quantités α, u_l et ε_l restent presque constantes à la traversée de l'onde de choc, ou bien elles affectent essentiellement les quantités α, u_l et ε_l alors que la masse volumique, la vitesse et l'énergie interne du gaz varient peu à la traversée de l'onde de choc [13]. La distinction entre ces deux types d'ondes est rendue possible par le faible couplage entre la phase gazeuse et la phase liquide lorsque α est proche de 1. C'est sur cette classification des ondes de choc que repose la méthode numérique utilisée dans [13] et [12] qui utilise un solveur de type Roe.
Cependant cette méthode est très coûteuse : les valeurs propres et les vecteurs propres de la matrice linéarisée de Roe ne sont pas explicitement connus; ils doivent être évalués numériquement. De plus, l'inversion de la matrice des vecteurs propres demande à chaque pas de temps et pour chaque flux d'espace la décomposition d'un vecteur sur les vecteurs propres.

L. Sainsaulieu dans [14] propose une méthode plus économique, qui tire parti du découplage partiel mis en évidence ci-dessus et qui nous ramène à un modèle dont on connaît explicitement les valeurs propres et les vecteurs propres de la matrice de convection. Il s'agit du modèle simplifié suivant (écrit ici en une dimension d'espace):

$$(VIII.2.2.i) \qquad \partial_t(\alpha\rho_g) + \partial_x(\alpha\rho_g u_g) = 0 \ ,$$

$$(VIII.2.2.ii) \qquad \partial_t(\alpha\rho_g u_g) + \partial_x(\alpha\rho_g u_g^2) + \partial_x(\alpha p_g) = 0 \ ,$$

$$(VIII.2.2.iii) \qquad \partial_t(\alpha\rho_g e_g) + \partial_x(\alpha\rho_g e_g u_g) + \partial_x(\alpha p_g u_g) = 0 \ ,$$

$$(VIII.2.2.iv) \qquad \partial_t((1-\alpha)\rho_l) + \partial_x((1-\alpha)\rho_l u_l) = 0 \ ,$$

$$(VIII.2.2.v) \qquad \partial_t(\rho_l u_l) + \partial_x\left(\frac{\rho_l u_l^2}{2} + \frac{\delta\theta_0(1-\alpha)^{\delta-1}}{(\delta-1)}\right) = 0 \ ,$$

$$(VIII.2.2.vi) \qquad \partial_t((1-\alpha)\rho_l\varepsilon_l^e) + \partial_x((1-\alpha)\rho_l u_l \varepsilon_l^e) = 0 \ .$$

où l'énergie spécifique interne effective ε_l^e est définie par:

$$\varepsilon_l^e = \varepsilon_l - \frac{\theta_0(1-\alpha)^{\delta-1}}{(\delta-1)\rho_l} \tag{VIII.12}$$

Seules les équations $(VIII.2.2.ii)$, $(VIII.2.2.iii)$ et $(VIII.2.2.v)$ de ce système sont approchées, les trois autres équations conservatives $(VIII.2.2.i)$, $(VIII.2.2.iv)$ et $(VIII.2.2.vi)$ étant exactement satisfaites par les solutions du système (VIII.2.1). Le nouveau modèle est obtenu par un développement limité pour les petites valeurs de $1-\alpha$ et dans le cas où le rapport $\dfrac{\rho_g}{\rho_l}$ est supposé petit [15]. On remarque que dans le système (VIII.2.2), le terme de pression p_g a été omis pour l'équation (VIII.2.2.v). En effet, les onde de taux de vide (affectant les variables du liquide) sont en général d'intensité beaucoup plus faible que celles du gaz si bien qu'elles ne peuvent être observées que lorsque la pression du gaz p_g est constante.

On remarque enfin que, si la pression p_g est donnée par la loi d'un gaz parfait, le premier système de trois équations est exactement le système de la dynamique des gaz dans lequel la masse volumique ρ_g du gaz a été remplacée par la quantité $\alpha\rho_g$.

L'avantage de cette méthode, qui consiste à remplacer le système (VIII.2.1) par le système plus simple (VIII.2.2), est donc de donner deux systèmes entièrement découplés dont l'un est celui de la dynamique des gaz.

VIII.3 Approximation numérique.

Nous considérons maintenant le système (VIII.2.2) écrit en deux dimensions d'espace.

$$(VIII.3.i) \qquad \partial_t(\alpha\rho_g) + \partial_x(\alpha\rho_g u_g) + \partial_y(\alpha\rho_g v_g) = 0 \ ,$$

$$(VIII.3.ii) \qquad \partial_t(\alpha\rho_g u_g) + \partial_x(\alpha\rho_g u_g^2) + \partial_y(\alpha\rho_g u_g v_g) + \partial_x(\alpha p_g) = -\Gamma_x \ ,$$

$$(VIII.3.iii) \qquad \partial_t(\alpha\rho_g v_g) + \partial_x(\alpha\rho_g u_g v_g) + \partial_y(\alpha\rho_g v_g^2) + \partial_y(\alpha p_g) = -\Gamma_y \ ,$$

$$(VIII.3.iv) \qquad \partial_t(\alpha\rho_g e_g) + \partial_x(\alpha\rho_g e_g u_g) + \partial_y(\alpha\rho_g e_g v_g) + \partial_x(\alpha p_g u_g) + \partial_y(\alpha p_g v_g) = -\Gamma_x u_l - \Gamma_y v_l \ ,$$

$$(VIII.3.v) \qquad \partial_t(1-\alpha) + \partial_x((1-\alpha)u_l) + \partial_y((1-\alpha)v_l) = 0 \ ,$$

$$(VIII.3.vi) \qquad \partial_t(u_l) + \partial_x\left(\frac{u_l^2}{2}\right) + \partial_y\left(\frac{u_l v_l}{2}\right) + \partial_x\left(\frac{\delta\theta_0(1-\alpha)^{\delta-1}}{\rho_l(\delta-1)}\right) = \frac{1}{(1-\alpha)\rho_l}\Gamma_x \ ,$$

$$(VIII.3.vii) \qquad \partial_t(v_l) + \partial_x\left(\frac{u_l v_l}{2}\right) + \partial_y\left(\frac{v_l^2}{2}\right) + \partial_y\left(\frac{\delta\theta_0(1-\alpha)^{\delta-1}}{\rho_l(\delta-1)}\right) = \frac{1}{(1-\alpha)\rho_l}\Gamma_y \ ,$$

$$(VIII.3.viii) \qquad \partial_t((1-\alpha)\varepsilon_l^e) + \partial_x((1-\alpha)\varepsilon_l^e u_l) + \partial_y((1-\alpha)\varepsilon_l^e v_l) = 0 \ .$$

Les vitesses du gaz et des gouttes sont respectivement $\mathbf{u}_g = (u_g, v_g)$ et $\mathbf{u}_l = (u_l, v_l)$. Le terme de traînée $\mathbf{\Gamma} = (\Gamma_x, \Gamma_y)$ s'écrit simplement :

$$\Gamma_x = \frac{(1-\alpha)\rho_l(u_g - u_l)}{\epsilon}, \quad \Gamma_y = \frac{(1-\alpha)\rho_l(v_g - v_l)}{\epsilon} \tag{VIII.13}$$

et le terme de relaxation ϵ est donné par :

$$\epsilon = \frac{16 r^2 \rho_l}{81 \mu_g} \tag{VIII.14}$$

On cherche à résoudre numériquement le problème de Cauchy pour le système (VIII.3) écrit sous forme condensée comme suit :

$$\begin{cases} \partial_t \mathbf{W} + \nabla.\mathbb{F}(\mathbf{W}) = \dfrac{\mathbf{R}(\mathbf{W})}{\epsilon} \ , \mathbf{W} \in \mathcal{E} \\ \mathbf{W}(0,x,y) = \mathbf{W}^0(x,y) \end{cases} \tag{VIII.15}$$

avec pour vecteur d'état $\mathbf{W}$:

$$\mathbf{W} = {}^t(\alpha\rho_g, \alpha\rho_g u_g, \alpha\rho_g v_g, \alpha e_g, (1-\alpha), u_l, v_l, (1-\alpha)\varepsilon_l^e) \tag{VIII.16}$$

et pour espace des états admissibles :

$$\mathcal{E} = \mathbb{R}_+ \text{ x } \mathbb{R} \text{ x } \mathbb{R} \text{ x } \mathbb{R}_+ \text{ x } [0,1] \text{ x } \mathbb{R} \text{ x } \mathbb{R} \text{ x } \mathbb{R}_+ \tag{VIII.17}$$

VIII.3.1 Hyperbolicité du système.

En l'absence de termes source, les équations de la phase gazeuse sont découplées de celles des gouttes. On peut donc réécrire chacun des deux sous-systèmes conservatifs de (VIII.3) sous la forme :

$$\partial_t \mathbf{W}_k + \partial_x \mathbf{F}_k(\mathbf{W}) + \partial_y \mathbf{G}_k(\mathbf{W}) = 0, \quad k = 1, 2 \tag{VIII.18}$$

où $\mathbf{W}_1$ est le vecteur d'état associé au gaz et $\mathbf{W}_2$ celui de la phase liquide.

– Pour le gaz :

On retrouve le système des équations d'Euler où la masse volumique ρ_g est remplaçée par la quantité $\alpha\rho_g$. Les expressions des flux $\mathbf{F}_1(\mathbf{W})$ et $\mathbf{G}_1(\mathbf{W})$ sont donnés en (VI.12) (l'énergie totale E vérifie : $E = \rho_g e_g$).

– Pour les gouttes :

La masse volumique ρ_l étant constante :

$$\mathbf{W}_2 = \begin{pmatrix} 1-\alpha \\ u_l \\ v_l \\ (1-\alpha)\,\varepsilon_l^e \end{pmatrix}.$$

$$\mathbf{F}_2(\mathbf{W}) = \begin{pmatrix} (1-\alpha)\,u_l \\ \dfrac{u_l^2}{2} + \dfrac{\delta\theta_0}{\rho_l}\dfrac{(1-\alpha)^{\delta-1}}{(\delta-1)} \\ \dfrac{u_l v_l}{2} \\ (1-\alpha)u_l\varepsilon_l^e \end{pmatrix}, \quad \mathbf{G}_2(\mathbf{W}) = \begin{pmatrix} (1-\alpha)v_l \\ \dfrac{u_l v_l}{2} \\ \dfrac{v_l^2}{2} + \dfrac{\delta\theta_0}{\rho_l}\dfrac{(1-\alpha)^{\delta-1}}{(\delta-1)} \\ (1-\alpha)v_l\varepsilon_l^e \end{pmatrix}.$$

Dans la suite on omet l'exposant e pour désigner l'énergie spécifique interne effective ε_l^e.

On montre aisément que, pour chacune des deux phases, le système est hyperbolique. En effet, considérons une combinaison linéaire des flux :

$$\mathcal{F}_k(\mathbf{W}, \boldsymbol{\eta}) = \eta_x \mathbf{F}_k(\mathbf{W}) + \eta_y \mathbf{G}_k(\mathbf{W}), \quad k = 1, 2 \tag{VIII.19}$$

où $\boldsymbol{\eta} = (\eta_x, \eta_y)$ est un vecteur non nul de $I\!R^2$.
Les matrices jacobiennes $\mathcal{A}_1$ de la phase gazeuse et $\mathcal{A}_2$ de la phase liquide sont définies par :

$$\mathcal{A}_k(\mathbf{W}, \boldsymbol{\eta}) = \eta_x A_k(\mathbf{W}) + \eta_y B_k(\mathbf{W}), \quad k = 1, 2 \tag{VIII.20}$$

avec

$$(A_k)_{k=1,2} = \frac{\partial \mathbf{F}_k}{\partial \mathbf{W}}(\mathbf{W}), \quad (B_k)_{k=1,2} = \frac{\partial \mathbf{G}_k}{\partial \mathbf{W}}(\mathbf{W}) \tag{VIII.21}$$

Nous avons vu dans la section VI.3.2 que le système d'équations d'Euler est hyperbolique. Le sous-système lié à la phase gazeuse l'est donc aussi et ses valeurs propres sont données par (VI.16). Nous avons donnons sans démonstration le résultat suivant concernant le sous-système de la phase liquide (l'expression de la matrice jacobienne $\mathcal{A}_2(\mathbf{W}, \boldsymbol{\eta})$ est donnée dans l'Annexe D) :

Proposition VIII.3.1 *Le sous-système lié à la phase liquide est hyperbolique. Ses vitesses caractéristiques sont :*

$$\begin{cases} \lambda_1 = \eta_x u_l + \eta_y v_l \\ \\ \lambda_2 = \lambda_1/2 \\ \\ \lambda_3 = \lambda_1 - c_l||\boldsymbol{\eta}|| \\ \\ \lambda_4 = \lambda_1 + c_l||\boldsymbol{\eta}|| \end{cases} \tag{VIII.22}$$

où la vitesse c_l est définie par : $c_l = \sqrt{\dfrac{\delta\theta}{(1-\alpha)\rho_l}}$

Remarque: Le nombre c_l est homogène à une vitesse mais n'a rien à voir avec la vitesse du son dans le liquide pur. La vitesse c_l est en général beaucoup plus petite que c_g.

VIII.3.2 Méthode de résolution.

Le système (VIII.3) est un système de lois de conservation avec un terme source. Il y a deux échelles de temps : l'une est liée à la convection, l'autre, beaucoup plus petite, est le temps de relaxation ϵ. Lorsque ϵ tend vers 0, le rapport entre ces deux échelles devient grand, et le système (VIII.3) est dit raide. Une méthode classique et efficace pour résoudre numériquement en temps un tel système est une méthode de pas fractionnaires (appelée aussi méthode de "splitting d'opérateurs"). Cette méthode conduit à traiter d'une part le système homogène découplé, puis à résoudre un système d'équations différentielles prenant en compte le terme source.

Soit $\mathbf{W}^n$ une approximation de la solution $\mathbf{W}(x,t)$ du système (VIII.3) au temps t^n. La méthode de splitting consiste à résoudre les deux sous-systèmes homogènes découplés de la phase gazeuse et de la phase liquide pendant un pas de temps Δt :

$$\begin{cases} \partial_t \mathbf{W}_k + \partial_x \mathbf{F}_k(\mathbf{W}) + \partial_y \mathbf{G}_k(\mathbf{W}) = 0 \quad k = 1,2 \\ \mathbf{W}_k(x,y,0) = \mathbf{W}_k^n \end{cases}$$

On obtient alors une solution approchée $\mathbf{W}_k(x, y, \Delta t)$.
La deuxième étape consiste à résoudre le système différentiel suivant pendant un pas de temps Δt :

$$\begin{cases} \partial_t \mathbf{V} = \dfrac{\mathbf{R}(\mathbf{V})}{\epsilon} \\ \mathbf{V}(x, y, 0) = \mathbf{W}(x, y, \Delta t) \end{cases}$$

où $\mathbf{W} = (\mathbf{W}_1, \mathbf{W}_2)$ désigne le vecteur d'état du sytème complet (gaz-liquide). On pose $\mathbf{W}^{n+1} = \mathbf{V}(x, y, \Delta t)$ et on obtient ainsi la solution approchée au temps $t^{n+1} = t^n + \Delta t$.

Nous envisageons aussi une "méthode couplée" qui consiste à ne pas séparer le traitement de la convection et celui des termes source. On résout directement le système pendant un pas de temps Δt :

$$\begin{cases} \partial_t \mathbf{W} + \partial_x \mathbf{F}(\mathbf{W}) + \partial_y \mathbf{G}(\mathbf{W}) = \dfrac{\mathbf{R}(\mathbf{W})}{\epsilon} \\ \mathbf{W}(x, y, 0) = \mathbf{W}^n \end{cases}$$

et on obtient la solution approchée $\mathbf{W}^{n+1} = \mathbf{W}(x, y, \Delta t)$ au temps t^{n+1}.

VIII.3.3 Formulation variationnelle.

La formulation variationnelle pour notre modèle diphasique est la même que celle des équations d'Euler. On utilise la même discrétisation du domaine Ω_h et le même type de conditions aux limites.
On considère le problème de Cauchy:

$$\begin{cases} \mathbf{W}_t \,+\, \nabla.\mathbb{F}(\mathbf{W}) = \chi \dfrac{\mathbf{R}(\mathbf{W})}{\epsilon} & (\mathbf{x}, t) \in \Omega \times \mathbb{R}^+ \\ \mathbf{W}(\mathbf{x}, t = 0) \,=\, \mathbf{W}^0(\mathbf{x}) & \mathbf{x} \in \Omega \end{cases} \tag{VIII.23}$$

avec $\chi = 1$ si on considère la "méthode couplée" et $\chi = 0$ pour la "méthode de splitting". On complète (VIII.23) avec des conditions aux limites. La formulation faible de (VIII.23) s'écrit pour chaque cellule C_i :

$$\int_{C_i} (\mathbf{W}_t + \nabla.\mathbb{F}(\mathbf{W}))\, d\mathbf{x} \;=\; \chi \int_{C_i} \frac{\mathbf{R}(\mathbf{W})}{\epsilon}\, d\mathbf{x} \tag{VIII.24}$$

En supposant $(\mathbf{W}_t)$ constant sur la cellule C_i et en appliquant la formule de Green, on obtient:

$$\begin{aligned} \text{Aire}\, C_i\, (\mathbf{W}_t)_i \;=\; & -\sum_{j\in K(i)} \int_{\partial C_{ij}} \mathbb{F}(\mathbf{W}).\boldsymbol{\nu_{ij}}\, d\sigma \quad < 1 > \\ & - \int_{\partial C_i \cap \Gamma_b} \mathbb{F}(\mathbf{W}).\boldsymbol{n}\, d\sigma \quad < 2 > \\ & - \int_{\partial C_i \cap \Gamma_\infty} \mathbb{F}(\mathbf{W}).\boldsymbol{n}\, d\sigma \quad < 3 > \\ & + \chi \int_{C_i} \frac{\mathbf{R}(\mathbf{W})}{\epsilon}\, d\mathbf{x} \quad < 4 > \end{aligned} \tag{VIII.25}$$

avec les notations déjà introduites dans le chapitre III.

VIII.3.4 Calcul des flux.

L'approximation du terme $\displaystyle\int_{\partial C_{ij}} \mathbb{F}(\mathbf{W}).\boldsymbol{\nu_{ij}}\ d\sigma$ a été détaillé dans la section VI.4.1 pour les équations d'Euler. Ici, nous utilisons le solveur de Roe (VI.18,VI.19) pour les deux sous-systèmes. L'état moyen $\tilde{\mathbf{W}}_1$ lié à la phase gazeuse est défini par les relations (VI.24), où la quantité $\tilde{\rho}$ est remplaçée par $\tilde{\alpha\rho}$.

En ce qui concerne l'écoulement liquide, la matrice jacobienne $\mathcal{A}_2(\tilde{\mathbf{W}}, \boldsymbol{\eta})$ ne satisfait pas la propriété (VI.20). On prend pour matrice $\mathcal{R}$ une matrice $\tilde{\mathcal{A}}_2(\mathbf{W}_i, \mathbf{W}_j, \boldsymbol{\eta})$, différente de $\mathcal{A}(\tilde{\mathbf{W}})$ mais toutefois très proche de celle-ci.
On construit l'état moyen $\tilde{\mathbf{W}}_2$ comme suit :

$$\mathbf{W}_i^2 = \begin{pmatrix} 1-\alpha_i \\ u_i \\ v_i \\ (1-\alpha_i)\varepsilon_i \end{pmatrix}, \quad \mathbf{W}_j^2 = \begin{pmatrix} 1-\alpha_j \\ u_j \\ v_j \\ (1-\alpha_j)\varepsilon_j \end{pmatrix}, \quad \tilde{\mathbf{W}}_2 = \begin{pmatrix} 1-\tilde{\alpha} \\ \tilde{u} \\ \tilde{v} \\ (1-\tilde{\alpha})\tilde{\varepsilon} \end{pmatrix},$$

avec:

$$1-\tilde{\alpha} = \frac{(1-\alpha_i)+(1-\alpha_j)}{2}\,,$$

$$\tilde{u} = \frac{u_i+u_j}{2}\,,$$

$$\tilde{v} = \frac{v_i+v_j}{2}\,,$$

$$\tilde{\varepsilon} = \frac{\varepsilon_i+\varepsilon_j}{2}\,,$$

L'expression de la matrice de Roe $\tilde{\mathcal{A}}_2(\mathbf{W}_i, \mathbf{W}_j, \boldsymbol{\eta})$ est donnée dans l'Annexe D.

VIII.3.5 Approximation d'ordre supérieur en espace.

L'extension à l'ordre deux, trois ou quatre en espace se fait, comme dans le cas linéaire (voir section III.3.3), à l'aide de β-schémas (III.10) appliqués aux deux sous-sytèmes du gaz et des gouttes. Dans le cas d'un maillage triangulaire, on utilise (III.12) comme approximation du gradient.

VIII.3.6 Traitement des conditions aux limites.

Comme pour le système d'équations d'Euler (voir paragraphe VI.4.2), la frontière Γ_h du domaine Ω_h se décompose en Γ_b, qui représente une paroi solide, et Γ_∞ constitué des frontières amont et aval de l'écoulement:

$$\Gamma_h = \Gamma_b \bigcup \Gamma_\infty \ .$$

On a à évaluer les deux intégrales :

$$\int_{\partial C_i \cap \Gamma_\infty} \mathcal{F}\left(\mathbf{W}, \mathbf{n}\right) d\sigma \ ,$$

$$\int_{\partial C_i \cap \Gamma_b} \mathcal{F}\left(\mathbf{W}, \mathbf{n}\right) d\sigma.$$

Traitement de la paroi

On applique une condition de glissement $u\, n_x + v\, n_y = 0$ sur la paroi Γ_b pour le gaz et le liquide. L'expression du flux sur le bord devient :

- <u>pour le gaz</u>

L'expression du flux $\mathbf{\Phi}^1_{ib}(\mathbf{W}_i, \boldsymbol{\eta})$ est donnée par (VI.26), où la pression p_i, supposée constante sur le segment de bord, est remplacée par αp_i.

- <u>pour le liquide :</u>

Nous obtenons :

$$\mathcal{F}_2(\mathbf{W}_i, \mathbf{n}) = \begin{pmatrix} 0 \\ n_x \dfrac{\delta\theta_0}{\rho_l} \dfrac{(1-\alpha_i)^{\delta-1}}{(\delta-1)} \\ n_y \dfrac{\delta\theta_0}{\rho_l} \dfrac{(1-\alpha_i)^{\delta-1}}{(\delta-1)} \\ 0 \end{pmatrix} \ ;$$

et

$$\mathbf{\Phi}_{ib}^{2}(\mathbf{W}_i,\boldsymbol{\eta}) = \int_{\partial C_i\cap\Gamma b} \mathcal{F}_2\left(\mathbf{W},\mathbf{n}\right) d\sigma \;=\; \begin{pmatrix} 0 \\ \eta_x \dfrac{\delta\theta_0}{\rho_l}\dfrac{(1-\alpha_i)^{\delta-1}}{(\delta-1)} \\ \eta_y \dfrac{\delta\theta_0}{\rho_l}\dfrac{(1-\alpha_i)^{\delta-1}}{(\delta-1)} \\ 0 \end{pmatrix} \quad ;$$

Conditions à l'infini.

Comme dans le cas des équations d'Euler, on se donne un champ $\mathbf{W}_\infty$ à l'extérieur du domaine Ω_h. Le traitement des conditions à l'infini pour le sous-système lié au gaz est explicité à la section VI.4.2.
En ce qui concerne le liquide, les flux physiques $\mathbf{F}_2(\mathbf{W})$ et $\mathbf{G}_2(\mathbf{W})$ n'étant pas homogènes d'ordre un, on approche $\displaystyle\int_{\partial C_i\cap\Gamma\infty} \mathcal{F}\left(\mathbf{W},\mathbf{n}\right) d\sigma$ par un flux de Roe évalué entre les états $\mathbf{W}_i$ et $\mathbf{W}_\infty$.

VIII.3.7 Intégration en temps.

Supposons connue $\mathbf{W}^n(x,y)$ une solution approchée au temps $t = t^n$ du système (VIII.3). Nous explicitons le calcul de $\mathbf{W}^{n+1}(x,y)$ la solution approchée au temps $t^{n+1} = t^n + \Delta t$, pour la "méthode de splitting" et la "méthode couplée".

Méthode de splitting.

– Schéma d'ordre un en temps.

Le schéma d'ordre un est obtenu de la manière suivante:

Etape 1: on résout pendant un pas de temps Δt les deux sous-systèmes découplés (gaz-liquide) de lois de conservation:

$$\begin{cases} \partial_t \mathbf{W}_k + \partial_x \mathbf{F}_k(\mathbf{W}) + \partial_y \mathbf{G}_k(\mathbf{W}) = 0 \quad k = 1,2 \\ \mathbf{W}_k(x,y,0) = \mathbf{W}_k^n \end{cases}$$

On obtient alors une solution approchée $\mathbf{W}_k^{n+1/2}(x,y) = \mathbf{W}_k(x,y,\Delta t)$.

Etape 2: on résout pendant un pas de temps Δt le système différentiel:

$$\begin{cases} \partial_t \mathbf{W} = \dfrac{\mathbf{R}(\mathbf{W})}{\epsilon} \\ \mathbf{W}(x,y,0) = \mathbf{W}^{n+1/2}(x,y) \end{cases}$$

On pose $\mathbf{W}^{n+1}(x,y) = \mathbf{W}(x,y,\Delta t)$.

Il est très facile de comprendre pourquoi ce schéma est d'ordre un en temps en considérant une équation simplifiée de la forme : $\partial_t u + (A+B)u = 0$, où A et B sont des matrices constantes. En appliquant la méthode de splitting ci-dessus et en effectuant des développements limités en Δt de la solution approchée, on montre que si A et B ne commutent pas, l'approximation en temps est seulement d'ordre un.

– <u>Schéma d'ordre deux en temps.</u>

On passe à l'ordre deux en utilisant la méthode de splitting de Strang [20, 16].

Etape 1 : on résout pendant un pas de temps $\Delta t/2$ le système d'équations différentielles ordinaires :

$$\begin{cases} \partial_t \mathbf{W} = \dfrac{\mathbf{R}(\mathbf{W})}{\epsilon} \\ \mathbf{W}(x,y,0) = \mathbf{W}^n(x,y) \end{cases}$$

On pose $\mathbf{W}^{n+1/3}(x,y) = \mathbf{W}(x,y,\Delta t/2)$.

Etape 2 : on résout pendant un pas de temps Δt les deux systèmes (gaz-liquide) de lois de conservation :

$$\begin{cases} \partial_t \mathbf{W}_k + \partial_x \mathbf{F}_k(\mathbf{W}) + \partial_y \mathbf{G}_k(\mathbf{W}) = 0 \quad k = 1,2 \\ \mathbf{W}_k(x,y,0) = \mathbf{W}_k^{n+1/3} \end{cases}$$

On pose $\mathbf{W}^{n+2/3}(x,y) = \mathbf{W}(x,y,\Delta t)$.

Etape 3 : on résout pendant un pas de temps $\Delta t/2$ le système différentiel ordinaire :

$$\begin{cases} \partial_t \mathbf{W} = \dfrac{\mathbf{R}(\mathbf{W})}{\epsilon} \\ \mathbf{W}(x,y,0) = \mathbf{W}^{n+2/3}(x,y) \end{cases}$$

On pose $\mathbf{W}^{n+1}(x,y) = \mathbf{W}(x,y,\Delta t/2)$.

Cette méthode est bien du second ordre en temps; on peut le vérifier en reprenant l'exemple linéaire $\partial_t u + (A+B)u = 0$. Même si A et B ne commutent pas, le schéma est d'ordre deux.

Un schéma symétrique existe, mais on a préféré effectuer deux fois l'opération la moins coûteuse.

– <u>Résolution du système différentiel.</u>

On veut résoudre numériquement le système :

$$\begin{cases} \partial_t \mathbf{W} = \dfrac{\mathbf{R}(\mathbf{W})}{\epsilon} \\ \mathbf{W}(x,y,0) = \mathbf{W}^0(x,y) \end{cases} \tag{VIII.26}$$

Dans le cas particulier du terme source du système (VIII.3), on peut intégrer le système (VIII.26) explicitement. Pour (x, y) donné, la solution $\mathbf{W}(x, y, t)$ de (VIII.26) ne dépend que de la condition initiale $\mathbf{W}^0(x, y)$ et de t. Il nous suffit donc de résoudre le système différentiel :

$$\begin{cases} \dfrac{d\mathbf{W}}{dt} = \dfrac{\mathbf{R}(\mathbf{W})}{\epsilon} \\ \mathbf{W}(0) = \mathbf{W}^0 \end{cases} \tag{VIII.27}$$

en chaque point (x, y) pour obtenir la solution de (VIII.26).

Soit $\mathbf{W}^0$ un état constant :

$$\mathbf{W}^0 = (\alpha\rho_g^0, \alpha\rho_g^0 u_g^0, \alpha\rho_g^0 v_g^0, \alpha\rho_g^0 e_g^0, (1-\alpha^0), u_l^0, v_l^0, (1-\alpha^0)\varepsilon_l^0)$$

On note c le rapport des densités de la phase liquide et du gaz:

$$c = \frac{(1-\alpha)\rho_l}{\alpha\rho_g} \tag{VIII.28}$$

Proposition VIII.3.2 *Le système (VIII.27) admet une solution unique pour tout $t > 0$. Les densités des deux phases, ainsi que l'énergie spécifique interne de la phase liquide restent constantes au cours du temps:*

$$\begin{aligned} &\alpha(t)\rho_g(t) = \alpha^0\rho_g^0 \\ &\rho_l(1-\alpha(t)) = \rho_l(1-\alpha^0) \\ &\varepsilon_l(t) = \varepsilon_l^0 \end{aligned}$$

Les vitesses du gaz et du liquide ainsi que l'énergie spécifique totale du gaz s'écrivent :

$$u_g(t) = \frac{u_g^0 + cu_l^0}{1+c} + \exp\left(-\frac{(1+c)t}{\epsilon}\right) \frac{c(u_g^0 - u_l^0)}{1+c}$$

$$v_g(t) = \frac{v_g^0 + cv_l^0}{1+c} + \exp\left(-\frac{(1+c)t}{\epsilon}\right) \frac{c(v_g^0 - v_l^0)}{1+c}$$

$$u_l(t) = \frac{u_g^0 + cu_l^0}{1+c} - \exp\left(-\frac{(1+c)t}{\epsilon}\right) \frac{(u_g^0 - u_l^0)}{1+c}$$

$$v_l(t) = \frac{v_g^0 + cv_l^0}{1+c} - \exp\left(-\frac{(1+c)t}{\epsilon}\right) \frac{(v_g^0 - v_l^0)}{1+c}$$

$$\begin{aligned} e_g(t) = e_g^0 - \frac{c}{2(1+c)^2}&(1 - \exp(-\frac{(1+c)t}{\epsilon})) \\ \Bigg[(u_g^0 - u_l^0)&\left\{(1-\exp(-\frac{(1+c)t}{\epsilon}))u_g^0 - (\exp(-\frac{(1+c)t}{\epsilon}) + 2c + 1)u_l^0\right\} \\ +(v_g^0 - v_l^0)&\left\{(1-\exp(-\frac{(1+c)t}{\epsilon}))v_g^0 - (\exp(-\frac{(1+c)t}{\epsilon}) + 2c + 1)v_l^0\right\}\Bigg] \end{aligned}$$

Preuve : Les équations de (VIII.26) sur $\alpha\rho_g$, $(1-\alpha)$, $(1-\alpha)\varepsilon_l$ sont :

$$\partial_t(\alpha\rho_g) = 0$$
$$\partial_t(1-\alpha)\rho_l = 0$$
$$\partial_t(1-\alpha)\varepsilon_l = 0$$

On en déduit que les densités $\alpha\rho_g$, $(1-\alpha)\rho_l$ et l'énergie interne $(1-\alpha)\varepsilon_l$ sont constantes, ainsi que c. Les équations sur la vitesse du gaz et du liquide sont obtenues en soustrayant l'équation de conservation de l'impulsion du gaz (VIII.3.ii) à celle des gouttes (VIII.3.vi) et en l'intégrant explicitement [4, 2]. On obtient alors :

$$(u_g - u_l)(t) = (u_g^0 - u_l^0)\exp\left(-\frac{(1+c)t}{\epsilon}\right) \qquad \text{(VIII.29)}$$

$$(v_g - v_l)(t) = (v_g^0 - v_l^0)\exp\left(-\frac{(1+c)t}{\epsilon}\right) \qquad \text{(VIII.30)}$$

En utilisant la conservation de l'impulsion totale, on obtient les expressions ci-dessus pour u_g, u_l, v_g, v_l.
L'équation sur l'énergie du gaz est en fait une quadrature à partir des expressions explicites de u_g et v_g •

On remarque, avec les expressions (VIII.29),(VIII.30) que le terme source $\dfrac{\mathbf{R(W)}}{\epsilon}$ tend à égaliser les vitesses du gaz et du liquide. De plus, quand $t \longrightarrow +\infty$, la vitesse du gaz $(u_g(t), v_g(t))$ et celle des gouttes $(u_l(t), v_l(t))$ tendent vers l'état (u_∞, v_∞) :

$$u_\infty = \frac{\alpha^0\rho_g^0 u_g^0 + (1-\alpha^0)\rho_l u_l^0}{\alpha^0\rho_g^0 + (1-\alpha^0)\rho_l} \quad , \quad v_\infty = \frac{\alpha^0\rho_g^0 v_g^0 + (1-\alpha^0)\rho_l v_l^0}{\alpha^0\rho_g^0 + (1-\alpha^0)\rho_l} \qquad \text{(VIII.31)}$$

qui est la vitesse moyenne de l'écoulement diphasique.

– Résolution par un schéma explicite.

Les systèmes différentiels des étapes 1 et 3 peuvent être intégrés explicitement. Pour l'étape 2, on utilise un schéma décentré d'ordre deux en temps et au moins d'ordre deux en espace. L'ordre deux en temps est obtenu à l'aide d'un schéma Runge-Kutta à deux pas.

La stabilité de ce schéma est conditionnée par un critère du type Courant-Friedrichs-Lewy (CFL) limitant les valeurs du pas de temps. On utilise comme condition :

$$\Delta t \leq \min_T \frac{h_{min}}{\lambda_{max}},$$

où h_{min} est la plus petite hauteur du triangle T, et λ_{max} la plus grande vitesse caractéristique du domaine.

Méthode couplée.

La précision en temps est importante dans la recherche de solutions instationnaires. Dans le cas de la méthode couplée, le terme source a une grande influence sur la stabilité du schéma. En particulier, lorsque le terme source devient raide, le pas de temps maximal que l'on peut choisir devient très petit [9].
Si on considère l'équation scalaire de convection modèle avec terme source dans le cas monodimensionnel :

$$u_t + c\,u_x - \lambda u = 0$$

avec c vitesse de transport et λ un réel positif ou négatif, et si l'on étudie par l'analyse de Fourier, la limite de stabilité du schéma explicite décentré d'ordre un :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_j^n - u_{j-1}^n}{\Delta x} - \lambda u_j^n = 0,$$

on obtient comme condition sur le pas de temps Δt :

$$\frac{c\Delta t}{\Delta x} \leq \frac{1}{1 - \dfrac{\lambda\Delta x}{2c}}. \tag{VIII.32}$$

On voit qu'une condition nécessaire de stabilité est $\lambda \leq 2c/\Delta x$. Dans notre cas, λ modélise l'effet des termes sources et on la prendra égale à une valeur propre de $\dfrac{\mathbf{R}(\mathbf{W})}{\epsilon}$. Dans le cas $\lambda \leq 0$, la présence du terme source a pour conséquence de restreindre la valeur maximale du pas de temps.
A titre d'exemple, on considère la propagation d'une onde sonore dans un milieu diphasique (les constantes physiques utilisées sont détaillées plus loin). Le pas d'espace vaut $\Delta x = 0.25$. Si le rayon des gouttes varie de 10^{-6} à 10^{-4} mètres, le module des valeurs propres varie de 100 à 10^6 sec^{-1}. La vitesse de propagation de l'onde V varie entre 1141 et 1360 m/s. En utilisant la formule (VIII.32), on aboutit aux valeurs de CFL suivantes :

r	$\lvert\lambda\rvert$	V	$\lvert\lambda\rvert\Delta x/V$	CFL
1.10^{-4}	100	1360	0.02	0.98
1.10^{-5}	4000	1146	0.87	0.7
5.10^{-6}	10^4	1142	2.1	0.5
1.10^{-6}	10^6	1141	210	0.01

TAB. VIII.1 – *Exemple numérique.*

On voit que pour des rayons de gouttes $\geq 5.10^{-6}$ mètres, les valeurs de CFL sont acceptables, par contre pour des rayons petits ($r = 1.10^{-6}$ mètres), le pas de temps devient excessivement petit (le CFL est 100 fois plus petit que la limite habituelle CFL = 1 en l'absence de termes source).

Nous cherchons à augmenter la valeur maximale du pas de temps Δt et aussi la précision en temps du schéma. Pour cela, nous allons mettre en oeuvre des méthodes de Runge-Kutta non linéaires d'ordre deux, trois ou quatre en temps. Nous utilisons les schémas (VI.9) décrits dans la section VI.5, où l'opérateur $\mathbf{D}(\mathbf{W}) = -\frac{1}{\mathrm{aireC_i}} \sum_{j \in K(i)} \Phi_{ij} + \frac{\mathbf{R}(\mathbf{W})}{\epsilon}$.

VIII.3.8 Résolution par un schéma implicite.

Les schémas implicites s'avèrent particulièrement intéressants pour résoudre des problèmes raides car ils n'ont pas, en général, de restriction sur le pas de temps. Nous nous intéressons aussi au calcul de solutions stationnaires obtenues à partir d'une approche instationnaire, afin de comparer nos deux méthodes ("méthode couplée" et "méthode de splitting"). Pour cela, nous utilisons les schéma implicites linéarisés décrits dans le paragraphe VI.6. Le schéma implicite d'ordre un en temps avec terme source s'écrit :

$$\frac{\mathbf{W}^{n+1} - \mathbf{W}^n}{\Delta t} + \mathbf{H}\left(\mathbf{W}^{n+1}\right) = \chi \frac{\mathbf{R}(\mathbf{W}^{n+1})}{\epsilon} , \tag{VIII.33}$$

avec $\chi = 1$ pour la "méthode couplée" et $\chi = 0$ pour la "méthode de splitting", et :

$$\Delta t = t^{n+1} - t^n, \quad \mathbf{W}^n = \mathbf{W}(x, y, t^n).$$

En linéarisant les flux convectifs et le terme source ($\mathbf{R}(\mathbf{W})$ est différentiable), on obtient le système suivant :

$$\left(\frac{Id}{\Delta t} + \mathbf{P}^n - \chi \frac{\mathbf{R}'\left(\mathbf{W}^n\right)}{\epsilon}\right) \delta \mathbf{W} = -\mathbf{H}\left(\mathbf{W}^n\right) + \chi \frac{\mathbf{R}\left(\mathbf{W}^n\right)}{\epsilon} . \tag{VIII.34}$$

où $\mathbf{P}^n$ approche le vrai jacobien $\mathbf{H}'\left(\mathbf{W}^n\right)$.

La linéarisation des flux convectifs pour un flux de Roe a déjà été décrite au chapitre VI. Il reste à expliciter la linéarisation des flux de bord pour le sous-système des gouttes; le traitement des flux de bord pour le sous-système du gaz étant identique à celui décrit dans la section VI.6.2.

Linéarisation sur le bord $\delta\Omega$:

Comme dans le cas explicite, le calcul sur les bords se décompose en deux parties :

- <u>Sur Γ_∞ :</u>

Pour le liquide, on linéarise le flux de Roe entre les états $\mathbf{W}_i$ et $\mathbf{W}_\infty$, on obtient :

$$\Phi_{i\infty}^{n+1} = \Phi_{i\infty}^{n} + \mathcal{A}(\mathbf{W}_i^n, \boldsymbol{\eta})\delta\mathbf{W}_i - \tilde{\mathcal{A}}^+(\mathbf{W}_i^n, \mathbf{W}_j^n, \boldsymbol{\eta})\delta\mathbf{W}_i$$

où $\tilde{\mathcal{A}}^+(\mathbf{W}_i^n, \mathbf{W}_j^n, \boldsymbol{\eta})$ est la partie positive de la matrice de Roe calculée entre les états $\mathbf{W}_i^n$ et $\mathbf{W}_\infty^n$.

– <u>Sur Γ_b :</u>

Sur la paroi Γ_b (bord glissant), on différencie le flux $\Phi_{ib}^2(\mathbf{W}, \boldsymbol{\eta})$ construit à partir de la condition de glissement $\mathbf{u}.\mathbf{n} = 0$, on obtient son jacobien $\mathcal{A}_b^2(\mathbf{W}, \boldsymbol{\eta})$.

$$\mathcal{A}_b^2(\mathbf{W}, \boldsymbol{\eta}) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \eta_x \frac{\delta\theta_0}{\rho_l}(1-\alpha)^{\delta-2} & 0 & 0 & 0 \\ \eta_y \frac{\delta\theta_0}{\rho_l}(1-\alpha)^{\delta-2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Linéarisation du terme source.

L' expression (VIII.34) de la méthode implicite linéarisée nous amène à calculer la jacobienne du terme source $\dfrac{\mathbf{R}'(\mathbf{W})}{\epsilon}$. Nous avons la proposition suivante.

Proposition VIII.3.3 *La matrice jacobienne du terme source est diagonalisable. Elle a huit valeurs propres réelles négatives ou nulles :*

$$\begin{cases} \mu_{1,6} = 0 \quad \textit{de multiplicite six}, \\ \mu_{7,8} = -\dfrac{1}{\epsilon}(1+c) \quad \textit{de multiplicite double}. \end{cases}$$

avec $c = \dfrac{(1-\alpha)\rho_l}{\alpha\rho_g}$ le rapport des densités des deux phases.

La preuve est immédiate, on donne simplement l'expression de la matrice jacobienne du terme source dans l'Annexe D.

Méthode de résolution.

L'algorithme implicite s'écrit sous la forme d'un schéma à deux étapes : une première étape explicite qui prend en compte les données physiques du problème à résoudre,

$$\text{RHS} = -\left(\sum_{j\in K(i)} \Phi_{ij}^n + \Phi_{i\infty}^n + \Phi_{ib}^n\right) + \chi \text{ aire } \mathrm{C_i}\frac{\mathbf{R}(\mathbf{W}_i^n)}{\epsilon} \qquad \text{(VIII.35)}$$

puis une étape implicite ù l'on résout le système suivant :

$$\begin{cases} \dfrac{\text{aire C}_\text{i}}{\Delta t}\delta\mathbf{W}_i + \displaystyle\sum_{j\in K(i)} \Big(\mathbf{A}(\mathbf{W}_i^n,\boldsymbol{\eta})\delta\mathbf{W}_i - \tilde{\mathbf{A}}^+(\mathbf{W}_i^n,\mathbf{W}_j^n,\boldsymbol{\eta})(\delta\mathbf{W}_i - \delta\mathbf{W}_j)\Big) \\ \qquad + \Big(\mathbf{A}(\mathbf{W}_i^n,\boldsymbol{\eta}_{i\infty})\delta\mathbf{W}_i - \tilde{\mathbf{A}}^+(\mathbf{W}_i^n,\mathbf{W}_\infty,\boldsymbol{\eta}_{i\infty})\delta\mathbf{W}_i\Big) \\ \qquad + \mathbf{A}_b(\mathbf{W}_i^n,\boldsymbol{\eta})\delta\mathbf{W}_i - \chi \text{ aire C}_\text{i}\dfrac{\mathbf{R}'(\mathbf{W}_i^n)}{\epsilon}\delta\mathbf{W}_i = \text{ RHS} \\ \mathbf{W}_i^{n+1} = \mathbf{W}_i^n + \delta\mathbf{W}_i \end{cases}$$

La méthode de résolution de notre système diphasique par un schéma implicite est identique à celle utilisée pour les équations d'Euler (voir section VI.6.3). Nous avons un système linéaire à résoudre à chaque pas de temps, qui s'écrit :

$$(\frac{\text{aire C}_\text{i}}{\Delta t}Id + \mathcal{B})\delta\mathbf{W} = RHS \tag{VIII.36}$$

Dans le cas de la "méthode de splitting" ($\chi = 0$), nous avons deux systèmes découplés à résoudre, un pour le gaz et l'autre pour le liquide. $\mathcal{B}$ est une matrice tridiagonale par blocs, formée de blocs 4x4 par noeud. Si $\chi = 1$, les deux phases liquide-gaz sont couplées et la matrice $\mathcal{B}$ du système est constituée de blocs 8x8 par noeud.

Pour résoudre le système (VIII.36), nous utilisons la méthode de Jacobi (VI.38).

Convergence vers la solution stationnaire.

On a vu que le schéma implicite décentré sans termes source appliqué à l'équation d'advection est inconditionnellement stable [18]. Dans notre cas, nous avons le résultat suivant :

Proposition VIII.3.4 *Le schéma implicite décentré d'ordre un en temps, appliqué à l'équation $u_t + c\,u_x = -Ku$, $(K \geq 0)$ est inconditionellement stable, pour des flux d'ordre un ou deux en espace.*

Preuve : Plaçons-nous pour simplifier dans le cas $c > 0$. Le schéma implicite décentré s'écrit :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_j^{n+1} - u_{j-1}^{n+1}}{\Delta x} = -Ku_j^{n+1}. \tag{VIII.37}$$

L'étude de stabilité se fait au moyen d'une analyse de Fourier. Soit u_j^n les modes de Fourier de la forme $u_j^n = exp\,(ij\xi_k)$, $\xi_k = \dfrac{2\pi k}{N+1}$, $(k = 0, ..., N)$, $i^2 = -1$.
u_j^{n+1} s'écrit en fonction de u_j^n : $u_j^{n+1} = g_k(\Delta t)u_j^n$, où $g_k(\Delta t)$ est le facteur d'amplification du schéma. Le schéma est stable si la condition de Von Neumann est vérifiée :

$$\max_{[k=0,n]} \mid g_k(\Delta t) \mid \leq 1. \tag{VIII.38}$$

On obtient après analyse de Fourier :

$$g_k(\Delta t) = \frac{1}{1 + \sigma\left(1 - exp(-i\xi_k)\right) + K\Delta t}, \quad \sigma = c\Delta t/\Delta x \tag{VIII.39}$$

et, pour $K \geq 0$, $\mid g_k(\Delta t) \mid \leq 1 \ \ \forall \xi_k \in [0, 2\pi]$.

Dans le cas d'un flux d'ordre au moins deux en espace obtenu à l'aide du β-schéma (III.10), le schéma s'écrit :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_{j+1/2,-}^{n+1} - u_{j-1/2,-}^{n+1}}{\Delta x} = -Ku_j^{n+1} \tag{VIII.40}$$

avec :

$$\begin{cases} u_{j-1/2,-} = u_{j-1} + \dfrac{1}{2}\left\{(1-\beta)(u_j - u_{j-1}) + \beta(u_{j-1} - u_{j-2})\right\} \\ u_{j+1/2,-} = u_j + \dfrac{1}{2}\left\{(1-\beta)(u_{j+1} - u_j) + \beta(u_j - u_{j-1})\right\} \end{cases}$$

L'analyse de Fourier donne :

$$g_k(\Delta t) = \frac{1}{1 + \sigma f(\beta) + K\Delta t}, \quad \sigma = c\Delta t/\Delta x \tag{VIII.41}$$

avec

$$f(\beta) = \frac{\beta}{2}exp(-2i\xi_k) - \frac{1+3\beta}{2}exp(-i\xi_k) + \frac{1-\beta}{2}exp(i\xi_k) + \frac{3\beta}{2} \tag{VIII.42}$$

Comme $\mathrm{Re}\left(f(\beta)\right) = \beta(1 - \cos(\xi_k))^2 \geq 0$, on a bien pour $K \geq 0$, $\mid g_k(\Delta t) \mid \leq 1$ $\forall \xi_k \in [0, 2\pi]\bullet$

On note RES^n la fonction d'erreur résiduelle à l'instant n. Pour le gaz, on prendra :

$$\mathrm{RES}_1^n = \frac{||\mathbf{G}(\rho^n) - \chi\dfrac{\mathbf{R}(\rho^n)}{\epsilon}||_1}{||\mathbf{G}(\rho^0) - \chi\dfrac{\mathbf{R}(\rho^0)}{\epsilon}||}$$

et pour le liquide :

$$\mathrm{RES}_2^n = \frac{||\mathbf{G}(1-\alpha^n) - \chi\dfrac{\mathbf{R}(1-\alpha^n)}{\epsilon}||_1}{||\mathbf{G}(1-\alpha^0) - \chi\dfrac{\mathbf{R}(1-\alpha^0)}{\epsilon}||}$$

où $||.||_1$ désigne la norme l_1.
On dira que le schéma a convergé vers la solution stationnaire lorsque $\max(\mathrm{RES}_1^n, \mathrm{RES}_2^n)$ est atteint.

Il est possible d'utiliser de grands pas de temps avec des lois de CFL faisant intervenir le résidu à chaque itération, du type :

$$CFL = \mathrm{MIN}(\mathrm{MAX}(y, z.kt, \frac{x}{\max(\mathrm{RES}_1^n, \mathrm{RES}_2^n)}), \mathrm{CFLmax}) \tag{VIII.43}$$

où x, y, z sont des réels à ajuster en fonction du type de l'écoulement, kt est le nombre d'itérations effectuées et CFLmax est le CFL à atteindre pour converger plus rapidement vers l'état stationnaire.

Schéma implicite dans le cas instationnaire.

Nous sommes amenés, pour modéliser des problèmes instationnaires (comme la propagation d'une onde sonore dans un milieu diphasique), à rechercher des schémas implicites précis en temps et en espace. Pour cela, nous nous intéressons au Θ-schéma suivant :

$$\frac{\mathbf{W}^{n+1} - \mathbf{W}^n}{\Delta t} + \Big(\Theta\, \mathbf{H}(\mathbf{W}^{n+1}) + (1-\Theta)\mathbf{H}(\mathbf{W}^n)\Big) = \chi \frac{\mathbf{R}(\mathbf{W}^{n+1})}{\epsilon}\ , \tag{VIII.44}$$

Pour la valeur $\Theta = 1/2$, on obtient le schéma de Cranck-Nicolson.

On peut montrer que le schéma de Cranck-Nicolson, en l'absence de termes source, appliqué à l'équation d'advection monodimensionnelle est d'ordre deux en temps. Si de plus, les flux sont calculés à l'aide du β-schéma (III.10), il est au moins d'ordre deux en espace. La preuve de résultat peut être trouvée dans [18]. Lorsque l'on prend en compte le terme source, nous avons le résultat suivant :

Proposition VIII.3.5 *Le schéma de Cranck-Nicolson appliqué à l'équation $u_t + c\, u_x = -Ku$, ($K \geq 0$) est inconditionnellement stable, pour des flux d'ordre un ou deux en espace.*

Preuve : On se place dans le cas $c > 0$. Le schéma d'ordre un en espace s'écrit :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{c}{2}\left(\frac{u_j^{n+1} - u_{j-1}^{n+1}}{\Delta x} + \frac{u_j^n - u_{j-1}^n}{\Delta x}\right) = -K u_j^{n+1}. \tag{VIII.45}$$

En proçédant par analyse de Fourier comme pour les études précédentes, on obtient comme coefficient d'amplification :

$$g_k(\Delta t) = \frac{1 - \frac{\sigma}{2}(1 - exp(-i\xi_k))}{1 + \frac{\sigma}{2}(1 - exp(-i\xi_k)) + K\Delta t}, \quad \sigma = c\Delta t/\Delta x$$

et on a bien pour $K \geq 0$, $\mid g_k(\Delta t) \mid \leq 1 \ \ \forall \xi_k \in [0, 2\pi]$.

Dans le cas d'un flux d'ordre au moins deux en espace, on obtient le schéma suivant :

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+\frac{c}{2}\left(\frac{u_{j+1/2,-}^{n+1}-u_{j-1/2,-}^{n+1}}{\Delta x}+\frac{u_{j+1/2,-}^{n}-u_{j-1/2,-}^{n}}{\Delta x}\right)=-Ku_j^{n+1} \qquad \text{(VIII.46)}$$

avec :

$$\left\{\begin{array}{l} u_{j-1/2,-}=u_{j-1}+\dfrac{1}{2}\left\{(1-\beta)(u_j-u_{j-1})+\beta(u_{j-1}-u_{j-2})\right\} \\ u_{j+1/2,-}=u_j+\dfrac{1}{2}\left\{(1-\beta)(u_{j+1}-u_j)+\beta(u_j-u_{j-1})\right\} \end{array}\right.$$

L'analyse de Fourier donne :

$$g_k(\Delta t)=\frac{1-\dfrac{\sigma}{2}f(\beta)}{1+\dfrac{\sigma}{2}f(\beta)+K\Delta t},\quad \sigma=c\Delta t/\Delta x \qquad \text{(VIII.47)}$$

où $f(\beta)$ est donnée par (VIII.42). Comme $\mathrm{Re}\,(f(\beta))=\beta(1-\cos(\xi_k))^2\geq 0$, on a bien pour $K\geq 0$, $\mid g_k(\Delta t)\mid\leq 1\ \ \forall\xi_k\in[0,2\pi]\bullet$

Cependant, dans notre cas, la matrice du système linéaire n'est pas calculée avec le vrai jacobien du flux, et elle est obtenue avec un schéma d'ordre un en espace. Il a été montré dans [18] que la condition de stabilité du schéma de Cranck-Nicolson est alors plus restrictive et le schéma n'est plus inconditionnellement stable. Néanmoins, on peut toujours l'utiliser en prenant $\Theta=\dfrac{1}{2}+\epsilon$, avec ϵ nombre réel petit, et on ne perd pas la précision temporelle d'ordre deux.

VIII.4 Simulation numérique.

VIII.4.1 Cas instationnaire.

On s'intéresse ici à la propagation d'ondes sonores harmoniques en temps dans un milieu diphasique composé d'un mélange de gaz et de particules initialement à l'équilibre. On suppose que toutes les particules ont le même rayon. Nous cherchons une solution analytique de ce problème, et nous comparons cette solution avec les solutions numériques du système (VIII.3) pour la "méthode de splitting" et pour la "méthode couplée".

Solution analytique du problème.

On se place dans un cadre monodimensionnel. On recherche une onde de la forme $\mathbf{U}=\mathbf{U}^0+\mathbf{U}'$, où :

$$\mathbf{U}'=\tilde{\mathbf{U}}\exp(i\omega t+ax),\quad \omega\in I\!\!R,\ a\in\mathbb{C}. \qquad \text{(VIII.48)}$$

$$||\tilde{\mathbf{U}}||<<||\mathbf{U}^0|| \qquad \text{(VIII.49)}$$

L'état $\mathbf{U}^0$ est un état d'équilibre constant pour le système (VIII.3) :

$$\mathbf{U}^0 = {}^t\left(\alpha\rho_g^0, 0, \alpha\rho_g^0 e_g^0, (1-\alpha^0)\rho_l, 0, (1-\alpha^0)\varepsilon_l^0\right) \tag{VIII.50}$$

La fonction $\mathbf{U}'$ est une perturbation de l'état constant $\mathbf{U}^0$. Si on linéarise le système (VIII.3) autour de $\mathbf{U}^0$, on obtient :

$$\partial_t \mathbf{U}' + \mathbf{A}(\mathbf{U}^0)\partial_x \mathbf{U}' = \frac{\mathbf{R}'(\mathbf{U}^0)}{\epsilon}\mathbf{U}' \tag{VIII.51}$$

On cherche à déterminer le nombre complexe a comme fonction de la pulsation ω tel que $\mathbf{U}'$ satisfasse (VIII.51). On pose :

$$a = \frac{-1}{c}(\tau + i\omega) \tag{VIII.52}$$

avec (τ, c) nombres réels.
En remplaçant $\mathbf{U}'$ par son expression (VIII.48) dans (VIII.51), on obtient le résultat suivant, démontré dans [2] :

Proposition VIII.4.1 *Sous les hypothèses suivantes, valides pour les écoulements diphasiques dispersés :*

$$\begin{aligned} &\mathbf{H1} \quad \rho_l \frac{\theta_0 (1-\alpha^0)^\delta}{\alpha p_g^0} << 1 \\ &\mathbf{H2} \quad |\frac{\tau}{\omega}| << 1, \textit{ qui reste vrai dès que } \ c_0 \equiv \frac{(1-\alpha^0)\rho_l}{\alpha\rho_g^0} << 1, \end{aligned} \tag{VIII.53}$$

soit une onde acoustique :

$$\mathbf{U} = \mathbf{U}^0 + \tilde{\mathbf{U}} \exp(-\frac{\tau}{c})\exp(i\omega(t - \frac{x}{c})) \tag{VIII.54}$$

solution de (VIII.51).

Alors, au premier ordre en $\frac{\tau}{\omega}$, τ et c vérifient :

$$\begin{aligned} \frac{\tau}{\omega} &= \frac{c_0(\omega\epsilon)}{2\,(1 + c_0 + (\omega\epsilon)^2)} + O\left((\frac{\tau}{\omega})^2\right) \\ c &= c_g^0 \sqrt{\frac{1 + (\omega\epsilon)^2}{1 + c_0 + (\omega\epsilon)^2}} \end{aligned} \tag{VIII.55}$$

où c_g^0 est la vitesse du son dans le gaz pour l'état d'équilibre $\mathbf{U}^0$.

Remarques : A partir de la proposition (VIII.4.1), on peut écrire une onde de pression dans un milieu diphasique sous la forme :

$$(\alpha p_g)(x,t) = (\alpha p_g)^0 + (\alpha\tilde{p}_g)\exp(-\frac{\tau}{c})\sin(\omega(t - \frac{x}{c})) \tag{VIII.56}$$

où c et τ représentent respectivement la vitesse du son dans le milieu diphasique et l'amortissement de l'onde.
L'expression (VIII.55) montre que pour une pulsation ω donnée, l'amortissement τ tend vers 0 quand ϵ tend vers 0 ou $+\infty$. F. Béreux a montré dans [2] qu'il existe une unique valeur de ϵ (et donc du rayon des gouttes r), tel que, à ω fixé, τ soit maximum :

$$\left\{\begin{aligned} \epsilon_{opt} &= \frac{1}{\omega}\sqrt{1+c_0} \\ r_{opt} &= \frac{81\mu_g}{16\rho_l}\sqrt{\frac{\sqrt{1+c_0}}{\omega}} \\ \tau_{opt} &= \frac{c_0\omega}{4\sqrt{1+c_0}} \end{aligned}\right. \tag{VIII.57}$$

Simulation numérique.

On considère la propagation d'une onde sinusoïdale dans un tube de longueur 100 mètres, rempli d'un mélange diphasique. On maintient toutes les variables à l'état d'équilibre $\mathbf{U}^0$, sauf la pression du gaz qui est initialisée au bord d'entrée du tube par :

$$(\alpha p_g)(x,t) = (\alpha p_g)^0(1 + 0.01\ \sin(\omega t)) \tag{VIII.58}$$

Les caractéristiques de l'écoulement sont données dans les tableaux (VIII.2),(VIII.3),(VIII.4). On choisit comme fréquence $f = 70$ Hz : il s'agit de l'ordre de fréquence du premier mode acoustique longitudinal dans une grande fusée. On arrête la simulation au bout d'un temps $t = 0.1$ seconde, l'onde acoustique a alors atteint la sortie du tube. Les calculs sont effectués sur un maillage bidimensionnel composé de 400 points en x et de 5 points en y, l'onde se propageant suivant la direction x.

ρ_l	1766 kg.m^{-3}
C_l	1375 J.kg^{-1}K^{-1}
γ	1.4
μ_g	8.85.10^{-5} kg.m^{-1}s^{-1}
θ_0	10000 Pa

TAB. VIII.2 – *Constantes physiques.*

C_l représente la capacité calorifique des particules et l'énergie interne vérifie :

$$\varepsilon_l = C_l\, T_l$$

ρ_g	3.78kg.m^{-3}
u_g	$0\ m.s^{-1}$
v_g	$0\ m.s^{-1}$
p_g	5.10^6 Pa

TAB. VIII.3 – *Etat initial* $\mathbf{U}^0$ *pour le gaz.*

$1-\alpha$	0.0009
u_l	$0\ m.s^{-1}$
v_l	$0\ m.s^{-1}$
T_l	3500 K

TAB. VIII.4 – *Etat initial* $\mathbf{U}^0$ *pour les particules.*

Nous savons que la pression exacte est donnée par (VIII.56). A partir de chaque profil de pression ($P^\star$) obtenu par une simulation numérique, nous déterminons les amortissement et vitesse approchés $\tau^\star$ et $c^\star$ et nous les comparons aux amortissement et vitesse théoriques donnés par (VIII.55). Pour cela, nous ajustons ($P^\star$) avec l'équation suivante :

$$\alpha p_g(x,t) = (\alpha p_g)^\star + (\alpha \tilde{p}_g)^\star \exp(-\frac{\tau^\star}{c^\star})\sin(\omega(t-\frac{x}{c^\star})) \tag{VIII.59}$$

La courbe (VIII.59) est solution de l'équation différentielle suivante :

$$F'' + 2(\frac{\tau^\star}{c^\star})F' + \frac{1+{\tau^\star}^2}{{c^\star}^2}(F-(\alpha p_g)^\star) = 0 \tag{VIII.60}$$

On utilise une méthode des moindres carrés, qui consiste à minimiser la norme L_2 de (VIII.60) et on obtient ainsi les valeurs $\tau^\star$ et $c^\star$ en fonction du rayon r.
Sur les figures (VIII.1) à (VIII.10), on représente les courbes d'amortissement et de vitesse du son, $\tau^\star$ et $c^\star$ en fonction du rayon des particules, pour la "méthode de splitting" et pour la "méthode couplée". On compare ces solutions avec les solutions exactes données par (VIII.55).

On fait varier r de 0 à 100 μm. On choisit un pas de temps identique pour tous les tests, en particulier, on a pris CFL = 0.5.

– Schémas explicites :

Les figures (VIII.1) à (VIII.4) permettent de comparer la "méthode de splitting" et la "méthode couplée". Sur les figures (VIII.1),(VIII.2) les schémas sont d'ordre un en temps et en espace, alors que sur les figures (VIII.3),(VIII.4) l'approximation est d'ordre deux.

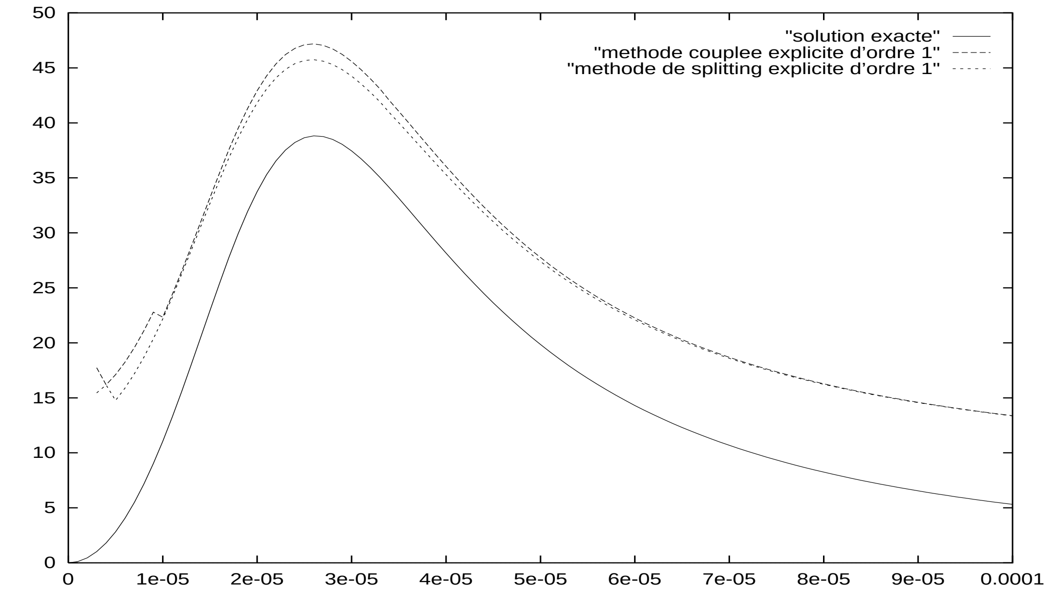


FIG. VIII.1 – *Amortissement pour les méthodes d'ordre un.*

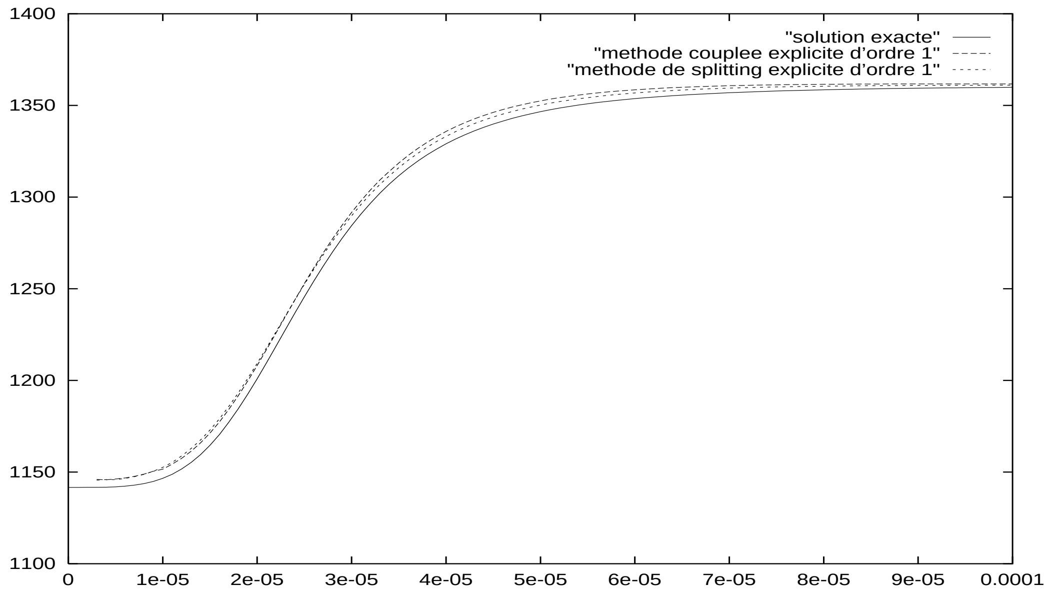


FIG. VIII.2 – *Vitesse du son pour les méthodes d'ordre un.*

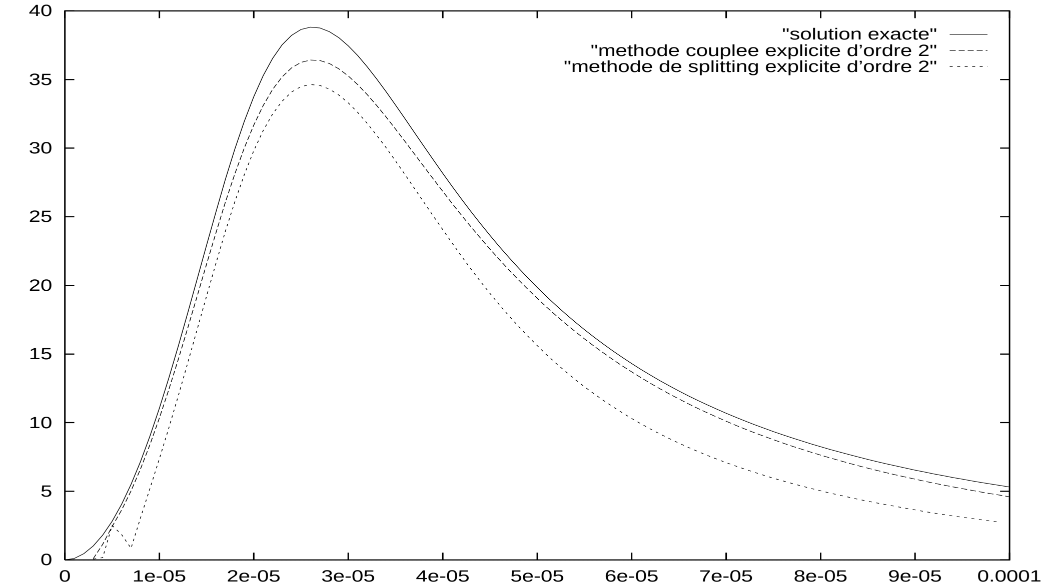


FIG. VIII.3 – *Amortissement pour les méthodes d'ordre deux.*

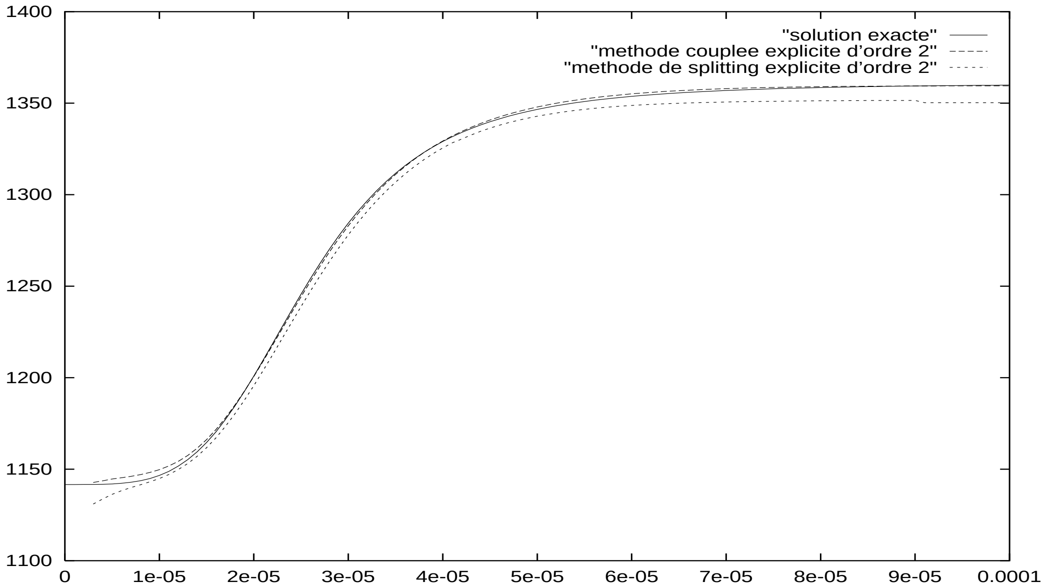


FIG. VIII.4 – *Vitesse du son pour les méthodes d'ordre deux.*

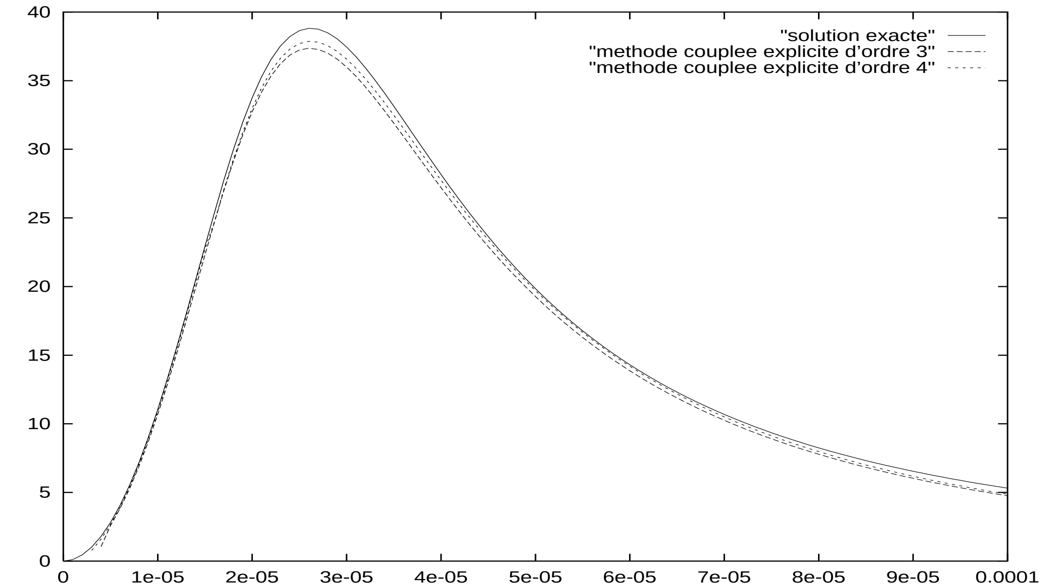


FIG. VIII.5 – *Amortissement pour les "méthodes couplées" d'ordre trois et quatre.*

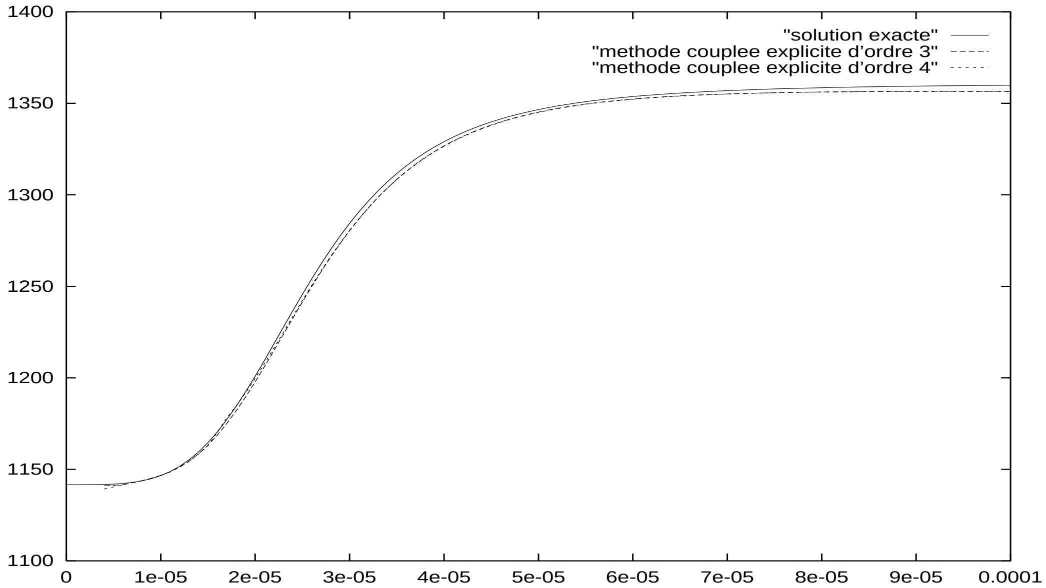


FIG. VIII.6 – *Vitesse du son pour les "méthodes couplées" d'ordre trois et quatre.*

On remarque que les courbes de vitesse (VIII.2),(VIII.4) sont satisfaisantes, sauf pour la "méthode de splitting" d'ordre deux où les valeurs s'éloignent de la solution exacte quand le rayon tend vers 0 ou vers $+\infty$. Par contre on observe de grandes différences entre les courbes d'amortissement numérique et les valeurs exactes. En particulier, les schémas d'ordre un (figure VIII.1) ont tendance à augmenter l'amortissement de la solution, aussi bien pour la "méthode de splitting" que pour la "méthode couplée".

Au contraire, pour une approximation d'ordre deux (figure VIII.3), l'amortissement est globalement moins élevé, pour les deux méthodes. Cela peut s'expliquer : les schémas d'ordre un en temps et en espace introduisent beaucoup de diffusion numérique et donc dissipent les solutions, tandis que les approximations d'ordre deux en temps (en particulier les schémas RK2) sont anti-dissipatives.
Pour les petits rayons $\leq 10^{-5}$ m, les courbes d'amortissement ne sont pas bonnes pour les schémas d'ordre un, et pour la "méthode de splitting" d'ordre deux : l'amortissement ne tend pas vers 0 avec le rayon des gouttes. Par contre, on remarque le bon comportement du schéma d'ordre deux pour la "méthode couplée" quand r tend vers 0 ou vers $+\infty$.

Les figures (VIII.5),(VIII.6) montrent les vitesses et les amortissements pour des schémas d'ordre trois et quatre en temps et en espace, avec la "méthode couplée". L'approximation en temps se fait au moyen des schémas Runge-Kutta non linéaires (VI.9). L'ordre trois en espace est obtenu avec les paramètres ($\beta = \frac{1}{3}$, $\gamma_c = 1$), et l'ordre quatre avec ($\beta = \frac{1}{3}$, $\gamma_c = 0$). On remarque une très bonne adéquation entre les courbes théoriques et la solution exacte, en particulier pour le schéma d'ordre quatre. Contrairement aux schémas d'ordre un et deux, les solutions obtenues restent très proches de la solution exacte, lorsque le rayon des gouttes diminue.

Cependant, avec le pas de temps choisi, nous avons des difficultés à obtenir des solutions pour des rayons $\leq 2.10^{-6}$m. Pour remédier à ce problème, nous nous tournons vers des schémas implicites.

– Schémas implicites :

Nous avons d'abord utilisé un schéma implicite d'ordre un en temps avec une approximation spatiale d'ordre trois. Mais ce schéma s'avère trop dissipatif (voir figure VIII.7) par-rapport aux schémas explicites et il a tendance à augmenter l'amortissement de la solution. Nous avons ensuite implémenté le schéma de Cranck-Nicolson, d'ordre deux en temps, avec une approximation spatiale d'ordre trois. Sur les figures (VIII.8) et (VIII.9) nous comparons les solutions approchées pour la "méthode couplée" et pour la "méthode de splitting". Les vitesses obtenues avec les deux méthodes sont très bonnes (voir figure VIII.9). En ce qui concerne l'amortissement (figure VIII.8), la "méthodes couplée" nous donne une très bonne solution; en particulier l'amortissement numérique tend bien vers 0 avec le rayon des gouttes. Avec la "méthode de splitting", la solution est meilleure que celle obtenue avec des schémas explicites. Cependant cette méthode introduit un amortissement excessif, en particulier pour les petits rayons.

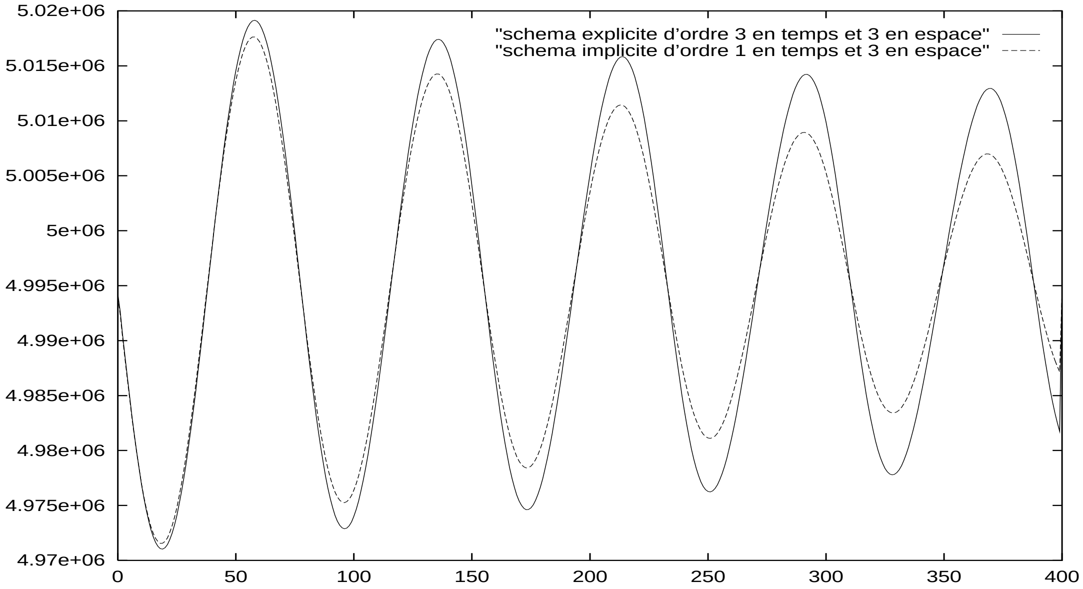


FIG. VIII.7 – *Comparaison entre le schéma implicite d'ordre un en temps et le schéma explicite d'ordre trois en temps avec la "méthode couplée" pour* $r = 1.10^{-4}$ *m.*

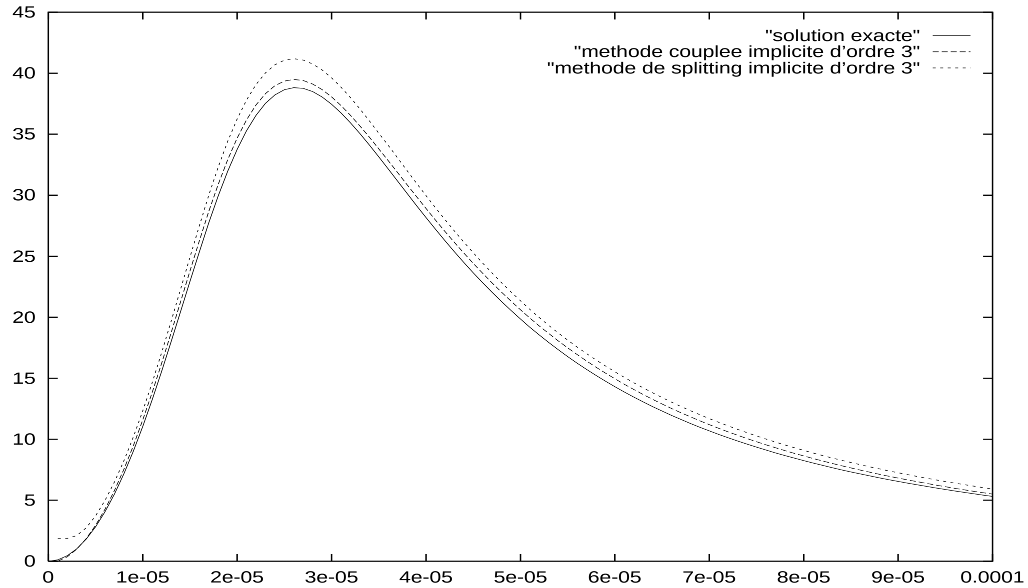


FIG. VIII.8 – *Amortissement pour les schémas de Cranck-Nicolson appliqués aux deux méthodes.*

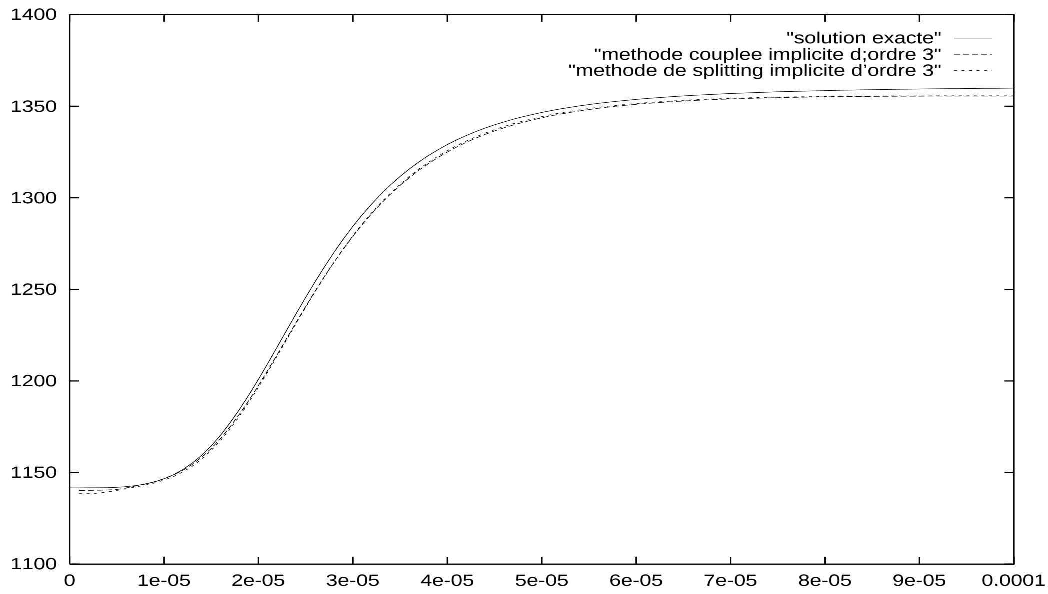


FIG. VIII.9 – *Vitesse du son pour les schémas de Cranck-Nicolson appliqués aux deux méthodes.*

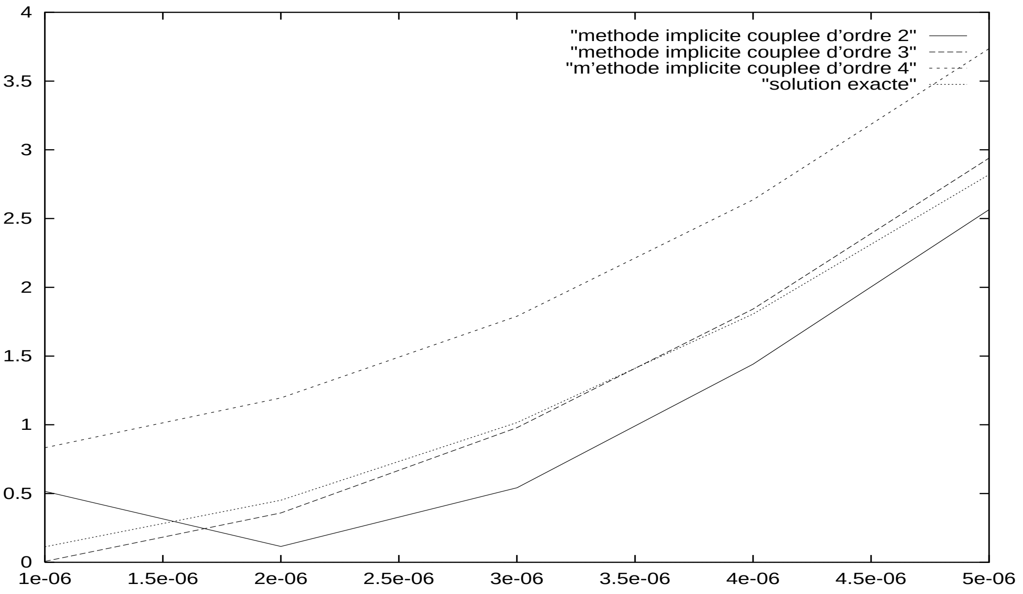


FIG. VIII.10 – *Amortissement pour les schémas de Cranck-Nicolson d'ordre deux, trois, quatre en espace appliqués à la "méthode couplée".*

La figure (VIII.10) montre l'amortissement obtenu avec la "méthode couplée" pour des petits rayons de gouttes. On a utilisé le schéma de Cranck-Nicolson et des approximations spatiales d'ordre deux, trois, quatre en espace. C'est le schéma décentré d'ordre trois, obtenu avec les paramètres $\beta = \frac{1}{3}$, $\gamma_c = 1$, qui approche le mieux la solution exacte. Le schéma d'ordre quatre, centré ($\beta = \frac{1}{3}$, $\gamma_c = 0$) a tendance à augmenter l'amortissement, alors qu'avec le schéma d'ordre deux ($\beta = \frac{1}{2}$, $\gamma_c = 1$) l'amortissement numérique est plus faible que l'amortissement exact, pour $r \geq 2.10^{-6}$ mètres.

On a vu que dans le cas linéaire, le schéma implicite était inconditionnellement stable. Dans le cas du schéma implicite d'ordre un en temps, nous avons pu aller jusqu'à CFLmax = 20 (il s'agit de la valeur maximale que l'on peut obtenir à $t = 0.1$ s), cependant, la solution est de plus en plus amortie quand on utilise de grands pas de temps. Avec le schéma de Cranck-Nicolson, en posant $\Theta = 0.5 + \epsilon$, nous obtenons les limites de CFL suivantes :

Θ	CFLMAX
0.5	3
0.51	4
0.53	20
0.6	20

TAB. VIII.5 – *Limites de CFL en fonction du paramètre Θ.*

– Interprétation physique :

On représente sur les figures (VIII.11),(VIII.12) les vitesses du gaz et des particules obtenues avec le schéma de Cranck-Nicolson pour la "méthode couplée". Lorsque le rayon est petit (voir figure VIII.11), l'inertie des particules est très faible, la force de traînée parvient très rapidement à égaliser les vitesses des deux milieux, et par conséquent, l'amortissement tend vers 0 avec le rayon des gouttes.
Pour de grands rayons (voir figure VIII.12), l'inertie des gouttes est importante et celles-ci sont peu affectées par les oscillations du gaz. L'amplitude de la vitesse des gouttes est très inférieure à celle du gaz et on observe un déphasage entre les deux phases. En fait, plus les gouttes sont grosses, plus le terme source est faible. Les deux systèmes tendent à devenir totalement découplés, et les solutions obtenues pour le gaz se rapprochent de celles que l'on obtiendrait pour le système de la dynamique des gaz avec les variables $(\alpha\rho_g, \alpha\rho_g u_g, \alpha\rho_g v_g, \alpha\rho_g e_g)$.

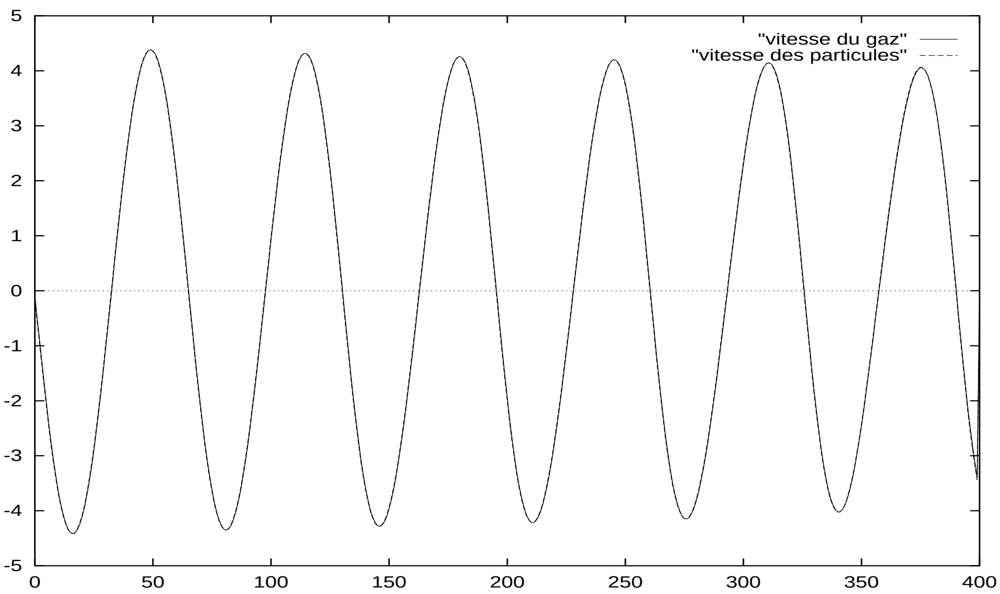


FIG. VIII.11 – *Vitesse du gaz et des particules pour* $r = 3.10^{-6}\,m$ *avec la "méthode couplée".*

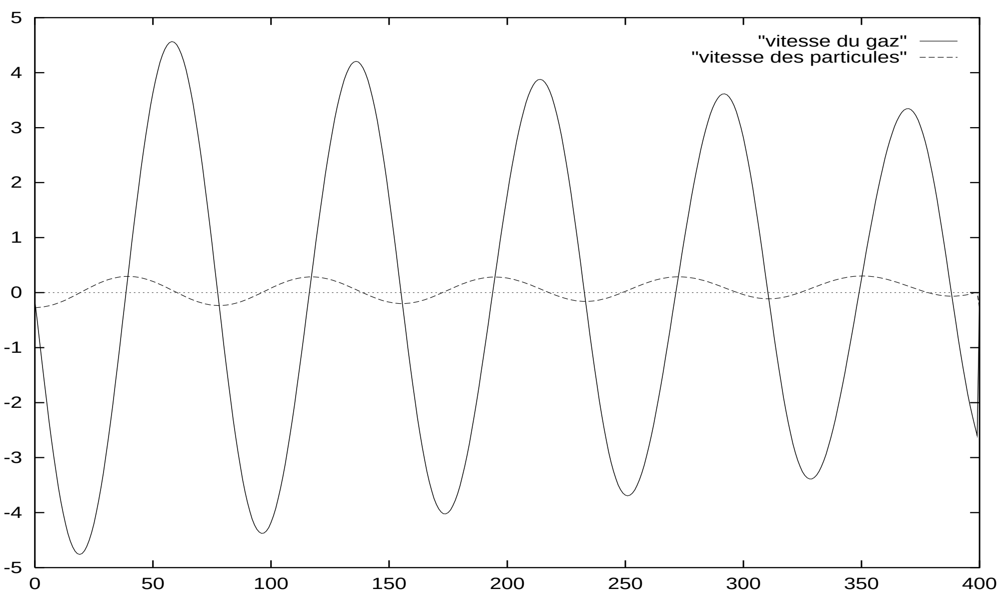


FIG. VIII.12 – *Vitesse du gaz et des particules pour* $r = 1.10^{-4}\,m$ *avec la "méthode couplée".*

VIII.4.2 Cas stationnaire.

Nous considérons maintenant deux situations d'écoulement d'un mélange diphasique dans une tuyère. Nous nous intéressons aux solutions stationnaires obtenues avec la "méthode couplée" et la "méthode de splitting". Pour cela, nous utilisons un schéma implicite d'ordre un en temps. La matrice du système linéaire (VIII.36) est calculée à l'aide d'un schéma d'ordre un en espace, tandis que pour la phase physique (VIII.35), on choisit un schéma d'ordre trois en espace (on prend $\beta = \frac{1}{3}$, $\gamma = 1$). Le maillage utilisé est un maillage triangulaire structuré comportant 1680 sommets et 3154 triangles (voir figure VIII.29).

Cas test 1 :

Le premier cas de calcul est présenté sur les figures VIII.30 à VIII.46. Pour chacune des grandeurs physiques, nous comparons la solution obtenue par la "méthode couplée" avec celle de la "méthode de splitting". Il s'agit d'un mélange constitué de gaz et de fines gouttelettes d'eau de rayon $r = 1.10^{-6}$ mètres. Le gaz et les gouttes sont injectés dans la tuyère à la vitesse de 20 mètres par seconde. On donne dans les tableaux (VIII.6),(VIII.7),(VIII.9) les caractéristiques de l'écoulement.

ρ_l	1000 kg.m^{-3}
C_l	4180 J.kg^{-1}K^{-1}
γ	1.4
μ_g	1.8 10^{-5} kg.m^{-1}s^{-1}
θ_0	100000 Pa

TAB. VIII.6 – *Constantes physiques.*

	entrée	sortie
ρ_g	3 kg.m^{-3}	.01 kg.m^{-3}
u_g	20 m.s^{-1}	0 m.s^{-1}
v_g	0 m.s^{-1}	0 m.s^{-1}
T_g	100 K	100 K

TAB. VIII.7 – *Conditions aux limites pour le gaz.*

Nous avons choisi la même loi de CFL, donc le même pas de temps pour les deux méthodes. Plus précisément, nous prenons, d'après (VIII.43), $x = y = z = 0.1$ et CFLmax = 40. La figure VIII.46 montre les résidus obtenus pour les deux méthodes

	entrée	sortie
$1-\alpha$	0.01	0.001
u_l	20 m.s^{-1}	0 m.s^{-1}
v_l	0 m.s^{-1}	0 m.s^{-1}
T_l	100 K	100 K

TAB. VIII.8 – *Conditions aux limites pour les gouttes.*

en fonction du temps. On voit que la "méthode couplée" converge légèrement plus rapidement vers la solution stationnaire lorsque le résidu est plus grand que 1.10^{-10}. Cette valeur du résidu est atteinte au bout de 550 itérations pour la "méthode couplée" et 600 pour la "méthode de splitting".

Physiquement, les isovaleurs de la masse volumique du gaz et de la vitesse du gaz sont très proches de celles que l'on aurait obtenu en considérant un écoulement de gaz "équivalent", c'est-à-dire un modèle diphasique avec comme seule vitesse, la vitesse moyenne du gaz et des gouttes. Le gaz est accéléré par la détente qui se forme à la sortie du col. Ici, les forces de traînée sont très importantes par-rapport à l'inertie des gouttes et les vitesses du gaz et des gouttes s'égalisent très rapidement (voir figures VIII.34, VIII.35, VIII.36, VIII.37). On observe sur les planches VIII.32 et VIII.33 que la densité des gouttes est plus importante à l'entrée et sur la partie supérieure de la tuyère avant le col. A cet endroit, le gaz étant plus comprimé, les gouttes occupent un plus grand volume.

Pour toutes les grandeurs physiques, nous observons des différences entre la "méthode couplée" et la "méthode de splitting". La concentration des gouttes à l'entrée de la tuyère et avant le col est plus importante pour la méthode de splitting. Les valeurs des vitesses du gaz et des gouttes sont plus grandes notamment à la sortie de la tuyère, et les champs des vitesses n'épousent plus la forme de la tuyère contrairement à la "méthode couplée" : le gaz et les gouttes ont tendance à continuer tout droit au niveau du col.

Ces anomalies proviennent de la dépendance en temps des vitesses du gaz et des gouttes pour la "méthode de splitting". Les expressions (VIII.29),(VIII.30) sur l'écart des vitesses entre les deux phases montrent que plus le pas de temps est grand, plus la vitesse relative entre le gaz et les gouttes diminue. La solution stationnaire obtenue avec la "méthode de splitting" dépend donc du pas de temps, et imposer la loi de CFL (VIII.43) revient en fait à augmenter à chaque pas de temps le terme source. Si on calcule la norme l_2 de la vitesse relative entre le gaz et les gouttes à l'état stationnaire, on obtient :

$$U_{\mathrm{moy}} = 1.\ 10^{-13} \quad \text{pour la "méthode de splitting".}$$

$$U_{\text{moy}} = 4.9\ 10^{-3} \quad \text{pour la “méthode couplée”.}$$

Si on utilise un pas de temps petit et constant (CFL=0.5), la “méthode de splitting” donne alors les mêmes résultats que la “méthode couplée”.

Cas test 2 :

Le second cas de calcul est un cas-test physique, il permet de mieux rendre compte des aspects diphasiques du modèle dans ce type de géométrie. Cette fois, les gouttes d'eau ont un rayon de 50 micromètres, le gaz est injecté à la vitesse de 100 mètres par seconde alors que les gouttes ont une vitesse de 20 mètres par seconde. On donne dans les tableaux (VIII.10),(VIII.11) les conditions limites d'entrée et de sortie pour les deux phases.

	entrée	sortie
$1-\alpha$	0.01	0.001
u_l	20 m.s^{-1}	0 m.s^{-1}
v_l	0 m.s^{-1}	0 m.s^{-1}
T_l	100 K	100 K

TAB. VIII.9 – *Conditions aux limites pour les gouttes.*

	entrée	sortie
ρ_g	3 kg.m^{-3}	.01 kg.m^{-3}
u_g	100 m.s^{-1}	0 m.s^{-1}
v_g	0 m.s^{-1}	0 m.s^{-1}
T_g	100 K	100 K

TAB. VIII.10 – *Conditions aux limites pour le gaz.*

	entrée	sortie
$1-\alpha$	0.005	0.001
u_l	20 m.s^{-1}	0 m.s^{-1}
v_l	0 m.s^{-1}	0 m.s^{-1}
T_l	100 K	100 K

TAB. VIII.11 – *Conditions aux limites pour les gouttes.*

Nous utilisons (VIII.43) comme loi de CFL avec $x = y = z = 0.1$ et CFLmax = 30 pour les deux méthodes. La figure VIII.47 représente les résidus en fonction du temps. La "méthode couplée" converge plus rapidement vers la solution stationnaire que la "méthode de splitting". Il faut 450 itérations avec la "méthode couplée" et 500 avec la "méthode de splitting" pour obtenir un résidu de 1.10^{-10}.

Les figures VIII.38 à VIII.45 montrent les isovaleurs des densités du gaz et des gouttes ainsi que les champs des vitesses des deux phases. Là encore, les isovaleurs de la densité et de la vitesse du gaz sont semblables à celle d'un écoulement gazeux équivalent. Par contre, les isovaleurs de la densité et du champ de vitesse des gouttes montrent un comportement propre aux gouttes. Comme dans l'exemple précédent, les gouttes ont tendance à se regrouper sur la partie supérieure de la tuyère. A cause de leur inertie, les gouttes du haut du convergent ainsi que celles du col sont déviées par l'écoulement du gaz et ont tendance à continuer tout droit : au niveau du col, leur vitesse n'est pas alignée avec l'axe de la tuyère comme pour le gaz.

On observe un écart significatif entre les vitesses du gaz et des gouttes, même pour la "méthode de splitting". Cette fois les deux méthodes donnent des résultats très proches aussi bien pour les densités que pour les champs des vitesses.

On obtient comme vitesse relative moyenne :

$$U_{\text{moy}} = 2.89 \quad \text{pour la "méthode de splitting".}$$

$$U_{\text{moy}} = 2.94 \quad \text{pour la "méthode couplée".}$$

Du fait de l'inertie des gouttes et de la grande différence de vitesse entre les deux phases à l'instant initial, les forces de traînée ne parviennent pas à égaliser les vitesses du gaz et des gouttes, même dans le cas de la "méthode de splitting" où le terme source augmente à chaque pas de temps. Ceci explique le comportement quasi-identique de l'écoulement avec les deux méthodes.

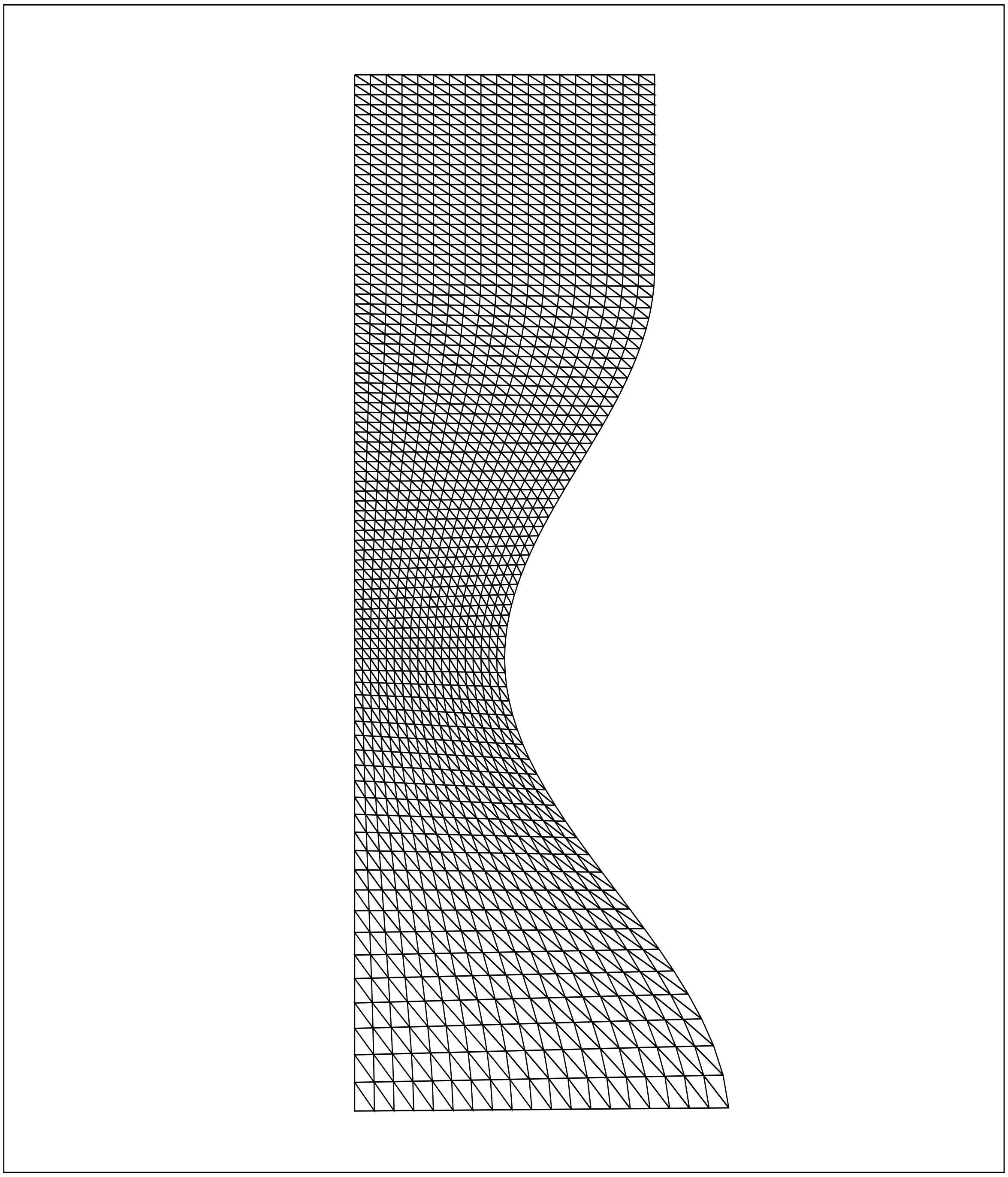


FIG. VIII.29 – *Maillage de la tuyère.*

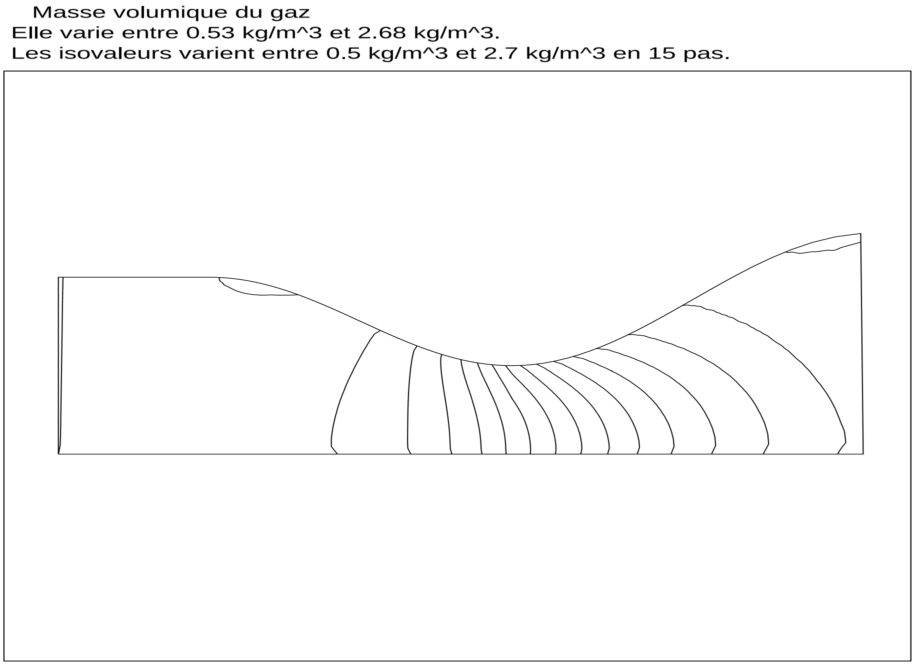


FIG. VIII.30 – *Masse volumique du gaz pour $r = 1.10^{-6}\,m$ avec la "méthode couplée".*

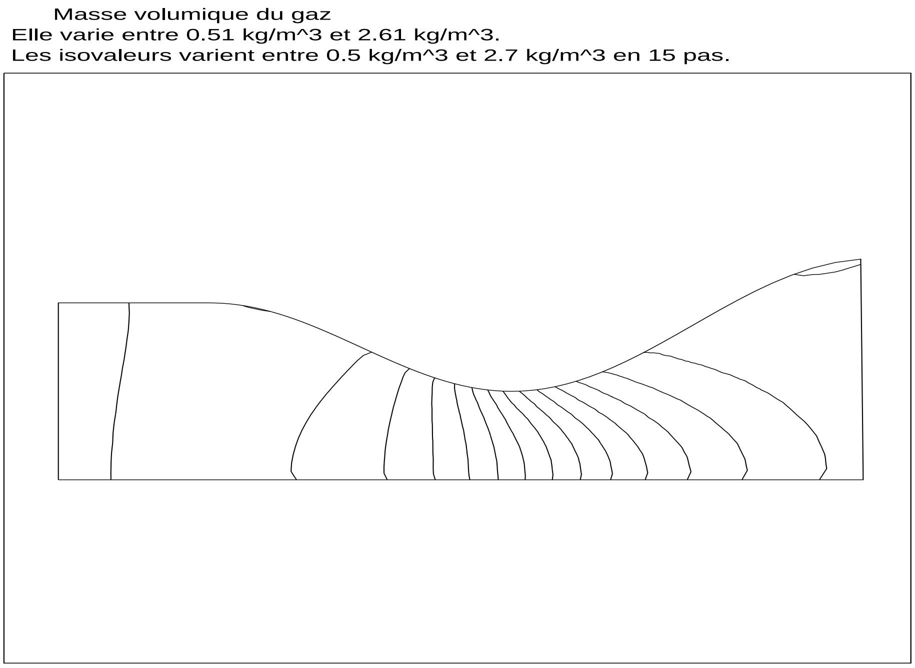


FIG. VIII.31 – *Masse volumique du gaz pour $r = 1.10^{-6}\,m$ avec la "méthode de splitting".*

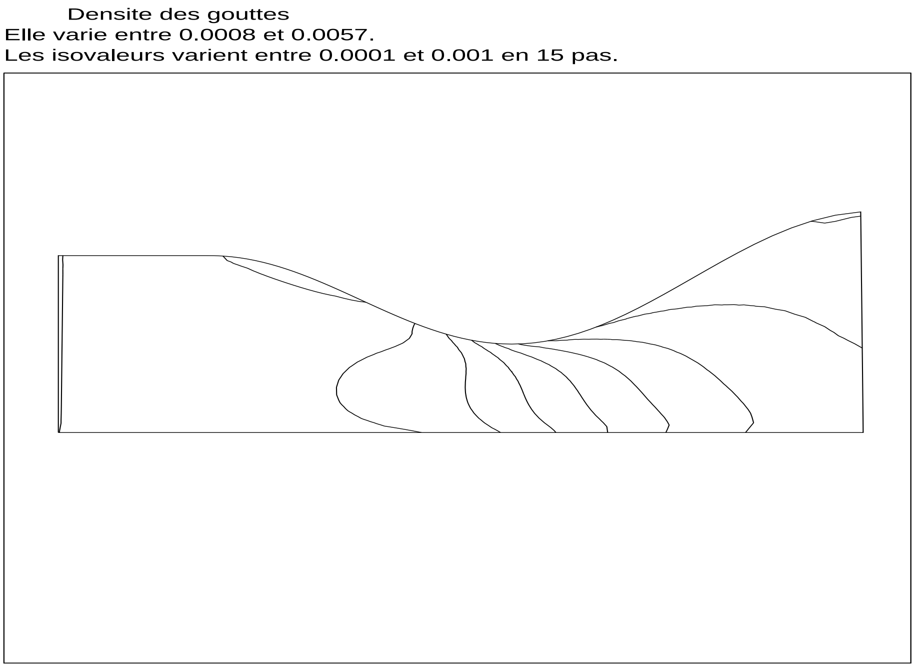


FIG. VIII.32 – *Densite des gouttes pour* $r = 1.10^{-6}\,m$ *avec la "méthode couplée".*

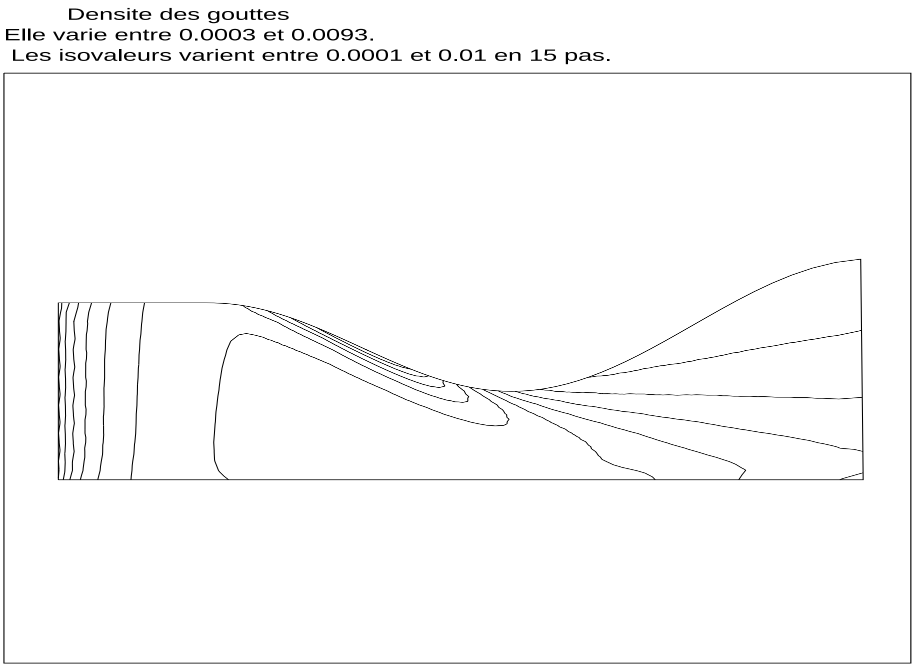


FIG. VIII.33 – *Densite des gouttes pour* $r = 1.10^{-6}\,m$ *avec la "méthode de splitting".*

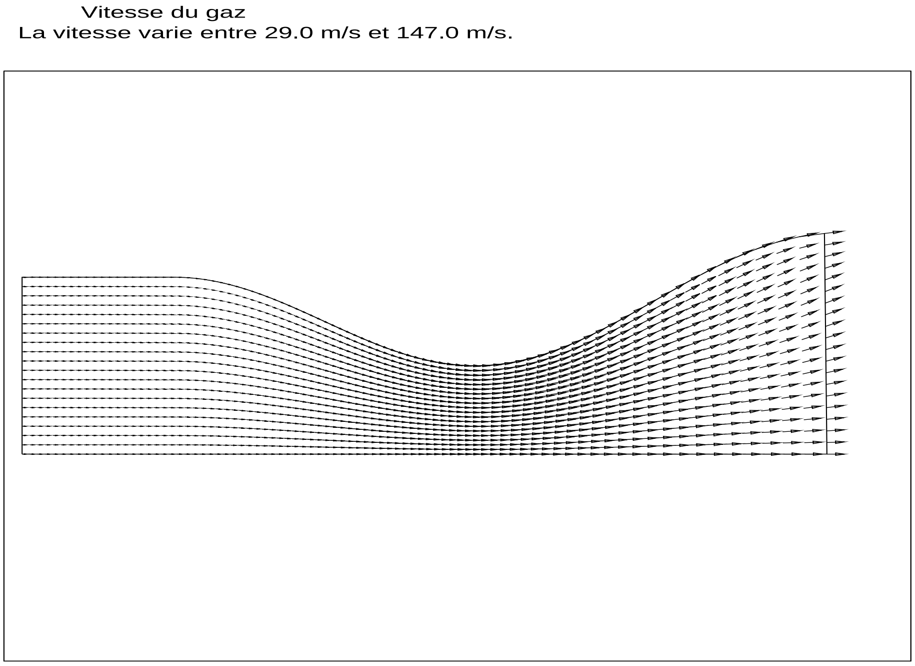


FIG. VIII.34 – *Champ de vitesse du gaz pour $r = 1.10^{-6}\,m$ avec la "méthode couplée".*

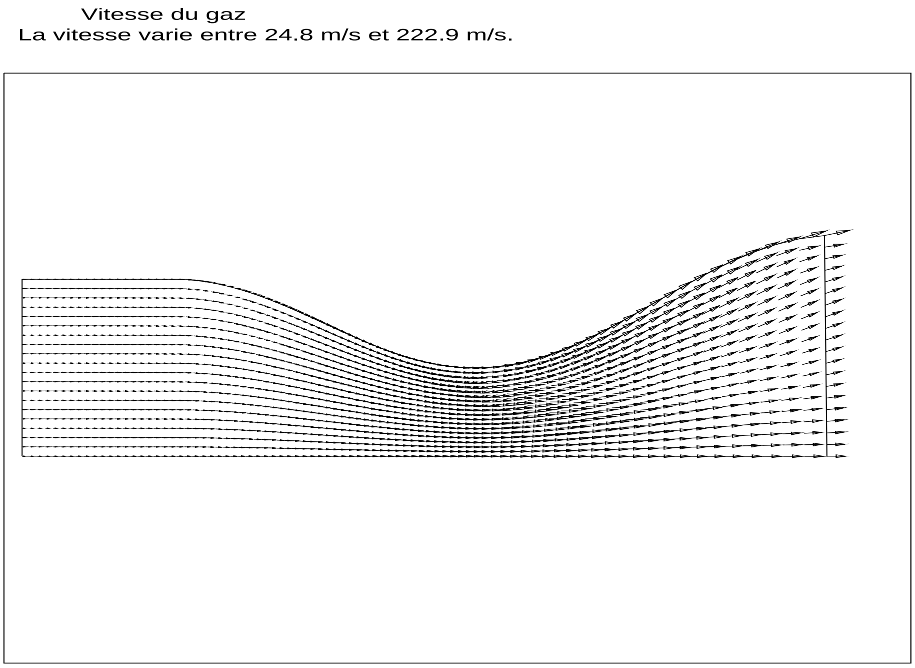


FIG. VIII.35 – *Champ de vitesse du gaz pour $r = 1.10^{-6}\,m$ avec la "méthode de splitting".*

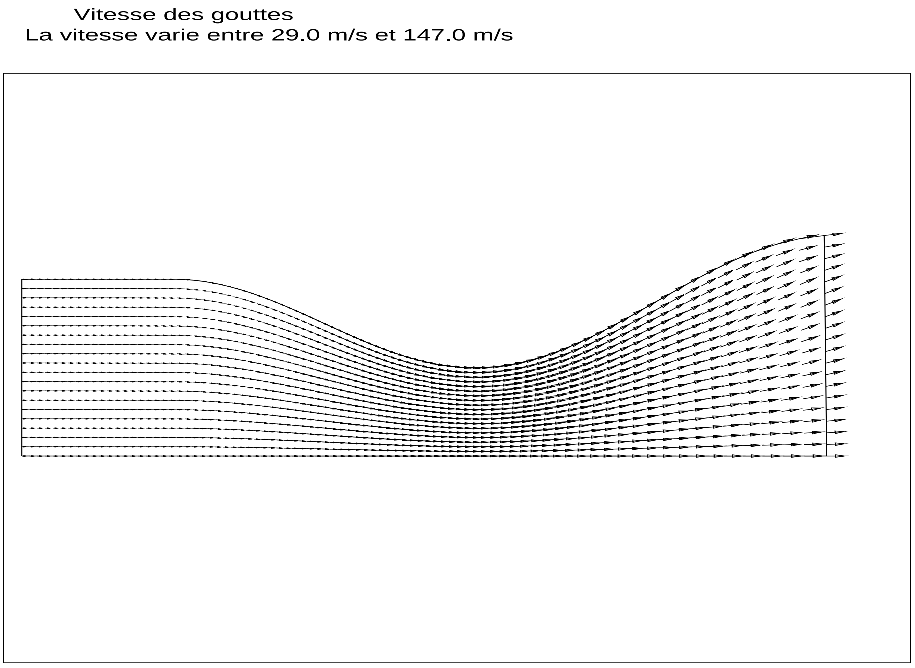


FIG. VIII.36 – *Champ de vitesse des gouttes pour $r = 1.10^{-6}\,m$ avec la "méthode couplée".*

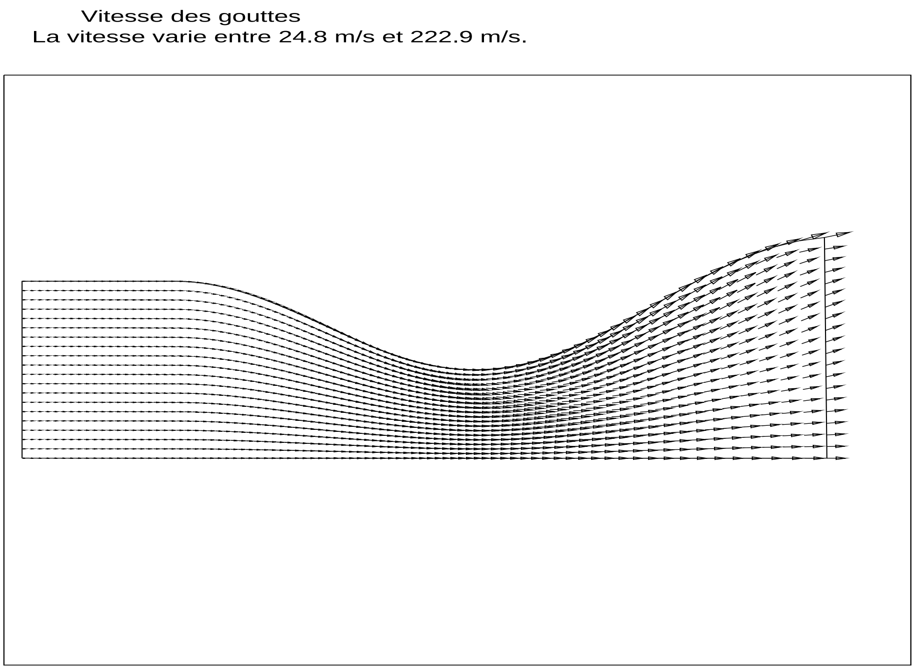


FIG. VIII.37 – *Champ de vitesse des gouttes pour $r = 1.10^{-6}\,m$ avec la "méthode de splitting".*

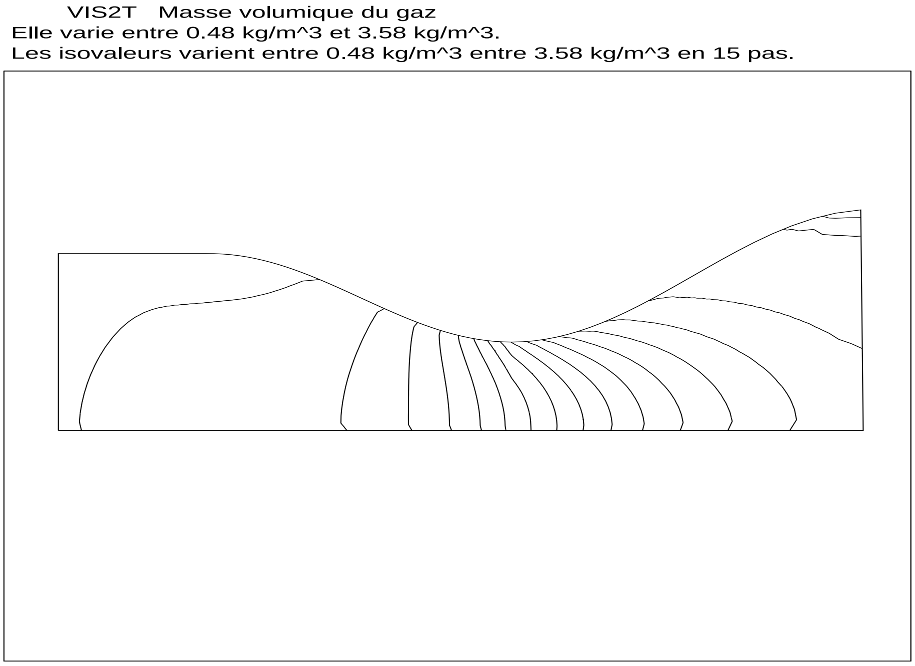


FIG. VIII.38 – *Masse volumique du gaz pour $r = 50.10^{-6}\, m$ avec la "méthode couplée".*

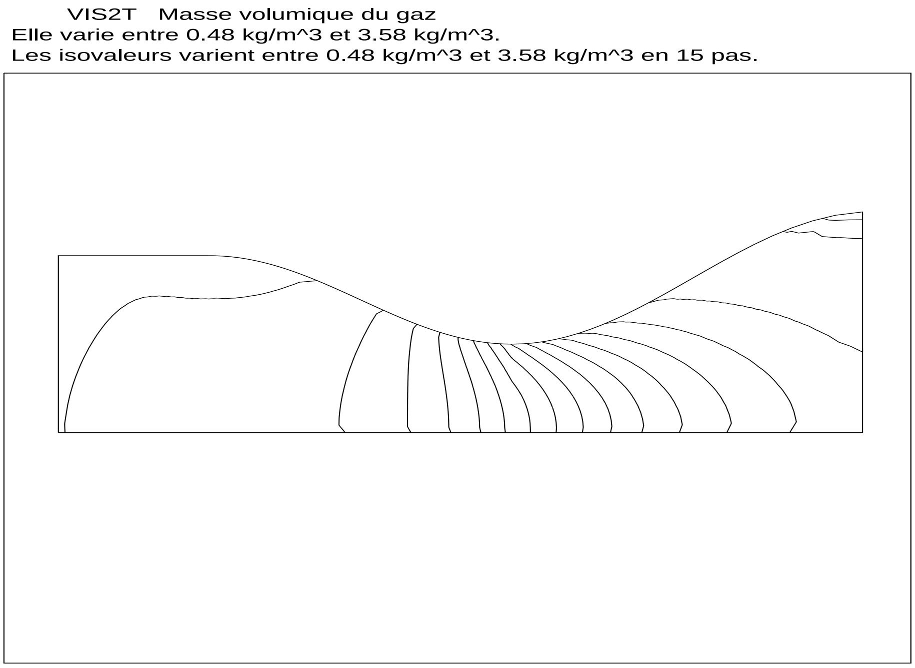


FIG. VIII.39 – *Masse volumique du gaz pour $r = 50.10^{-6}\, m$ avec la "méthode de splitting".*

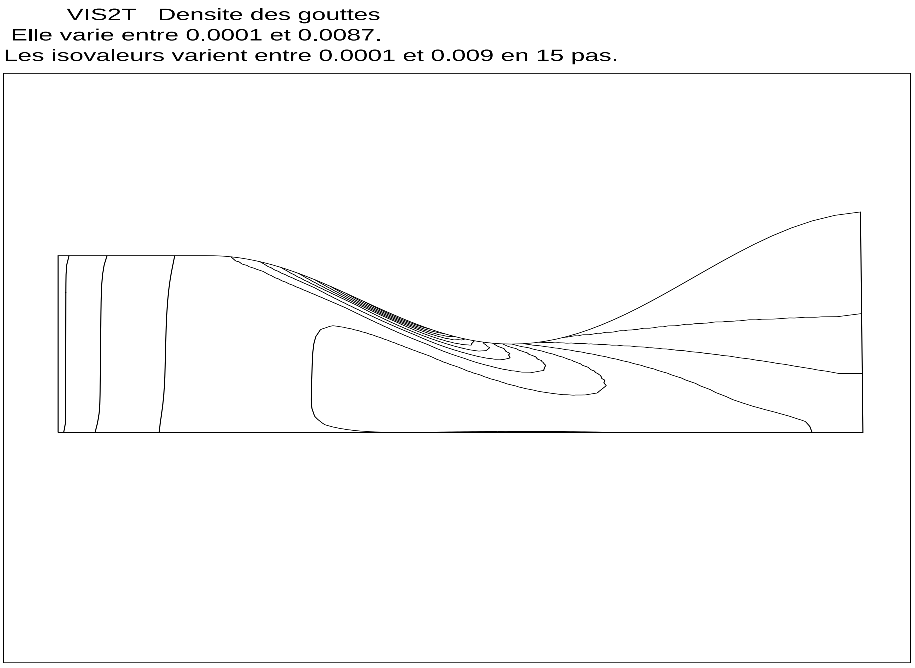


FIG. VIII.40 – *Densite des gouttes pour* $r = 50.10^{-6}$ *m avec la "méthode couplée".*

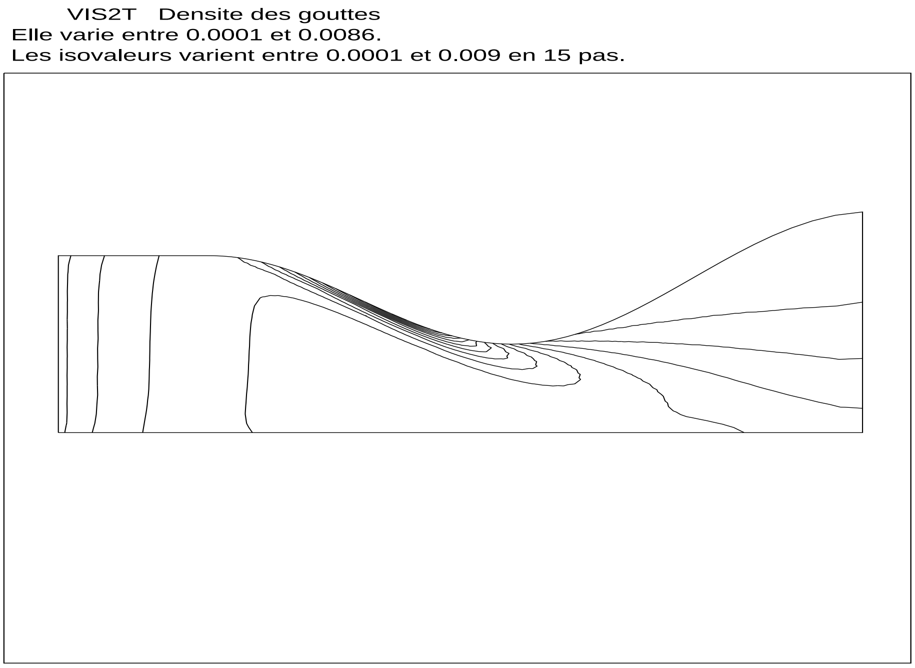


FIG. VIII.41 – *Densite des gouttes pour* $r = 50.10^{-6}$ *m avec la "méthode de splitting".*

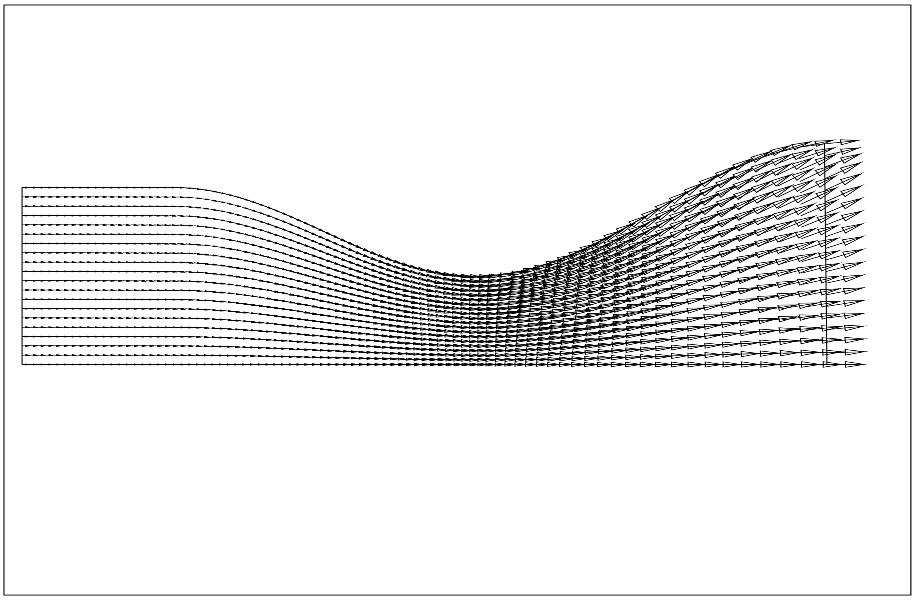


FIG. VIII.42 – *Champ de vitesse du gaz pour* $r = 50.10^{-6}$ *m avec la "méthode couplée".*

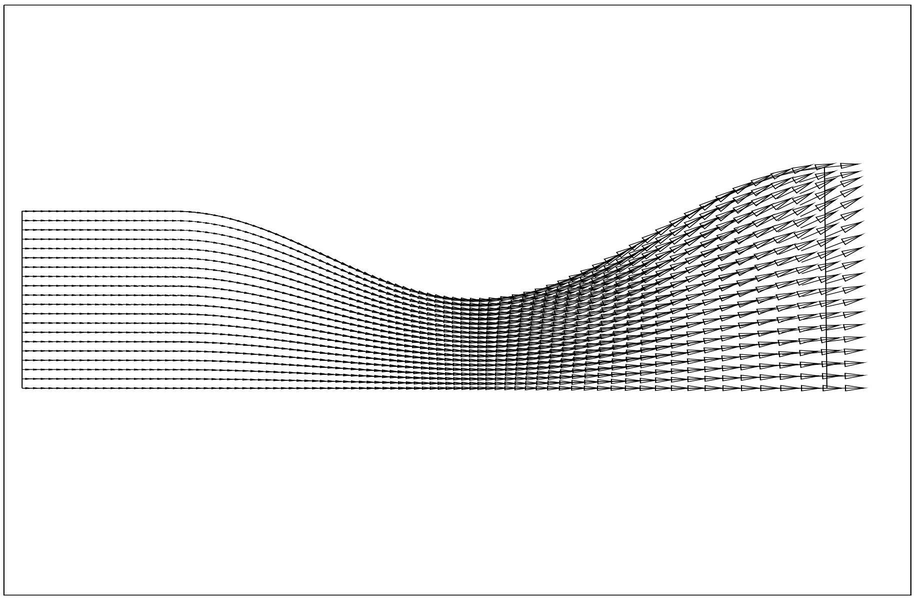


FIG. VIII.43 – *Champ de vitesse du gaz pour* $r = 50.10^{-6}$ *m avec la "méthode de splitting".*

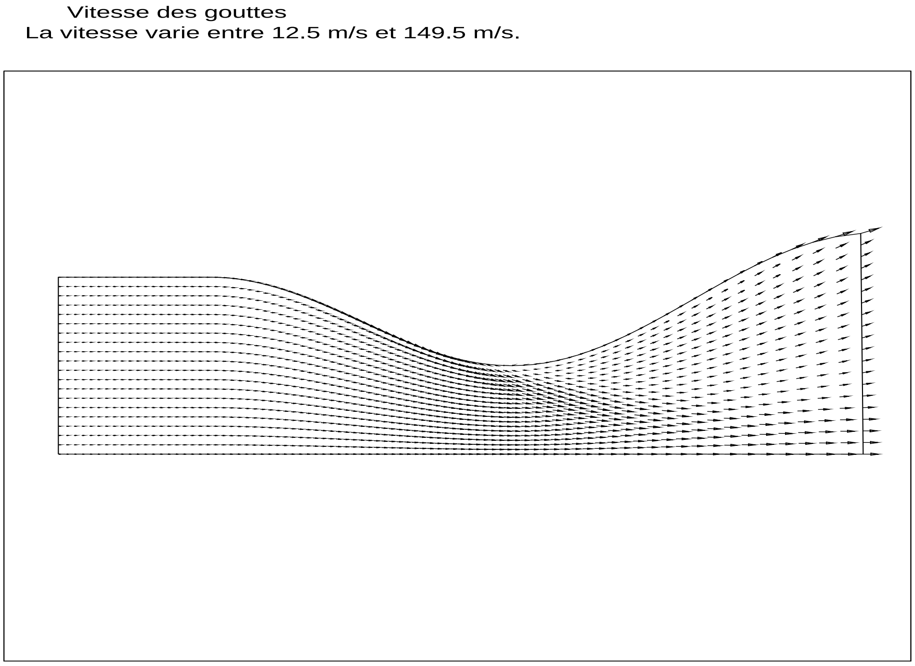


FIG. VIII.44 – *Champ de vitesse des gouttes pour* $r = 50.10^{-6}\,m$ *avec la "méthode couplée".*

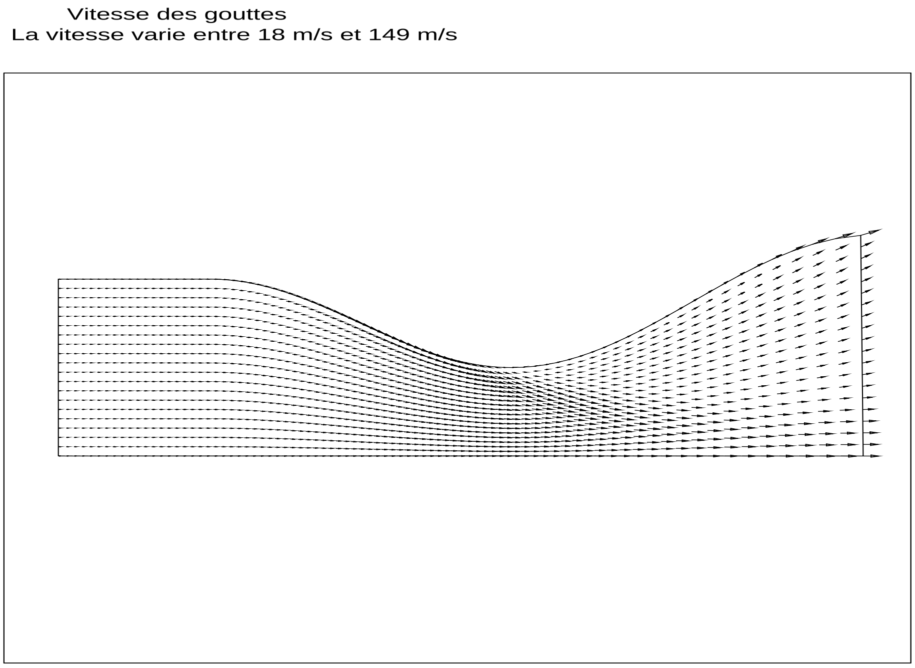


FIG. VIII.45 – *Champ de vitesse des gouttes pour* $r = 50.10^{-6}\,m$ *avec la "méthode de splitting".*

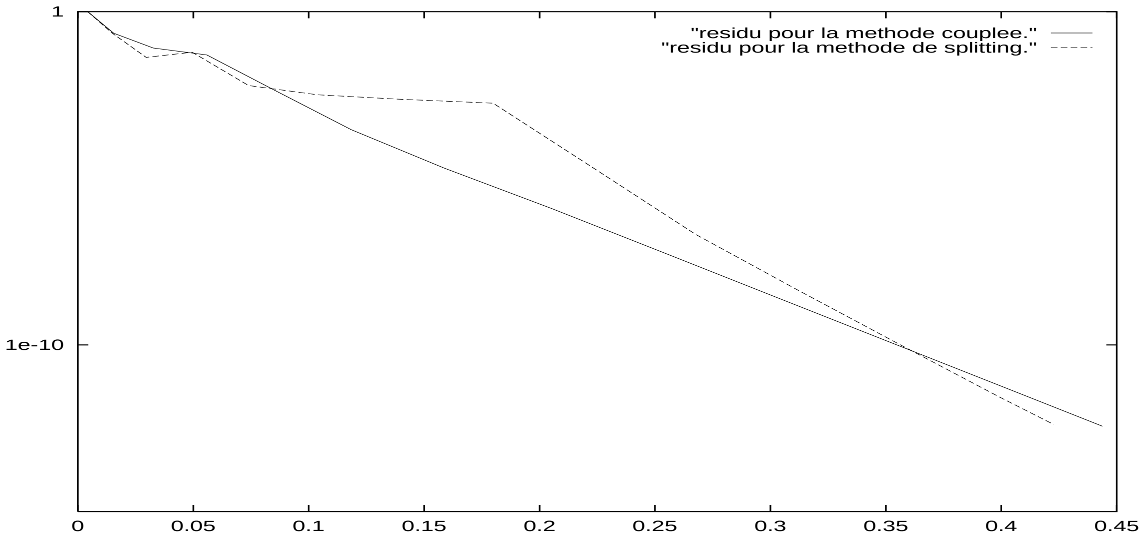


FIG. VIII.46 – *Residus obtenus pour* $r = 1.10^{-6}\, m$ *avec les deux méthodes.*

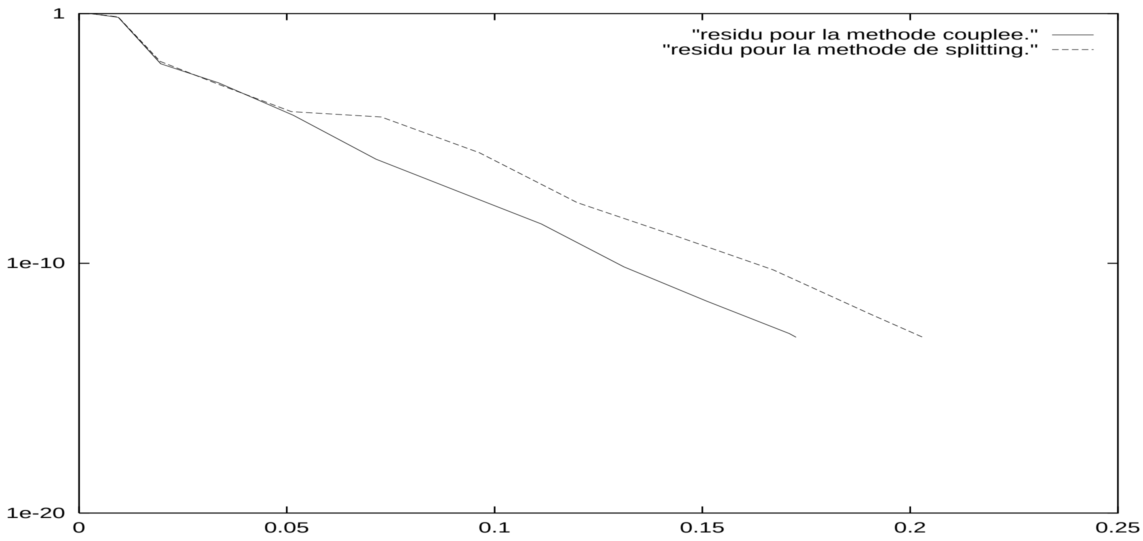


FIG. VIII.47 – *Residus obtenus pour* $r = 50.10^{-6}\, m$ *avec les deux méthodes.*

Temps de calcul et encombrement mémoire :

Nous reprenons le cas-test 1 avec le même maillage et le même schéma implicite que précédemment, et nous mesurons le temps CPU pour les deux méthodes jusqu'au temps d'arrêt $t = 0.7057$ seconde. On voit sur le tableau VIII.12 que le gain de temps de calcul est divisé par 2.25 pour la "méthode de splitting". De même, l'encombrement mémoire est 2.3 fois moins grand que pour la "méthode couplée". En effet, dans le cas d'un schéma implicite, nous sommes amenés à inverser la matrice du système (VIII.36) : avec la "méthode couplée", nous inversons une seule matrice constituée de blocs 8x8 par noeud, tandis que pour la "méthode de splitting", nous résolvons deux systèmes linéaires avec chacun une matrice constituée de blocs 4x4 par noeud. Le coût de l'inversion est donc plus élevé pour la "méthode couplée" que pour la "méthode de splitting".
Nous mesurons maintenant le temps de calcul pour la "méthode de splitting" en choisissant un nombre de Courant $CFL = 0.8$, de manière à avoir la bonne valeur de U_{moy} (c'est-à dire le même écart de vitesse que pour la "méthode couplée"). On observe que le temps nécessaire pour obtenir la solution stationnaire à $t = 0.7057$ seconde est 3.7 fois plus grand que pour la "méthode couplée".

	Temps CPU en secondes	Temps moyen par itération en secondes
Méthode de splitting avec CFL=40	4762	4.7
Méthode couplée avec CFL=40	10765	11.8
Méthode de splitting avec CFL=.8	39632	1.5

TAB. VIII.12 – *Coûts CPU pour les deux méthodes.*

VIII.5 Annexe D.

VIII.5.1 Matrice jacobienne des équations d'Euler.

Nous donnons ici les expressions de la matrice jacobienne $\mathcal{A}(\mathbf{W},\boldsymbol{\eta})$ et de la matrice de passage $\mathbf{T}$.

$$\mathcal{A}(\mathbf{W},\boldsymbol{\eta}) = \begin{pmatrix} 0 & \eta_x & \eta_y & 0 \\ \eta_x(\frac{\gamma-3}{2}u^2 + \frac{\gamma-1}{2}v^2) - \eta_y uv & \eta_x(3-\gamma)u + \eta_y v & \eta_x(1-\gamma)v + \eta_y u & (\gamma-1)\eta_x \\ \eta_y(\frac{\gamma-1}{2}u^2 + \frac{\gamma-3}{2}v^2) - \eta_x uv & \eta_y(1-\gamma)u + \eta_x v & \eta_y(3-\gamma)v + \eta_x u & (\gamma-1)\eta_y \\ (-\eta_x u - \eta_y v)a & \eta_x b + \eta_y(1-\gamma)uv & \eta_x(1-\gamma)uv + \eta_y c & \gamma(u\eta_x + v\eta_y) \end{pmatrix}$$

avec

$$\begin{cases} a = (\frac{\gamma E}{\rho} + (\gamma-1)(u^2+v^2)) \\ b = (\frac{\gamma E}{\rho} - (\gamma-1)(3u^2+v^2)) \\ c = (\frac{\gamma E}{\rho} - (\gamma-1)(u^2+3v^2)) \end{cases}$$

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & \frac{1}{2c^2} & \frac{1}{2c^2} \\ u & \eta_y & \frac{u+c\eta_x}{2c^2} & \frac{u-c\eta_x}{2c^2} \\ v & -\eta_x & \frac{v+c\eta_y}{2c^2} & \frac{u-c\eta_y}{2c^2} \\ \frac{u^2+v^2}{2} & \eta_y u - \eta_x v & \frac{H+c(\eta_x u+\eta_y v)}{2c^2} & \frac{H-c(\eta_x u+\eta_y v)}{2c^2} \end{pmatrix}$$

où H représente l'enthalpie par unité de masse et vérifie :

$$H = \frac{E+p}{\rho} = \gamma\varepsilon + \frac{1}{2}(u^2+v^2)$$

VIII.5.2 Matrice jacobienne $\mathcal{A}_2(\mathbf{W},\boldsymbol{\eta})$ du sous-système des gouttes.

$$\mathcal{A}_2(\mathbf{W},\boldsymbol{\eta}) = \begin{pmatrix} \eta_x u + \eta_y v & \eta_x(1-\alpha) & \eta_y(1-\alpha) & 0 \\ \eta_x X & \eta_x u + \eta_y \frac{v}{2} & \eta_y \frac{u}{2} & 0 \\ \eta_y X & \eta_x \frac{v}{2} & \eta_x \frac{u}{2} + \eta_y v & 0 \\ 0 & \eta_x(1-\alpha)\varepsilon & \eta_y(1-\alpha)\varepsilon & \eta_x u + \eta_y v \end{pmatrix}$$

où $X = \frac{\delta\theta_0}{\rho_l}(1-\alpha)^{\delta-2}$.

La matrice des vecteurs propres $\mathbf{T}_2$ est telle que :

$$\mathbf{T}_2 = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & \eta_y & \dfrac{b\eta_x - gu}{d} & \dfrac{b\eta_x + gu}{a} \\ 0 & -\eta_x & \dfrac{b\eta_y - gv}{d} & \dfrac{b\eta_y + gv}{a} \\ 1 & 0 & \varepsilon & \varepsilon \end{pmatrix}$$

avec :

$$\begin{cases} b = \dfrac{2X}{\sqrt{1-\alpha}} \\ g = \sqrt{X}||\boldsymbol{\eta}|| \\ a = \lambda_1\sqrt{1-\alpha} + 2g(1-\alpha) \\ d = \lambda_1\sqrt{1-\alpha} - 2g(1-\alpha) \end{cases}$$

La matrice de Roe $\tilde{\mathcal{A}}_2(\mathbf{W}_i, \mathbf{W}_j, \boldsymbol{\eta})$ ne diffère de la matrice jacobienne $\mathcal{A}_2(\tilde{\mathbf{W}}, \boldsymbol{\eta})$ que par le terme :

$$\tilde{X} = \begin{cases} \dfrac{\delta\theta_0}{\rho_l(\delta-1)} \dfrac{[(1-\alpha)^{\delta-1}]}{[(1-\alpha)]} & \text{si} \quad \alpha_\mathrm{i} \neq \alpha_\mathrm{j} \\ \dfrac{\delta\theta_0}{\rho_l}(1-\tilde{\alpha})^{\delta-2} & \text{si} \quad \alpha_\mathrm{i} = \alpha_\mathrm{j} \end{cases}$$

VIII.5.3 Matrice jacobienne $\dfrac{\mathbf{R}'(\mathbf{W})}{\epsilon}$ du terme source.

$$\frac{\mathbf{R}'(\mathbf{W})}{\epsilon} = \frac{1}{\epsilon} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ cu_g & -c & 0 & 0 & -\rho_l(u_g - u_l) & (1-\alpha)\rho_l & 0 & 0 \\ cv_g & 0 & -c & 0 & -\rho_l(v_g - v_l) & 0 & (1-\alpha)\rho_l & 0 \\ c\mathbf{u}_g.(2\mathbf{u}_g - \mathbf{u}_l) & -c(2u_g - u_l) & -c(2v_g - v_l) & 0 & -\rho_l\mathbf{u}_g.(\mathbf{u}_g - \mathbf{u}_l) & c\alpha\rho_g u_g & c\alpha\rho_g v_g & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\dfrac{u_g}{\alpha\rho_g} & \dfrac{1}{\alpha\rho_g} & 0 & 0 & 0 & -1 & 0 & 0 \\ -\dfrac{v_g}{\alpha\rho_g} & 0 & \dfrac{1}{\alpha\rho_g} & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

avec $\mathbf{u}_g$ (resp. $\mathbf{u}_l$) le vecteur vitesse de la phase gazeuse (resp. liquide) et $c = \dfrac{(1-\alpha)\rho_l}{\alpha\rho_g}$ le rapport des densités des deux phases.

Chapitre IX

CONCLUSION.

Dans cette deuxième partie, nous nous sommes intéressés à la résolution numérique de problèmes non linéaires hyperboliques et conservatifs, comme l'équation de Burgers, le système des équations d'Euler, puis un modèle simplifié d'écoulement diphasique. Pour tous ces modèles, les méthodes mises en oeuvre ont été de type Volumes Finis associées à des solveurs de type Godunov, en particulier les schémas basés sur des linéarisées de Roe. La discrétisation du domaine de calcul a été faite à l'aide de maillages en triangles, structurés ou non. Nous avons repris la technique des β-schémas avec des fonctions de flux numériques centrées ou décentrées pour étendre l'approximation spatiale à un ordre supérieur. Comme nous l'avions vu dans le cas linéaire, avec une condition initiale régulière, le β-schéma avec $\beta = \dfrac{1}{3}$ est d'ordre trois en espace si le flux est décentré, et d'ordre quatre avec un flux centré. Nous nous sommes intéréssés à des calculs instationnaires, comme celui du tube à choc de Sod pour les équations d'Euler ou la propagation d'une onde sonore dans un mélange diphasique, mais aussi au calcul de solutions stationnaires, pour un écoulement d'air autour d'un profil d'aile, ou pour un écoulement diphasique dans une tuyère. Avec ces différentes situations, nous avons utilisé, soit des schémas explicites d'ordre trois et quatre en temps, soit des méthodes implicites linéarisées utilisant la méthode de Jacobi comme algorithme de résolution.

Dans un premier temps, nous avons cherché à construire des limiteurs d'ordre élevé afin de rendre nos β-schémas TVD, sans perdre de précision dans les zones où la solution est régulière. En ce qui concerne le schéma utilisant un flux décentré, le nouveau limiteur s'est avéré robuste, stable et précis comparé aux limiteurs classiques de Van Albada et de Spekreijse. Dans le cas du tube à choc, les solutions présentent moins de diffusion numérique avec le nouveau limiteur et le nombre de Courant vaut 1.05 (contre 0.96 quand on utilise celui de Spekreijse). Nous avons aussi présenté un calcul de solutions stationnaires pour un écoulement transsonique autour d'un profil d'aile NACA 0012. Avec le nouveau

limiteur, nous avons obtenu la solution stationnaire facilement même pour un maillage très fin comportant 12284 noeuds, tandis qu'avec le limiteur de Spekreijse, il nous a été impossible d'obtenir une solution convergée avec la méthode implicite choisie (la norme L^2 du résidu reste constante au cours du temps).
Le β-schéma centré ne peut pas être limité sans aucune information sur le décentrage. Nous utilisons une fonction de flux décentrée qui se réduit à un flux centré lorsque le schéma n'est pas limité. Avec $\beta = \frac{1}{3}$, le schéma est TVD, stable (le nombre de Courant est de 1.28 avec un schéma d'ordre quatre en temps), mais il s'avère cher puisqu'il nécessite une intégration d'ordre quatre en temps pour des questions de précision ainsi que l'utilisation d'une fonction de flux décentrée.

Parmi les perspectives immédiates, nous envisageons de poursuivre le calcul de solutions stationnaires pour d'autres types d'écoulements, en particulier des écoulements supersoniques, dans le but de comparer la solution convergée avec le nouveau limiteur à celle obtenue avec le limiteur de Spekreijse.

Nous nous sommes ensuite intéressés à la construction d'une méthode de volumes finis pour la simulation numérique d'un système eulérien simplifié modélisant un écoulement diphasique. Le modèle mathématique utilisé est un système hyperbolique conservatif où les deux phases gaz-particules sont couplées par un terme source qui dépend d'un paramètre ϵ. Pour ce type d'écoulement, la méthode classique est une méthode de pas fractionnaires (ou “splitting”) [13, 2].
Cependant, cette méthode a plusieurs faiblesses. Si on considère la propagation d'une onde sonore dans le mélange diphasique, cette méthode introduit un amortissement excessif de l'onde lorsque le terme source devient raide. Les résultats ne sont pas en accord avec la théorie : l'amortissement ne tend pas vers 0 lorsque ϵ tend vers 0.

De plus, cette méthode n'est pas fiable pour calculer des solutions stationnaires, puisque la solution obtenue dépend du temps. Si on utilise un schéma implicite avec de grands pas de temps pour accélérer la convergence, cela revient en fait à augmenter l'importance des forces de traînée et la solution stationnaire peut être beaucoup modifiée.

C'est pourquoi nous avons proposé une “méthode couplée”, plus naturelle, qui consiste à ne pas séparer la résolution de la partie convective avec celle du terme source. Dans le cas de la propagation de l'onde sonore, cette méthode s'avère plus précise pour une grande échelle de valeurs de ϵ. En particulier, l'amortissement diminue avec le rayon des gouttes. On obtient une très bonne adéquation entre les résultats numériques et théoriques avec des schémas explicites de type Runge-Kutta, d'ordre trois et quatre en temps et en espace, ainsi qu'avec le schéma de Cranck-Nicolson pour une approximation spatiale d'ordre trois.

Dans le cas stationnaire, nous pouvons utiliser de grands pas de temps sans que la solution stationnaire soit perturbée, contrairement à la “méthode de splitting”.

Cependant, l'inconvénient majeur de la “méthode couplée” réside dans le coût en termes de temps de calcul et d'occupation mémoire. Une perspective intéressante de ce travail serait de trouver des moyens pour réduire ces deux facteurs, par exemple en parallélisant cette méthode.

Les deux cas-tests effectués sont des cas-tests connus, qui ont permis de comparer nos deux méthodes, et de mesurer leur aptitude à prendre en compte le couplage entre les deux phases. Il serait constructif maintenant de s'intéresser à des cas-test industriels, comme par exemple, l'étude d'une couche limite acoustique d'un écoulement dans un tube bidimensionnel, avec une injection pariétale et une sortie oscillante. Ce cas-test utilisé à l'ONÉRA, permet de modéliser de façon satisfaisante l'écoulement de gaz dans une fusée lorsque celui-ci engendre une couche limite acoustique près de la paroi d'injection [2].

Une autre perspective serait d'étendre le modèle simplifié à un modèle qui prenne mieux en compte les phénomènes physiques. Ainsi, dans la modélisation du terme source, nous avons supposé que toutes les gouttes étaient sphériques et avaient localement le même rayon. De même, nous avons négligé les phénomènes de vaporisation, de coalescence ou de fractionnement des gouttes, ainsi que les échanges d'énergie par conduction thermique. Peut-être pourrions-nous envisager une modélisation du terme source qui intègre quelques-unes de ces caractéristiques, comme c'est le cas pour les modèles cinétiques, sans que nous perdions le caractère hyperbolique et conservatif de notre modèle.

Bibliographie

[1] ANDERSON D. A., TANNEHILL J. C., PLETCHER R. H., *Computational Fluid Mechanics and Heat Transfer*, Vol. 1, Hemisphere Publishing Corporation (1984).

[2] BÉREUX F., *Modélisation et analyse d'écoulements diphasiques dans les moteurs à poudre.* Thèse de Doctorat de l'Ecole Polytechnique (1995).

[3] BURMAN E., SAINSAULIEU L., *Numerical analysis of two operator splitting methods for an hyperbolic system of conservation laws with stiff relaxation terms*, Rapport CERMICS no 28, (1994).

[4] DEPEYRE S., *Méthode de volumes finis pour des écoulements diphasiques dispersés*, Rapport de Stage de DESS Informatique et Sciences de l'Ingénieur (1993).

[5] DEPEYRE S., ISSAUTIER D., *Application aux schémas volumes finis d'une méthode de pénalisation des contraintes pour le système de Maxwell*, Rapport CERMICS no 39 (1995) et accepté pour publication dans "Mathematical Modelling and Numerical analysis".

[6] DEPEYRE S., LARROUTUROU B., CARPENTIER R., *Méthodes numériques décentrées d'ordre élevé en deux dimensions d'espace*, Rapport CERMICS no 41 (1995).

[7] DERVIEUX A., *Steady Euler simulations using unstructured meshes*, Cours au Von Karman Institute Lectures Series 85-04 (1985).

[8] FÉZOUI L., LANTÉRI S., LARROUTUROU B., OLIVIER C., *Résolution numérique des équations de Navier-Stockes pour un fluide compressible en maillage triangulaire*, Rapport de recherche INRIA (Mai 1989).

[9] GLINSKY N., *Simulation numérique d'écoulements hypersoniques réactifs hors-équilibre chimique*, Thèse de l'université de Nice-Sophia-Antipolis (1990).

[10] GODLEWSKI E., RAVIART P. A., *Hyperbolic Systems of Conservation Laws*, Vol. 1, Masson (1991).

[11] LARROUTUROU B., *Modélisation mathématique et numérique pour les sciences de l'ingénieur*, Cours de Majeure de l'Ecole Polytechnique (1995).

[12] RAVIART P. A., SAINSAULIEU L., *Mathematical and numerical modelling of two-phase flows*, Computing methods in applied sciences and engineering, Glowinsky éd., pp. 119-132, Nova Science Publisher, New-York, (1991).

[13] SAINSAULIEU L., *Modélisation, analyse mathématique et numérique d'écoulements diphasiques constitués d'un brouillard de gouttes*, Thèse de Doctorat de l'Ecole Polytechnique (1991).

[14] SAINSAULIEU L., LARROUTUROU B., *Modélisation eulérienne des écoulements diphasiques dispersés et résolution par une méthode décentrée*, Rapport CERMICS no 9, (1992).

[15] SAINSAULIEU L., *Finite-volume approximation of two-phase fluid flows based on an approximate roe-type riemann solver*, Rapport CERMICS no 10, (1992).

[16] SCHATZMANN M., *Higher order alternate directions methods*, Computer methods in applied mechanics and engineering, pp. 219-225, (1994).

[17] SMOLLER J., *Shock Waves and Reaction-Diffusion Equations*, Springer Verlag, Heidelberg, pp. 337-358, (1983).

[18] STÈVE H., *Schémas implicites linéarisés décentrés pour la résolution des équations d'Euler en plusieurs dimensions*, Thèse de l'université de Nice-Sophia-Antipolis (1988).

[19] STOUFFLET B., *Résolution numérique des équations d'Euler des fluides parfaits compressibles par des schémas implicites en éléments finis*, Thèse de Doctorat de l'Université Paris VI (1984).

[20] STRANG G., *Accurate partial difference methods II: non linear problems*, Numer. Math., (1964).

[21] VAN LEER B., *Flux vector splitting for the Euler equations*, Lecture Notes in Physics, Vol 170, pp 405-512 (1982)

[22] VILA J. P., *Approximation numérique des lois de conservation hyperboliques non linéaires*, Cours de l'Ecole Supérieure en Sciences Informatiques de Nice-Sophia-Antipolis (1992).

RÉSUMÉ

Nous nous sommes intéressés à la construction et à l'étude d'une classe de schémas d'ordre trois ou quatre en temps et en espace, basés sur des formulations "β-schémas" de type volumes finis ou éléments finis, pour des maillages bidimensionnels en rectangles ou en triangles. Nous considérons dans un premier temps des problèmes hyperboliques linéaires, comme l'équation d'advection et le système de Maxwell. Une étude de stabilité et de précision, à l'aide des équations équivalentes a été présentée, afin de comparer les schémas et de retenir les plus précis. En particulier, pour le système de Maxwell, une condition nécessaire et suffisante de stabilité a été démontrée pour le schéma décentré d'ordre un, sur un maillage en rectangles. Nous avons aussi proposé une nouvelle formulation du système de Maxwell, en rajoutant un terme de viscosité dans les équations, afin que nos schémas prennent mieux en compte les relations de divergence. Une étude de stabilité a permis de déterminer le paramètre de viscosité n'introduisant aucune contrainte supplémentaire sur le pas de temps, et nous avons montré à l'aide de résultats numériques, pourquoi la nouvelle formulation était meilleure. Dans la deuxième partie, nous nous sommes intéressés à des modèles hyperboliques non linéaires, comme les équations d'Euler. Nous avons cherché à construire des limiteurs d'ordre élevé afin de rendre nos schémas positifs. En particulier, nous avons présenté un nouveau limiteur d'ordre trois, qui s'est avéré stable et robuste, pour des calculs de tube à choc et d'écoulements transsoniques stationnaires. Nous avons finalement considéré un modèle eulérien d'écoulement diphasique, hyperbolique et conservatif, comportant un terme source raide. La méthode classique d'intégration en temps est une méthode de pas fractionnaires; toutefois, elle comporte plusieurs faiblesses, et nous avons proposé une méthode "couplée", qui s'avère plus précise lorsque le rayon des particules devient petit.

Mots clés: volumes finis – éléments finis – équations de Maxwell – stabilité – équations équivalentes – équations d'Euler – écoulements diphasiques – termes source raides

ABSTRACT

The aim of this work is the study and the construction of a class of third or fourth-order time-and-space accurate schemes based on finite volume or finite element formulations, using two-dimensional rectangular or triangular meshes. We have first considered linear hyperbolic models, like the advection equation and the Maxwell system. An analysis of the stability and the modified equations of the schemes has been presented, in order to compare and to draw the most efficient schemes. In the case of the first-order scheme using a rectangular mesh, applied to the Maxwell system, a necessary and sufficient stability condition has been be proved. We also have proposed a constrained formulation of the Maxwell system, in order to better satisfy the divergence conditions. A stability study allowed us to determine an optimum viscosity parameter which does not introduce any restriction on the time-step. Numerical results showed that this new formulation was well-adapted to the numerical conservation of the divergence. In the second part, we have considered non linear hyperbolic problems, like the inviscid Euler equations. We have aimed to construct high-order accurate limiters, in order to obtain positive schemes. We have presented a third-order accurate limiter, which is shown to be stable and robust, for shock tube tests and transsonic steady flows computations. Finally, we have considered a conservative hyperbolic eulerian two-phase model, where both phases are connected with a source term that may be stiff. A classical way to deal with this is a time-splitting method; unfortunately, it can lead to poorly accurate solutions, that is why we have proposed a "coupled method", which shows a good behaviour for small values of the particles radius.

Key words: finite volume – finite element – Maxwell equations– stability – modified equations – Euler equations – two-phase flows – stiff source terms